



UNIVERSIDADE FEDERAL FLUMINENSE  
ESCOLA DE ENGENHARIA  
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E DE  
TELECOMUNICAÇÕES

**FELIPE MUNARO LIMA**

**Modelo Baseado em Dados para Projeção do  
Número Mensal de Interrupções em Conjuntos  
Elétricos Utilizando Informações Públicas**

NITERÓI

2023

UNIVERSIDADE FEDERAL FLUMINENSE  
ESCOLA DE ENGENHARIA  
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E DE  
TELECOMUNICAÇÕES

**Felipe Munaro Lima**

**Modelo Baseado em Dados para Previsão do Número Mensal de  
Interrupções em Conjuntos Elétricos Utilizando Informações  
Públicas**

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica e de Telecomunicações da Universidade Federal Fluminense, como requisito parcial para obtenção do título de Mestre em Engenharia Elétrica e de Telecomunicações. Área de concentração: Modelagem de Sistemas

Orientador:

Andre Abel Augusto

Coorientador:

Vitor Hugo Ferreira

NITERÓI

2023

Ficha catalográfica automática - SDC/BEE  
Gerada com informações fornecidas pelo autor

L732m Lima, Felipe Munaro  
Modelo Baseado em Dados para Projeção do Número Mensal  
de Interrupções em Conjuntos Elétricos Utilizando  
Informações Públicas / Felipe Munaro Lima. - 2024.  
122 f.: il.

Orientador: Andre Abel Augusto.  
Coorientador: Vitor Hugo Ferreira.  
Dissertação (mestrado)-Universidade Federal Fluminense,  
Escola de Engenharia, Niterói, 2024.

1. Electrical Power Distribution. 2. Electrical Power  
Quality. 3. Outages Forecast. 4. Data-Based Modeling. 5.  
Produção intelectual. I. Augusto, Andre Abel, orientador.  
II. Ferreira, Vitor Hugo, coorientador. III. Universidade  
Federal Fluminense. Escola de Engenharia. IV. Título.

CDD - XXX

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

Felipe Munaro Lima

Modelagem Computacional para Projeção do Número Mensal de Interrupções em Conjuntos Elétricos Utilizando Informações Públicas

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica e de Telecomunicações da Universidade Federal Fluminense, como requisito parcial para obtenção do título de Mestre em Engenharia Elétrica e de Telecomunicações. Área de concentração: Modelagem de Sistemas

BANCA EXAMINADORA

---

Prof. André Abel Augusto, D.Sc. – Orientador, UFF

---

Prof. Vitor Hugo Ferreira, D.Sc. – Coorientador - UFF

---

Prof. Bruno Soares Moreira César Borba, D.Sc. – UFF

---

Prof. Leonardo Willer de Oliveira, D.Sc. – UFJF

Niterói

Dezembro de 2023

*À minha família.*

# Agradecimentos

Agradeço a todos que ajudaram na construção desse trabalho. E que dedicaram uma porção de seu tempo para possibilitar que sua realização fosse possível. Em especial a meus pais, sem os quais, não conseguiria nem ter iniciado, nem terminado esse trabalho, e aos meus orientadores pelas incontáveis revisões e conversas. Todos os erros que cometi durante o processo foram meus, mas os acertos divido com eles. Também agradeço ao laboratório da UFF FRIENDS que possibilitou a utilização do seu espaço para a confecção de parte da pesquisa, bem como ao instituto INERGE que prestou o auxílio financeiro como patrocinador deste projeto. Sem eles a confecção deste trabalho também não seria possível. Agradeço ainda a todo o corpo docente e administrativo do curso de Pós Graduação da Universidade Federal Fluminense (UFF) que possibilitou o intercâmbio de ideias que contribuíram em grande parte para esta dissertação.

# Resumo

As concessionárias de distribuição de energia elétrica são responsáveis pelo fornecimento deste insumo a um elevado número de consumidores, dispersos geograficamente. Elas são encarregadas do planejamento, projeto, implantação e operação das redes elétricas de distribuição na sua área de concessão. Uma das mais importantes obrigações das concessionárias de energia elétrica (também conhecidas como “distribuidoras”) é garantir o fornecimento ininterrupto de energia elétrica aos consumidores, considerando os padrões de qualidade de energia estabelecidos pela Agência Nacional de Energia Elétrica (ANEEL), sob o risco de penalizações que impactam sua receita anual. Nesse contexto, a previsão da quantidade de interrupções permite não apenas antever o desempenho da rede de distribuição, segundo os indicadores de qualidade do serviço prestado acompanhados pela ANEEL, como estimar a demanda de serviços em campo e, conseqüentemente, a composição da força de trabalho. Diversos trabalhos na literatura dedicam-se a investigar a projeção da quantidade de interrupções no fornecimento de energia, explorando diversas variáveis de impacto, como fatores meteorológicos, tipologia de rede, aspectos geográficos, dentre outros. Todavia, são escassos os trabalhos que consideram particularidades geográficas, ambientais e econômicas do Brasil e a acessibilidade dos dados. Sendo assim, este trabalho apresenta uma metodologia para projeção da quantidade mensal de interrupções em conjuntos elétricos, utilizando somente dados de domínio público. A metodologia foi avaliada levando-se em consideração um modelo de regressão linear múltipla, tanto no conjunto de dados de interrupções para o cenário geral brasileiro como, mais especificamente, para quatro distribuidoras selecionadas (Energisa Sergipe, Cemig, Copel e CPFL). Em todos os testes realizados, houve a melhoria das métricas de desempenho do modelo em relação ao método utilizado como comparação. Constatou-se uma redução no erro médio quadrático (*Mean Square Error* – MSE), no Erro Percentual Absoluto Médio (*Mean Absolute Percentage Error* – MAPE), e um aumento no coeficiente de determinação  $R^2$ .

**Palavras-chaves:** Distribuição de Energia Elétrica, Qualidade de Energia Elétrica, Previsão de Interrupções, Modelagem Baseada em Dados.

# Abstract

Electric utilities are responsible for supplying energy to a large number of consumers, spread across vast geographic locations. They are responsible for planning, designing, implementing and operating the electrical distribution networks in their concession area. One of the most important necessities for utilities (also known as energy distribution companies) is to ensure that the service is provided uninterruptedly to the consumers, considering the energy supply quality targets defined by the National Electric Energy Agency (ANEEL), at risk of penalties that impact the annual revenue. In this context, the outages forecast allows not only to foresee the performance of the distribution network, according to the quality indicators of the service provided monitored by ANEEL, but also estimate the field services demand, and consequently, the composition of the workforce. Different works in the literature are dedicated to investigating the outages forecast in energy supply, exploring several impact variables, such as meteorological factors, network typology, geographic aspects, among others. However, there are few studies that consider geographic, environmental and economic particularities of Brazil and data accessibility. Therefore, this work presents a methodology for projecting the monthly number of outages in electrical sets of consumers units, using only public domain data. The methodology is evaluated considering a multiple linear regression model on the outage's datasets for both the general Brazilian scenario and for four selected utilities (Energisa Sergipe, Cemig, Copel and CPFL). In all tests performed, there was an optimization in the metrics of performance of the model compared to the baseline. Can be observed a reduction in the mean square error (MSE), in the Mean Absolute Percentage Error (MAPE), and an increase in the coefficient of determination  $R^2$ .

**Keywords:** Electrical Power Distribution, Electrical Power Quality, Outages Forecast, Data-Based Modeling.



# Lista de Figuras

FIGURA 1 - DIVISÃO DO MAPA DO BRASIL POR CONJUNTOS DE UNIDADES CONSUMIDORAS A PARTIR DOS DADOS DE INFORMAÇÕES GEOGRÁFICAS CONSIDERADAS PELO SISTEMA SIGEL – ANEEL - 2023 [65] .....	3
FIGURA 2 - HISTÓRICO DA QUANTIDADE DE INTERRUPÇÕES NOS CONJUNTOS BEZERROS, BH ADELAIDE, BH ATALAIA E BH BARRO PRETO NO PERÍODO COMPREENDIDO ENTRE JANEIRO DE 2021 E DEZEMBRO DE 2022. INFORMAÇÕES RETIRADAS DO SITE DA ANEEL 2023. [66] .....	4
FIGURA 3 - FLUXOGRAMA GERAL DA METODOLOGIA PROPOSTA .....	28
FIGURA 4 - EXEMPLO DO GEOPROCESSAMENTO PARA O CASO EM QUE É REALIZADA A UNIÃO ENTRE AS INFORMAÇÕES DE MUNICÍPIOS E CONJUNTOS DE UNIDADES CONSUMIDORAS. NA ILUSTRAÇÃO À ESQUERDA, EM VERDE, PODEM SEREM VISTOS OS CENTROIDES EXTRAÍDOS DOS POLÍGONOS QUE DEFINEM OS CONJUNTOS DE UNIDADES CONSUMIDORAS. JÁ NA ILUSTRAÇÃO À DIREITA, AS INFORMAÇÕES DOS CENTROIDES SÃO CORRELACIONADAS COM AS REGIÕES EM AZUL, QUE DEFINEM CADA UM DOS MUNICÍPIOS NO TERRITÓRIO BRASILEIROS. ....	32
FIGURA 5 - RELAÇÃO ENTRE A MÁXIMA PRESSÃO ATMOSFÉRICA MÉDIA DIÁRIA (ESQUERDA) E A MÉDIA PRESSÃO ATMOSFÉRICA MÉDIA DIÁRIA (DIREITA) COM A QUANTIDADE DE INTERRUPÇÕES MENSAL POR CADA CONJUNTO ELÉTRICO (SIGEL). ....	37
FIGURA 6 - RELAÇÃO ENTRE A MÁXIMA UMIDADE RELATIVA DO AR MÉDIA DIÁRIA (ESQUERDA) E A MÉDIA UMIDADE RELATIVA DO AR MÉDIA DIÁRIA (DIREITA) COM A QUANTIDADE DE INTERRUPÇÕES MENSAL POR CADA CONJUNTO ELÉTRICO (SIGEL). ....	37
FIGURA 7 - RELAÇÃO ENTRE A MÉDIA UMIDADE RELATIVA DO AR MÍNIMA DIÁRIA (ESQUERDA) E A MÁXIMA UMIDADE RELATIVA DO AR MÍNIMA DIÁRIA (DIREITA) COM A QUANTIDADE DE INTERRUPÇÕES MENSAL POR CADA CONJUNTO ELÉTRICO (SIGEL). ....	38
FIGURA 8 - RELAÇÃO ENTRE A LATITUDE (ESQUERDA) E A LONGITUDE (DIREITA) COM A QUANTIDADE DE INTERRUPÇÕES MENSAL POR CADA CONJUNTO ELÉTRICO (SIGEL). ....	38
FIGURA 9 - RELAÇÃO ENTRE A MÁXIMA TEMPERATURA MÉDIA DIÁRIA E A MÉDIA TEMPERATURA MÉDIA DIÁRIA COM A QUANTIDADE DE INTERRUPÇÕES MENSAL POR CADA CONJUNTO ELÉTRICO (SIGEL). ....	39
FIGURA 10 - RELAÇÃO ENTRE A MÉDIA TEMPERATURA MÁXIMA DIÁRIA (ESQUERDA) E A MÁXIMA TEMPERATURA MÁXIMA DIÁRIA COM A QUANTIDADE DE INTERRUPÇÕES MENSAL POR CADA CONJUNTO ELÉTRICO (SIGEL). ....	39
FIGURA 11 - RELAÇÃO ENTRE A MÁXIMA TEMPERATURA MÍNIMA DIÁRIA (ESQUERDA) E A MÉDIA TEMPERATURA MÍNIMA DIÁRIA COM A QUANTIDADE DE INTERRUPÇÕES MENSAL POR CADA CONJUNTO ELÉTRICO (SIGEL). ....	40
FIGURA 12 - RELAÇÃO ENTRE A MÉDIA VELOCIDADE DO VENTO MÉDIA DIÁRIA (ESQUERDA) E A MÁXIMA VELOCIDADE DO VENTO MÉDIA DIÁRIA (DIREITA) COM A QUANTIDADE DE INTERRUPÇÕES MENSAL POR CADA CONJUNTO ELÉTRICO (SIGEL). ....	40
FIGURA 13 - RELAÇÃO ENTRE A MÉDIA RAJADA DE VENTO MÁXIMA DIÁRIA (ESQUERDA) E A MÁXIMA RAJADA DE VENTO MÁXIMA DIÁRIA (DIREITA) COM A QUANTIDADE DE INTERRUPÇÕES MENSAL POR CADA CONJUNTO ELÉTRICO (SIGEL). ...	41
FIGURA 14 - RELAÇÃO ENTRE A MÁXIMA TEMPERATURA DO PONTO DE ORVALHO MÉDIA DIÁRIA (ESQUERDA) E A MÉDIA TEMPERATURA DO PONTO DE ORVALHO MÉDIA DIÁRIA (DIREITA) COM A QUANTIDADE DE INTERRUPÇÕES MENSAL POR CADA CONJUNTO ELÉTRICO (SIGEL). ....	41
FIGURA 15 - RELAÇÃO ENTRE A MÉDIA PRECIPITAÇÃO TOTAL DIÁRIA (ESQUERDA) E A MÁXIMA PRECIPITAÇÃO TOTAL DIÁRIA COM A QUANTIDADE DE INTERRUPÇÕES MENSAL POR CADA CONJUNTO ELÉTRICO (SIGEL). ....	42
FIGURA 16 - RELAÇÃO ENTRE A PRECIPITAÇÃO TOTAL MENSAL E O NÚMERO DE UNIDADES CONSUMIDORAS NO CONJUNTO COM A QUANTIDADE DE INTERRUPÇÕES MENSAL POR CADA CONJUNTO ELÉTRICO (SIGEL). ....	42
FIGURA 17 - RELAÇÃO ENTRE O RAIO E A ÁREA COM A QUANTIDADE DE INTERRUPÇÕES MENSAL POR CADA CONJUNTO ELÉTRICO (SIGEL). ....	43
FIGURA 18 - RELAÇÃO ENTRE A TRANSFORMAÇÃO LOGARÍTMICA DA QUANTIDADE DE INTERRUPÇÕES E O PIB COM A QUANTIDADE DE INTERRUPÇÕES MENSAL POR CADA CONJUNTO ELÉTRICO (SIGEL). ....	43
FIGURA 19 - RELAÇÃO ENTRE O PERCENTUAL DE COBERTURA ARBÓREA E A POPULAÇÃO ESTIMADA EM 2021 COM A QUANTIDADE DE INTERRUPÇÕES MENSAL POR CONJUNTO ELÉTRICO (SIGEL) .....	44
FIGURA 20 - HISTOGRAMA CONTENDO A DISPERSÃO DOS CASOS DE INTERRUPÇÕES MENSAIS NOS CONJUNTOS ELÉTRICOS. ....	51
FIGURA 21 - QUANTIDADE DE INTERRUPÇÕES NO FORNECIMENTO DE ENERGIA ELÉTRICA LEVANTADOS PELA ANEEL, NOS ANOS DE 2020 A 2022, POR DISTRIBUIDORA DE ENERGIA. ....	52
FIGURA 22 - QUANTIDADE DE INTERRUPÇÕES POR ESTADO BRASILEIRO NO CONJUNTO DE DADOS FINAL, NOS ANOS DE 2020 A 2022 – DADOS ANEEL – 2023 .....	53
FIGURA 23 - HISTOGRAMA DO TEMPO DURAÇÃO DE INTERRUPÇÕES EM HORAS, PARA O ANO DE 2021. ....	55
FIGURA 24 - DENSIDADE ACUMULADA DO TEMPO DE DURAÇÃO DE INTERRUPÇÕES EM HORAS, PARA O ANO DE 2021. ....	55

FIGURA 25 – APLICAÇÃO DO MÉTODO ELBOW PARA IDENTIFICAÇÃO DO NÚMERO DE CLUSTER .....	56
FIGURA 26 – RESULTADO DO MÉTODO K-MEANS COM A DEFINIÇÃO DE 3 CLUSTERS ( $k=3$ ) EM RELAÇÃO AO INTERVALO DE TEMPO MÉDIO DE DURAÇÃO DOS EVENTOS DE INTERRUPÇÃO E A QUANTIDADE TOTAL DE INTERRUPÇÕES MENSAIS.....	57
FIGURA 27 – RESULTADO DO MÉTODO K-MEANS COM A DEFINIÇÃO DOS 4 CLUSTER ( $k=4$ ) EM RELAÇÃO AO INTERVALO DE TEMPO MÉDIO DE DURAÇÃO DOS EVENTOS DE INTERRUPÇÃO E A QUANTIDADE TOTAL DE INTERRUPÇÕES MENSAIS.....	58
FIGURA 28 – HISTOGRAMA COM OS RESULTADOS DA FERRAMENTA DE PROJEÇÃO EM NÍVEL NACIONAL, UTILIZANDO O MODELO <i>BASELINE</i> DE PREVISÃO SAZONAL SIMPLES (ESQUERDA) E O MODELO DE REGRESSÃO MÚLTIPLA (DIREITA).....	63
FIGURA 29 - DISPERSÃO DOS RESULTADOS DA FERRAMENTA DE PROJEÇÃO, APRESENTANDO A QUANTIDADE DE INTERRUPÇÕES REAIS, NO EIXO HORIZONTAL, E PROJETADO, NO EIXO VERTICAL. CADA COLUNA APRESENTA, A ESQUERDA, O MODELO <i>BASELINE</i> E A DIREITA O MODELO DE REGRESSÃO LINEAR MÚLTIPLA. EM CADA LINHA TÊM-SE OS RESULTADOS DE CADA DISTRIBUIDORA, SEGUINDO A SEGUINTE ORDEM, ENERGISA SERGIPE, CEMIG, COPEL, CPLF.....	66
FIGURA 30 - HISTOGRAMA DOS RESÍDUOS DE PREVISÃO DO MODELO DE REGRESSÃO LINEAR MÚLTIPLA, PARA OS CASOS DE TESTES COM CADA DISTRIBUIDORA.....	69

# Lista de Tabelas

TABELA 1 - CARACTERÍSTICAS UTILIZADAS COMO ENTRADA DO MODELO – CARACTERÍSTICAS 1 A 8.....	34
TABELA 2 – CARACTERÍSTICAS UTILIZADAS COMO ENTRADA DO MODELO – CARACTERÍSTICAS 9 A 21 .....	35
TABELA 3 - CARACTERÍSTICAS UTILIZADAS COMO ENTRADA DO MODELO – CARACTERÍSTICAS 22 A 37 .....	36
TABELA 4 - QUANTIDADE DE CONJUNTOS MAPEADOS VS. FALTANTES.....	50
TABELA 5 – RELAÇÃO ENTRE AS FONTES DE DADOS PÚBLICAS E AS INFORMAÇÕES QUE FORAM EXTRAÍDAS .....	50
TABELA 6 – QUANTIDADE DE INTERRUPÇÕES VS. TEMPO PARA RESTABELECIMENTO DO SERVIÇO. ....	56
TABELA 7 – RESULTADOS DA APLICAÇÃO DO TESTE DE HIPÓTESE PARA A ESTATÍSTICA F.....	60
TABELA 8 – SELEÇÃO DE ATRIBUTOS REALIZADA A PARTIR DA APLICAÇÃO DA ESTATÍSTICA T.....	61
TABELA 9 - RELAÇÃO ENTRE A QUANTIDADE DE REGISTROS UTILIZADOS PARA TESTE E TREINO .....	62
TABELA 10 - RESULTADOS DAS INFERÊNCIAS <i>BASELINE</i> (MÉTODO DE PREVISÃO SAZONAL SIMPLES) VERSUS A APLICAÇÃO DO MODELO DE REGRESSÃO LINEAR MÚLTIPLA PARA A PROJEÇÃO EM NÍVEL NACIONAL. ....	63
TABELA 11 – RELAÇÃO ENTRE A QUANTIDADE DE REGISTROS UTILIZADOS PARA TESTE E TREINO.....	65
TABELA 12 - RESULTADOS DAS INFERÊNCIAS <i>BASELINE</i> (MÉTODO DE <i>NAIVE</i> SAZONAL) VS. A APLICAÇÃO DO MODELO DE REGRESSÃO LINEAR MÚLTIPLA PARA DISTRIBUIDORA ENERGISA SERGIPE, CEMIG E COPEL. ....	68
TABELA 13 - COMPARAÇÃO DOS RESULTADOS DA INFERÊNCIA DO MODELO APÓS A SELEÇÃO DE ATRIBUTOS PARA OS CASOS DE TESTE REALIZADOS.....	71

# Lista de Abreviaturas e Siglas

ABCs	<i>Aerial Bundled Cables</i> (Cabos Aéreos Agrupados)
ANEEL	Agência Nacional de Energia Elétrica
ABM	<i>Agent-based Model</i> (Modelo Baseado em Agente)
ARIMA	<i>Autoregressive Integrated Moving Average</i> (Modelo Autoregressível Integrado de Média Móvel)
BPNNs	<i>Belief Propagation Neural Networks</i> (Redes de Retro Propagação de Confiança)
CELPE	Companhia de Eletricidade de Pernambuco
CEMIG	Companhia Energética de Minas Gerais
COPEL	Companhia Paranaense de Energia
CNN	<i>Convolutional Neural Network</i> (Rede Neural Convolutacional)
DSOs	<i>Distribution System Operators</i> (Operadores da Rede de Distribuição)
COVID-19	Doença Causada pelo Coronavírus Sars-cov-2
DEC	Duração Equivalente de Interrupções por Consumidor

ECM	<i>Error-correction Model</i> (Modelo de Correlação de Erro)
FEC	Frequência Equivalente de Interrupções por Consumidor
GS	<i>Gaussian Process</i> (Processo Gaussiano)
GRNN	<i>General Regression Neural Network</i> (Rede Neural de Regressão Generalizada)
GIS	<i>Geographic Information System</i> (Sistema de Informações Geográficas)
IBGE	Instituto Brasileiro de Geografia e Estatística
INMET	Instituto Nacional de Meteorologia
INPE	Instituto Nacional de Pesquisas Espaciais
KNN	<i>K-Nearest Neighbors</i> (K-ésimos Vizinhos mais Próximos)
MODIS	<i>MODerate-resolution Imaging Spectroradiometer</i> (Espectrorradiômetro de imagem de resolução moderada)
NASA	<i>National Aeronautics and Space Administration</i> (Administração Nacional da Aeronáutica e Espaço)
NGO	Otimização Açor-Nortenho ( <i>Northern Goshawk Optimization</i> )
O&M	Operação e Manutenção

ONS	Operador Nacional do Sistema Elétrico
PNNs	<i>Probabilistic Neural Networks</i> (Redes Neurais Probabilísticas)
PIB	Produto Interno Bruto
RVM	<i>Relevance Vector Machine</i> (Máquina de Vetor de Relevância)
SONN	<i>Self Organizing Neural Network</i> (Rede Neural Auto Organizável)
SIGEL	Sistema de Informações Georreferenciadas do Setor Elétrico
SIN	Sistema Interligado Nacional
SEB	Sistema Elétrico Brasileiro
SED	Subestação de Distribuição
SVMs	<i>Support Vector Machines</i> (Máquinas de Vetor de Suporte)
SAIDI	<i>System Average Interruption Duration Index</i> (Índice de duração média de interrupção do sistema)
SAIFI	<i>System Average Interruption Frequency Index</i> (Índice de frequência média de interrupção do sistema)
DHS	<i>United States Department of Homeland Security</i> (Departamento de Segurança Interna dos Estados Unidos)

VECM                    *Vector Error-Correction Model* (Modelo Vetorial de Correção de Erro)

WP                      *Workforce Planning* (Planejamento da Força de Trabalho)

# Sumário

<b>Capítulo 1 - Introdução</b> .....	1
<b>1.1 Considerações Iniciais</b> .....	1
<b>1.2 Objetivos</b> .....	3
<b>1.3 Contribuições deste Trabalho</b> .....	5
<b>1.4 Estrutura do Trabalho</b> .....	5
<b>Capítulo 2 - Revisão da Literatura</b> .....	8
<b>2.1 Previsão de Interrupções em Sistemas Elétricos</b> .....	8
<b>2.2 Aplicações no Planejamento e Operação de Redes de Distribuição</b> .....	11
<b>2.2.1 Regulamentação das Interrupções no Serviço de Fornecimento de Energia</b> 11	
<b>2.2.2 Planejamento da Força de Trabalho</b> .....	5
<b>2.3 Causas de Interrupções na Rede de Elétrica</b> .....	7
<b>2.3.1 Condições Climáticas</b> .....	7
<b>2.3.2 Áreas Litorâneas</b> .....	10
<b>2.3.3 Vegetação e Áreas Arborizadas</b> .....	10
<b>2.3.4 Características Socioeconômicas</b> .....	11
<b>2.3.5 Características Geográficas</b> .....	12
<b>2.4 Reestabelecimento de Energia</b> .....	12
<b>Capítulo 3 - Conceituação Teórica</b> .....	14
<b>3.1 Método de Previsão Sazonal Simples</b> .....	14
<b>3.2 Regressão Linear Múltipla</b> .....	15
<b>3.2.1 Estimação por Mínimos Quadrados</b> .....	16
<b>3.3 Métricas de Qualidade do Ajuste e Avaliação da Qualidade da Previsão</b> ...	17
<b>3.3.2 Erro Percentual Absoluto Médio (MAPE)</b> .....	18
<b>3.4.2 Coeficiente de Determinação (<math>R^2</math>)</b> .....	19
<b>3.4.2 Erro Médio Quadrático (MSE)</b> .....	20
<b>3.4 Teste de Hipótese</b> .....	21
<b>3.4.1 Teste F</b> .....	21
<b>3.4.2 Teste T</b> .....	22
<b>3.5 Clusterização</b> .....	24
<b>3.5.1 K-médias (<i>K-means</i>)</b> .....	24
<b>3.5.1.1 Método do Cotovelo (<i>Elbow Method</i>)</b> .....	25



<b>Capítulo 4 - Metodologia</b> .....	27
<b>4.1</b> <b>Formulação Geral da Metodologia Proposta</b> .....	27
<b>4.2</b> <b>Geoprocessamento</b> .....	29
<b>4.3</b> <b>Clusterização</b> .....	32
<b>4.4</b> <b>Avaliação das Relações de Causa e Efeito entre as Características e as Interrupções</b> .....	33
<b>4.5</b> <b>Processamento dos Dados e Aplicação do Modelo</b> .....	45
<b>4.6</b> <b>Validação do Modelo via Estatística F</b> .....	48
<b>4.7</b> <b>Seleção de Características via Estatística T</b> .....	48
<b>Capítulo 5 - Resultados e Discussões</b> .....	49
<b>5.1</b> <b>Conjunto de Dados</b> .....	49
<b>5.2</b> <b>Descrição Estatística do Conjunto de Dados Utilizado</b> .....	50
<b>5.2.1</b> <b>Avaliação dos Registros de Interrupções</b> .....	50
<b>5.3</b> <b>Resultados da Clusterização</b> .....	54
<b>5.4</b> <b>Resultados da Aplicação do Teste com a Estatística F</b> .....	59
<b>5.5</b> <b>Resultados da Seleção de Características com a Estatística T</b> .....	60
<b>5.6</b> <b>Resultados da Projeção da Quantidade de Interrupções em Nível Nacional</b> 62	
<b>5.7</b> <b>Resultado das Projeções da Quantidade de Interrupções para Distribuidoras Específicas Seleccionadas</b> .....	64
<b>5.7.1</b> <b>Diagnóstico dos Resíduos de Previsão</b> .....	69
<b>5.7.2</b> <b>Resultados após a Seleção de Características do Modelo</b> .....	70
<b>Capítulo 6 - Conclusão e Trabalhos Futuros</b> .....	73
Bibliografia	75
<b>Anexo A – Correlação entre os atributos e a quantidade de interrupções</b> .....	85
<b>Anexo B – Atributos considerados após a aplicação da estatística T.</b> .....	86

# Capítulo 1 - Introdução

## 1.1 Considerações Iniciais

A continuidade no fornecimento de energia elétrica é uma preocupação constante de autoridades, Operadores de Redes de Distribuição (*Distribution System Operators – DSOs*) e consumidores. À medida que cresce a dependência da sociedade da eletricidade, interrupções do sistemas de energia elétrica tem impactos sociais, políticos e econômicos cada vez maiores. Eventos como os apagões no Amapá (2020), afetando a vida econômica de 13 dos 16 municípios do estado [1], no Nordeste (2023) [2], gerado por uma sobrecarga no sistema de transmissão do Ceará levando este à um colapso, Enel-RJ [3] e Enel-SP [4], mostram o impacto de interrupções na rede elétrica, uma vez que cidades inteiras ficaram sem energia. Levando em consideração a importância da continuidade no fornecimento, não é surpresa que diversos órgãos de estado, como por exemplo o Departamento de Segurança Interna dos Estados Unidos (*United States Department of Homeland Security – DHS*), reconheçam o setor energético como crítico, ou como descrito pelo DHS, “excepcionalmente crítico, porque fornece uma função facilitadora em todos os setores que apresentam infraestruturas críticas” [5, 6], devido ao papel vital para diversos setores, tanto públicos (como por exemplo hospitais, transportes e segurança) como privados (como por exemplo empresas). Assim, uma falha pode afetar diretamente diferentes sistemas que apresentam o papel de fornecer energia elétrica para o funcionamento destes setores [7, 8].

Nesse contexto, a crescente frequência e duração de eventos de impacto na rede elétrica e em seus equipamentos, causados por fenômenos climáticos extremos, tornaram-se um risco para a segurança operativa dos sistemas de energia elétrica [8]. As durações e frequências das interrupções, especialmente aquelas induzidas por tais eventos, são grandes preocupações das companhias de distribuição de energia. A sua inadequação aos padrões previamente estabelecidos pelos órgãos reguladores e fiscalizadores pode levar a penalidades por parte desses. Portanto, elas devem ser minimizadas ou ao menos mantidas em níveis aceitáveis.

A previsão da quantidade de interrupções é importante tanto para as companhias de distribuição quanto para as entidades que têm atuação direta ou indireta no acompanhamento e fiscalização do serviço que é prestado no sistema elétrico. Esta previsão pode servir de insumo e de apoio à tomada de decisão para diversos processos e estudos que têm o potencial

de mitigar os impactos da queda do fornecimento de energia. São exemplos desses estudos, a avaliação da manutenção dos indicadores de qualidade do fornecimento de energia (definidos pela regulamentação vigente) [9], a alocação da força de trabalho para ação em casos de interrupções na rede elétrica [10], análise da confiabilidade de redes de distribuição [11] e a estruturação de processos de reestabelecimento do fornecimento elétrico [12].

Uma variedade de trabalhos na literatura vem investigando a construção de ferramentas para avaliação de danos na rede sob a influência de eventos externos graves, especialmente relacionados a condições meteorológicas [13–15]. Têm sido utilizadas para este fim técnicas como: lógica nebulosa (lógica *fuzzy*) [16], regressão linear e não-linear, método baseados em inferência bayesiana e métodos qualitativos de tomada de decisão [17]. Alguns trabalhos na literatura debruçam-se sobre o restabelecimento de energia após a ocorrência de interrupções [11-14].

O planejamento de recursos operacionais é um exemplo recorrente de processo em que é necessário o entendimento da quantidade de interrupções no fornecimento de eletricidade para os consumidores. Dentre os tipos de estudos, podem ser consideradas tarefas como: gestão de equipamentos, programação para atuação em contingência e alocação da força de trabalho (*Workforce Planning* – WP) para atendimento em campo [18]. O planejamento da força de trabalho voltado para ações de operação e manutenção (WP O&M) tem sido estudado no setor elétrico nos últimos anos, principalmente pelas distribuidoras de energia. Estes estudos se refletem tanto na maior qualidade do serviço prestado quanto na melhoria dos indicadores coletivos de Duração Equivalente de Interrupções (DEC) - ou, utilizando o seu equivalente internacional, *System Average Duration Index* (SAIDI) - e Frequência Equivalente de Interrupções (FEC) - ou, em seu equivalente internacional, *System Average Interruption Frequency Index* (SAIFI). No caso do Brasil, a Companhia Paranaense de Energia (COPEL) [19], por exemplo, usa o O&M WP para aperfeiçoar o dimensionamento e envio de equipes de campo de um determinado órgão até os locais de ocorrência através de um modelo matemático. A Companhia Energética de Pernambuco (CELPE) [20], por sua vez, realiza estudos para a logística das equipes dos chamados de emergência a fim de aumentar a capacidade de atendimento. A Companhia Energética de Minas Gerais (CEMIG) [21] utiliza um sistema para gerenciar a logística das equipes de campo, mas sem dar mais detalhes sobre a solução aplicada. Outras distribuidoras, como a AES Sul [22] e Elektro [23], investigaram o WP O&M a partir de soluções analíticas e modelos matemáticos. O grupo de distribuidoras brasileiro, Energisa,

utiliza um modelo de previsão e otimização da alocação da força de trabalho baseado em padrões espaço-temporais como principal entrada [24, 25], sendo este modelo implementado em onze distribuidoras deste grupo.

Mesmo com o planejamento realizado pelas distribuidoras, os serviços de restauração de energia são essencialmente intensivos em mão de obra, sendo a força de trabalho que executa essas tarefas altamente remunerada e com necessidade do domínio de habilidade em diferentes proficiências sobre o funcionamento do sistema. Além disto, há uma alta variabilidade na quantidade de ocorrências de eventos de interrupções ao longo do ano, o que dificulta o agendamento de tarefas, a contratação de mão de obra e a projeção de demanda necessária para realizar o trabalho [10].

Levando em consideração todos os estudos que foram realizados por entidades e companhias, e principalmente o impacto negativo de uma falha no serviço de fornecimento de energia elétrica, projetar a quantidade de interrupções é uma tarefa essencial e de grande importância. Esta projeção é uma necessidade tanto das distribuidoras quanto dos demais agentes que são afetados direta ou indiretamente pela falta do fornecimento de energia elétrica.

## 1.2 Objetivos

Dado o contexto apresentado na seção 1.1 e a importância da projeção da quantidade de paradas não planejadas no fornecimento de energia elétrica, esse trabalho tem o objetivo de abordar o problema de previsão da quantidade de interrupções ocorridas no sistema elétrico brasileiro.

Visando desenvolver uma metodologia que seja independente dos dados corporativos de uma distribuidora em particular, este estudo utiliza somente informações coletadas de fontes públicas, de acessibilidade irrestrita.

A ANEEL disponibiliza o número de interrupções por conjunto de unidades consumidoras (também descrito no decorrer do estudo, como conjuntos elétricos). Esta é uma agregação referente a conjuntos de clientes das companhias de distribuição a partir de critérios previamente definidos por esta instituição. O acompanhamento dos indicadores de DEC e FEC também é realizado nesta agregação. Dessa forma, o modelo proposto gera as projeções da quantidade de interrupções levando em consideração este agrupamento. Apesar da disponibilidade de dados horários de algumas das fontes que foram utilizadas nesta

dissertação, a previsão da quantidade de interrupções na rede elétrica é uma atividade de médio a longo prazo e, portanto, o horizonte de previsão definido para a ferramenta apresenta a agregação mensal.

Na literatura, não foi encontrado um estudo que contemplasse ao mesmo tempo, tanto todos os atributos meteorológicos, as informações acerca da geografia territorial brasileira e os indicadores socioeconômicos considerados neste trabalho, e que, além disto, se baseasse apenas em fontes de livre acesso. Tendo a necessidade de um método de comparação para a avaliação dos resultados que foram gerados pela ferramenta construída, foi utilizada uma técnica simplificada como ponto de partida para as previsões. Esta consiste em um modelo de previsão sazonal simples, sendo este, utilizado como comparativo na avaliação da aplicação de diversas metodologias para resolução de problemas de séries temporais, e também em sistemas de potência [26]. Na metodologia aplicada para a projeção da quantidade de intermissões não planejadas, foi utilizado um modelo de regressão múltipla. Este modelo descreve a relação entre as variáveis coletadas em cada um dos conjuntos de unidades consumidoras, com as quantidades de interrupções que foram observadas na rede elétrica nestes conjuntos. Escolheu-se esse método devido a sua simplicidade, compreensibilidade (os parâmetros do modelo quantificam a interdependência, do ponto de vista estatístico, entre cada variável de entrada e a saída desejada), e pelos resultados promissores obtidos com a sua aplicação.

É certo que já existe uma vasta discussão acerca da relação entre variáveis de impacto e eventos de paradas não planejadas na rede elétrica levando em consideração diversos casos de redes específicas e em diversas regiões. Porém, na literatura ainda são escassos trabalhos que:

- Busquem traçar a projeção da quantidade de eventos de interrupções para o caso específico do Sistema de Potência Nacional (Sistema Interligado Nacional - SIN) levando em consideração variáveis meteorológicas, regionais e socioeconômicas;
- Utilizem para isso apenas fontes de dados de livre acesso, fornecidas por entidades governamentais.

Este trabalho busca, além de realizar a previsão da quantidade de interrupções no sistemas de potência nacional, também vir a encorajar outros estudos, que como este, busquem avaliar o impacto de fenômenos naturais e de outros eventos para o serviço fornecimento de energia.

## 1.3 Contribuições deste Trabalho

Dado os objetivos que foram expostos na seção 1.2, este trabalho, contribui com as demais referências da literatura de forma que descreve e propõem uma metodologia única para a previsão da quantidade de interrupções em conjuntos elétricos brasileiros. Ainda, além de unir um conjunto de informações de fontes de livre acesso em uma base única que é utilizada na metodologia proposta no Capítulo 4, realiza a coleta destes dados apenas de fontes de acesso público. Com isso, possibilita também que outros trabalhos possam vir a utilizar uma metodologia parecida com esta e tragam resultados comparativos com os que foram expostos na seção 5 deste trabalho, otimizando, assim, a tarefa proposta nesta dissertação (referente à projeção da quantidade de interrupções em conjuntos de unidades consumidoras).

Ainda, de posse dos resultados gerados (capítulo 5) por este estudo, é possível realizar a avaliação da qualidade do fornecimento de energia nos conjuntos que foram estudados e traçar planos para a otimização dos indicadores que são utilizados para acompanhá-la.

As informações deste estudo, assim como descrito na seção 1.1, podem ser utilizadas para diversos outros, como por exemplo, o planejamento da força de trabalho. Sendo assim, tendo-se a projeção da quantidade de interrupções é possível realizar o atendimento aos serviços necessários para restauração da energia de forma mais eficaz, minimizando assim, custos e reduzindo-se o tempo de atendimento aos eventos de paradas não planejadas no fornecimento de energia elétrica.

## 1.4 Estrutura do Trabalho

O desenvolvimento desta dissertação foi organizado como descrito a seguir. No atual capítulo (Cap. 1) é apresentada a motivação, o objetivo e a estrutura deste trabalho (Secs. 1.1, 1.2, 1.3). No capítulo 2, é realizada uma revisão da literatura, levando em consideração pesquisas fundamentais tanto de estudos relacionados a previsão de interrupções, como também, de temas correlatos em que esta projeção é utilizada, como o planejamento da força de trabalho. Seguindo o sumário, os principais conteúdos apresentados neste capítulo são:

- Um estudo e pesquisa do conhecimento fundamental, aplicações, diretrizes e o escopo de vários métodos de previsão de interrupções (Sec. 2.1).

- Uma lista de referências incluindo recursos relacionados à previsão de interrupções retirados de livros, periódicos acadêmicos, artigos de conferências, *reports* técnicos, dentre outros recursos (Sec. 2.1).
- Uma exposição acerca dos estudos em que a informação da projeção da quantidade de interrupções é utilizada (Sec. 2.2).
- Uma apresentação das principais causas relacionadas a casos de interrupções em sistemas elétricos (Sec. 2.3).
- Uma discussão acerca do restabelecimento de energia após os casos de interrupções (Sec. 2.4).

No capítulo 3, é apresentada a conceituação teórica dos métodos utilizados para o desenvolvimento deste trabalho, abordando a teoria que serve de base para a construção da metodologia que será apresentada no capítulo subsequente (Cap. 4).

No capítulo 4, é descrita a metodologia aplicada no desenvolvimento da ferramenta proposta para previsão de quantidade de interrupções, bem como do método de previsão e suporte à tomada de decisão que foi desenvolvido. O capítulo 4, de forma mais detalhada, apresenta:

- Um sistema de geoprocessamento das informações que realiza a união dos dados de diferentes fontes em uma única base, levando em consideração a proximidade geográficas. A base final apresenta o índice formado pelos atributos de conjuntos de unidades consumidoras, mês e ano (Sec. 4.2).
- A apresentação do método utilizado para agrupamento das informações por quantidade de interrupções e duração destes eventos. Sendo abordada, a criação de um novo atributo, que é incluído no modelo representando o pertencimento de cada conjunto aos grupos (*clusters*) definidos (4.3).
- A apresentação de um relatório contendo todos os parâmetros e características capturadas para a construção do modelo, além do estudo da relação de cada característica com as quantidades de interrupções observadas (Sec. 4.4).
- O processamento de dados para a criação do modelo de previsão e suporte à tomada de decisão que realiza a estimativa da quantidade de interrupções, por conjunto de unidades consumidoras (Sec. 4.5).

- Um processo de avaliação do modelo, que testa a sua significância estatística, com a execução do teste de hipóteses com a estatística F (Sec. 4.6).
- A seleção das características de entrada do modelo utilizando o teste de hipóteses baseado na estatística T (Sec. 4.7).

O capítulo 5 apresenta os resultados da aplicação da metodologia, tanto para a base completa, contendo todas as distribuidoras, como para companhias específicas no Brasil (Energisa Sergipe, Cemig, Copel e CPFL).

Todo os scripts que realizam cada uma das etapas descritas foram escritos em linguagem de programação Python. Ao final do processo, são geradas todas as saídas de previsões para cada mês e para cada um dos conjuntos de unidades consumidoras avaliados nos testes.



## Capítulo 2 - Revisão da Literatura

Este capítulo apresenta uma ampla revisão literária sobre a previsão da quantidade de ocorrências de interrupções em sistemas de energia elétrica e temas correlatos. Na seção 2.1, será discutido sobre os principais métodos utilizados para previsão de intermissões em sistemas de potência. Em seguida, na seção 2.2, serão abordadas aplicações da previsão da quantidade de interrupções no planejamento e operação de redes de distribuição de energia elétrica. A seção 2.3 expõe os principais fatores que impactam na indisponibilidade do fornecimento de energia. Estes, em parte, serão utilizados no decorrer deste trabalho como insumos para a projeção da quantidade de interrupções nos grupos que serão avaliados. Por fim, na seção 2.4, aborda-se o restabelecimento de energia.

### 2.1 Previsão de Interrupções em Sistemas Elétricos

Quando se detecta uma interrupção na rede de uma concessionária, ela deve agir rapidamente para restabelecer o fornecimento de energia elétrica aos clientes afetados, isolando a parte defeituosa da rede de distribuição, reparando-a, e por fim reconectando-a à rede elétrica. Assim sendo, a investigação de técnicas de identificação de falhas em redes de distribuição tem sido extensivamente investigada na literatura. De acordo com Ferreira *et al.* [27] os métodos para diagnóstico de falhas podem ser agrupados em duas abordagens: clássica e inteligente. Abordagens clássicas empregam conceitos da teoria de circuitos tradicional para construir algoritmos de detecção, classificação e localização de falhas. Os métodos inteligentes utilizam técnicas de inteligência computacional, aprendizado de máquinas, sistemas inteligentes, modelagem baseada em dados [28], lógica *fuzzy*, redes neurais artificiais [29] e sistemas multiagentes [30] para realizar a análise das falhas. Behkam *et al.* utilizou uma Rede Neural de Regressão Generalizada (*Generalized Regression Neural Network* - GRNN) para detecção de faltas em enrolamentos de transformadores de potência [31]. Da mesma forma, Zimnickas *et al.* implementou uma proteção para motores de indução utilizando uma Rede Neural Convolutiva (*Convolutional Neural Network* – CNN) para identificação de defeitos mecânicos e elétricos na máquina [32]. Klomjit e Ngaopitakkul realizaram a comparação entre Redes Neurais Probabilísticas (*Probabilistic Neural Networks* – PNNs), Redes de Retro Propagação de Confiança (*Belief Propagation Neural*

*Networks* – BPNNs) e Máquinas de Vetor de Suporte (*Support Vector Machines* – SVMs) para detecção de defeitos [33]. Skowron aplicou uma Rede Neural Auto Organizável (*Self Organized Neural Network* – SONN) na identificação de falhas em motores de indução [34]. O trabalho de Patel et al. [35] sugere que a Máquina de Vetor de Relevância (*Relevance Vector Machine* – RVM) desempenha melhor que PNNs e SVMs para diagnóstico de falhas em transformadores. Métodos da teoria de sistemas moderna mostram o uso sistemático da matemática e modelagem de sinais, métodos de identificação e estimação, e de inteligência computacional [36]. Bhasker [37] propôs o uso de Transformada *Wavelet* e Redes Neurais Artificiais para classificação de interrupções. Outros trabalhos realizaram a modelagem de métodos de diagnósticos de falhas em sistemas elétricos utilizando lógica *fuzzy* [38] e Filtro de Kalman [39].

A maioria das falhas que resultam em interrupções originam-se de eventos que ocorrem regularmente e sem periodicidade definida. Exemplos destes, são: chuvas intensas e ventos forte. Desta forma, métodos de previsão da quantidade de defeitos servem de apoio para as empresas se programarem para intervenções que podem se originar destes eventos.

Muitos estudos abordam a previsão do número de interrupções em sistemas isolados, fontes de geração de energia, para um fenômeno natural específico, ou mesmo em cenários de ocorrências de eventos atípicos – como durante a pandemia de COVID-19, em que houve a alteração substancial do comportamento da geração e do consumo de energia elétrica. Em [40], Madurai Elavarasan R. *et al.* aborda este tema mais detalhadamente. Neste sentido, o trabalho de Leva *et al.* utilizou um método baseado em um conjunto (*ensemble*) de redes neurais artificiais para traçar o relacionamento entre variáveis meteorológicas e interrupções em sistemas fotovoltaicos [41]. Já Stéfano Frizzo, utilizou um modelo baseado em *Wavelet* Neuro-*Fuzzy* para previsão de interrupções devido a isoladores elétricos [42]. Pelo estudo ele mostrou que é possível atingir uma alta acurácia, mesmo comparando essa técnica com modelo de séries temporais clássicos, para problemas específicos de intermissões no sistema de potência. Yusha Hu et. al utilizou um método de Decomposição em Modelo Empírico (EMD) juntamente com um modelo baseado em Otimização por Enxame de Partículas (PSO) e Máquina de Vetor Suporte para, primeiramente, realizar a previsão da carga elétrica consumida levando em consideração um cenário industrial e, após, identificar casos de eventos de interrupção [43]. Ele chegou a conclusão, com base nos resultados do método, que os tempos de paradas não programadas em empreendimentos industriais poderia sofrer uma redução a partir da aplicação da metodologia que foi proposta. Branco N. W. *et al.* fez uso

da Transformada *Wavelet* e de um modelo de rede neural recorrente do tipo *Long Short-Term Memory* (LSTM) para previsão de falhas em distribuidoras de energia elétrica no Brasil durante o primeiro ano da pandemia COVID-19 (2020), mostrando que a aplicação de uma técnica de processamento de sinais, juntamente à uma rede neural pode otimizar o desempenho do modelo de previsão [44]. Levando em consideração outros algoritmos propostos, vários foram testados com a combinação de previsores baseando-se tanto em técnicas simples quanto complexas. Neste contexto, Ribeiro *et al.* [45], utilizou uma combinação entre um método baseado em *bootstrapping*, modelos de Processo Gaussiano (*Gaussian Process* - GS), SVMs, K-ésimos Vizinhos mais Próximos (*K-Nearest Neighbors* - KNN) e Florestas Aleatórias (*Random Forests*) para previsão da geração de energia eólica, atingindo um MAPE entre 7% e 25% melhor em comparação com estes modelos aplicados de forma única. Em relação à técnicas de regressão, Reed [13], em 2008, utilizou um modelo de regressão linear para previsão de interrupções em sistemas elétricos devido a incidência de ventos fortes, constatando que a avaliação da causa das interrupções, realizada pelas equipes de campo, é um insumo importante para o melhor desempenho do método aplicado. Liu *et al.* [46], utilizou modelos lineares generalizados (*Generalized Linear Models* – GLMs) e modelos lineares generalizados mistos para a previsão de intermissões devido a ciclones, chegando a conclusão de que o número de transformadores no alimentador é uma variável de alto impacto na previsão da quantidade de interrupções. Han *et al.* [47, 48] utilizou modelos generalizados aditivos (*Generalized Additive Models* – GAMs), mostrando que este tipo de abordagem pode aumentar o desempenho da tarefa de previsão da quantidade de intermissões na rede, em relação à outras abordagem de regressão linear. Nateghi *et al.* [49], por sua vez, construiu, baseado no trabalho anterior, um modelo que utiliza florestas aleatórias para atingir melhor exatidão .

Além dos casos em que foram desenvolvidos modelos para avaliar a quantidade de interrupções em cenários específicos, outros estudos realizam a avaliação da quantidade de interrupções em um horizonte temporal levando em consideração um conjunto de variáveis explicativas, em sua maioria relacionadas às condições meteorológicas. Yang F. *et al.* [50] utilizou um modelo *Gradient Boosting* para previsão de interrupções em sistemas elétricos devido a fenômenos climáticos encontrando que as variáveis de precipitação e rajada de vento desempenham um papel significativo na incerteza da previsão do modelo, enquanto a temperatura e o vento máximo sustentável um papel menor. Também foi observado que a incerteza das variáveis meteorológicas no estudo cresce quanto maior a janela de previsão.

Wanik *et al.* [51] estudou 89 eventos de interrupção causados pelo clima em diferentes estações, no estado de Connecticut, Estados Unidos, e empregou vários métodos de aprendizado de máquina, incluindo a árvore de gradiente aumentada, para prever as interrupções de energia naquela localidade, tendo o modelo alcançado um MAPE de 57,2%. Cerrai *et al.* [52] investigou diversos modelos de aprendizado de máquina para previsão de interrupções, obtendo para o melhor modelo um erro percentual absoluto médio (MAPE) de 65% para todo o conjunto de dados e de 80% para os eventos extratropicais. Os resultados das pesquisas de Wanik *et al.* e Cerrai *et al.* foram ambos provenientes de simulações numéricas de previsões meteorológicas. Nas últimas décadas outros métodos também foram utilizados para previsão do número de interrupções [53-59].

Apesar da vasta literatura sobre os métodos utilizados para previsão do número de interrupções em sistemas de potência, são escassos os trabalhos que abordam a previsão no contexto do sistema elétrico brasileiro, e também os que consideram restrições de acessibilidade dos dados empregados na construção de previsores.

## **2.2 Aplicações no Planejamento e Operação de Redes de Distribuição**

### **2.2.1 Regulamentação das Interrupções no Serviço de Fornecimento de Energia**

No Brasil, a Agência Nacional de Energia Elétrica (ANEEL) é responsável por acompanhar a qualidade do serviço provido pelas concessionárias de energia elétrica brasileiras. O seu escopo de atuação cobre tanto conformidades legais, como serviços comerciais e indicadores de continuidade de fornecimento. Esses indicadores, por sua vez, são verificados e enviados periodicamente pelas concessionárias a este órgão regulador para que seja verificada e avaliada a continuidade do abastecimento de energia aos consumidores. Conforme mencionado na seção 1.1, os indicadores de qualidade do fornecimento de energia (DEC e FEC, ou em suas siglas internacionais, SAIDI e SAIFI) representam a quantidade de vezes em que a unidade consumidora ficou sem o fornecimento de eletricidade (FEC ou SAIFI) e a duração (DEC ou SAIDI) desses eventos no período analisado (mês, trimestre ou ano), respectivamente [60].

O anexo VIII da normativa da ANEEL nº 948, de dezembro de 2021 [61] estabelece os critérios de qualidade do serviços prestado pelas distribuidoras com relação à continuidade do fornecimento de energia, bem como os procedimentos aplicáveis em caso de seus descumprimentos. As disposições previstas neste módulo são aplicáveis a todas as concessionárias de serviço público de distribuição.

Se a apuração dos indicadores de FEC e DEC, como descritos no módulo 8 do PRODIST [62], ultrapassar os limites anuais definidos pela ANEEL para o ano civil de referência, considera-se descumprida a norma estabelecida. Ainda, o descumprimento de qualquer um dos limitantes estabelecidos, pelo período de um ano, torna obrigatória a apresentação por parte da concessionária de um Plano de Resultados que deve ser aprovado pela ANEEL. A não adequação dos valores de DEC e FEC por dois anos, ou 3 dos cinco anos anteriores (sendo a análise feita anualmente em janelas de 5 anos corridos até o final do contrato) leva a concessionária a distribuidora a limitar o pagamento de dividendos e de juros sobre o capital próprio, isoladamente ou em conjunto, a 25% (vinte e cinco por cento) do lucro líquido, diminuído ou acrescido de valores definidos na normativa. Além de outras restrições, como a vedação da celebração de novos atos e negócios jurídicos entre a concessionária e suas partes relacionadas e a abertura de um processo administrativo punitivo, pela ANEEL, voltado a aplicação da penalidade de declaração de caducidade de concessão ou normas supervenientes que vierem a suceder esta última. .

A ANEEL, por meio do PRODIST [62], considera o DEC e o FEC como:

$$DEC = \frac{\sum_{i=1}^{Cc} DIC(i)}{Cc} \quad (1)$$

$$FEC = \frac{\sum_{i=1}^{Cc} FIC(i)}{Cc} \quad (2)$$

em que o *DIC* e o *FIC* são, respectivamente, as durações e frequências de interrupções por unidade consumidora individual, *i* é o índice de unidades consumidoras faturadas do conjunto atendidas em BT ou MT. A variável *Cc* se refere ao número total de unidades consumidoras faturadas do conjunto, também atendidas por BT ou MT, no período em que foi realizada a apuração.

Além dos indicadores, a ANEEL também disponibiliza as divisões territoriais da rede elétrica em conjuntos de unidades consumidoras no sistema de informações georreferenciadas do setor elétrico (SIGEL). Os indicadores de continuidade FEC e DEC

são acompanhados também nas bases de dados da ANEEL nesta mesma hierarquia de agregação.

O conceito de conjunto elétrico ou áreas geoeletricas assemelha-se ao da linha de restabelecimento de energia adotado na Itália [63] no que diz respeito o particionamento da rede *a priori*. Em ambos os casos, os agrupamentos representam configurações de rede em que o fornecimento de energia está ainda regulado em seus limites normais [64]. Todavia, no caso da Itália, as áreas geoeletricas são configurações mínimas de rede contendo uma ou mais unidades de geração, e permitindo assim o restabelecimento de cargas prioritárias de forma mais rápida e segura. No caso brasileiro, os conjuntos elétricos são considerados agregações de clientes, aprovados pela ANEEL, e pertencente a uma mesma área de concessão ou permissão, sendo estes definidos por Subestação de distribuição (SED). Em casos de redes MT das distribuidoras que não possuam subestação com o primário em AT o conjunto é composto pelas redes em MT de sua propriedade até o ponto de conexão com o agente supridor.

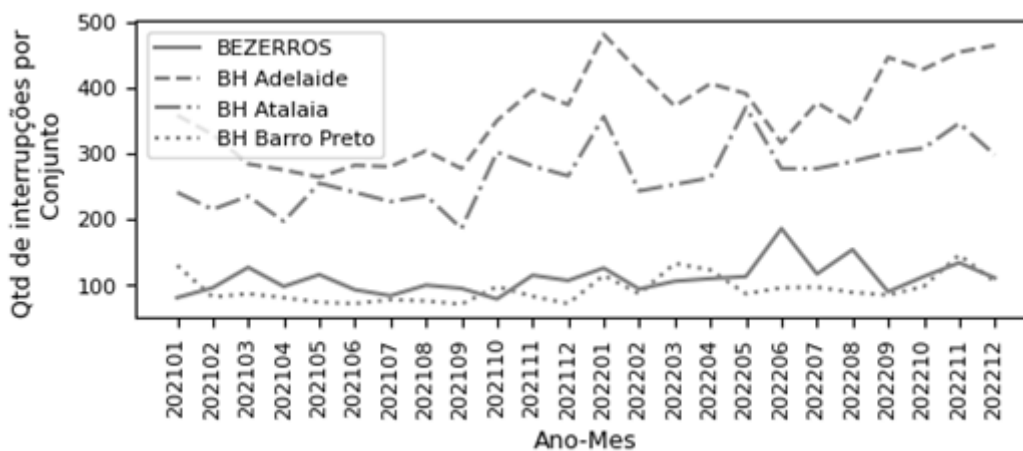
Nesta dissertação adotar-se-á o conceito de áreas geoeletricas como os conjuntos elétricos de unidades consumidoras devido à disponibilidade dos dados de interrupções e ao acompanhamento dos indicadores de qualidade DEC E FEC, por parte da ANEEL, nesta agregação. A ANEEL disponibiliza publicamente um conjunto de dados que relaciona os conjuntos de unidades consumidoras aos municípios da federação. Porém, como essa relação não é biunívoca (ou seja, para alguns municípios existe mais de um conjunto de unidades consumidoras em sua região geográfica e, por outro lado, alguns conjuntos atendem mais de um município), é mantida a agregação que é disponibilizada nos dados de interrupções, ou seja, por conjunto. A figura 1 apresenta o mapa do Brasil contendo as divisões territoriais por conjuntos elétricos que são consideradas pelo sistema SIGEL da ANEEL.



**Figura 1 - Divisão do mapa do Brasil por conjuntos de unidades consumidoras a partir dos dados de informações geográficas consideradas pelo sistema SIGEL – ANEEL - 2023 [65]**

A Figura 2 apresenta a evolução mensal, no período de 2021 a 2022, da quantidade de interrupções mensais em quatro conjuntos de unidades consumidoras presentes no mapa da Figura 1, denominados Bezerros, BH Adelaide, BH Atalaia e BH Barro Preto. Observa-se que há uma variação expressiva dependendo do conjunto de unidades consumidoras que é analisado. Neste exemplo da figura 2, os conjuntos de Bezerros e BH Barro Preto apresentam as suas quantidades de interrupções variando entre 0 e 200 eventos, no período que foi analisado. Nos conjuntos de BH Adelaide e BH Atalaia, por sua vez, nota-se que a quantidade de interrupções é na faixa de 200 a 500 eventos. Além dos dados de interrupções

no fornecimento, a ANEEL também disponibiliza no conjunto de dados informações referentes à duração de cada um desses eventos [66].



**Figura 2 - Histórico da quantidade de interrupções nos conjuntos Bezerros, BH Adelaide, BH Atalaia e BH Barro Preto no período compreendido entre janeiro de 2021 e dezembro de 2022. Informações retiradas do site da ANEEL 2023. [66]**

A evolução das interrupções, dos procedimentos de restauração de fornecimento e da regulação no Brasil se relacionam intimamente com a evolução do Sistema Elétrico Brasileiro (SEB). De acordo com [64], a região Sul do país, até meados de 1970, era dependente do abastecimento da região Sudeste por uma rede de 230 kV. Essa configuração causava um elevado número de distúrbios severos, levando ao desenvolvimento de um esquema de alívio de carga para prevenir o colapso do sistema. Enquanto as distribuidoras na região Sul sentiam a necessidade de descentralização dos processos de reestauração de energia, outras distribuidoras no Brasil adotaram uma operação envolvendo processos mais centralizados. Furnas (uma companhia de transmissão e geração de energia responsável por grande parte do sistema de transmissão da região Sudeste e Centro Oeste do Brasil), foi uma exceção.

Em 1982, começaram-se a adotar procedimentos mais descentralizados. Três distúrbios de grande porte na região sudeste do Brasil causaram interrupções drásticas nos anos de 1984 e 1985. A maior parte das indústrias brasileiras, neste período, estava localizada na região Sudeste. O montante interrompido nesses eventos foi próximo de 10.000 MW. Em 1989, dessa forma, uma força tarefa foi criada para definir padrões e critérios de operações de rede que foram mais tarde implementados por todas as distribuidoras de energia no Sul, Sudeste e Centro-Oeste.



Após diversos blecautes, como os de 1999 e 2002, inicializados no sistema de 440kV, a ANEEL solicitou ao Operador Nacional do Sistema Elétrico (ONS) a realização da revisão de todos os processos de restauração do abastecimento em casos de interrupções [67]. Atualmente, os processos operacionais após um desligamento são de responsabilidade dos operadores e companhias de distribuição, sendo estes, orientados de forma legal pelos órgãos reguladores regionais e nacionais [68]. Dessa forma, as distribuidoras desempenham um papel importante na definição dos procedimentos de restabelecimento de energia em casos de interrupções. A Resolução Normativa N° 1000 da ANEEL, de dezembro de 2021 [69], define estes papéis, incluindo as normas para o atendimento a interrupções no sistema de distribuição.

Dado que uma falha em um sistema elétrico pode afetar a distribuição de energia e resultar em uma interrupção a qualquer momento, o desenho dos sistemas elétricos também evoluiu no decorrer dos anos, sendo construídos com redundância de componentes.

### **2.2.2 Planejamento da Força de Trabalho**

Além do aumento de desempenho das distribuidoras em relação aos indicadores de qualidade do fornecimento de energia, que, em níveis acima do limite estabelecido, podem gerar multas, outra aplicação da estimação da quantidade de interrupções é o planejamento da alocação da força de trabalho. Esta programação visa garantir a pronta disponibilidade de mão de obra para os serviços de operação e de manutenção da rede elétrica. A previsão de interrupções constitui um dos insumos necessários para projetar a demanda de serviços emergenciais e conseqüentemente a mão de obra necessária para atendê-los. Embora o planejamento da força de trabalho seja prática corrente nas empresas, nem todas elas fazem o planejamento considerando o histórico de dados disponível.

Perdas no fornecimento de energia, assim como discutido na seção 1.1, geram impacto em diferentes setores sociais. Além disto, de acordo com Liu C. [70], eventos extremos de interrupções podem causar múltiplas quedas do fornecimento elétrico em diversos pontos do sistema de potência. Devido a importância da continuidade do abastecimento de eletricidade, uma grande quantidade de pesquisas se propõem a investigar a melhor forma de realizar o planejamento da alocação de equipes de campo. Neste sentido, em 2013, Perrier *et al.* [71, 72] fez um extenso estudo acerca de métodos de previsão para logística da resposta a emergências em sistemas elétricos. Wan H. *et al.* [73] realizou um

levantamento do estado da arte no qual analisou cerca de 300 literaturas relevantes, destas sendo selecionadas 289, sobre o problema de previsão de demanda de força de trabalho. Ward [74] discorreu sobre algumas categorias de previsão de força de trabalho: entrada gerencial, razões históricas, processo analítico, métodos estatísticos e análise de cenários. Liu C. [70] modelou a incerteza de carga elétrica com um algoritmo de Monte Carlos, buscando considerar diferentes cenários, e realizou a aplicação de um método de Otimização por Enxame de Partículas modificado, baseado no comportamento da espécie Açor-Nortenho (*Northern Goshawk Optimization - NGO*), para resolver o melhor esquema de alocação de recursos humanos. Berg A. e Haubrich H.-J. [75] propuseram um método para alocação de recursos operacionais de sistemas elétricos buscando a redução dos custos associados as tarefas à serem realizadas. No estudo foram consideradas diferentes variáveis de incerteza, como por exemplo a quantidade de interrupções na rede elétrica. Safaei N. [76] desenvolveu uma ferramenta para planejamento da força de trabalho tanto interna, quanto externa à uma distribuidora localizada no Canada. O objetivo do estudo foi determinar a combinação ótima de recursos de forma a reduzir os custos operacionais e a duração dos eventos de interrupções para os clientes, sendo utilizado para isto, um modelo que combina um algoritmo de simulação aleatória de eventos e um sistema especialista que desempenha a alocação das equipes. Foi concluído pelo estudo, além de outras descobertas, que de junho a setembro, a simulação evidenciou um aumento da quantidade de interrupções, e que eventos meteorológicos específicos levam ao aumento das quedas do fornecimento de energia, em determinados meses específicos. M.-J. Yao e K. J. Min [77] desenvolveram um modelo para resolver o problema de despacho ótimo de equipes de campo para restauração do serviço de energia. Eles desenvolveram um método que, após a estimação da quantidade de interrupções, define o melhor despacho das equipes de campo e a melhor estratégia para o planejamento de curto e longo prazo de alocação das equipes. J. S. Wu *et al.* [78], avaliando um cenário mais específico, construiu um modelo para alocação de equipes de atendimento à serviços emergências levando em consideração casos de múltiplas falhas, devido a eventos meteorológicos extremos. Para isto, ele utilizou uma modelagem baseada em lógica fuzzy, definindo a prioridade de envio das equipes de campo para a restauração da energia. Langevin A. *et al.* [79] fez um levantamento de diferentes técnicas de otimização do atendimento à serviços no sistema de potência, levando em consideração características distintas a serem otimizadas, como por exemplo o intervalo de tempo que os clientes ficam sem energia, os custos de envio das equipes, o custo do material a ser utilizado para a

realização do serviço, o custo do transporte, dentre outras. Duffuaa, S.O. *et al.* [80] empreendeu a modelagem do planejamento e agendamento da força de trabalho em uma planta da Arábia Saudita. Choudhari S. e Gajjar H. [81], de forma semelhante desenvolveram uma metodologia baseada na probabilidade de eventos de interrupções na rede e, após, executaram um algoritmo para otimização do serviço de manutenção nos possíveis casos esperados. Por sua vez, Borba *et al.* [18] realizou um extenso estudo acerca de métodos de otimização para o planejamento da força de trabalho em companhias elétricas de distribuição de energia.

A maior parte desses estudos teve como uma das informações principais a quantidade de interrupções ocorridas na rede elétrica. Sendo este um insumo fundamental para que sejam definidos os planos de envio das equipes e realização do reparo necessário.

## **2.3 Causas de Interrupções na Rede de Elétrica**

Levando em consideração que os eventos de interrupções não programadas no sistema elétrico são casos não planejados de corte no fornecimento de energia, é importante avaliar os agentes externos que levam a suas ocorrências. Dessa forma, ao relacionar esses fatores aos casos de interrupções, é possível posteriormente, tendo embasamento nos métodos discutidos nas seções 2.1, aplicar uma inferência estatística para projeção de casos futuros de interrupções, utilizando este conhecimento tanto para os casos que foram abordados na seção 2.2, quanto para outros de importância para as concessionárias.

Dessa forma, o estudo dos atributos que podem causar as paradas não planejadas no fornecimento de energia é de suma importância. Diversos trabalhos investigaram essa relação propondo análises e modelos para abordá-la e quantificá-la de maneira matemática [99-103]. Os influenciadores apresentados nas seções subsequentes desempenham um papel importante no estudo da quantidade de interrupções, dado que são as principais causas abordadas em pesquisas, e que levam a estas ocorrências.

### **2.3.1 Condições Climáticas**

Primeiramente, condições climáticas desempenham um papel importante na análise de defeitos e necessidade de serviços em sistemas elétricos, sendo estas assim, um dos principais temas abordados nos últimos anos. Diversas distribuidoras no mundo têm

construído modelos para compreender e quantificar o impacto meteorológico na operação das suas redes, em especial nos casos de interrupções do fornecimento de energia [82].

Além do impacto em interrupções no sistema elétrico, as modificações no clima também têm sido uma das maiores forças motoras para a modernização da infraestrutura do sistema elétrico, com a inclusão de sistemas com maior confiabilidade [83]. Mukherjee *et al.* [84] desenvolveu um modelo quantitativo para análise de risco de incidentes na rede elétrica. Ele estimou que aproximadamente 53% dos casos que foram observados aconteceram devido a condições climáticas.

Geralmente, nos estudos realizados para a concepção da infraestrutura dos sistemas, geralmente são utilizados vários anos das medições de condições meteorológicas. Dessa forma, é buscado levar em consideração a tolerância a possíveis eventos extremos dentro do padrão histórico geral. Isso pode ser um problema pois, como evidenciado por Burillo em [82], primeiramente, as condições climáticas apresentam variações sazonais dependendo dos anos e em segundo lugar o clima como um todo está se modificando.

Graças ao avanço na modelagem climática global, pesquisadores agora são capazes de prever as condições futuras, embasando o planejamento para eventos em condições extremas com maior confiabilidade do que em épocas anteriores. Todavia, a extração de estimativas quantitativas que representem a taxa de variação das condições meteorológicas, e que possam ser utilizadas com confiabilidade, continua um desafio, dado que um evento com impacto maior do que os observados infelizmente tem a possibilidade de ocorrer.

Dias mais com temperaturas elevadas podem ocasionar aumento no pico de demanda, estressando os componentes do sistema. Por exemplo, transformadores projetados considerando a redução da temperatura à noite podem ser incapazes de manter-se em sobrecarga em noites mais quentes [85]. O efeito direto do aumento da temperatura média no sistema de transmissão e distribuição é limitar ou reduzir a máxima taxa de dissipação de energia dos equipamentos e aumentar as perdas. Além disso, variações de temperatura podem modificar os parâmetros da linha, como por exemplo a resistência, reatância e admitância, alterando o estado de operação do sistema [86].

Ventos de altas velocidades podem resultar em interrupções nas linhas, causadas por detritos sendo soprados contra os cabos e estruturas da rede elétrica, árvores sendo levadas às linhas de energia elétrica e, também em casos extremos, quedas de postes e torres [87]. Em uma análise realizada na América do Norte por Hines *et al.* [88], aproximadamente

metade dos eventos de maior perda de fornecimento tiveram como causa principal ventos, tempestades, furacões ou tornados.

Gases ionizados produzidos por raios perto das linhas de transmissão podem causar curtos-circuitos, desconectando a linha de energia devido atuação do sistema de proteção. Geralmente essas faltas podem ser consideradas como temporárias, sendo extintas com o desligamento e reconexão do sistema. Assim o serviço pode ser recomposto rapidamente. Porém, a o surto de tensão que surge com o raio pode propagar-se pela rede e causar danos a equipamentos, como transformadores, por exemplo. Em 2006, nos Estados Unidos, o EPRI estimou que as despesas com reparo de equipamentos foram aproximadamente 1 bilhão de dólares por ano [89].

Chuvas fortes geralmente são associadas com fortes ventos e raios, os quais são maiores causadores de faltas do que a própria chuva. Chuvas fortes podem causar curtos-circuitos em isoladores. Felizmente, mudanças no projeto de isoladores passaram a reduzir esse risco [90]. A manutenção preditiva também pode minimizar o risco de defeitos nestes equipamentos [87].

Os sistemas de transmissão e distribuição geralmente não usam água para resfriamento, portanto, não seriam afetados, por exemplo, por uma seca prolongada. No entanto, um longo período seco pode causar o ressecamento do solo, o que reduz a condutividade térmica deste, reduzindo a capacidade de condução de carga de cabos subterrâneos em mais de 50% em relação a capacidade nominal [91]. Para cabos em baixa tensão esse efeito é ainda mais pronunciado, sendo a sua capacidade de condução de corrente nominal maior do que cabos em média tensão. A seca do subsolo também pode reduzir a condutividade elétrica, o que pode levar a necessidade de mudanças no sistema de aterramento.

Um efeito mais significativo da falta de chuvas é a seca da vegetação que pode induzir incêndios. Nesse contexto, linhas aéreas podem iniciar incêndios de grandes proporções durante o período de seca caso a vegetação entre em contato com os condutores [92, 93]. A fumaça do fogo pode causar a ruptura dielétrica do ar, ionizando-o, e produzindo por fim arcos elétricos nas redondezas dos condutores.

Dado a importância das condições climáticas para diversos fenômenos, como o de interrupção no fornecimento de energia, os dados relacionados aos fenômenos meteorológicos são acompanhados, no Brasil, pelo Instituto Nacional de Meteorologia

(INMET), sendo essas informações adquiridas a partir da mensuração aferida por unidades de medição espalhadas por regiões no território nacional.

### **2.3.2 Áreas Litorâneas**

Além das condições meteorológicas, alguns fatores específicos existentes em determinadas localidades também influenciam nas interrupções, dado que impactam os equipamentos da rede, podendo levar a falhas. Os cabos aéreos isolados com polietileno reticulado, por exemplo, substituíram os condutores de cobre convencionais sendo resistentes ao furto de eletricidade, oferecendo menores custos de manutenção e reduzindo as perdas no condutor. Entretanto, em áreas costeiras, cabos aéreos agrupados — *Aerial Bundled Cables* (ABCs) — sofrem frequentemente com quebra de isolação, causando interrupções inesperadas no fornecimento de energia [94]. Os cabos ABC são reportados tendo uma vida útil mais curta quando expostos a condições ambientais severas em áreas costeiras [95, 96]. Um dos principais efeitos que ocorre nestes cabos, devido principalmente à corrosão dos isoladores, é o efeito Corona [97].

A corrosão em áreas costeiras é um tópico relevante, sendo este abordado em diferentes estudos [98-103]. Em 2007 Ambler e Bain [104] conduziram um dos primeiros estudos rigorosos sobre a salinidade da atmosfera marítima e a corrosão metálica na Nigéria. O Brasil apresenta 279 municípios (5% do total nacional) litorâneos, em 17 estados. Os efeitos da maresia nestas regiões tem um grande impacto, sendo que, o desgaste das estruturas do sistemas elétricos pode levar a ruptura de cabos condutores de energia ou a queda de postes. O Instituto Brasileiro de Geografia e Estatística (IBGE), sendo o órgão brasileiro que fornece diversas informações geográficas nacionais, disponibiliza um conjunto de dados relativos aos municípios costeiros (sendo estes defronte ao mar) no Brasil.

### **2.3.3 Vegetação e Áreas Arborizadas**

Outra causa importante de eventos de interrupções na rede elétrica, que está relacionada com padrões regionais específicos, é a vegetação. Devido a presença de árvores e galhos próximos a rede, pode haver o contato entre estes elementos naturais com a linha elétrica por diferentes motivos. Esse contato, por sua vez, pode levar a um curto-circuito,

dado que o objeto funciona como um condutor de eletricidade para a terra, contendo baixa resistência elétrica em relação aos cabos da rede.

O IBGE disponibiliza um conjunto de dados que apresenta a classificação dos tipos de vegetação encontrados no território brasileiro. A tipologia, por sua vez, é definida a partir de critérios fisionômicos-ecológicos que são definidos dentro de uma hierarquia de formações e esquematizados segundo uma chave de classificação iniciada a partir de duas grandes classes: florestal e campestre.

Além destes dados, o Instituto Nacional de Pesquisas Espaciais (INPE) disponibiliza uma base de informações que sistematiza, a partir de diferentes fontes, dados que foram organizados para o exercício de modelagens em diversas aplicações. Essa base de dados, além de outras características, apresenta informações sobre a porcentagem de cobertura arbórea no Brasil por ponto geográfico (latitude e longitude). Dessa forma, é possível avaliar, por suas coordenadas, regiões que tem maior ou menor cobertura de vegetação. Assim, quanto maior a porcentagem coberta, maior a probabilidade de contato com a rede elétrica. Além disso, estes dados também podem servir de apoio a equipes de manutenção para a poda de árvores que estão próximas da rede elétrica. Estas informações podem ser consultadas no conjunto gerado pelo sensor Espectrorradiômetro de Imagem de Resolução Moderada (*MODerate-resolution Imaging Spectroradiometer – MODIS*), que está a bordo do satélite Terra da Administração Nacional da Aeronáutica e Espaço dos Estados Unidos (*National Aeronautics and Space Administration – NASA*) e tem 50 metros de resolução [105]. O conjunto de informações é formado pela latitude e longitude de diversas regiões do Brasil e a porcentagem de área arborizada nestas mesmas regiões.

### **2.3.4 Características Socioeconômicas**

A população e os índices de atividade econômica, como o Produto Interno Bruto (PIB), são dados relevantes para os eventos de interrupção. Estes, diferentemente de questões como a proximidade de áreas costeiras e a vegetação, são fatores mais relacionados à interação e distribuição dos sistemas produtivos e de consumo do território sendo, portanto, indiretos. A maior quantidade de unidades consumidoras também implica em maior demanda de energia e a consequente necessidade de uma rede mais robusta. Analogamente, uma atividade industrial mais intensiva utiliza mais energia elétrica. Além de demanda elevada de energia elétrica, essas regiões têm uma maior criticidade em casos de interrupções.

### **2.3.5 Características Geográficas**

Dado a sua extensão territorial, quando abordado o tema de características regionais é importante levar em consideração que o Brasil apresenta uma estrutura geográfica bem diversificada compreendendo uma área de 8.510.000 km<sup>2</sup>. O IBGE acompanha e disponibiliza além de informações referentes a características socioeconômicas, dados relativos à geografia regional. Exemplos destes dados são as classificações em microrregiões e macrorregiões.

Essas estruturas podem evidenciar a presença de elementos do quadro natural ou de relações econômicas particulares de relevância para a análise da quantidade de interrupções. Por exemplo, a região metropolitana de uma determinada região pode apresentar uma quantidade maior de unidades consumidoras do que uma região rural. De outra forma, uma região do agreste pode evidenciar uma presença maior de vegetação em relação a uma região do sertão. As divisões podem resultar em características sociais, geográficas e a articulação espacial, sendo constituídas por municípios limítrofes [106, 107].

Além destas características, as informações da divisão federativa, em municípios, distritos e subdistritos compõem todos os anos um relatório realizado pelo IBGE com a divisão oficial territorial. Os municípios foram estabelecidos pela constituição de 1988. Já os distritos são definidos como sendo unidades administrativas menores dentro de cada município, disciplinados pelas constituições estaduais que estabelecem critérios para criação, alteração ou supressão. Por fim, os subdistritos são unidades geográficas que dividem integralmente o território do distrito ou do município.

## **2.4 Reestabelecimento de Energia**

Dados os fatores que foram abordados na seção 2.3, o sistema de potência está sujeito à interrupções a qualquer momento. Estas intermissões no fornecimento elétrico, dependendo da criticidade, podem influenciar diretamente no esforço para ser realizada a restauração do serviço de abastecimento de energia.

Em 2012, o furacão Isaac atingiu o estado de Lousiana nos Estados Unidos, causando a perda do serviço de eletricidade para cerca de 1 milhão de consumidores, alguns por mais do que 10 dias. Este fenômeno, motivou diferentes questões acerca do restabelecimento energético, principalmente em questões de comunicação e conscientização das ocorrências



anteriormente, durante e após a sua eclosão e à preparação para casos de interrupções de grande porte [108].

A reestauração do sistema de potência, após casos de interrupção vai além da reconexão de unidades de geração e elementos de transmissão. Recompôr a rede elétrica é considerado um problema de otimização combinatória multi-objetivo, com variáveis e restrições não-lineares [109]. Sendo assim, o planejamento do processo utilizado nesta tarefa é um elemento crucial que determina o seu sucesso [110].

Esse processo, por sua vez, requer a determinação do status do sistema e dos equipamentos, a preparação da rede para o restabelecimento sistemático, a determinação da sequência de componentes a serem reenergizados, a reenergização dos componentes, e a interligação do sistema [111].

Dessa forma, na reconexão é buscado o retorno do abastecimento de energia considerando-se as condições e os padrões de funcionamento anteriores à interrupção. O objetivo, sendo assim, é o reestabelecimento garantindo-se a estabilidade do sistema. Se, por outro lado, este se tornar instável, pode gerar variações significantes nos parâmetros de tensão, ângulo e frequência. A estabilidade de tensão em um processo de restabelecimento está diretamente ligada à capacidade de serem recuperadas as tensões em todas as barras. Já a estabilidade angular está ligada à capacidade de que as máquinas síncronas resistam a um distúrbio e continuem em regime de operação.

No decorrer dos anos, diversos estudos foram propostos buscando a criação de processos de reconexão de forma automática, minimizando casos de instabilidade [112]. Dessa forma também, casos de ocorrências de erros humanos são mistigados. A projeção da quantidade de interrupções pode ser utilizada como insumo de forma a possibilitar a simulação de diferentes cenários na rede e a preparação para a execução do reestabelecimento de energia de forma mais eficiente. Uma vez esse processo sendo otimizado, o indicador de DEC é reduzido, dado que o melhor desempenho nas etapas que são necessárias na reconexão levam à minimização do tempo desta tarefa, e portanto, a duração das interrupções também é menor.

## Capítulo 3 - Conceituação Teórica

Neste capítulo, são apresentados os principais referenciais teóricos dos métodos utilizados neste estudo para resolução das tarefas propostas (Cap. 1) e discutidas (Cap. 2) nos capítulos anteriores. As seções que se seguem apresentam a base teórica da metodologia que será apresentada no capítulo 4, servindo como guias para a explicação conceitual das técnicas utilizadas. Serão descritos, primeiramente, os modelos de previsão que foram utilizados para a projeção da quantidade de interrupções nos conjuntos de unidades consumidoras. É apresentado, dessa forma, o algoritmo de previsão simples sazonal simples ou ingênuo (Sec 3.1), adotado como padrão de referência e em seguida (Sec 3.2) o modelo de regressão múltipla, empregado neste trabalho como o método de previsão proposto. A seguir serão descritos a qualidade do ajuste dos modelos de previsão, sendo também discutidas as métricas de avaliação da aderência dos modelos e comparação entre os previsores (Sec. 3.3). Na seções 3.4 a 3.6, serão discutidos os testes de hipóteses T e F empregados no trabalho. Estes testes são utilizados para validação da relevância estatística do modelo proposto, assim como, de cada uma das características de entrada utilizadas. Por fim, as últimas seções do capítulo tratarão das técnicas de clusterização utilizadas para agrupar as informações de intervalo de tempo e quantidade de interrupções. Os métodos *k*-médias (*k-means*) (Sec 3.5.1) e o método *Elbow* (Sec 3.5.2) serão abordados.

### 3.1 Método de Previsão Sazonal Simples

Considere uma série temporal  $y_1, \dots, y_t$ . O método de previsão simples ou ingênuo (do inglês *naive*) caracteriza-se por considerar que o valor previsto  $\hat{y}_t$  é dado pelo último valor observado  $y_{t-1}$ . O método de previsão simples sazonal incorpora o conhecimento sobre as sazonalidades envolvidas, de forma que as observações *a posteriori* são equivalentes às observações passadas, considerando um ou mais períodos de sazonalidade (por exemplo, hora do dia e dia da semana). Por ser uma abordagem de modelagem simples é aplicável em diferentes cenários, principalmente em finanças. Matematicamente, o método de previsão simples sazonal pode ser descrito da forma que segue:

$$\hat{y}_t = y_{t-m} \quad (1)$$

onde  $m$  é o período de sazonalidade.

## 3.2 Regressão Linear Múltipla

A regressão linear múltipla é uma técnica amplamente utilizada em problemas de aprendizagem [113]. Dado uma observação na forma  $(x_{i1}, \dots, x_{iN}, y_i)$ , onde  $x_{i1}, \dots, x_{iN}$  são as entradas do modelo e  $y_i$  a saída, o modelo de regressão múltipla é dado por:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_N x_{iN} + \varepsilon_i, \quad i=1, 2, \dots, n \quad (2)$$

onde  $\beta_0, \dots, \beta_N$  e  $\varepsilon_i$  são respectivamente os parâmetros e o erro do modelo.

O modelo descrito pela expressão (2) possui as seguintes premissas:

1. Os erros  $\varepsilon_i$  são mutualmente independentes, identicamente distribuídos e possuem valor esperado nulo, isto é  $E[\varepsilon_i]=0$ ;  $Var[\varepsilon_i]=\sigma_\varepsilon^2$ , onde  $E[X]$  e  $Var[X]$  são respectivamente o valor esperado e a variância da variável aleatória  $X$  [128];
2. A variável aleatória  $\varepsilon_i$  independe estatisticamente do vetor aleatório  $X_{i1}, \dots, X_{iN}$ ;

Considere uma amostra  $\{(x_{11}, \dots, x_{1N}, y_1), (x_{21}, \dots, x_{2N}, y_2), \dots, (x_{n1}, \dots, x_{nN}, y_n)\}$ .

Aplicando a expressão (2) a cada tupla na amostra resulta no sistema de  $n$  equações a seguir:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \dots + \beta_N x_{1N} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \dots + \beta_N x_{2N} + \varepsilon_2 \\ &\cdot \\ &\cdot \\ &\cdot \\ y_n &= \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \dots + \beta_N x_{nN} + \varepsilon_n \end{aligned} \quad (3)$$

Usando a notação matricial, esse conjunto de expressões pode ser escrito compactamente:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (4)$$

onde

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{nx1} ; \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1N} \\ 1 & x_{21} & x_{22} & \dots & x_{2N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nN} \end{bmatrix}_{nx(N+1)} ;$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_N \end{bmatrix}_{Nx1} \quad \text{e} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon \\ \varepsilon \\ \vdots \\ \varepsilon_n \end{bmatrix}_{nx1} ;$$

Uma vez determinado os parâmetros do modelo (4), a saída prevista é dada por:

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_N x_{iN} \quad (5)$$

### 3.2.1 Estimação por Mínimos Quadrados

A forma usual de obtenção dos parâmetros de um modelo de regressão linear é utilizando o critério dos mínimos quadrados. De acordo com ele, a melhor estimativa do vetor de parâmetros  $\hat{\boldsymbol{\beta}}$  é aquela que soluciona o seguinte problema de otimização:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = [\mathbf{y} - \mathbf{X}\boldsymbol{\beta}]^T [\mathbf{y} - \mathbf{X}\boldsymbol{\beta}] \quad (6)$$

O problema (6) é usualmente solucionado utilizando-se as equações normais [114], sendo o vetor de parâmetros (coeficientes)  $\boldsymbol{\beta}$  solução da equação normal proposta, em forma matricial, como:

$$[\mathbf{X}^T \mathbf{X}] \hat{\boldsymbol{\beta}} = \mathbf{X}^T \mathbf{y} \quad (7)$$

### 3.3 Métricas de Qualidade do Ajuste e Avaliação da Qualidade da Previsão

Uma vez construídos os modelos de previsão, existem diferentes técnicas para avaliar a qualidade do ajuste e posteriormente dos resultados gerados. Dependendo da indústria, da natureza do problema de previsão e de preferências pessoais do desenvolvedor, diferentes métodos podem ser aplicados. Os métodos variam em complexidade de cálculo e alguns são combinações de outras medidas básicas. Em relação à previsão, não há um consenso sobre a melhor métrica a ser utilizada para a mensuração do erro e da adequação do ajuste [115]. Levando em consideração esse contexto, são apresentadas algumas métricas de avaliação que podem ser utilizadas e que posteriormente serão aplicadas no teste do modelo proposto nessa dissertação.

#### 3.3.1 Análise dos Resíduos

Uma forma simples de avaliar a qualidade do modelo ajustado é analisar os resíduos de estimação. Considere a saída observada  $y_t$  e a saída prevista  $\hat{y}_t$ , calculada pelo modelo. O resíduo de estimação é dado por:

$$e_t = y_t - \hat{y}_t \quad (8)$$

onde  $e_t$  o resíduo ou erro do modelo.

Para um modelo adequadamente ajustado espera-se que os resíduos não sejam correlacionados e tampouco tenham valor esperado nulo. A segunda condição indica que o modelo não é enviesado, enquanto a primeira sinaliza que as saídas previstas são independentes. Se o modelo de previsão possuir as características listadas, isto sugere que o mesmo ou não capturou toda dinâmica presente nos dados, ou os dados não atendem às premissas do modelo, sendo necessário, portanto, a escolha de outros modelos ou processamento da base de dados.

### 3.3.2 Erro Percentual Absoluto Médio (MAPE)

Diferentes métricas para avaliação das previsões de modelos matemáticos podem ser construídas a partir dos resíduos. As diferenças entre elas estão relacionadas basicamente às necessidades distintas de avaliação do resultado. Por exemplo, algumas evidenciam o erro de forma percentual (como o erro percentual médio absoluto –MAPE), já outras o erro absoluto (como o erro médio absoluto – MAE) [116], ou mesmo o desvio dos erros (como o MDE) [117]. O erro MAPE, por exemplo, é uma métrica frequentemente utilizada para comparar o desempenho da previsão. Apesar de sua ampla utilização em problemas de previsão há diferentes críticas em relação ao uso do MAPE, a saber:

1. O MAPE é assimétrico e não padronizado: um ajuste perfeito, ou seja, todos os valores projetados são iguais aos valores observados, resulta em um MAPE igual a zero. No entanto, não há restrições para o seu limite superior (quanto maior a distância entre os valores projetados e o real, maior também será o valor do MAPE).
2. Para uma observação que tem o valor de zero, a fórmula do MAPE resulta em uma divisão por zero, causando inconsistência numérica.
3. O MAPE pode ser enviesado por valores extremos (*outliers*): neste caso o valor do MAPE é influenciado por esse ponto em relação aos demais registros da amostra. (Medidas de erro que avaliem a dispersão dos resíduos podem ser utilizadas em conjunto com o MAPE para análise de valores extremos. Uma análise dos resíduos é também frequentemente utilizada em problemas de regressão para avaliar casos que influenciem o MAPE [Referência])

Em contrapartida, as vantagens de se utilizar o MAPE são a sua fácil interpretação e a sua larga utilização, principalmente em problemas de negócios [118]. O MAPE, então, é calculado pela média das diferenças percentuais entre as observações realizadas e a previsão do modelo.

$$MAPE = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i} \cdot 100 \quad (9)$$

onde  $n$  é o número de observações.

### 3.4.2 Coeficiente de Determinação ( $R^2$ )

A qualidade do ajuste do modelo também é usualmente avaliada através do coeficiente de determinação  $R^2$ . Este indicador quantifica o grau de explicabilidade dos valores previstos em relação à dispersão dos valores observados [119]. O coeficiente de determinação  $R^2$  pode ser definido da forma:

$$R^2 = \frac{SQE}{SQT} \quad (10)$$

onde SQE é a soma dos quadrados explicados e SQT a soma dos quadrados totais.

De acordo Stock e Watson [114], define-se a soma dos quadrados explicados (SQE) pela expressão:

$$SQE = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad (11)$$

onde  $\bar{y}$  é a média das saídas observadas.

A soma dos quadrados totais (SQT) pode ser expressa como:

$$SQT = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (12)$$

Alternativamente, o coeficiente de determinação pode ser descrito em termos de uma fração da variância de  $y_i$ , que não é explicada por  $X$ , dessa forma:

$$R^2 = 1 - \frac{SQR}{SQT} \quad (13)$$

Sendo SQR a soma dos quadrados dos resíduos, dada por  $\sum_{i=1}^n e_i^2$ .

A expressão (13) significa que o coeficiente de determinação  $R^2$  exprime o quanto o modelo explica a variação observada de  $y$ . Se a relação entre o somatório dos quadrados dos resíduos (SQR) e o somatório dos resíduos totais (SQT) é perto de 1 (e, portanto, o  $R^2$ , próximo de 0) assume-se que os valores previstos pelo modelo não conseguem explicar bem a variação observada em  $Y$ . Seria desejável que o valor de  $R^2$  fosse maior. Alternativamente, se a relação entre SQR e SQT é pequena, então o ajuste está bom e o  $R^2$  é grande.

Tendo isto exposto, o valor de  $R^2$  perfeito, no cenário em que todas as observações estão ajustadas, seria  $Y_i = \hat{Y}$  e  $e_i = 0$ . Isso significa que o somatório do quadrado dos erros é igual a zero e que  $R^2=1$ . Porém, este caso não é interessante para a projeção, dado que há uma variabilidade intrínseca dos dados e o modelo estaria estritamente ajustado ao conjunto em que foi treinado. No outro caso extremo, em que o somatório dos quadrados da regressão é igual a zero, quer dizer que a previsão explica em nada a variação em  $Y$  e, nesse caso,  $R^2=0$ .

### 3.4.2 Erro Médio Quadrático (MSE)

Outra métrica frequentemente utilizada na avaliação de estimadores em problemas de regressão é o erro médio quadrático (MSE), dada por [120]:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (14)$$

Pode-se notar que o MSE é a norma  $\ell_2$  (isto é, a norma quadrática) do resíduo de estimação. Por conseguinte, o MSE é uma métrica sensível a registros extremos (outliers) [121]. Um estimador ideal teria MSE igual a zero, sendo todas as previsões iguais aos valores observados. Esse ajuste, assim como já discutido (sec. 3.3.3), não é interessante em termos de projeção, sendo que, para conjuntos de dados diferentes do que modelo foi estimado o resultado da previsão seria insatisfatório pois os parâmetros do modelo estariam demasiadamente ajustados à base em que foi criado.



## 3.4 Teste de Hipótese

O teste de hipótese é uma ferramenta estatística conhecida que fornece uma metodologia objetiva para tomar decisões utilizando um conjunto de regras (métodos probabilísticos). É comum se chegar a diferentes conclusões analisando e investigando os dados, porém o teste de hipótese é um método estatístico de tomada da decisão.

Esse método utiliza a probabilidade de que determinada premissa formulada não possua evidência estatística que permita a sua negação. A hipótese de que a premissa avaliada não pode ser aceita como verdadeira estatisticamente, é chamada de hipótese nula, sendo sua contrária chamada de hipótese alternativa. A decisão para rejeitar a hipótese nula e, portanto, aceitar a validade da hipótese alternativa leva em consideração a seguinte regra: se o valor da estatística de teste avaliada é menor do que o limiar de rejeição considerando um determinado nível de significância, se rejeita a hipótese nula e, portanto, se aceita a hipótese alternativa. O nível de significância ( $\alpha$ ) é a probabilidade de rejeição da hipótese nula, quando ela é verdadeira. Este, em grande maioria das aplicações, é considerado como tendo valor de 0,05. O valor de  $p$  (*p-value*), por sua vez, é a probabilidade de se obter um efeito pelos menos tão extremo quanto aquele em seus dados amostrais, assumindo-se que a hipótese nula seja verdadeira. Quando o *p-value* é menor ou igual ao nível de significância, deve-se rejeitar a hipótese nula.

### 3.4.1 Teste F

A estatística  $F$  é uma métrica que é frequentemente utilizada em testes de hipóteses para avaliação de modelos, sendo que pelo cálculo do seu valor é possível avaliar a significância do modelo de regressão múltipla que foi estimado. Considerando a soma dos quadrados explicados (SQE), definido na equação (15), e a soma dos quadrados dos resíduos (SQR), o cálculo da estatística  $F$  pode ser realizado utilizando-se a seguinte equação:

$$R(f) = \frac{SQE}{SQR} = \frac{\sum_{i=1}^{n_i} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n_i} (y_i - \hat{y}_i)^2} \quad (15)$$

Em que  $n$  é quantidade de amostras. Nota-se que neste caso está sendo calculado o quanto pode ser explicado da variação de  $y$ , em relação ao erro do modelo.

Após o cálculo da estatística  $F$ , é então, realizado o seguinte teste de hipóteses:

- $H_0$  (hipótese nula): Não há diferença entre o modelo e a utilização de uma constante para a inferência, ou seja, todos os coeficientes são identicamente nulos.
- $H_A$  (hipótese alternativa): Há diferença estatisticamente significativa entre a aplicação do modelo e a inferência utilizando apenas uma constante. Ou seja, o desempenho o modelo estatisticamente significativo e superior na estimação do comportamento da variável resposta e ao menos um coeficiente é não-nulo.

Com o teste estatístico dado pela estatística  $F$ , o valor do limitante em que aceita-se a hipótese alternativa ( $p$ -value) é sempre obtido respondendo-se a seguinte pergunta: Qual o valor crítico na distribuição  $F$ , em que pode-se dizer que a ocorrência da hipótese alternativa é estatisticamente significativa. Ou seja, em que valor de  $F$ , levando em consideração a confiança esperada pela definição do valor de  $p$ , pode ser aceita a hipótese alternativa e se rejeitar a hipótese nula.

### 3.4.2 Teste T

Outra métrica utilizada comumente em problemas de regressão é a estatística  $T$ . A distribuição da estatística  $t$ , além de outras aplicações, é utilizada para realizar inferências sobre os coeficientes da regressão. Esta distribuição é simétrica em relação ao zero, com os valores da cauda à esquerda podendo ser obtidos dos correspondentes valores da cauda à direita [122].

Se  $X$  é uma variável aleatória de distribuição normal, e se  $\chi^2$  é uma variável aleatória com distribuição qui-quadrado com  $\nu$  graus de liberdade de  $X$ , então a variável aleatória constituída pela razão  $X / (\chi^2 / \nu)^{1/2}$  tem uma distribuição  $t$  de Student com  $\nu$  graus de liberdade. A distribuição  $t$  de Student é simétrica e, como a distribuição normal, tem média zero e índice de assimetria e curtose também nulos. A variância da distribuição é  $\nu/(\nu-2)$ , que converge para 1, assim que  $\nu \rightarrow \infty$ . Em amostras grandes, de outra forma, a distribuição  $t$  é

geralmente substituída por uma distribuição normal. A distribuição  $t$  para 10 graus de liberdade é comparada com a distribuição normal padrão.

A distribuição  $t$  é a distribuição amostral que é associada com a estatística de teste definida como:

$$t = \frac{\bar{x} - \mu}{\hat{\sigma}_{\bar{x}}} = \frac{\bar{x}}{\hat{\sigma}_{\bar{x}}} \quad (16)$$

em que é assumido na equação (16) que, a média da população é igual a zero e o desvio padrão da média estimado da população é igual a:

$$\hat{\sigma}_{\bar{x}} = \frac{s_{\bar{x}}}{\sqrt{N-1}} \quad (17)$$

Em que  $s$  é a variância amostral. A distribuição  $t$  é similar a distribuição  $z$ , podendo ser derivada desta pela aplicação de uma transformação, sendo  $z = \frac{\bar{x} - \mu}{\hat{\sigma}_{\bar{x}}}$ , e definindo a relação entre  $t$  e  $z$  por,  $t = z \sqrt{n - 1}$ . A única diferença entre estas é que a distribuição  $t$  é baseada no desvio padrão estimado da população, derivado do desvio padrão amostral, e a distribuição  $z$  é derivada do desvio padrão da população [123].

Para a aplicação particular da estatística  $t$  para o teste da significância estatística dos coeficientes de modelos de regressão, sobre um conjunto de premissas usualmente referenciadas como condições de Gauss-Markov, esse teste pode ser utilizado. O teste de significância, para este caso, é similar com aquele evidenciado para a média amostral. Começa-se assumindo uma hipótese inicial nula, ou seja, que o coeficiente a ser avaliado é igual a zero. Dado entendimento a priori do erro amostral, sabe-se que ao retirar diferentes amostras de uma população chega-se a um intervalo de valores de diferentes coeficientes de regressão calculados. É provado estatisticamente que os mínimos quadrados amostrais proveem uma estimativa não enviesada dos coeficientes de regressão [123].

Da mesma maneira que foi realizado no caso da média amostral, pode-se avaliar cada coeficiente do modelo de regressão em relação ao seu erro padrão amostral da forma:

$$t = \frac{b_{yx} - \beta_{yx}}{\hat{\sigma}_b} = \frac{b_{yx}}{\hat{\sigma}_b} \quad (18)$$

Neste teste, avalia-se para cada coeficiente do modelo de regressão  $b_{yx}$ , se o seu valor é semelhante a um determinado coeficiente definido,  $\beta_{yx}$ . Em que é assumido que os coeficientes  $\beta_{yx}$  são nulos. Neste caso, o teste  $t$  é simplesmente a razão entre o coeficiente de regressão e o erro padrão deste. Em repetidas amostras de tamanho  $N$  da população sobre a hipótese nula, este teste estatístico segue uma distribuição  $t$ , com  $N-2$  graus de liberdade. Estes, neste caso, são iguais a 2, porque 2 graus de liberdade são perdidos no cálculo dos parâmetros da equação de regressão: um para os coeficientes de regressão e outro para o intercepto.

## 3.5 Clusterização

### 3.5.1 K-médias (*K-means*)

Técnicas de agrupamento são outro ferramental comumente utilizado em modelagens matemáticas. Estas se caracterizam por agruparem os registros com características semelhantes, tendo como finalidade a descoberta de padrões de forma não supervisionada ou a estimação não paramétrica da distribuição dos dados. O *k-means* é um método de agrupamento utilizado para classificar os dados de acordo com métricas de distância ou mais genericamente, de similaridade entre estes [124]. Neste método, primeiramente se inicia a quantidade de clusters e se assume cada ponto pertencente a um desses clusters. Existem, na literatura, diferentes variantes deste algoritmo [125].

Assumindo  $x = \{x_1, x_2, \dots, x_n\}$  como uma conjunto de  $n$  objetos no  $R^n$  e um número  $k$  de partições  $C = (C_1, C_2, \dots, C_k)$ . O algoritmo *k-means* realiza a busca por uma partição de  $x$  em  $k$  clusters de forma a minimizar a função objetivo  $J$ :

$$\min_{C, Z} J(C, Z) = \sum_{i=1}^k \sum_{l=1}^k d(x_l, z_i)^2 \quad (19)$$

sendo  $Z = \{\bar{x}_{C_1}, \bar{x}_{C_2}, \dots, \bar{x}_{C_k}\}$  os centroides de cada cluster e  $d(l, z)^2$  uma medida de distância entre o objeto  $l$  e a classe  $z$  do  $l$ -ésimo cluster. O valor ótimo da função  $J(C, Z)$ ,

descrita pela equação (19), é normalmente resolvido realizando-se a atualização iterativa, calculando-se a cada etapa o novo centróide e os registros correspondentes a cada grupo. Esse método se repete até ser atingida a convergência. O seguinte algoritmo pode ser aplicado até ser alcançado o ótimo (local ou global) para a configuração  $(C, Z)$ :

Passo i: Escolher  $\{z_1^{(0)}, z_2^{(0)}, \dots, z_k^{(0)}\}$  aleatoriamente. Fazer  $C \leftarrow \{z_1^{(0)}, z_2^{(0)}, \dots, z_k^{(0)}\}$  e  $t \leftarrow 0$ ;

Passo ii: Obter  $C, Z = \underset{C, Z}{\operatorname{argmin}} J(C, Z)$  e fazer  $t \rightarrow t + 1$ ;

Passo iii: Se o critério de parada seja atendido, encerre. Senão, retornar ao Passo ii;

Existem, na literatura, diversas variações do método do *k-means* [126]. Muitos autores incluindo Diday [127, 128, 129] e Diday *et al.* [130] utilizaram o *k-means* generalizado (19) de várias formas. Estas diferenciam-se entre si pelas técnicas utilizadas para minimizar a função objetivo (distorção –  $J$ ). Por exemplo, utilizando a distância de Mahalanobis, distâncias  $Lq$ , ou Minkowski. Em particular, a distância  $L1$  é utilizada por Vinod [131], Massart *et al.* [132] e Spath [133].

### 3.5.1.1 Método do Cotovelo (*Elbow Method*)

Tendo-se a necessidade de escolha do valor de  $k$ , representado pelo número de clusters para o método *k-means*, o método do cotovelo (*Elbow*) é utilizado para definir a quantidade ótima de clusters a partir do cálculo da soma dos erros quadrados e da ideia de que, em um determinado ponto de valores avaliados de  $k$ , o aumento do número de clusters não aumentará significativamente a qualidade da informação. O processo para essa avaliação envolve a análise da porcentagem de variação explicada (ou distorção, que pode ser representada pela função de custo  $J$  da equação (17) como uma função do número de clusters. Os primeiros clusters vão adicionar muita informação, porém, em algum ponto o ganho marginal vai cair drasticamente e ocasionar um ângulo perceptível no gráfico semelhante ao formato de um cotovelo (*elbow*). Na medida que é iterado o valor de  $k$ , o número de observações em cada cluster diminui, desta forma, as distâncias intraclusters são reduzidas e, portanto, o índice  $J$  também diminui. Logo, o número de clusters é escolhido como este ponto. O método pode ser inicializado levando em consideração  $k=2$  (ou seja, dois clusters),

e iterado por um passo 1, calculando, em cada iteração, a distorção até o patamar em que é definido o melhor valor de  $k$  [134].

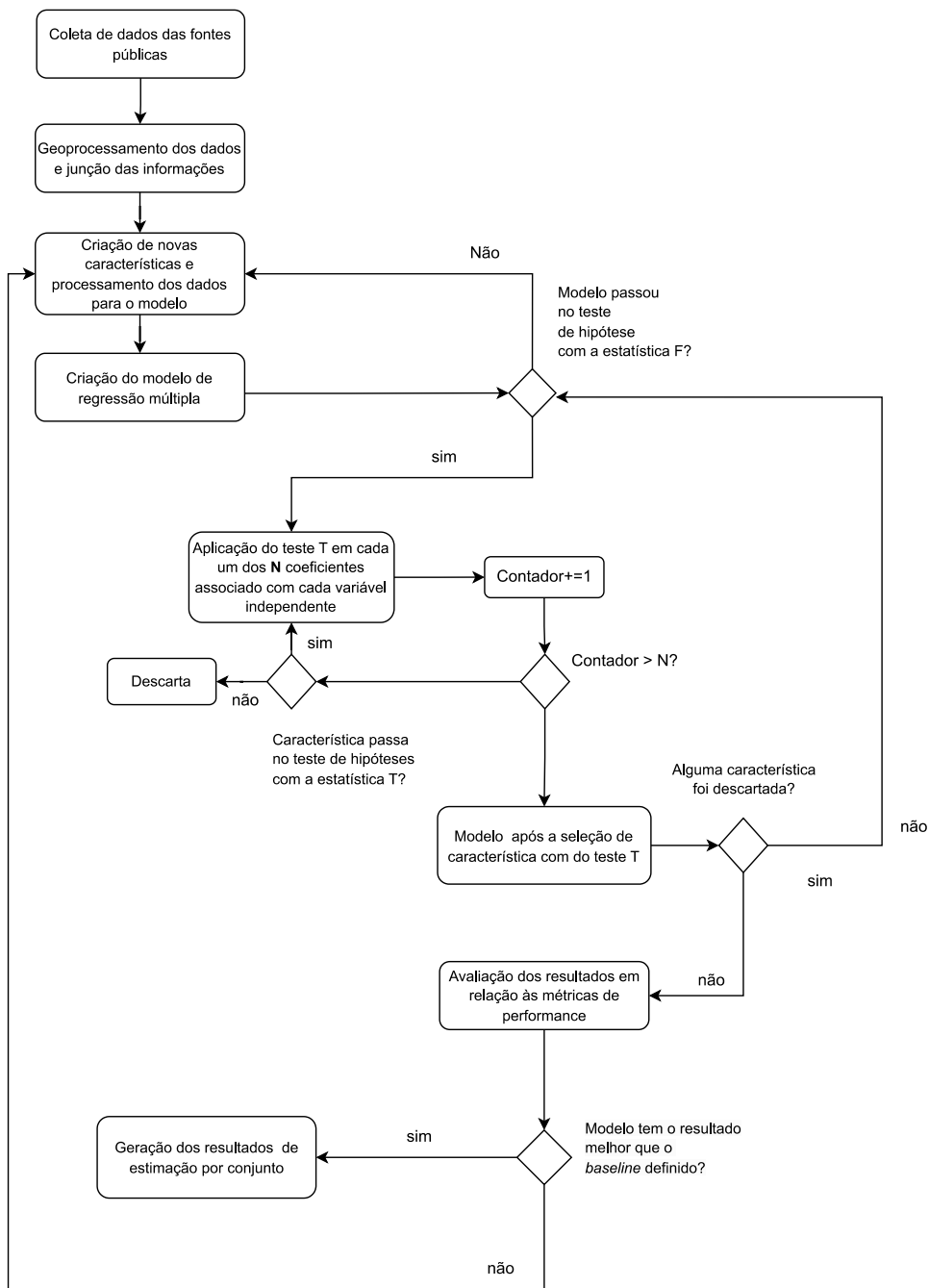
## Capítulo 4 - Metodologia

Após a introdução (Cap. 1), a revisão da literatura (Cap. 2) e a conceituação teórica das técnicas que foram utilizadas (Cap. 3), neste capítulo faz uma análise da metodologia proposta neste trabalho. A apresentação geral da metodologia proposta é realizada na seção 4.1, na qual também é estruturada em sub etapas, que são detalhadas nas seções subsequentes (4.2 a 4.7).

### 4.1 Formulação Geral da Metodologia Proposta

A Figura 3 apresenta um fluxograma com os principais passos da referida metodologia. Primeiramente, realiza-se um geoprocessamento das informações de forma a unificá-las em um conjunto de dados único, agregado por conjunto de unidades consumidoras. Além da divisão geográfica das informações, faz-se também uma agregação temporal por ano e mês. Dessa forma, ao final deste processamento, é obtido um conjunto de dados em que cada um dos registros (linhas da matriz de dados) contém as informações sobre um conjunto de unidades consumidoras em um ano e mês específicos do período que foi considerado (Sec. 4.2).

Tendo a base de dados sido unificada por suas informações geográficas e temporais, aplica-se um algoritmo de agrupamento em relação aos atributos de intervalo de tempo de cada interrupção e quantidade de interrupções (ambas as métricas agregadas por conjunto) criando-se assim um atributo para o conjunto de dados (Sec. 4.3). Esse processamento busca a produção de novas informações que apresentem relação com os indicadores de DEC e FEC acompanhados pela ANEEL, buscando melhorar assim, o desempenho do modelo.



**Figura 3 - Fluxograma geral da metodologia proposta**



Em seguida, analisa-se as relações entre cada característica na base de dados e as informações de interrupções, permitindo que sejam descritos os relacionamentos entre estas informações para o embasamento da criação do modelo de estimação (Sec. 4.4). Concluída a análise entre as informações sobre os eventos e os casos de interrupções, são então realizados os tratamentos como, a transformação dos valores categóricos em numéricos, substituição de valores ausentes por estatísticas do conjunto de dados e a padronização dos valores da matriz de dados final. Estas manipulações são necessárias para a aplicação do estimador. Os tratamentos são descritos de forma detalhada na seção 4.5. Nesta seção também é apresentado o modelo que foi criado, bem como a metodologia utilizada em sua aplicação.

Nas seções 4.6 e 4.7, apresenta-se uma descrição da aplicação da estatística  $F$  para avaliação do modelo, e da estatística  $T$ , para seleção de variáveis. Após a aplicação desses testes, compara-se o último resultado em relação ao método de referência, e então, são gerados os resultados projetados de interrupções por cada conjunto e mês. O modelo de comparação é construído levando em consideração um método de projeção sazonal simples, em que se é estimado que as mesmas quantidades de casos de interrupções mensais irão se repetir de um ano para o seguinte.

## 4.2 Geoprocessamento

Inicialmente, é necessário correlacionar as informações dos eventos causadores e dos casos de interrupções com a classificação em que os indicadores de qualidade do fornecimento de energia são acompanhados pela ANEEL. Dessa forma, todos os atributos devem ser agregados de forma a cada registro pertencer à um único conjunto de unidades consumidoras, em um mês específico do período analisado. Dado que esses atributos apresentam identificadores geográficos que podem ser relacionados às coordenadas de latitude e longitude, realiza-se um geoprocessamento no conjunto de dados, buscando unificar as informações por esses pontos referenciais. Esse geoprocessamento busca garantir que cada uma das características estudadas: meteorológicas, regionais e socioeconômicas, estejam relacionadas com cada um dos conjuntos de unidades consumidoras, representados por regiões do território brasileiro definidas pelo Sistema de Geolocalização da ANEEL (SIGEL).

Essas regiões nas quais se busca realizar a relação, por sua vez, não são idênticas. Em outras palavras, as regiões que definem geograficamente cada conjunto de unidades consumidoras, são diferentes dos territórios que representam, por exemplo, os municípios e as unidades de medição do INMET. Cada mensuração apresenta uma projeção diferente no sistema euclidiano de coordenadas geográficas. Sendo alguns atributos representados por pontos de coordenadas (*lat*, *long*) e outros representados por regiões da forma ( $v_1, v_2, \dots v_n$ ), em que *lat* e *long* e *v* são, respectivamente, as coordenadas de latitude, longitude e os vértices dos polígonos que descrevem as regiões espaciais. Sendo assim, é necessário ser realizado um tratamento nos dados e definidas regras para a unificação dessas informações.

Primeiramente em relação às informações meteorológicas do INMET e de porcentagem de cobertura arbórea do INPE (MODIS), sendo estas representadas como pontos no sistema euclidiano, verifica-se se as coordenadas de latitude e longitude de cada uma das unidades de medição do INMET (*lat*, *long*) pertencem à região geográfica (polígono) de cada um dos conjuntos de unidades consumidoras  $C=\{v_1, v_2, \dots v_n\}$ . Caso seja verificado o pertencimento, as medições destas unidades são consideradas como as medidas meteorológicas dos respectivos conjuntos. Nesta verificação caso o ponto tenha as suas coordenadas coincidindo com as coordenadas das bordas (areastas) das regiões que estão sendo avaliadas, não é considerado que a medida pertence ao conjunto. Desta forma, são apenas considerados casos que estão contidos na localidade interior às regiões.

Em alguns casos específicos é identificado que mais de uma unidade de medição se relaciona à um conjunto específico. Todavia, para fins de construção do conjunto de variáveis independentes, e visto que para alguns conjuntos somente uma medição é encontrada, essas informações devem ser agregadas a fim de construir um registro único por conjunto. Assim, são consideradas os valores máximo e médio para agregação por conjunto e por mês de todos os atributos climáticos. Para a variável relacionada com a precipitação acumulada, são calculados também os valores totais mensais desta característica.

Para os atributos que foram consultados do IBGE, procede-se um geoprocessamento similar ao que foi realizado com as mensurações pontuais. Em grande parte, essas informações apresentam a dimensão de municípios (exemplos de informação dessa natureza são os atributos socioeconômicos). Sendo assim, transformam-se as informações dos conjuntos de unidades consumidoras em pontos a partir do cálculo de seus centróides (*lat*, *long*) e então avalia-se se cada um dos pontos estão contidos nas regiões  $M=\{v_1, v_2, \dots v_n\}$  que representam os territórios municipais. Caso estejam contidos, as informações que estão

na agregação de município são consiradas como os atributos referentes ao conjunto que foi avaliado.

De forma geral, as regras para a definição do processamento em todos os casos pode ser generalizada com o seguinte algoritmo:

Passo i: Definir o vetor  $A=\{v_{a1}, v_{a2}, \dots, v_{an}\}$ , que representa o polígono em que se pretende avaliar o pertencimento. Definir o vetor  $B=\{v_{b1}, v_{b2}, \dots, v_{bn}\}$  do polígono que se pretende avaliar se está contido em  $A$ , sendo  $v_{an}$  e  $v_{bn}$  os  $n$  vértices dos polígonos;

Passo ii: Caso o vetor  $B$  não conter apenas um elemento, transformar o vetor  $B$  para que apenas possua 1 elemento (um ponto);

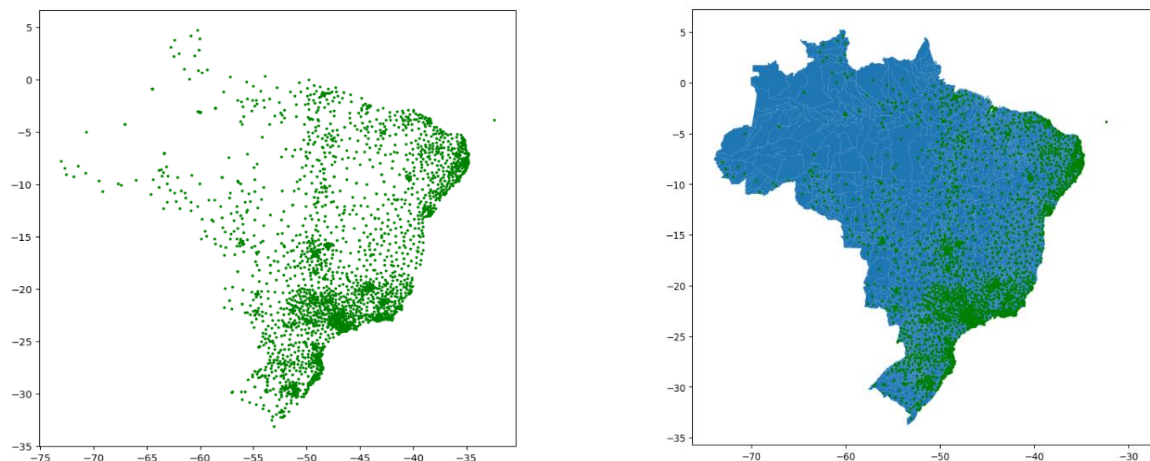
Passo iii: Avaliar se  $B$  está contido em  $A$ , se não está contido ou se está em suas arestas;

Passo iii: Se  $B$  estiver contido em  $A$ , considerar que os atributos relativos a  $A$  e  $B$  são semelhantes;

Passo iv: Repetir os passos i, ii e iii para todos os conjuntos de unidades consumidoras e atributos do conjunto de dados.

Levando em consideração estas heurísticas de geoprocessamento, são unificadas todas as informações da base de dados.

A Figura 4 apresenta o exemplo ilustrativo da união das informações que estão na agregação de município (dados socioeconômicos, geografia, classificação da vegetação e áreas litorâneas) com as informações de conjuntos elétricos. Na figura à esquerda, em verde, podem ser vistos os centroides dos conjuntos de unidades consumidoras. Na figura à direita, em azul, a divisão do território brasileiro em cada uma das regiões municipais e os centroides (em verde) sobrepostos nas regiões que representam esses territórios. Após a sobreposição, é definida a relação entre as informações dos municípios e cada um dos conjuntos de unidades consumidoras. A biblioteca geopandas do python foi utilizada para desempenhar este processamento geográfico.



**Figura 4 - Exemplo do geoprocessamento para o caso em que é realizada a união entre as informações de municípios e conjuntos de unidades consumidoras. Na ilustração à esquerda, em verde, podem ser vistos os centroides extraídos dos polígonos que definem os conjuntos de unidades consumidoras. Já na ilustração à direita, as informações dos centroides são correlacionadas com as regiões em azul, que definem cada um dos municípios no território brasileiros.**

### 4.3 Clusterização

Após a unificação da base de dados, a avaliação dos atributos que à compõem buscando, a partir do conhecimento acerca do problema, realizar a definição de tratamentos a serem realizados, é de suma importância. Estes processamentos buscam melhorar o desempenho do estimador na tarefa definida. Dessa forma, é sabido que dentre os indicadores que são acompanhados pela ANEEL para casos de penalizações, o DEC tem relação com o intervalo de tempo em que o serviço leva para ser restabelecido após a ocorrência de uma interrupção e o FEC se relaciona com o número de interrupções ocorridas. Paradas mais graves no fornecimento impactam tanto o DEC (demora para restabelecimento do fornecimento) quanto o FEC (quantidade elevada de unidades consumidoras impactadas ou alta recorrência de interrupções no período analisado) — Neste caso, é importante ressaltar que não estão sendo considerados eventos de paradas extremas do fornecimento de energia (como é o caso de *blackouts*) visto que a ANEEL não considera dias críticos para a contabilização dos indicadores (módulo 8 do PRODIST [66]) — Sendo assim, os eventos críticos podem ser caracterizados como: tendo um i) alto DEC e alto FEC. Além destes casos, podemos ter também cenários em que o ii) DEC é alto e o FEC é baixo, isto é, uma

quantidade baixa de interrupções, porém demora no reestabelecimento de energia, ou iii) FEC alto e DEC baixo, sendo uma frequência grande de ocorrências e o retorno do abastecimento de forma rápida, ou ainda o iv) DEC e o FEC, ambos baixos. Neste último caso, além da quantidade de ocorrência ser pequena o reestabelecimento do serviço é feito de forma rápida, sendo este, o melhor caso.

Dessa forma, realiza-se um agrupamento considerando, primeiramente, essas 4 possibilidades de estados de DEC e FEC. A definição dos pontos de corte é feita de forma empírica, levando em consideração a distribuição dos atributos, sendo nesta, levadas em consideração as características de intervalo de tempo e quantidades de interrupções (sendo esses os atributos utilizados no cálculo do DEC e FEC). Nesta divisão inicial, para a quantidade de interrupções considera-se o valor da média como o ponto de corte e para o intervalo de tempo das interrupções, é avaliado o ponto em a sua distribuição acumulada têm o seu crescimento suavizado, também sendo este o momento em que a derivada segunda da distribuição tem um menor valor.

Após essa avaliação inicial, é então aplicado um algoritmo de clusterização buscando a seleção dos conjuntos que têm maior similaridade levando em consideração a distância intracluster. Para este agrupamento foi utilizado o método de  $k$ -médias e do cotovelo, assim como foram descritos nas seções 3.5.1 e 3.5.1.1 deste trabalho. Após a aplicação do método de separação e classificação do conjunto de dados em grupos, essa informação é incluída no conjunto de variáveis independentes como uma nova característica. Ou seja, quatro novas colunas para identificar (1 ou 0) o pertencimento do conjunto ao respectivo cluster.

## **4.4 Avaliação das Relações de Causa e Efeito entre as Características e as Interrupções**

Tendo as informações de todas as fontes sido unificadas em relação a suas posições geográficas em uma base única de atributos e quantidade de interrupções – considerando os índices de conjuntos elétricos de distribuição, mês e ano – e o agrupamento do conjunto de dados em relação as características de impacto para o DEC e FEC sido realizado, é importante ser executado um estudo acerca dos efeitos que essas características têm sobre as interrupções ocorridas. Esta análise é realizada para avaliar se os atributos coletados têm coerência com o que é esperado pelo levantamento da literatura realizado na seção 2.3. Para isso, é feita uma avaliação de cada característica em relação aos eventos de interrupção

observados. Esta análise busca auxiliar na seleção das variáveis independentes para posterior aplicação ao modelo.

A relação entre a nomenclatura das características de entrada no modelo, de forma resumida, pode ser vista nas tabelas 1, 2 e 3. As tabelas 1, 2 e 3 buscam facilitar o entendimento de cada variável e das características de entrada no modelo.

É importante ressaltar que a base de medições meteorológicas foi consultada no INMET em agregação diária, contendo o valor mínimo, médio, máximo e o total dependendo da medida avaliada. Após a coleta dos dados, foi realizado um tratamento que faz a agregação em métricas mensais (vide seção 4.1), calculando-se a média e máxima de cada uma das medidas e o valor total mensal da precipitação.

**Tabela 1 - Características utilizadas como entrada do modelo – Características 1 a 8**

	Características utilizadas como entrada do modelo	Significado
1	MAX_PRECIPITACAO TOTAL, DIARIO (AUT)(mm)	Máxima Precipitação Total Mensal
2	MAX_PRESSAO ATMOSFERICA MEDIA DIARIA (AUT)(mB)	Máxima Pressão Atmosférica Média Diária
3	MED_PRESSAO ATMOSFERICA MEDIA DIARIA (AUT)(mB)	Média Pressão Atmosférica Média Diária
4	MAX_TEMPERATURA DO PONTO DE ORVALHO MEDIA DIAR...	Máxima Temperatura do Ponto de Orvalho Média Diária
5	MED_TEMPERATURA DO PONTO DE ORVALHO MEDIA DIAR...	Média Temperatura do Ponto de Orvalho Média Diária
6	MAX_TEMPERATURA MAXIMA, DIARIA (AUT)(°C)	Máxima Temperatura Mensal
7	MAX_TEMPERATURA MEDIA, DIARIA (AUT)(°C)	Máxima temperatura Média Diária
8	MAX_TEMPERATURA MINIMA, DIARIA (AUT)(°C)	Máxima temperatura Mínima Diária

**Tabela 2 – Características utilizadas como entrada do modelo – Características 9 a 21**

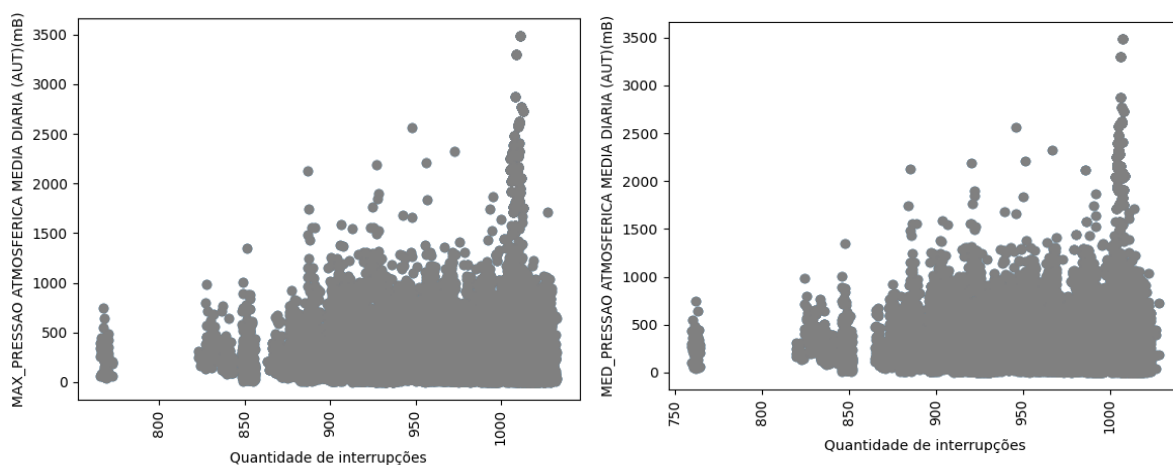
9	MAX_UMIDADE RELATIVA DO AR, MEDIA DIARIA (AUT)(%)	Máxima umidade Relativa do ar Média Diária
10	MAX_UMIDADE RELATIVA DO AR, MINIMA DIARIA (AUT)(%)	Máxima Umidade Relativa do Ar, Mínima Diária
11	MAX_VENTO, RAJADA MAXIMA DIARIA (AUT)(m/s)	Máxima Rajada de Vento Mensal
12	MAX_VENTO, VELOCIDADE MEDIA DIARIA (AUT)(m/s)	Máxima Velocidade do Vento Média Diária
13	MED_PRECIPITACAO TOTAL, DIARIO (AUT)(mm)	Média Precipitação Total Diária
14	MED_TEMPERATURA MAXIMA, DIARIA (AUT)( $^{\circ}\text{C}$ )	Média Temperatura Máxima Diária
15	MED_TEMPERATURA MEDIA, DIARIA (AUT)( $^{\circ}\text{C}$ )	Média Temperatura Média Diária
16	MED_TEMPERATURA MINIMA, DIARIA (AUT)( $^{\circ}\text{C}$ )	Média Temperatura Mínima Diária
17	MED_UMIDADE RELATIVA DO AR, MEDIA DIARIA (AUT)(%)	Média Umidade Relativa do ar Média Diária
18	MED_UMIDADE RELATIVA DO AR, MINIMA DIARIA (AUT)(%)	Média Umidade Relativa do ar Mínima Diária
19	MED_VENTO, RAJADA MAXIMA DIARIA (AUT)(m/s)	Média Rajada do Vento Máxima Diária
20	MED_VENTO, VELOCIDADE MEDIA DIARIA (AUT)(m/s)	Média Velocidade do Vento Média Diária
21	TOTAL_PRECIPITACAO TOTAL, DIARIO (AUT)(mm)	Total Precipitação Mesal

**Tabela 3 - Características utilizadas como entrada do modelo – Características 22 a 37**

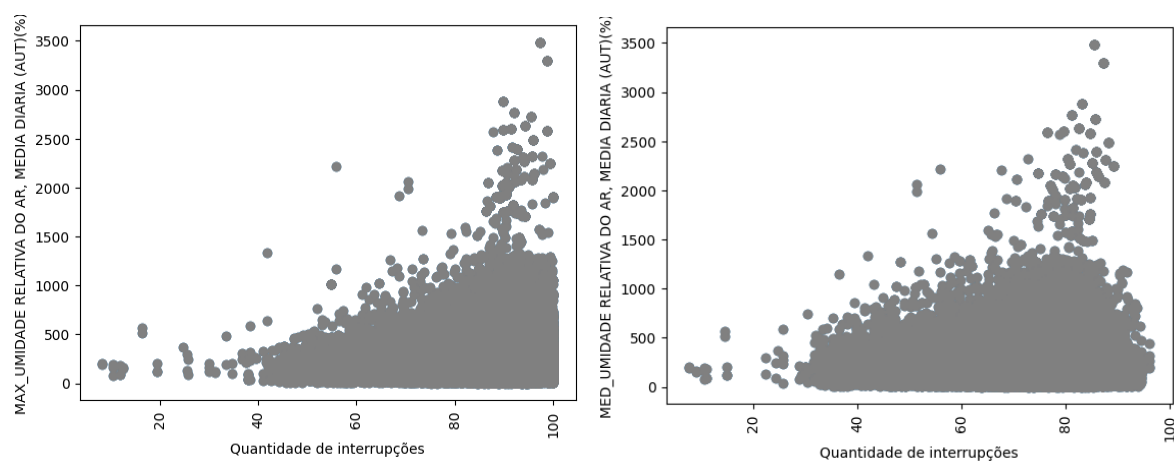
22	Pop estimada 2021	População estimada (2021)
23	Produto Interno Bruto, a preços correntes (R\$ 1.000)	PIB
24	Produto Interno Bruto per capita, a preços correntes (R\$ 1,00)	PIB / per capita
25	MUN_MINIMUM_BOUND_RADIO	Raio do Município que contém o Conjunto de Unidades Consumidoras
26	AREA	Área do Município que contém o Conjunto de Unidades Consumidoras
27	MUN_NUMBER_OF_BOUNDARYS	Número de Vértices do Polígono do Município que Forma o Conjunto de Unidades Consumidoras
28	NUMERO_UNIDADES_CONSUMIDORAS	Numero de Unidades Consumidoras no Conjunto
29	FLAG_INTERRUPCOES	Quantidade de Interrupções
30	FLAG_INTERRUPCOES_lognorm	Quantidade de Interrupções em escala logarítmica
31	FLAG_LIT	Indicador de Estado Litorâneo
32	Amazônia Legal	Indicador de Amazônia Legal
33	Semiárido	Indicador de Semiárido
34	veg_class	Classificação do Tipo de Vegetação (IBGE)
35	Labels	Clusters
36	percent_tree_cover	Porcentagem de Área Arborizada
37	UF	Estado



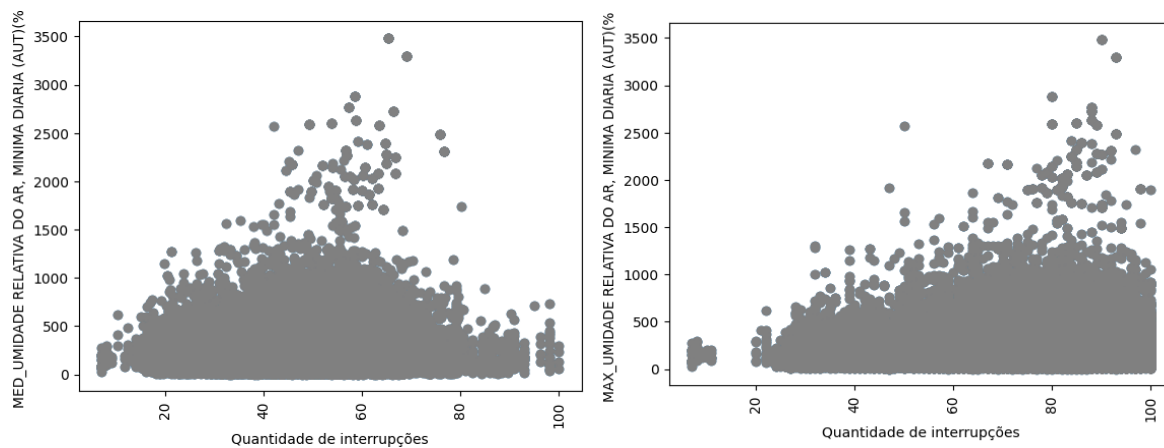
Os atributos foram avaliados como sendo representativos (ou o seu contrário), inicialmente, a partir do estudo de suas correlações com as quantidades de interrupções apresentado na tabela do anexo, A1. Nas figuras de 5 a 19 pode ser visualizada a dispersão entre algumas informações quantitativas no conjunto de dados e a quantidade de interrupções.



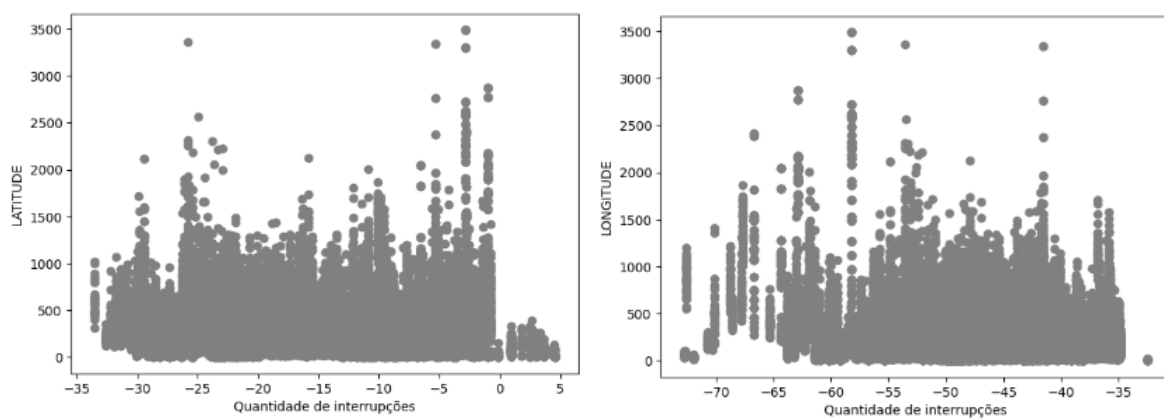
**Figura 5 - Relação entre a Máxima Pressão Atmosférica Média Diária (esquerda) e a Média Pressão Atmosférica Média Diária (direita) com a quantidade de interrupções mensal por cada conjunto elétrico (SIGEL).**



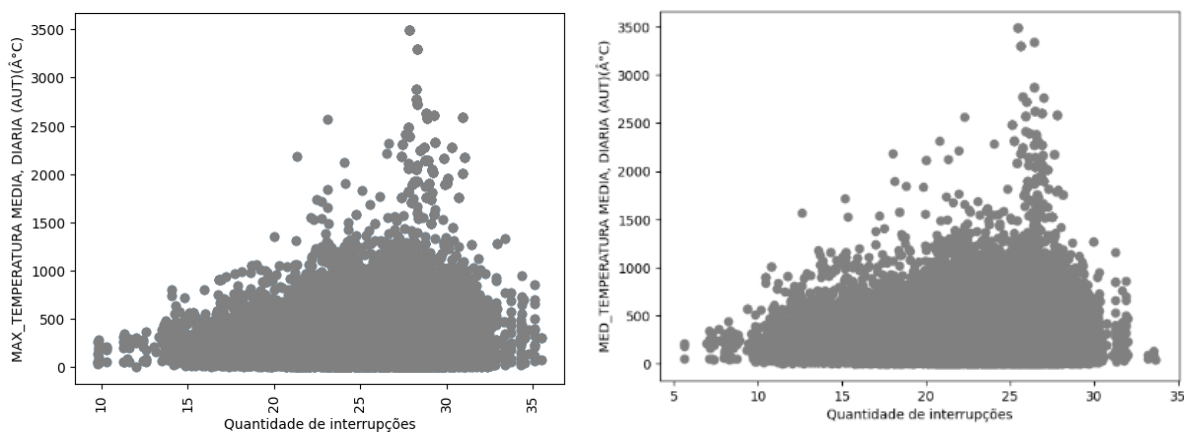
**Figura 6 - Relação entre a Máxima Umidade Relativa do Ar Média Diária (esquerda) e a Média Umidade Relativa do Ar Média Diária (direita) com a quantidade de interrupções mensal por cada conjunto elétrico (SIGEL).**



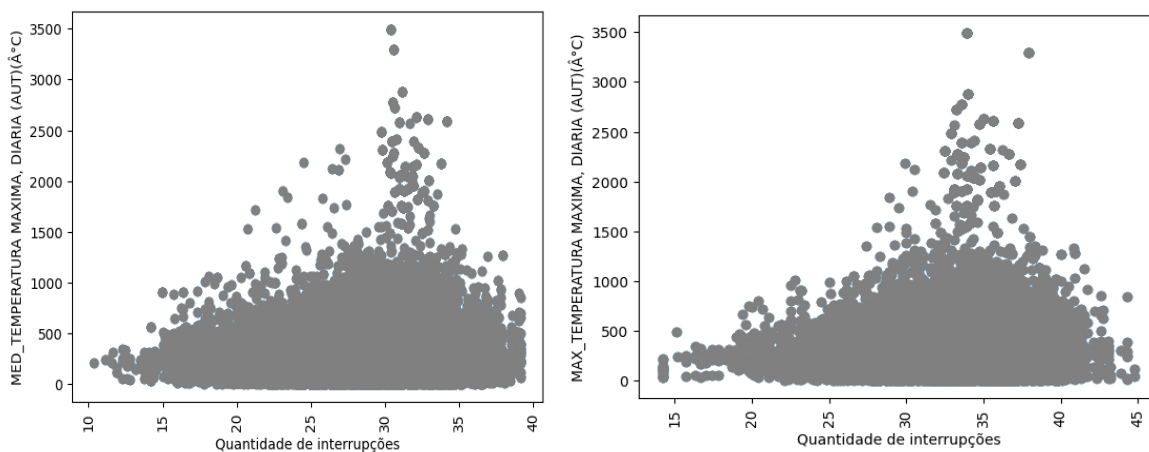
**Figura 7 - Relação entre a Média Umidade Relativa do Ar Mínima Diária (esquerda) e a Máxima Umidade Relativa do Ar Mínima Diária (direita) com a quantidade de interrupções mensal por cada conjunto elétrico (SIGEL).**



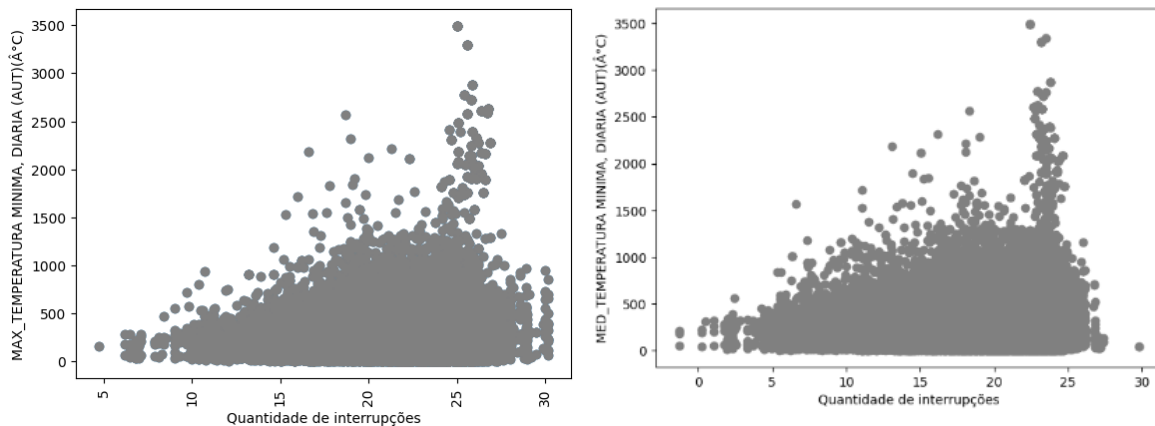
**Figura 8 - Relação entre a Latitude (esquerda) e a Longitude (direita) com a quantidade de interrupções mensal por cada conjunto elétrico (SIGEL).**



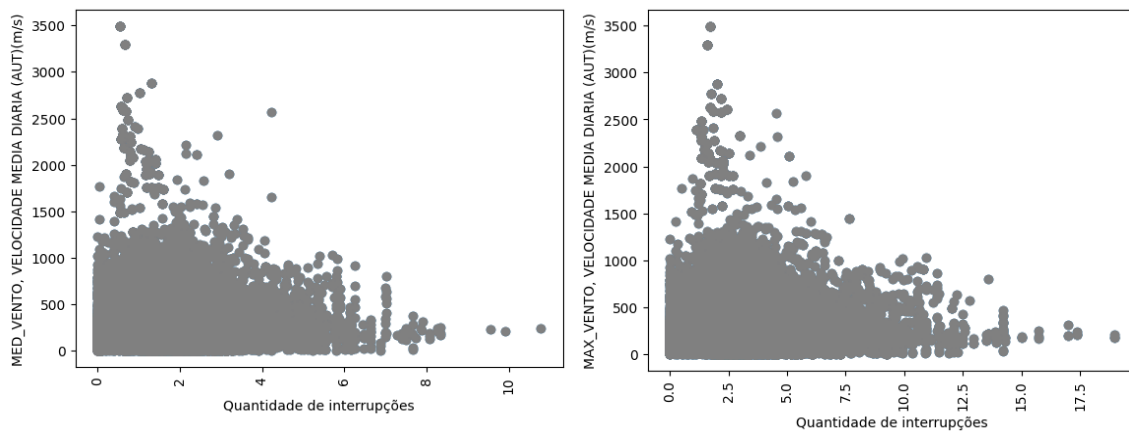
**Figura 9 - Relação entre a Máxima Temperatura Média Diária e a Média Temperatura Média Diária com a quantidade de interrupções mensal por cada conjunto elétrico (SIGEL).**



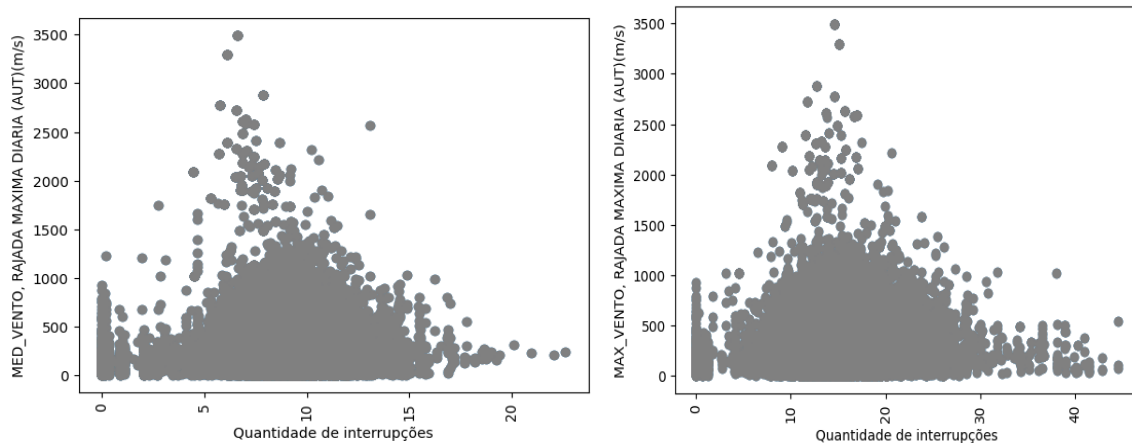
**Figura 10 - Relação entre a Média Temperatura Máxima Diária (esquerda) e a Máxima Temperatura Máxima Diária com a quantidade de interrupções mensal por cada conjunto elétrico (SIGEL).**



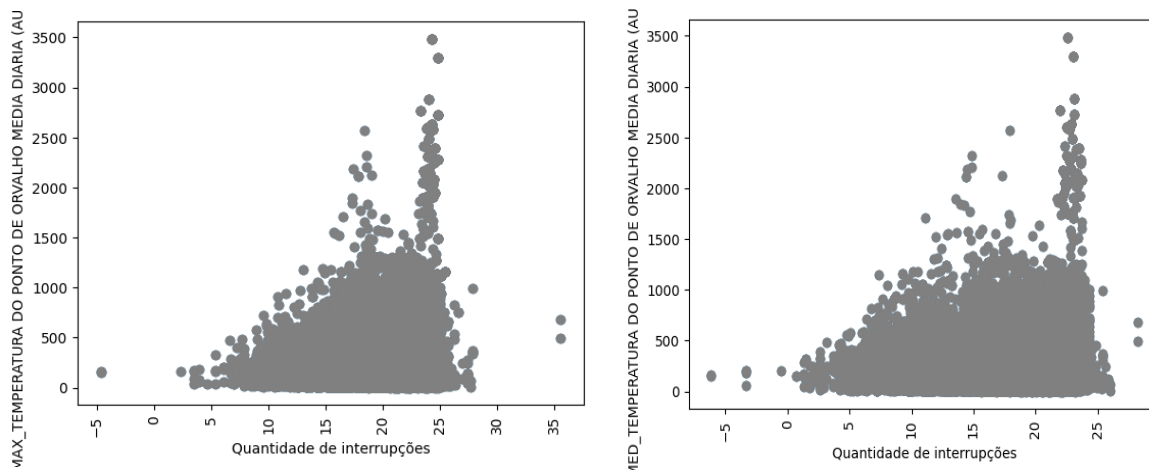
**Figura 11 - Relação entre a Máxima Temperatura Mínima Diária (esquerda) e a Média Temperatura Mínima Diária com a quantidade de interrupções mensal por cada conjunto elétrico (SIGEL).**



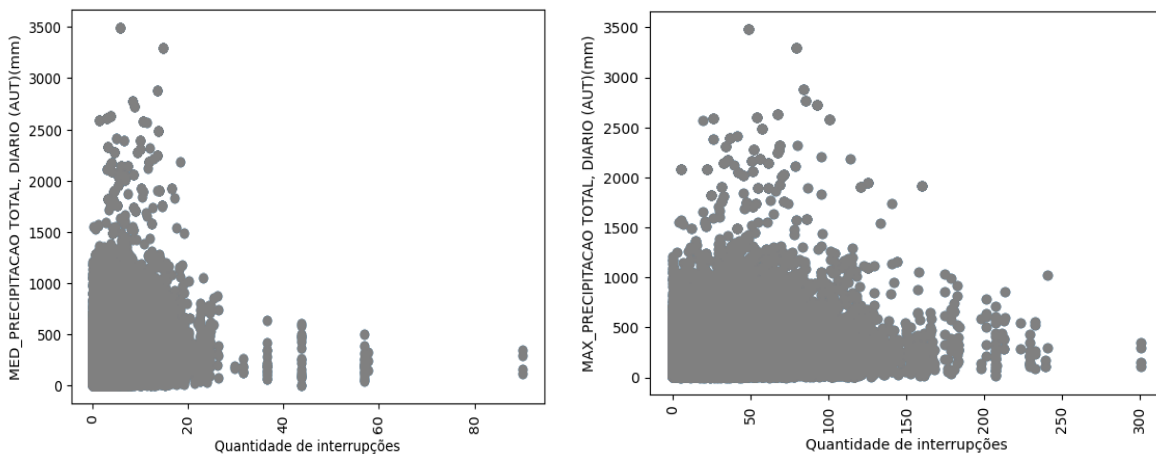
**Figura 12 - Relação entre a Média Velocidade do Vento Média Diária (esquerda) e a Máxima Velocidade do Vento Média Diária (direita) com a quantidade de interrupções mensal por cada conjunto elétrico (SIGEL).**



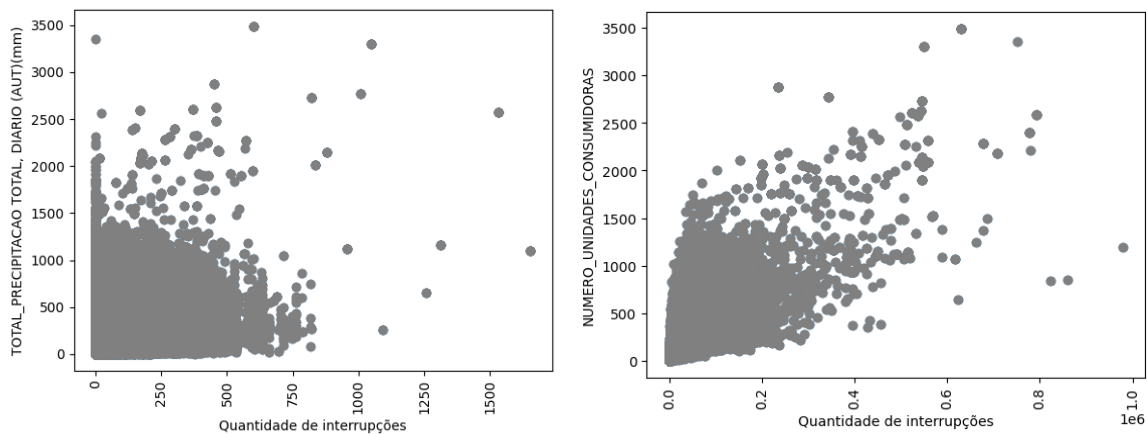
**Figura 13 - Relação entre a Média Rajada de Vento Máxima Diária (esquerda) e a Máxima Rajada de Vento Máxima Diária (direita) com a quantidade de interrupções mensal por cada conjunto elétrico (SIGEL).**



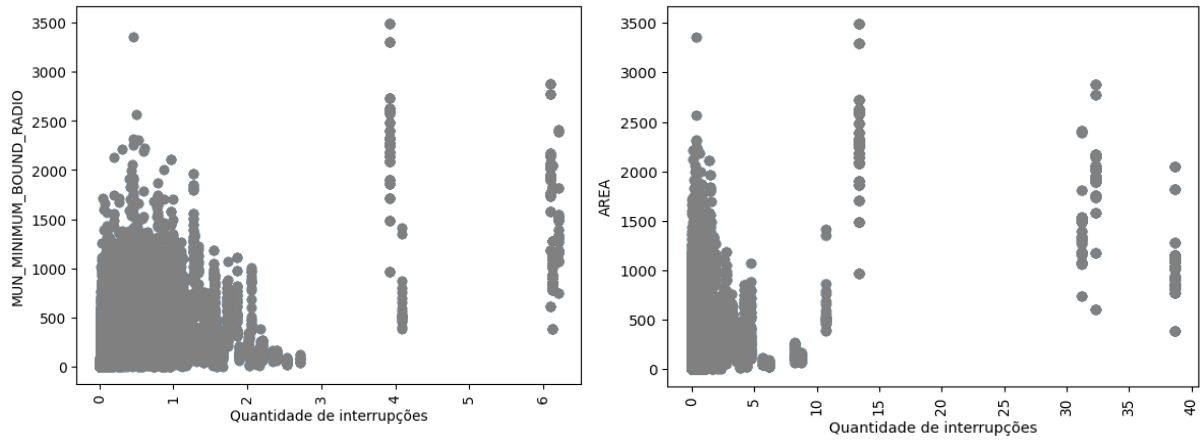
**Figura 14 - Relação entre a Máxima Temperatura do Ponto de Orvalho Média Diária (esquerda) e a Média Temperatura do Ponto de Orvalho Média Diária (direita) com a quantidade de interrupções mensal por cada conjunto elétrico (SIGEL).**



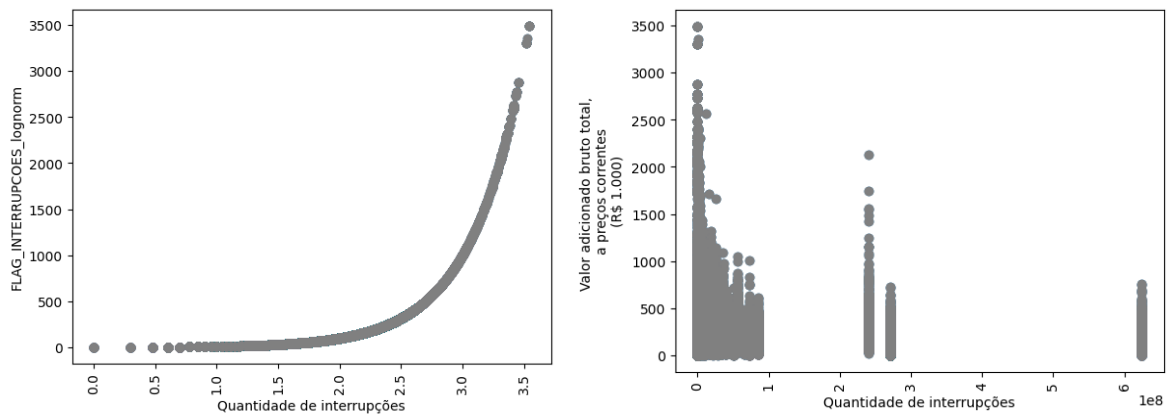
**Figura 15 - Relação entre a Média Precipitação Total Diária (esquerda) e a Máxima Precipitação Total Diária com a quantidade de interrupções mensal por cada conjunto elétrico (SIGEL).**



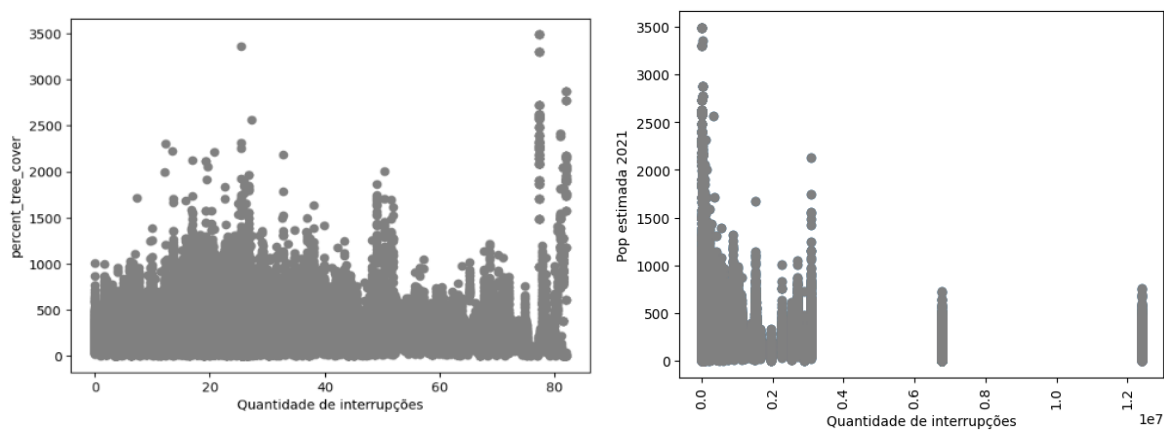
**Figura 16 - Relação entre a Precipitação Total Mensal e o Número de Unidades Consumidoras no Conjunto com a quantidade de interrupções mensal por cada conjunto elétrico (SIGEL).**



**Figura 17 - Relação entre o Raio e a Área com a quantidade de interrupções mensal por cada conjunto elétrico (SIGEL).**



**Figura 18 - Relação entre a Transformação Logarítmica da Quantidade de Interrupções e o PIB com a quantidade de interrupções mensal por cada conjunto elétrico (SIGEL).**



**Figura 19 - Relação entre o Percentual de Cobertura Arbórea e a População Estimada em 2021 com a quantidade de interrupções mensal por conjunto elétrico (Sigel)**

Pela figura 5 é possível ver que, ao ser aumentada a pressão atmosférica, também são aumentados os casos de interrupções. Isto também pode ser observado em relação a umidade relativa do ar nas figuras 6 e 7. O número de consumidores, como observado na figura 16, também é um fator que tem correlação com a quantidade de interrupções, desta forma, conjuntos com mais unidades consumidoras tendem a apresentar um número maior de interrupções, conforme abordado na seção 2.2.4.

Os atributos foram avaliados como sendo representativos (ou o seu contrário), inicialmente, a partir do estudo de suas correlações com as quantidades de interrupções apresentado na tabela A1 do anexo. Esta tabela mostra e confirma as relações que foram observadas pelas figuras (5 a 19). Observa-se que a precipitação, que pelas figuras 15 e 16, não parece ter uma correlação positiva tão evidente com a quantidade de interrupções, apresenta um dos índices de correlação mais relevantes, como pode ser confirmado pela tabela A1 do anexo. Esse resultado vai de acordo com o que é descrito na literatura e discutido na seção (2.3.1).

Nas seção 4.7 é apresentado o teste de hipóteses com a estatística T. Com este, tem-se o objetivo ser realizada uma seleção de atributos levando em consideração um teste de



hipóteses que avalia a significância de cada um destes. Sendo está a técnica utilizada neste trabalho para selecionar as características de entrada do modelo.

## 4.5 Processamento dos Dados e Aplicação do Modelo

Após a análise da relação dos atributos com a quantidade de interrupções buscando ser verificado o tipo de relação presente entre estes, é então realizado o pré-processamento final das informações. Este tratamento busca preparar a base de dados para a aplicação do modelo. Para o estimador é necessário que as informações estejam todas na forma numérica. Além disso, outros tratamentos, como o filtro de registros discrepantes, são realizados, sendo retirados 10% dos conjuntos que contém as menores quantidades de interrupções (esse tratamento foi aplicado devido a presença de valores de quantidades de interrupções muito baixos no conjunto de dados, o que enviesa as métricas de desempenho utilizadas). Uma discussão mais aprofundada sobre estes casos é realizada no capítulo 5, seção 5.6). Nenhum filtro foi aplicado para os valores altos.

Sendo assim, para informações que apresentam características categóricas, são completadas as classes ausentes com uma nova classificação indicando a ausência da categoria (“Sem classificação”). Estas variáveis também são submetidas a codificação 1-de-N (representação binária), em que cada classe, de cada um dos atributos da base de dados, é representada por uma nova coluna contendo um indicador 1 caso o registro pertença ou 0 caso o não pertencimento à classe avaliada. Por exemplo, o atributo de tipo de vegetação que contém 39 categorias diferentes, é processado e representado por 38 colunas na base de dados, sendo cada uma apresentando o valor 1, caso o registro pertença a categoria e 0, caso contrário. Após esse tratamento, uma das colunas é definida como a referência e então retirada do conjunto de dados. Desta forma, os registros relacionados a essas colunas são representados como um vetor em que todas as outras categorias são zero, da forma,  $[0, 0, 0, \dots, 0]$ . Esse processamento final é realizado tendo-se o objetivo de evitar a multicolinearidade, sendo diminuído o número de atributos no conjunto de dados final que podem conter informações correlacionadas.

Já para os dados numéricos, em casos de ausência é utilizado o valor da média completando a informação ausente e é realizada a padronização dos dados levando em consideração os valores máximos e mínimos. Esse tratamento busca deixar todas as características em uma escala de 0 a 1 para a estimação do modelo. Este processamento é

importante para remover qualquer viés de escala entre as variáveis durante o processo de estimação dos parâmetros e realização dos testes de hipótese.

Ainda, é aplicado um processamento na base de dados buscando criar uma variável que represente o comportamento de períodos passados no estimador construído com modelo de regressão múltipla. Dessa forma, uma nova coluna é incluída na matriz de dados contendo a quantidade de interrupção do ano anterior, levando em consideração o mesmo mês e o mesmo conjunto.

Após esses tratamentos, estrutura-se uma nova base de dados contendo tanto as características processadas quanto a quantidade de interrupções por conjunto elétrico. Esta base é então dividida em duas porções, uma para a estimação dos parâmetros do modelo e outra para avaliação dos resultados. São realizadas diferentes divisões e o teste com diferentes configurações do conjunto de dados, levando em consideração a natureza estocástica do problema, sendo que, no teste contendo todas as distribuidoras, utilizam-se os 2 meses finais do conjunto de dados para avaliação, e nos testes com distribuidoras específicas que foram selecionadas, utiliza-se o último ano para validação. Ao final, o modelo é aplicado, sendo a sua formulação dada pela equação (20) em que cada coeficiente de regressão é representado pelas informações coletadas e geoprocessadas:

$$y_i = \beta_0 + \beta_1 x_{atr_{i1}} + \beta_2 x_{atr_{i2}} + \dots + \beta_N x_{atr_{iN}} + \varepsilon_i \quad (20)$$

Sendo que cada  $x_{atr_i}$  representa uma mensuração em cada conjunto de unidades consumidoras, para os  $N$  atributos do conjunto de dados e cada  $y_i$  a quantidade de interrupções observada nesses mesmos conjuntos. Os coeficientes de regressão  $\beta$  são as variáveis a serem estimadas para cada atributo do modelo.

Para cada conjunto de dados é aplicado então o Algoritmo de Mínimos Quadrados (sec. 3.2.1), tendo o objetivo de serem estimados os coeficientes  $\beta$  do modelo, que minimizam o erro de estimação em relação aos casos reais de quantidades de interrupções que foram observadas. Após a aplicação, as estimações  $\hat{y}_i$  do modelo podem ser geradas com a equação (21):

$$\hat{y}_i = \beta_0 + \beta_1 x_{atr_{i1}} + \beta_2 x_{atr_{i2}} + \dots + \beta_N x_{atr_{iN}} \quad (20)$$

Ao final são gerados modelos diferentes (com coeficientes  $\beta$  diferentes) para cada um dos conjunto de dados, levando em consideração o caso em que são utilizadas as informações de todas as distribuidoras e o caso em que são utilizadas partições do conjunto geral contendo somente os registros de cada uma das distribuidoras que foram selecionadas para o teste. Desta forma, como se diferem os dados históricos em cada caso, os coeficientes de regressão também o são dessemelhantes.

Os modelos propostos são então comparado com um método simplificado (utilizado como *baseline*), considerando que todas as interrupções históricas no ano anterior e nos mesmos conjuntos elétricos e meses de referência irão se repetir no futuro nas mesmas datas. Esta estratégia de previsão, baseada nas sazonalidades envolvidas, é denominado de modelo ingênuo. E como descrito na seção 1.2, é um método utilizado comumente como comparação para modelos baseados em séries temporais.

Tanto o método ingênuo quanto o modelo final construído são avaliados levando em consideração as mesmas métricas, quais sejam:  $R^2$ , MAPE e MSE. Estas foram escolhidas para avaliação do estimador devido às características de distribuição dos dados de interrupções. Sendo o  $R^2$  uma métrica comumente aceita para avaliação da qualidade do ajuste de modelos de regressão, sua utilização é a primeira escolha para o problema. O MAPE é uma métrica de erro percentual que permite uma avaliação de fácil entendimento e direta dos resultados do modelo, porém este sofre impacto (enviesamento) na presença de registros discrepantes. O MSE é uma métrica modular que penaliza registros discrepantes com valores altos e, assim como o MAPE, também é enviesado nestes casos, porém, como o conjunto de dados apresenta muitos registros extremos abaixo da média geral de quantidade de interrupções, esta métrica permite uma avaliação melhor do que o MAPE dos resultados do modelo, levando em consideração estes casos.

O melhor resultado do modelo de regressão múltipla em relação ao método ingênuo mostra que o método é eficaz para a projeção da quantidade de interrupções em relação ao método de comparação simples que é comumente utilizado. Porém, a garantia de que o modelo criado é suficiente para a realização da tarefa que é proposta, depende da necessidade e dos limitantes imposto em cada caso de uso e também da avaliação crítica dos valores das métricas de desempenho que foram definidas.

## 4.6 Validação do Modelo via Estatística F

Com a base de dados tratada para o modelo e após a avaliação de seus atributos em relação a quantidade de interrupções, a estatística  $F$  – assim como abordado no capítulo 3 (seção 3.4) – é utilizada para validação da representatividade estatística do modelo de regressão múltipla. A hipótese nula, dessa forma, assim como discutido na seção 3.4, é definida considerando todos os coeficientes do modelo nulos. Já a hipótese alternativa é de que existe pelo menos um desses coeficientes que não é estatisticamente nulo. Dessa forma, pode-se dizer que o estimador é representativo pelo menos para um dos coeficientes do modelo. Essa estatística é calculada tanto para o estimador completo, contendo todas as distribuidoras do Brasil, quanto para os casos de companhias específicas avaliadas.

## 4.7 Seleção de Características via Estatística T

Após a aplicação do teste de hipóteses para avaliação da significância estatística do modelo com a estatística  $F$ , é feita então uma seleção de variáveis com um teste estatístico utilizando-se a estatística  $T$ . Após a aplicação do teste, são selecionados apenas os atributos que resultem em *p-value* menor do que 0,05. Apenas as variáveis selecionadas são utilizadas no modelo final.

O teste com a estatística  $T$  é feito de forma recursiva. Isso porque ao se retirar uma variável do modelo altera-se a relação entre as variáveis de entrada remanescentes e a variável de saída. Essa remoção pode excluir informações relevantes do modelo, vindo a causar problemas, como favorecer a multicolinearidade, por exemplo. Dessa forma, primeiramente, são selecionadas somente as variáveis consideradas relevantes (coeficiente não-nulo) segundo a aplicação do teste  $T$ . Os coeficientes desse novo modelo são estimados novamente e este modelo é então avaliado segundo o teste  $F$ . Caso o modelo seja considerado estatisticamente relevante segundo o teste  $F$ , é novamente aplicada a estatística  $T$  nas variáveis que foram selecionadas no teste anterior. Caso não exista nenhuma variável retirada do modelo na nova iteração, o processo é finalizado e o conjunto de atributos final é definido. Caso ainda exista alguma variável a ser retirada na nova iteração, o processo é refeito, recalculando-se a estatística  $F$  para verificar a validade do novo modelo.

A seleção de característica utilizando-se a estatística  $T$  foi realizada em todos os testes menos para o modelo em nível nacional.

## Capítulo 5 - Resultados e Discussões

Após a discussão realizada nos capítulos anteriores com a apresentação, tanto do problema a ser abordado quanto da metodologia definida para seu enfrentamento, neste capítulo serão discutidos os resultados gerados após a aplicação da metodologia proposta no Capítulo 4. É realizada também uma discussão e avaliação crítica dos resultados, visando verificar a sua relação com o que se era esperado levando em consideração literatura pesquisada e que foi apresentada no Capítulo 2.

### 5.1 Conjunto de Dados

Levando em consideração a metodologia que foi apresentada no capítulo 4, o conjunto de dados utilizado [135] para a aplicação desta e realização dos testes foi coletado de fontes de informações domínio público.

Após ser realizada a coleta e levando em consideração as informações presentes em todas as fontes, foram mapeados 2.923 conjuntos de unidades consumidoras. Para o estudo foram considerados os dados de todas as fontes no período que compreende os anos de 2020, 2021 e 2022, tendo sido extraídas as informações para todos os meses destes anos.

A Tabela 4 evidencia, de forma resumida, a relação entre a quantidade total de conjuntos levantados no sistema SIGEL e a quantidade de conjuntos considerados na base deste trabalho. Foram avaliados, aproximadamente, 80% dos conjuntos do sistema SIGEL, sendo que os 20% restantes não foram considerados por razões referentes a disponibilidade e a qualidade dos dados nas fontes em que foram coletados.

Com isso, apesar de existirem 3.636 definidos no sistema da ANEEL, apenas 3.125 apresentam informações geográficas após a coleta de dados. Destes, apenas 2.922 apresentam relação de 1 para 1 entre o seu identificador e a sua descrição. Na base de dados de interrupções da ANEEL, nos anos de 2020, 2021 e 2022 são observados respectivamente dados para 2969, 2965, 2948 conjuntos de unidades consumidoras. Após unir estas informações, as bases de dados de 2020, 2021 e 2022 apresentaram 2.914, 2.830 e 2.733 conjuntos, respectivamente, em cada ano. A base final, considerando todos estes anos, apresenta 2.923 conjuntos de unidades consumidoras. Na Tabela 4 ainda é possível ver o sumário que descreve a quantidade de conjuntos na base do sistema da ANEEL e a

quantidade de conjuntos que as informações foram consideradas ausentes. Na Tabela 5, é descrita a relação entre as fontes de dados utilizadas e as informações que foram coletadas de cada uma dessas fontes.

**Tabela 4 - Quantidade de conjuntos mapeados vs. faltantes**

Sumário dos Dados	Valores
Conjuntos Total Base SIGEL	3636 conjuntos
% de Conjuntos faltantes na base final	20% (713 conjuntos)

**Tabela 5 – Relação entre as fontes de dados públicas e as informações que foram extraídas**

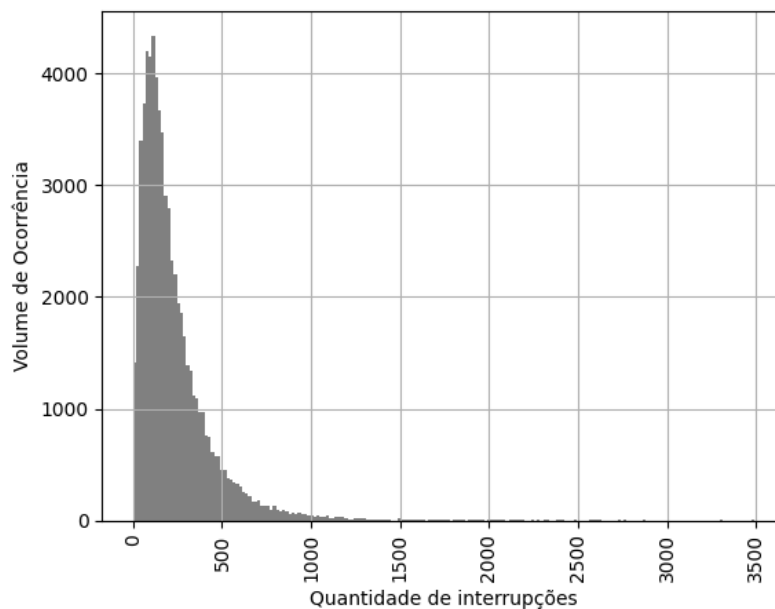
Fonte	Informações Extraídas
ANEEL	Interrupções
INMET	Meteorologia
SIGEL (ANEEL)	Conjuntos de Unidades Consumidoras
IBGE	Geografia
IBGE	Áreas Litorâneas
IBGE	Vegetação
IBGE	Socioeconômicos
INPE (MODIS)	Porcentagem de Área Arbórea

## 5.2 Descrição Estatística do Conjunto de Dados Utilizado

### 5.2.1 Avaliação dos Registros de Interrupções

Sendo o objetivo deste estudo a projeção da quantidade de interrupções por conjuntos de unidades consumidoras (como discutido na sec. 4.1), a primeira pergunta a ser respondida é: qual é o comportamento da quantidade de interrupções de energia no conjunto de dados, pela chave de conjunto de clientes, no período definido? Assim, é possível entender se essas informações apresentam uma dispersão muito alta, em que intervalo se encontra o volume maior de ocorrências e se a série apresenta alguma padrão atípico (como por exemplo, uma

distorção de cauda longa para algum dos extremos). Buscando-se responder estas perguntas, é realizado primeiramente um histograma da quantidade de interrupções. Este, por sua vez, é apresentado na Figura 20.



**Figura 20 - Histograma contendo a dispersão dos casos de interrupções mensais nos conjuntos elétricos**

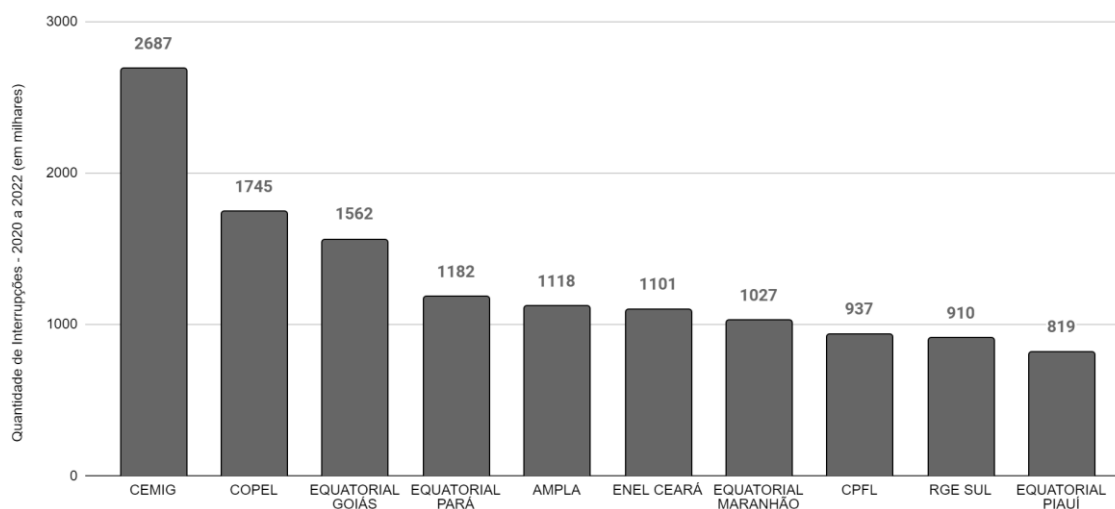
Pode-se notar que os dados de interrupção apresentam uma cauda longa para a direita, o que também pode ser validado pela mensuração da curtose (definida por Fisher [136]), que tem valor de aproximadamente 28,3. Isso evidencia que as interrupções no conjunto de dados apresentam divergência em relação a uma distribuição gaussiana (distribuição normal).

Diversas alternativas podem ser aplicadas nesse caso. A usualmente adotada é a transformação logarítmica. Esta transformação foi aplicada nos dados durante os testes, porém foi observado um resultado pior nas métricas de performance utilizadas ( $R^2$ , MAPE e MSE) em relação ao estimador sem a aplicação desta transformação.

Outro aspecto importante é que o maior volume de ocorrências está disposto entre os intervalos de 0 a 1000 interrupções. É possível notar também que há registros extremos chegando ao valor de 3500 interrupções pelo gráfico da Figura 20. Ainda, parece haver um pico de ocorrências entre o intervalo de 0 a 500, sendo que o maior volume de conjuntos tem ocorrências no começo desse intervalo. Sendo este comportamento evidenciado pelo ponto em que o eixo y chega no valor 4000.

Além da quantidade geral de interrupções e sua distribuição, uma análise importante é a investigação destes eventos para determinadas hierarquias de classificação importantes no contexto do conjunto de dados em que é realizada a análise. Essa investigação é importante tanto porque as previsões são realizadas levando em consideração testes em concessionárias específicas em determinadas regiões do país, quanto para o maior conhecimento sobre o conjunto de dados.

Com isso, os dados na Figura 21 apresentam a quantidade de interrupções contabilizada na base de dados da ANEEL considerando o ranking das 10 distribuidoras de energia com maior quantidade de eventos nos anos de 2020 a 2022. Devido à natureza estocástica dos fenômenos de interrupções estudados, essas métricas apresentam variações dependendo do ano avaliado. Por exemplo, foi observado para a distribuidora CEMIG, um aumento de aproximadamente 12% do número de interrupções entre os anos 2021 e 2022. Dessa forma, é esperado que se tenham flutuações nos valores apresentados em períodos diferentes do que foi realizado o estudo. Por exemplo levando em consideração anos anteriores, ou períodos posteriores ao que foi avaliado (2020 a 2022).



**Figura 21 - Quantidade de interrupções no fornecimento de energia elétrica levantadas pela ANEEL, nos anos de 2020 a 2022, por distribuidora de energia.**

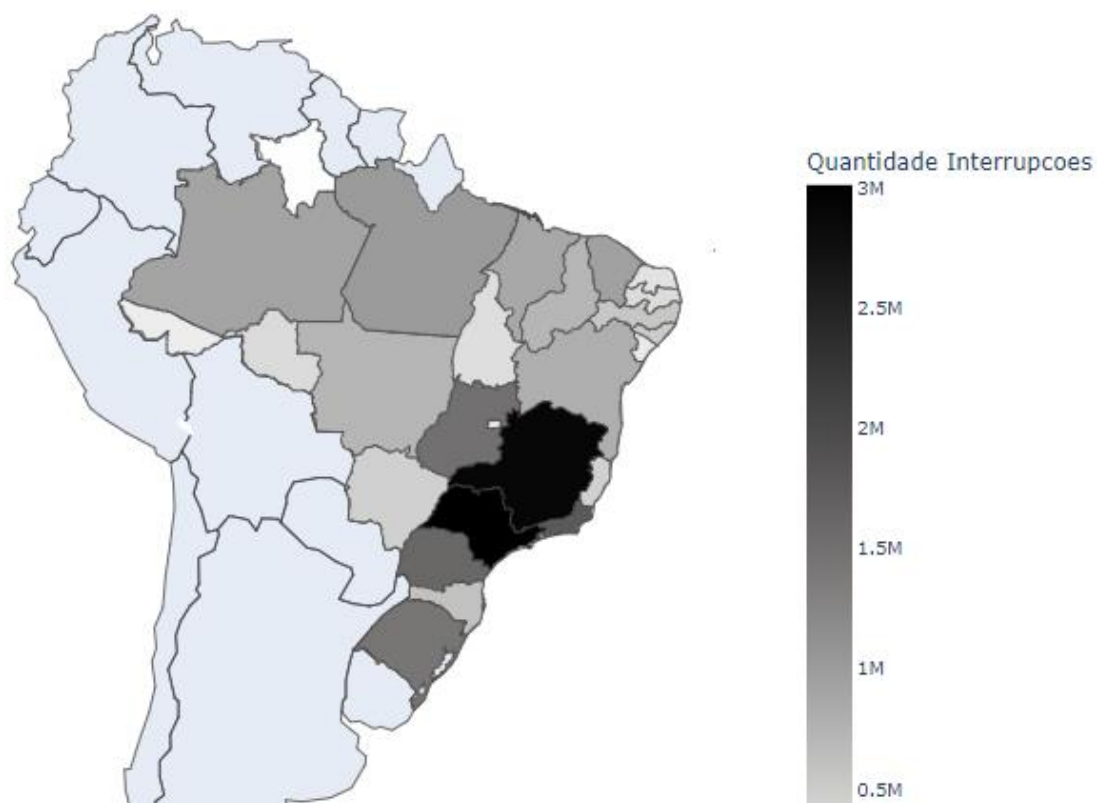
Após a avaliação da quantidade de interrupções por companhias de distribuição, e levando em consideração a diversidade dos diferentes atributos de impacto nas paradas do fornecimento em relação a região de avaliação territorial do Brasil (já discutida mais detalhadamente na seção 2.3), é feita a avaliação regional.



Na figura 22, nota-se que para os estados do Sudeste nota-se uma maior quantidade de casos, por exemplo, em Minas Gerais e em São Paulo (aproximadamente 1,96 e 2,04 milhões de interrupções cada).

Ainda, dos conjuntos avaliados que foram avaliados, aproximadamente 20% do total de interrupções no período analisado tem concentração no estado de São Paulo e 12% no estado de Minas Gerais. Sendo estas regiões estaduais as que apresentam maior quantidade de interrupções. Demais estados, como Bahia, Rio de Janeiro, Goiás, Rio Grande do Sul, Paraná, dentre outros, apresentam uma faixa de 1 a 7 % do total na amostra.

Levando em consideração todas as regiões do território brasileiro, o conjunto de dados apresenta um total de 22.886.945 interrupções (para os anos de 2020, 2021 e 2022).



**Figura 22 - Quantidade de interrupções por estado brasileiro no conjunto de dados final, nos anos de 2020 a 2022 – Dados ANEEL – 2023**

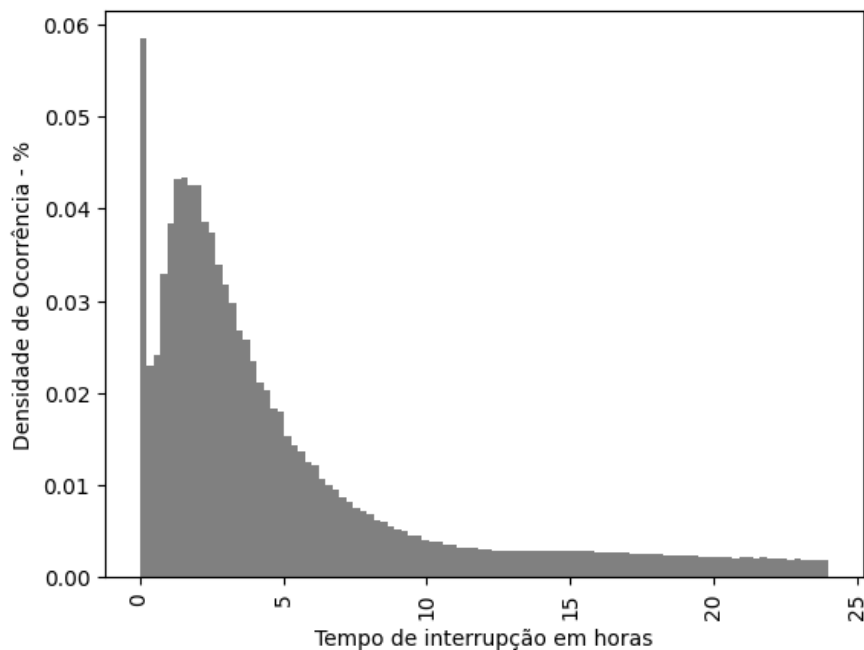
### 5.3 Resultados da Clusterização

Uma vez sido realizada a avaliação do conjunto de dados utilizado, como descrito na seção 5.2, a inclusão no estimador de informações relacionadas com as grandezas de DEC e FEC pode aprimorar o desempenho da ferramenta. Dessa forma, são aplicados o algoritmo de agrupamento dos dados em conjuntos (*clusters*). Para representação dos conjuntos neste agrupamento, utilizam-se as informações mensais de DEC e FEC durante o período de estudo.

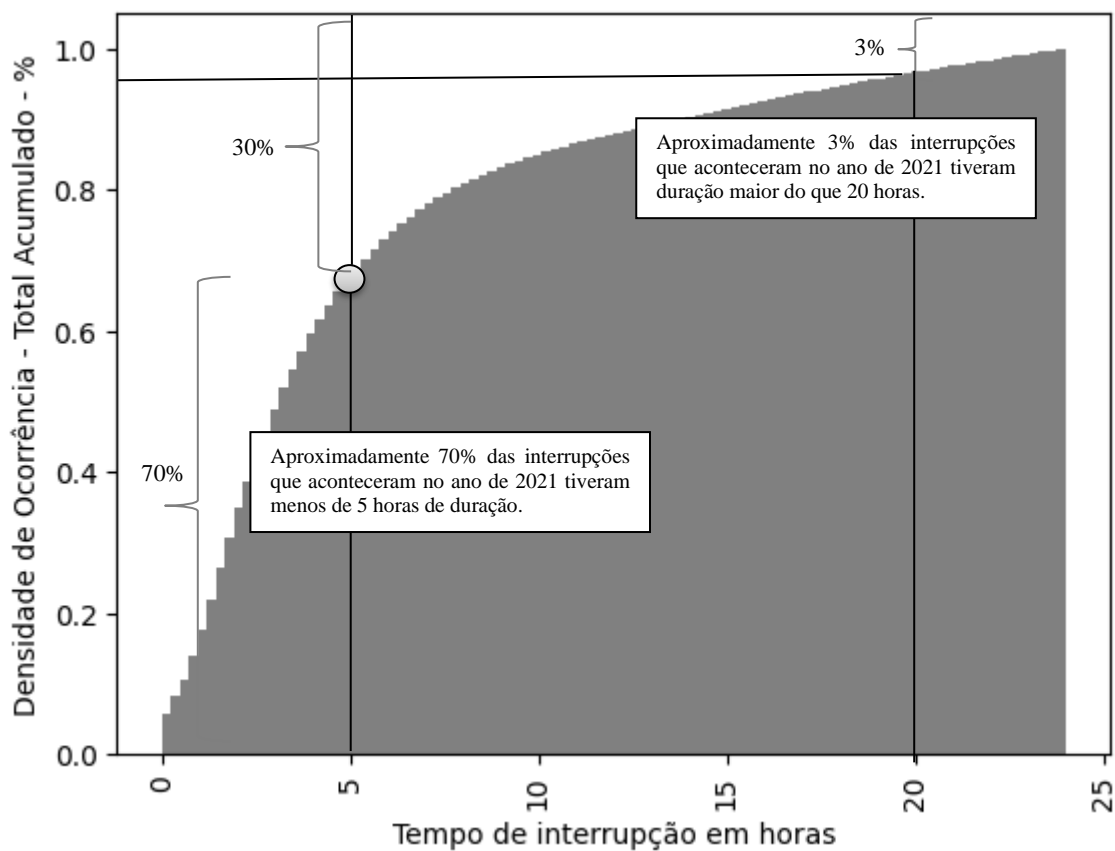
Para fins de visualização da distribuição dos dados nesse espaço bidimensional (DEC e FEC), primeiramente foi realizada uma divisão em 4 grupos (como descritos na seção 4.2), levando em consideração o critério de divisão de 5 horas para a duração dos eventos de interrupção e a média geral para a quantidade destas ocorrências. Esses critérios de separação foram definidos de forma empírica a partir de testes com o conjunto de dados.

Nas figuras 23 e 24 que são referentes, respectivamente, ao histograma da duração das interrupções e ao histograma acumulado desta mesma grandeza, é possível observar que 70% das interrupções ocorreram com duração entre 0 e 5 horas em 2021 (parte do gráfico em que há um crescimento elevado). Após o valor de 5 horas é possível notar que o gráfico parece começar a apresentar um crescimento menor na curva acumulada da Figura 24. Após esse período é possível perceber uma suavização da curva com um crescimento mais lento, contemplando aproximadamente 30% dos casos, que são os que têm mais do que 5 horas de duração. Nota-se também, pelas figuras 23 e 24, que alguns eventos tiveram mais de 20 horas de duração, acompanhando-se o eixo horizontal do tempo de interrupções. Avaliando em relação ao acumulado, na Figura 24, esses eventos contabilizam aproximadamente 3% do total de casos. Na Tabela 6 é exposta uma análise levando em consideração as divisões em relação aos pontos de corte de 5 horas para o retorno do serviço e a média geral da quantidade de interrupções.

Após ser realizada a análise inicial, é aplicado o algoritmo k-médias visando identificar os agrupamentos segundo os próprios dados. Para definir a quantidade de clusters é utilizado o método do cotovelo (*elbow*) (Sec. 3.8). Após a aplicação do método de *Elbow*, pode ser observado o resultado mostrado na Figura 25.



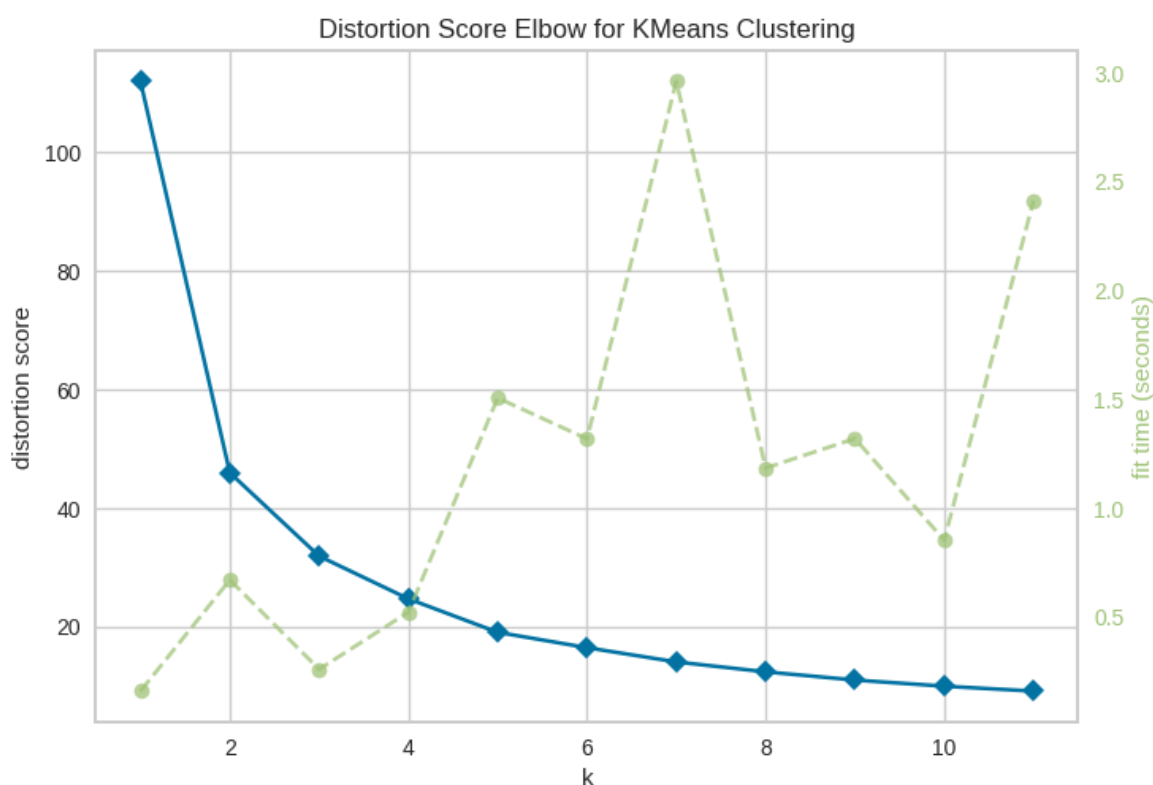
**Figura 23 - Histograma do tempo duração de interrupções em horas, para o ano de 2021.**



**Figura 24 - Densidade acumulada do tempo de duração de interrupções em horas, para o ano de 2021.**

**Tabela 6 – Quantidade de interrupções vs. tempo para restabelecimento do serviço.**

	Tempo de retorno do serviço < 5 horas	Tempo de retorno do serviço > 5 horas
Quantidade de Interrupções maior que a média geral da amostra	445	641 (grupo de mais afetado)
Quantidade de Interrupções menor do que a média geral da amostra	851	1011

**Figura 25 – Aplicação do Método Elbow para identificação do número de Cluster**

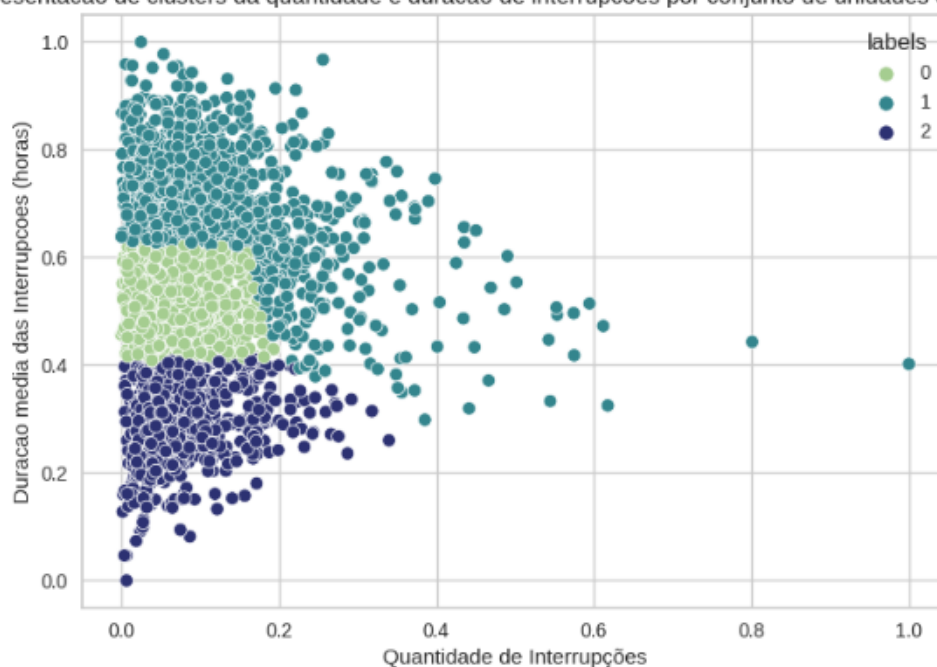
Como pode ser notado, até o um número de *clusters* igual a 4 — ponto em que o eixo horizontal  $k$  tem o valor igual a 4 — há uma redução significativa do valor da distorção (eixo vertical), tendo como base a análise da curva em azul, no gráfico da Figura 25. Após quatro *clusters* a taxa de diminuição se mostra menos acentuada.

Os pontos em que  $k=3$  e  $k=4$  parecem, pelo gráfico, serem as escolhas em que a curva forma um “cotovelo” e, assim, a distorção começa a decair de forma mais lenta. A curva em

verde marca o tempo em segundos para execução do modelo para cada valor de  $k$  e tanto para  $k=3$ , como  $k=4$ , o tempo de execução é menor do que 2 segundos. Considerando que o tempo não é um limitante, foi optado pela escolha do valor de  $k=4$  para o número de clusters como resultado do método. Dado que não há uma convenção aceita de forma geral para um critério de parada em relação a taxa de diminuição da distorção na aplicação do método de *elbow*, a escolha de 4 *clusters* foi selecionada como o número ideal de grupamentos.

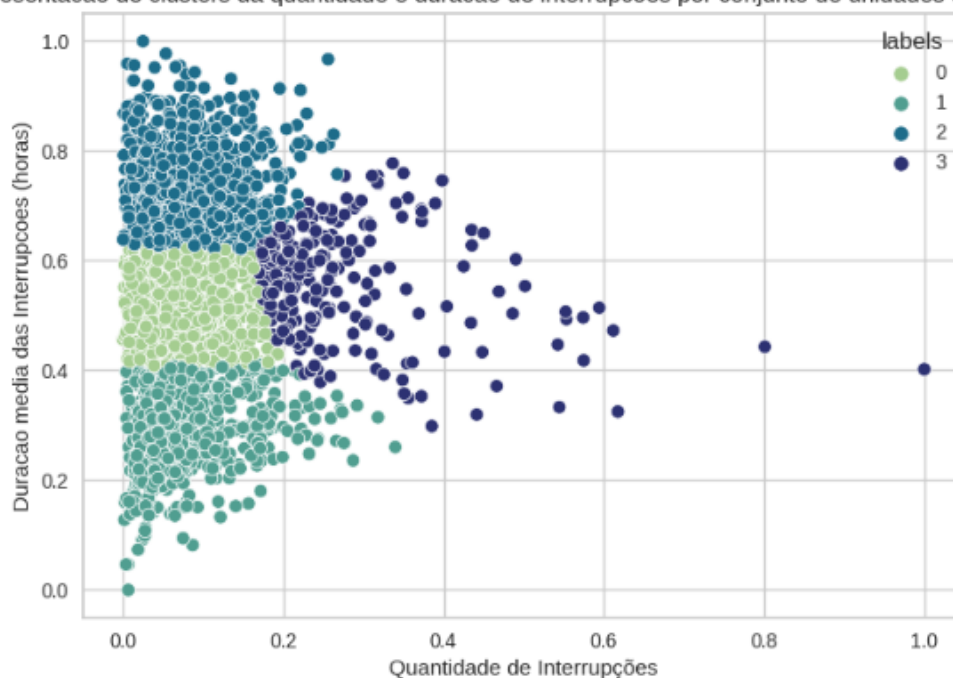
Os resultados da aplicação do  $k$ -médias sendo definidos 3 e 4 clusters são evidenciados, respectivamente, nas figuras 26 e 27. A clusterização foi aplicada considerando o conjunto de interrupções ocorridas durante o ano de 2021.

Representação de clusters da quantidade e duração de interrupções por conjunto de unidades consumidoras



**Figura 26 – Resultado do método  $k$ -means com a definição de 3 clusters ( $k=3$ ) em relação ao intervalo de tempo médio de duração dos eventos de interrupção e a quantidade total de interrupções mensais.**

Representação de clusters da quantidade e duração de interrupções por conjunto de unidades consumidoras



**Figura 27 – Resultado do método *K-means* com a definição dos 4 cluster ( $k=4$ ) em relação ao intervalo de tempo médio de duração dos eventos de interrupção e a quantidade total de interrupções mensais.**

Concluída a fase de clusterização, os conjuntos de unidades consumidoras recebem um rótulo de associação com o respectivo cluster ao qual pertence. A informação que determina cada cluster é considerada como uma nova característica para o modelo e utilizada no seu treinamento como uma nova variável independente, sinalizando para o modelo de regressão múltipla sobre a diferença entre os conjuntos no que diz respeito ao valor do DEC e do FEC.

Como pode ser notado pela figura 27, com a utilização de quatro clusters, os agrupamentos 0, 1 e 2 são horizontalmente separados, estabelecendo os níveis baixo, médio e alto de intervalo de duração dos eventos de parada no fornecimento elétrico. O grupo 3 é verticalmente separado dos grupos 0, 1 e 2, o que sugere os patamares de quantidade de interrupção baixa e alta. Ou seja, o intervalo de tempo desempenhou um papel mais significativo na escolha dos clusters para a execução do algoritmo, neste conjunto de dados.

Nos testes é avaliada a projeção do modelo para os casos que a informação de cluster é incluída como uma nova característica e para o caso em que a base de dados é dividida em conjuntos distintos, dando origem a um modelo de projeção do número de interrupções para cada cluster. No primeiro caso, foi observado um resultado mais satisfatório em relação as

métricas de performance. Dessa forma, essas informações foram incluídas como um novo atributo do estimador. Sendo elas categóricas, suas representações na matriz de dados foram feitas a partir de 3 colunas ( $k-1$  colunas, assim como descrito em 4.4) contendo os valores 1 ou 0 a partir do pertencimento ou não do registro ao cluster. Após a inclusão do novo atributo, levando em consideração o fluxo da ferramenta, é realizado o teste de hipóteses utilizando-se a estatística  $F$ , e após a seleção de característica com a estatística  $T$ . Este último, tem por objetivo selecionar as variáveis representativas em relação a quantidade de interrupções nos conjuntos de unidades consumidoras.

Levando em consideração especificamente o teste com a estatística  $T$  para as informações referente aos agrupamentos (o teste com todas as características será mais bem explorado na seção 4.5), os *clusters* 1 e 3 foram os mais relevantes. Nestes foi observado o p-valor menor do que 0,05 após a execução da iteração do teste de hipóteses com a estatística  $T$ . Especificamente para a distribuidora Copel, somente o *cluster* 3 foi selecionado no final do teste.

## **5.4 Resultados da Aplicação do Teste com a Estatística F**

Após a aplicação da clusterização e uma vez que os dados tenham sido processados seguindo a aplicação dos tratamentos discutidos na seção 4.4, são então realizados os testes de hipóteses para a aplicação do modelo de regressão. Primeiramente, com isso, é avaliada a estatística  $F$ . Os resultados da aplicação do teste de hipóteses com a estatística  $F$  para o modelo de regressão linear múltipla, considerando todos os atributos de entrada coletados e definidos na seção 4.3 são expostos na tabela 7. O modelo foi estimado considerando os dados dos conjuntos de todas as distribuidoras brasileiras (a nível nacional) e de distribuidoras específicas.

**Tabela 7 – Resultados da Aplicação do Teste de hipótese para a estatística F.**

Teste de Hipóteses utilizando a estatística $F$	Estatística $F$	$p$ -value
Modelo em Nível Nacional (contendo todas as distribuidoras)	2036	0,00
Energia Sergipe	2229	0,00
Cemig	1959	0,00
Copel	2241	0,00
CPFL	2120	0.00

Tendo os estimadores apresentado valores calculados para estatística  $F$  entre 1959 e 2241 e o p-valor igual a zero, pode-se então rejeitar a hipótese nula e dar por aceita a hipótese alternativa que supõe a existência de pelo menos um coeficiente não-nulo, ou seja, pelo menos uma variável independente contribuí para a projeção do número mensal de interrupções por conjunto. Sendo assim, os estimadores treinados com as variáveis descritas nas tabelas 4, 5 e 6 podem ser considerados válidos estatisticamente em relação ao teste de hipóteses que foi aplicado.

## **5.5 Resultados da Seleção de Características com a Estatística T**

Consequente a aplicação da estatística  $F$ , o modelo de regressão múltipla é então reduzido em relação a quantidade de atributos de previsão com a aplicação do teste de hipóteses com a estatística  $T$ , buscando-se avaliar a significância estatística de cada atributo. Os resultados da quantidade de atributos antes e após a aplicação do teste de hipóteses com a estatística  $T$  são expostos na Tabela 8. Como discutido na seção 4.6, a aplicação da seleção de características utilizando a estatística  $T$  foi feita apenas para os modelos das distribuidoras específicas, não sendo aplicada ao modelo em nível nacional.



**Tabela 8 – Seleção de Atributos realizada a partir da aplicação da estatística T.**

Distribuidora	Quantidade de Atributos		
	Modelo de Regressão Múltipla Simples	Modelo de Regressão Múltipla após a seleção (p-value < 0.05)	Diferença Percentual
Energisa Sergipe	112	69	<b>- 38%</b>
Cemig	112	64	<b>- 45%</b>
Copel	112	63	<b>- 44%</b>
CPFL	112	65	<b>- 45%</b>

As variáveis selecionadas após a aplicação do teste de hipóteses utilizando-se a estatística T são então utilizadas na construção de um novo modelo.

Em relação aos gráficos apresentados na seção 4.3, que representam as relações entre as variáveis numéricas de entrada no modelo com as quantidades de interrupções por conjunto de unidades consumidoras é possível avaliar estes gráficos em relação aos coeficientes do estimador. Com isso, como foi avaliado a característica de pressão atmosférica que apresentou uma aparente relação de aumento da quantidade de interrupções, teve um coeficiente positivo para todas as distribuidoras, sendo este um atributo com um dos maiores impactos em relação as variáveis meteorológicas. A média mensal das temperaturas máximas diárias também obteve um coeficiente positivo, sendo este um atributo com o coeficiente relevante para todas as distribuidoras. A máxima rajada de vento, que aparentemente pelas figuras da seção 4.3 parece apresentar uma relação negativa com a quantidade de interrupções, mostrou um coeficiente tanto positivo quanto também significativo em toda as previsões. Esse resultado, levando-se em consideração o que foi exposto no capítulo 2 era esperado em relação a este fenômeno meteorológico. Sendo ventos mais fortes levam a maior quantidade de detritos, principalmente de árvores e galhos, caindo sobre a rede elétrica causando interrupções. Em relação ao tipo de vegetação, a classificação campinara, com características campestres e florestadas teve uma influência positiva para

todas as companhias distribuidoras. Já a vegetação ombrófila aberta de Terras Baixas apresentou um coeficiente negativo em todos os testes.

As tabelas do Anexo B (B1, B2, B3 e B4), expõem as características que foram consideradas nestes testes, e selecionadas após a aplicação da estatística T.

## 5.6 Resultados da Projeção da Quantidade de Interrupções em Nível Nacional

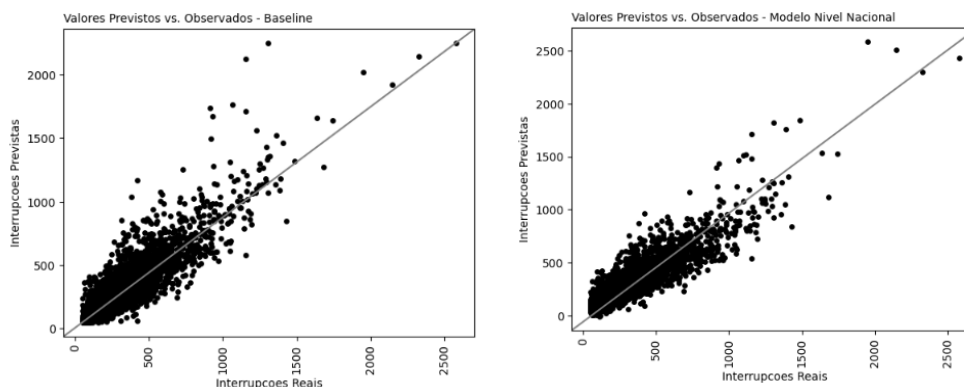
Tendo sido realizados os testes de hipóteses e a validação da significância tanto do modelo como dos seus atributos, o estimador é então testado com a utilização de diferentes bases de dados.

Em relação ao teste inicial, em nível nacional, o modelo é, em primeiro momento, avaliado levando em consideração os dados de todas as distribuidoras do Brasil, tanto para o ajuste de seus parâmetros, como para a avaliação da qualidade da inferência realizada. O ajuste do estimador é realizado com utilização dos dados referentes ao período de janeiro de 2020 a outubro de 2022. Já na avaliação utilizam-se os dados dos meses de novembro e dezembro de 2022. Dessa forma, é buscado avaliar o desempenho da ferramenta em todo o conjunto de dados. A Tabela 9, mostra a quantidade de registros utilizados para o ajuste (treino) e avaliação (teste) do estimador.

**Tabela 9 - Relação entre a quantidade de registros utilizados para teste e treino**

	Treino	Teste
Quantidade de Conjuntos Considerados	2703	2535
Quantidade de Registros Considerados	52702	4978

Na Figura 28, é possível avaliar a dispersão dos dados observados em relação aos projetados pelos modelos, para os casos de aplicação dos métodos de previsão sazonal simples (*baseline* de comparação) e de regressão múltipla, respectivamente. Na Tabela 10, são apresentados os resultados da inferência levando em consideração as métricas de avaliação abordadas.



**Figura 28 – Histograma com os resultados da ferramenta de projeção em nível nacional, utilizando o modelo *baseline* de previsão sazonal simples (esquerda) e o modelo de regressão múltipla (direita).**

**Tabela 10 - Resultados das inferências *baseline* (Método de Previsão Sazonal Simples) versus a aplicação do modelo de Regressão Linear Múltipla para a projeção em nível nacional.**

Ferramenta	$R^2$	MAPE	MSE
Baseline (Método de Naive Sazonal)	0,79	25,58	10035,31
Modelo de Regressão Linear Múltipla	0,82	26,39	8307,98
Diferença Aproximada	+4%	+1 p.p.	-17%

É possível notar, pela Figura 28, que houve uma diminuição da dispersão dos resultados, principalmente na região abaixo da diagonal que descreve os valores em que o resultado previsto é igual ao projetado. Esse resultado representa uma maior acurácia devido ao aumento no valor das interrupções previstas, já que todos os resultados nesta região representam os casos em que o modelo previu uma quantidade menor de interrupções do que o observado. É possível também validar esse resultado levando em consideração a métrica de ajuste  $R^2$  (coeficiente de determinação), sendo que essa métrica avalia o quanto da dispersão observada nos dados pode ser explicada pelo modelo. Também é possível ver uma diminuição no MSE do estimador, sendo que quanto menor essa medida menor a variância do modelo.

Foi notado nos testes que, apesar do MAPE ser uma métrica usual para avaliação de resultados de modelos de regressão, em alguns casos, ele foi enviesado pela quantidade de interrupções no conjunto. A partir da avaliação, é constatado que alguns registros apresentam baixa quantidade de interrupções historicamente e a regressão estima um resultado relevantemente superior ao observado em valor absoluto. Desta forma, por ser uma métrica percentual, o MAPE tem um valor muito alto nesses casos. Esse comportamento já não acontece no *baseline* com a utilização do modelo de previsão sazonal simples, dado que no caso deste, são repetidos os valores de anos anteriores. Por exemplo, no teste com os dados da distribuidora CEMIG, no conjunto Carlos Chagas 2, um dos valores observado de interrupções foi de 57 ocorrências, enquanto o modelo realizou a inferência de 121 casos no mesmo mês. Tendo assim, um resíduo de 113%. Este valor é bem superior à média dos resultados observados.

## **5.7 Resultado das Projeções da Quantidade de Interrupções para Distribuidoras Específicas Seleccionadas**

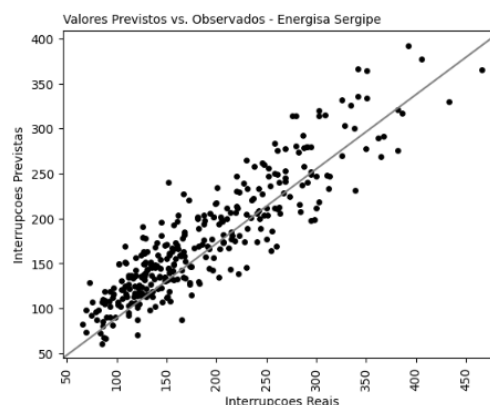
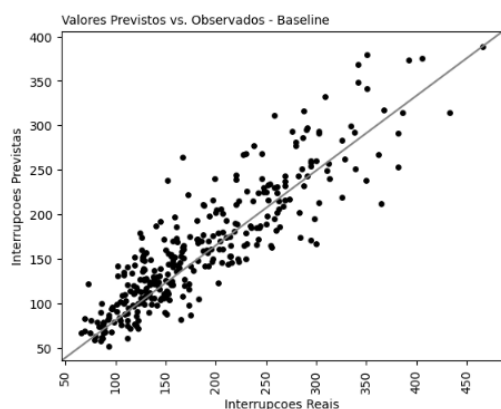
Após a projeção em nível nacional com todas as distribuidoras do Brasil, foram selecionadas determinadas concessionárias para realização de um teste regional. Dado que o objetivo da ferramenta é a previsão para concessionárias específicas tanto a seleção de atributos utilizando-se a estatística  $T$ , quanto o teste final de avaliação do modelo após esta, é realizado apenas para os testes com os dados das concessionárias.

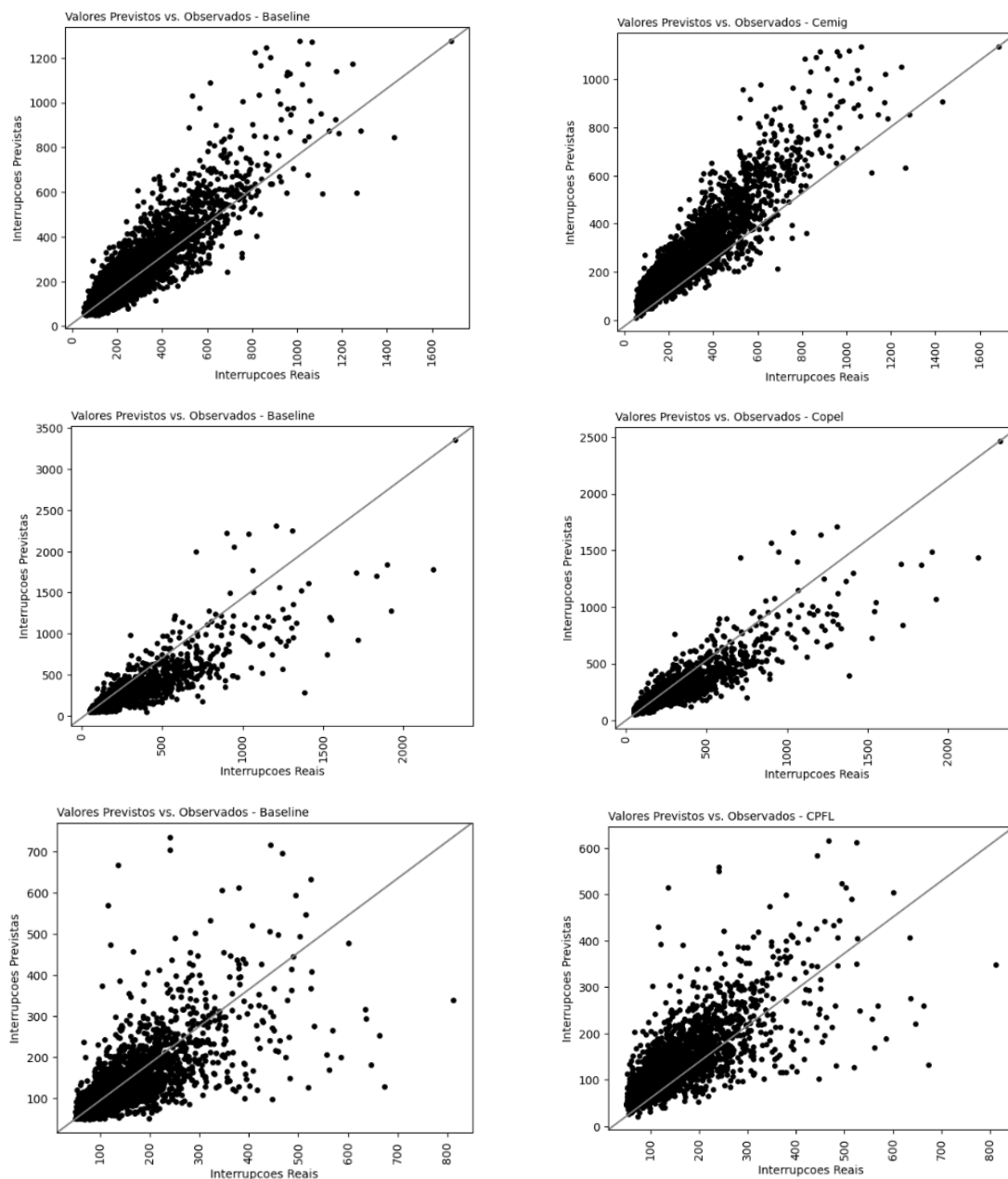
Sendo assim, é então realizada a avaliação da ferramenta levando em consideração as companhias de distribuição Energisa Sergipe, Cemig, Copel e CPFL. Os registros relacionados aos conjuntos atendidos pelas respectivas distribuidoras são retirados da base de dados utilizada para estimativa do modelo. De posse do modelo, em cada caso, são avaliados os resultados da inferência do modelo em comparação com o observado levando em consideração as mesmas métricas de avaliação utilizadas para o primeiro teste da ferramenta em nível nacional. A relação entre os registros selecionados para treino e para teste é exposta na tabela 11. Dado que cada distribuidora apresenta uma quantidade específica de conjuntos elétricos, a relação muda dependendo da concessionária de avaliação.

**Tabela 11 – Relação entre a quantidade de registros utilizados para teste e treino.**

Distribuidora	Nível avaliação	Treino	Teste
Energisa Sergipe	Quantidade de conjuntos considerados	2687	27
	Quantidade de Registros Considerados	57033	324
Cemig	Quantidade de Conjuntos Considerados	2422	292
	Quantidade de Registros Considerados	50849	3437
Copel	Quantidade de Conjuntos Considerados	2586	128
	Quantidade de Registros Considerados	54837	1408
CPFL	Quantidade de Conjuntos Considerados	2536	178
	Quantidade de Registros Considerados	53698	2021

A figura 29 apresenta os resultados de dispersão entre os valores previstos e observados, comparando-se o *baseline* (modelo de previsão sazonal simples) e a regressão múltipla. Cada linha apresenta o resultado para uma das distribuidoras seguindo a mesma ordem que foi considerada na tabela 12, sendo esta, Energisa Sergipe, Cemig, Copel e, por fim, a distribuidora CPFL.





**Figura 29 - Dispersão dos resultados da ferramenta de projeção, apresentando a quantidade de interrupções reais, no eixo horizontal, e projetado, no eixo vertical. Cada coluna apresenta, a esquerda, o modelo *baseline* e a direita o modelo de regressão linear múltipla. Em cada linha têm-se os resultados de cada distribuidora, seguindo a seguinte ordem, Energisa Sergipe, Cemig, Copel, CPLF.**

É possível observar, pela Figura 29, que em todos os testes há uma diminuição da dispersão entre os registros de previsão e observados, comparando-se os resultados dos testes com o modelo *baseline* (subfiguras à esquerda) e com o modelo de regressão (subfiguras à direita). Também é possível ver uma tendência nos resultados do modelo, uma vez que os

dados estão descentralizados da diagonal principal que descreve os pontos em que os resultados de previsão são iguais aos observados.

Para a distribuidora CPFL alguns dados apresentam alta dispersão em relação ao previsto e realizado. Esse resultado sugere que para alguns registros o modelo não tem a capacidade de inferência para informações históricas de ajuste, nem para o caso do *baseline*, nem para o modelo construído. Esses casos podem ser registros atípicos ou casos regionais que se encontram na área de distribuição da concessionária e que tem influência de alguma variável explicativa que não foi considerada na modelagem. Foi feita uma análise da dispersão dos dados para esta distribuidora para o ano de 2022, tentando avaliar a presença de registros atípicos. Nesta investigação foi constatado que a dispersão no conjunto de dados não tem alta discrepância para esta distribuidora, em relação ao total geral. A média da quantidade de interrupções por conjunto, neste caso, é de 126, tendo o valor de 181 para o terceiro quartil, e o desvio padrão fica em 85 interrupções. Já a base total apresenta um desvio padrão de aproximadamente 220 interrupções. Em relação a segunda hipótese, do modelo não apresentar variáveis explicativas suficientes, o teste T realizado na seção 5.3.2 mostra a seleção de variáveis com significância estatística comprovada, porém outras fontes de dados seriam necessárias para confirmar a correlação com outros fatores explicativos.

Na tabela 12, são mostrados os resultados das métricas de desempenho para cada uma das distribuidoras nos testes que foram realizados. Em todos os testes realizados houve uma melhora na performance do modelo confrontando com o resultado *baseline*, tendo-se a redução tanto do MSE quanto do MAPE, como também o aumento do  $R^2$ . Para os casos das distribuidoras Energisa Sergipe, Copel e CPFL foi observado um aumento de mais 10% no  $R^2$ , sendo que especificamente na distribuidora CPFL o aumento foi de 27%. Esses resultados evidenciam a capacidade do modelo de inferência para diferentes casos de estudo e a sua eficácia no ajuste em relação ao resultado comparativo.

Levando em consideração os resultados de cada distribuidora em específico, a distribuidora CPFL apresentou uma métrica de ajuste  $R^2$  baixa em comparação com as demais companhias que foram testadas, tendo tido o valor de 0,42 para esta medida. Esse resultado pode ser explicado, como já citado, pela dificuldade de ajuste das características avaliadas em relação a quantidade de interrupções nesta distribuidora, tanto pelo modelo *baseline*, quanto pelo modelo de regressão. Todavia, importante destacar que mesmo neste caso, pode ser verificada uma melhora nas métricas de avaliação com a aplicação do modelo construído.

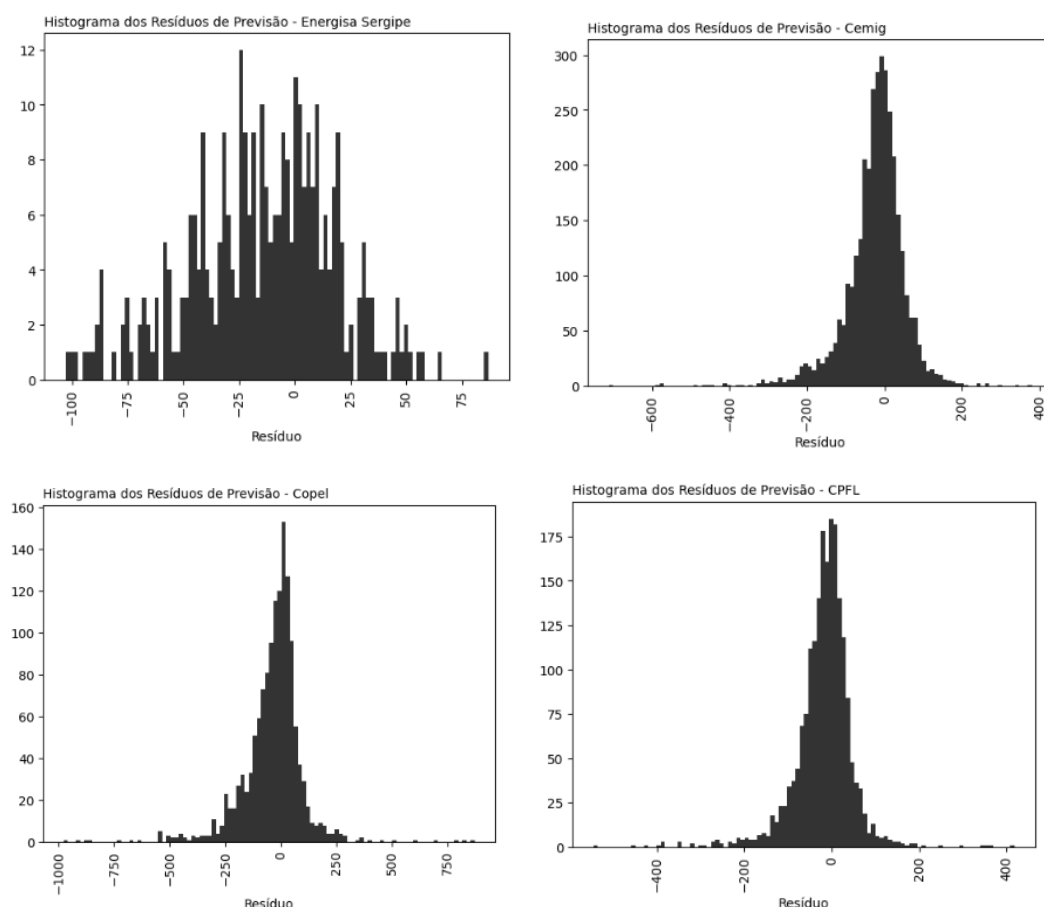
**Tabela 12 - Resultados das inferências *baseline* (método de *Naive Sazonal*) vs. a aplicação do modelo de Regressão Linear Múltipla para distribuidora Energisa Sergipe, Cemig e Copel.**

Distribuidora	Ferramenta	R2	MAPE	MSE
Energisa Sergipe	<i>Baseline</i> (Método de <i>Naive Sazonal</i> )	0,70	17,88	1823,76
	Modelo de Regressão Linear Múltipla	0,79	15,00	1258,46
	Variação Aproximada para o <i>baseline</i>	<b>+11%</b>	<b>-3 p.p.</b>	<b>-31%</b>
Cemig	<i>Baseline</i> (Método de <i>Naive Sazonal</i> )	0,77	22,10	7658,61
	Modelo de Regressão Linear Múltipla	0,81	20,47	6186,72
	Variação Aproximada para o <i>baseline</i>	<b>+5%</b>	<b>-2 p.p.</b>	<b>-19%</b>
Copel	<i>Baseline</i> (Método de <i>Naive Sazonal</i> )	0,64	28,65	26364,88
	Modelo de Regressão Linear Múltipla	0,74	27,18	18885,02
	Variação Aproximada para o <i>baseline</i>	<b>+16%</b>	<b>-1 p.p.</b>	<b>-28%</b>
CPFL	<i>Baseline</i> (Método de <i>Naive Sazonal</i> )	0,33	26,92	5454,95
	Modelo de Regressão Linear Múltipla	0,42	26,49	4684,76
	Variação Aproximada para o <i>baseline</i>	<b>+27%</b>	<b>-0 p.p.</b>	<b>-14%</b>



## 5.7.1 Diagnóstico dos Resíduos de Previsão

Os resíduos obtidos após a previsão utilizando o modelo de regressão linear múltipla são mostrados na Figura 30.



**Figura 30 - Histograma dos resíduos de previsão do modelo de regressão linear múltipla, para os casos de testes com cada distribuidora.**

**A cauda dos histogramas à esquerda parece um pouco mais longa, apesar da forma se aparentar a uma distribuição normal, evidenciando uma leve tendência no modelo.**

A partir da figura 30, é possível avaliar que há uma semelhança gráfica dos histogramas com a distribuição normal, apresentando, todas as curvas, uma leve cauda para a esquerda. Contudo, para se verificar esta premissa, um teste estatístico é necessário. Dessa forma, pela métrica de assimetria é possível avaliar se a distribuição tem semelhança com uma curva normal. A distribuição pode ser considerada aproximadamente simétrica (normal) contendo os seus valores de assimetria entre -0,5 e 0,5, já moderadamente simétrica, tendo os valores contidos no intervalo entre -1 e -0,5 e 0,5 e 1 e, fortemente simétrica, com os valores menores que -1 e maiores que 1 [137]. No gráfico de todas as

distribuidoras é obtida uma assimetria entre os valores de -0,52 e -1,32. Sendo o menor valor do módulo da assimetria, 0,52, apresentado pela distribuidora Energisa Sergipe. Dado que a distribuição pode ser considerada moderadamente simétrica, tendo os seus valores de curtose contidos no intervalo entre -1 e -0,5 e 0,5 e 1, e, fortemente simétrica com os valores menores que -1 e maiores que 1 [137], então as distribuições encontradas podem ser classificadas como moderadas e fortemente simétricas. Dessa forma, o modelo apresenta um desvio entre a quantidade de interrupções observadas em relação as projetadas tendo uma pequena tendência para a esquerda (prevendo um valor maior do que o realizado).

### **5.7.2 Resultados após a Seleção de Características do Modelo**

Após o teste de hipóteses com a estatística  $T$ , discutido na seção 5.5, em que são selecionados os atributos do estimador, é realizada então uma nova execução apenas com as características selecionadas. Esse processo é realizado com o objetivo de comparação entre o desempenho do modelo completo e reduzido (após a seleção de atributos).

É válido ressaltar que, sendo o teste com a estatística  $T$  comumente utilizado em modelagens de problemas de regressão, os testes anteriores foram uma etapa adicional para o estudo, tendo objetivo comparativo. Como o teste é válido estatisticamente para a otimização do desempenho da ferramenta, considerando as premissas discutidas (Sec. 3.5), a apresentação dos resultados antes e posteriormente a sua aplicação, são utilizados no estudo para avaliação da otimização do desempenho. Na tabela 13, pode ser vista a comparação entre os modelos.

**Tabela 13 - Comparação dos resultados da inferência do modelo após a seleção de atributos para os casos de teste realizados.**

Distribuidora	Ferramenta	$R^2$	MAPE	MSE
Energisa Sergipe	Baseline (Método de Previsão Sazonal Simples)	0,70	17,88	1823,76
	Modelo de Regressão Linear Múltipla	0,79	15,00	1258,46
	Modelo com aplicação da Seleção de Atributos	<b>0,78</b>	<b>14,98</b>	<b>1310,09</b>
	Variação Aproximada do Modelo de Regressão Múltipla para este após a Seleção de Atributos	<b>-1%</b>	<b>0 p.p.</b>	<b>4%</b>
Cemig	Baseline (Método de Previsão Sazonal Simples)	0,77	22,10	7658,61
	Modelo de Regressão Linear Múltipla	0,80	20,25	6406,96
	Modelo com aplicação da Seleção de Atributos	<b>0,82</b>	<b>20,27</b>	<b>5954,78</b>
	Variação Percentual do Modelo de Regressão Múltipla para este após a Seleção de Atributos	<b>2%</b>	<b>0 p.p.</b>	<b>-7%</b>
Copel	Baseline (Método de Previsão Sazonal Simples)	0,64	28,65	26364,88
	Modelo de Regressão Linear Múltipla	0,74	27,18	18885,02
	Modelo com aplicação da Seleção de Atributos	<b>0,74</b>	<b>27,32</b>	<b>18822,82</b>
	Variação Percentual do Modelo de Regressão Múltipla para este após a Seleção de Atributos	<b>0%</b>	<b>0 p.p.</b>	<b>0%</b>

CPFL	Baseline (Método Previsão Sazonal Simples)	0,33	26,92	5454,95
	Modelo de Regressão Linear Múltipla	0,42	26,49	4684,76
	Modelo com aplicação da Seleção de Atributos	<b>0,43</b>	<b>26,38</b>	<b>4632,11</b>
	Variação Percentual do Modelo de Regressão Múltipla para este após a Seleção de Atributos	<b>2%</b>	<b>0 p.p.</b>	<b>-1%</b>

É possível notar que há uma melhora no  $R^2$  em relação ao modelo *baseline* e um resultado semelhante ao modelo de regressão linear múltipla completo (contendo todas as variáveis anteriores a seleção). Ainda em relação ao  $R^2$ , o resultado é aproximadamente 2 p.p. superior ao modelo de Regressão Linear Múltipla anteriormente utilizado, porém sendo uma variação pequena e considerando o ganho de simplificação do modelo com a redução de 45% do tamanho das variáveis de modelagem, como discutido na seção 4.9, a seleção é considerada válida. Outra consideração é que este resultado é exposto apenas para algumas distribuidoras específicas do território brasileiro. Dessa forma, para outras distribuidoras o modelo pode ter um potencial de generalização maior.

## Capítulo 6 - Conclusão e Trabalhos Futuros

Levando em consideração a discussão que foi realizada nos capítulos desta dissertação, este trabalho apresentou um modelo para projeção do número mensal de interrupções por conjunto de unidades consumidoras. Essa projeção apresenta estrutura geográfica que é definida pelas informações do sistema SIGEL, utilizando fontes de dados públicas. É possível portanto, concluir que, após os testes realizados a ferramenta apresentou resultados promissores levando em consideração sua comparação com o *baseline* nas distribuidoras de teste. Foi observado, em todos os casos, a diminuição do MSE e do MAPE e a melhoria do coeficiente de determinação  $R^2$ .

Também é possível estudar a relação das características de entrada com cada previsão, verificando a confirmação positiva ou não das premissas levantadas de causa e efeito para cada informação de estudo no modelo e, assim, entender quais os atributos influenciam com menor ou maior grau o fenômeno estudado.

Foi observado no decorrer dos testes que alguns conjuntos de unidades consumidoras apresentam baixa quantidade de interrupções, fato este que leva a um MAPE alto, trazendo viés para essa medida em alguns dos testes. Foi aplicado um filtro para valores considerados muito baixos no software. Uma investigação mais profunda destes casos foi apresentada na seção 5.5.

Outro ponto de destaque, foram os resultados da divisão do conjunto de dados em clusters a partir da quantidade de interrupções e do tempo de cada interrupção. Esta clusterização mostrou-se eficaz como atributo de entrada no modelo e na melhoria do desempenho da ferramenta, tendo, após o teste de hipótese com a estatística t, alguns dos clusters mostrado relação significativa com as previsões.

A ferramenta tem potencial para ser utilizada com dados de todas as distribuidoras brasileiras, realizando as inferências e possibilitando tanto o estudo, quanto o planejamento prévio da quantidade de interrupções em sua área de concessão.

Atualmente, as fontes de dados são apenas de organizações públicas. Apesar da qualidade e do volume de dados dessas fontes, é esperado que em utilizações futuras, sendo realizada a integração com as informações internas das distribuidoras, os resultados apresentem melhora significativa. Além disso, tendo disponibilizadas além das interrupções pela ANEEL, também as ordens de serviços executados pelas distribuidoras, é esperado que estas possam ser utilizadas como informações de previsão para o modelo, aumentando assim

a abrangência da ferramenta. Outrossim, a utilização dos históricos de ordens de serviço executadas por conjunto ou por localidade viabiliza o desenvolvimento de previsores para a quantidade de serviços de diversas naturezas, apoiando os estudos de alocação de bases operacionais e dimensionamento de equipes de campo.

Para trabalhos futuros, é esperado que sejam incluídas novas informações de fontes públicas e privadas na ferramenta e que sejam realizados testes com demais distribuidoras. As informações relativas à estrutura da rede elétrica, contendo a quantidade de equipamentos, alimentadores, extensão de linhas, etc., assim como discutido na seção 2.1, pode ser acrescentada para o aumento do desempenho do estimador. Ainda é possível avaliar a inclusão de dados contendo um histórico maior ou mesmo de informações de outros países. Dado que, em testes de cenários realizados apenas com o treinamento em regiões específicas, a ferramenta mostrou resultados inferiores. Sendo assim, a premissa de que o fenômeno de interrupções esteja mais correlacionado com as características meteorológicas, vegetação, e demais causas externas do que com regionais é importante ser avaliada.

## Bibliografia

[1]	Delgado F., Victor J. (2019) "Crise energética em Roraima, deterioração da Venezuela e a posição brasileira", <i>periodicos.fgv.br</i> , <a href="https://periodicos.fgv.br/rce/article/view/80512/76865">https://periodicos.fgv.br/rce/article/view/80512/76865</a> , (accessed Jan. 12, 2024).
[2]	Gov (2023) "Ministro diz que Brasil tem sistema elétrico seguro e pede ação da PF na apuração", <i>gov.br</i> , <a href="https://www.gov.br/secom/pt-br/assuntos/noticias/2023/08/ministro-diz-que-brasil-tem-sistema-eletrico-seguro-e-pede-acao-da-pf-na-apuracao">https://www.gov.br/secom/pt-br/assuntos/noticias/2023/08/ministro-diz-que-brasil-tem-sistema-eletrico-seguro-e-pede-acao-da-pf-na-apuracao</a> , (accessed Jan. 12, 2024).
[3]	Gov (2023) "Após apagões, prefeitos do Rio se unem contra Enel", <i>angra.rj.gov.br</i> , <a href="https://www.angra.rj.gov.br/noticia.asp?vid_noticia=66540&amp;indexsigla=imp">https://www.angra.rj.gov.br/noticia.asp?vid_noticia=66540&amp;indexsigla=imp</a> , (accessed Jan. 12, 2024).
[4]	CNN (2023) "Enel diz que reestabeleceu energia de 70% dos clientes após segundo apagão em SP", <i>cnnbrasil.com.br</i> , <a href="https://www.cnnbrasil.com.br/nacional/enel-diz-que-restabeleceu-energia-de-70-dos-clientes-apos-segundo-apagao-em-sp/">https://www.cnnbrasil.com.br/nacional/enel-diz-que-restabeleceu-energia-de-70-dos-clientes-apos-segundo-apagao-em-sp/</a> , (accessed Jan. 12, 2024).
[5]	Office of the Press Secretary. Presidential Policy Directive/Ppd-21 (2013) Critical Infrastructure Security and Resilience; The White House: Washington, DC, USA.
[6]	J. Shi, J. Wan, H. Yan, and H. Suo (2011) "A survey of cyber-physical systems," in <i>Wireless Communications and Signal Processing (WCSP)</i> , 2011 International Conference on. IEEE, pp. 1-6.
[7]	E. Zio. (2009) "Reliability engineering: Old problems and new challenges," <i>Reliability Engineering &amp; System Safety</i> , vol. 94, no. 2, pp. 125-141.
[8]	Securing The U.S. Electrical Grid, (2014) "Understanding the Threats to the Most Critical of Critical Infrastructure, While Securing a Changing Grid"; Center for the Study of the Presidency & Congress: Washington, DC, USA.
[9]	Pessanha, J. F. M., Souza, R. C., & Laurencel, L. da C. (2007). "Um modelo de análise envoltória de dados para o estabelecimento de metas de continuidade do fornecimento de energia elétrica. <i>Pesquisa Operacional</i> ", 27(1), 51–83. <a href="https://doi.org/10.1590/s0101-74382007000100004">https://doi.org/10.1590/s0101-74382007000100004</a> .
[10]	N. Safaei, D. Banjevic, A.K.S. Jardine (2011) "Bi-objective workforce-constrained maintenance scheduling: a case study" <i>J. Oper. Res. Soc.</i> , 62, pp. 1005-1018, <a href="https://doi.org/10.1057/jors.2010.51">10.1057/jors.2010.51</a> .
[11]	R. J. Campbell. (2012) "Weather-related power outages and electric system resiliency". Congressional Res. Service, Washington, DC, USA, Tech. Rep. 7-5700.
[12]	Maravilha L. André. (2018) "Scheduling maneuvers for the restoration of electric power distribution networks: Formulation and heuristics" <i>Electr. Power Syst. Res.</i> , 163, pp. 301-309, <a href="https://doi.org/10.1016/j.epsr.2018.06.020">10.1016/j.epsr.2018.06.020</a> .
[13]	Reed DA. (2008) "Electric utility distribution analysis for extreme winds. <i>Journal of Wind Engineering and Industrial Aerodynamics</i> ", 2008; 96(1): 123–140.
[14]	Liu H, Davidson RA, Rosowsky DV, Stedinger JR. (2005) "Negative binomial regression of electric power outages in hurricanes" <i>Journal of Infrastructure Systems</i> ; 11(4): 258–267.
[15]	Han SR, Guikema SD, Quiring SM. (2009) "Improving the predictive accuracy of hurricane power outage forecasts using generalized additive models. <i>Risk Analysis</i> "; 29(10): 1443–1453.

[16]	S. Frizzo Stefenon et al., (2020) "Electrical insulator fault forecasting based on a Wavelet Neuro-Fuzzy System," <i>Energies</i> , vol. 13, no. 2, p. 484.
[17]	M. Yue, T. Toto, M. P. Jensen, S. E. Giangrande, and R. Lofaro (2018) "A Bayesian approach-based outage prediction in electric utility systems using radar measurement data," <i>IEEE Trans. Smart Grid</i> , vol. 9, no. 6, pp. 6149–6159.
[18]	Borba, B. S. M. C. <i>et al.</i> (2019) "A review on optimization methods for workforce planning in electrical distribution utilities" <i>Computers &amp; industrial engineering</i> , 135, pp. 286–298. doi: 10.1016/j.cie.2019.06.002.
[19]	Steiner, M., Einhardt da Silva Costa, C., Maria Berthodi Costa, D., Andretta Filho, É., & Zambenedetti, V. (2009) "Técnicas da pesquisa operacional aplicadas à logística de atendimento aos usuários de uma rede de distribuição de energia elétrica" (Vol. 1).
[20]	Moura, L., & Melo, M. (2012) "Otimização da logística das equipes do plantão através da aplicação da teoria das filas", <i>Caso Celpe</i> (pp. 1–9). Retrieved from: <a href="http://www.mfap.com.br/pesquisa/arquivos/20121217141222-1222.pdf">http://www.mfap.com.br/pesquisa/arquivos/20121217141222-1222.pdf</a> .
[21]	Paragon (n.d.). Projeto Cemig. Retrieved March 13, 2018, from Paragon website: <a href="http://www.paragon.com.br/cases/cemig/">http://www.paragon.com.br/cases/cemig/</a> .
[22]	Garcia, Vinicius Jaques, Bernardon, D. P., De Araujo Bassi, O., Abaide, A., De Melo Reck, W., Gundel, M., & Fernandes, J. (2012) "Emergency work orders in electric distribution utilities: From business process definition to quantitative improvements" (pp. 1–5). <a href="https://ieeexplore.ieee.org/document/6398676">https://ieeexplore.ieee.org/document/6398676</a> .
[23]	G. Franzese, L., Fioroni, M., Pinheiro, L., & Eustachio Soares, J. (2006) "Allocating Field Service Teams with Simulation in Energy/Utilities Environment", (pp. 516–520). <a href="https://dx.doi.org/10.1145/1218112.1218210">https://dx.doi.org/10.1145/1218112.1218210</a> .
[24]	Ferreira, V.H.; Correa, R.L.d.S.; Colombini, A.C.; Fortes, M.Z.; Mello, F.L.d.; Araujo, F.C.C.d.; Pereira, N.R. (2021) "Big Data Analytics for Spatio-Temporal Service Orders Demand Forecasting in Electric Distribution Utilities", <i>Energies</i> 14, no.23:7991. <a href="https://doi.org/10.3390/en14237991">https://doi.org/10.3390/en14237991</a> .
[25]	Ferreira, V.H. et al., "SIDECSistema Inteligente para Dimensionamento de Equipes de Campo – Relatório Etapa 2", Universidade Federal Fluminense-HOC Soluções em TI, 2018.
[26]	F. Martinez Alvarez, A. Troncoso, J. C. Riquelme, and J. S. Aguilar Ruiz (2011) "Energy time series forecasting based on pattern sequence similarity," <i>IEEE Trans. Knowl. Data Eng.</i> , vol. 23, no. 8, pp. 1230–1243.
[27]	Ferreira, V. H. et al. (2016) "A survey on intelligent system application to fault diagnosis in electric power system transmission lines," <i>Electric power systems research</i> , 136, pp. 135–153. doi: 10.1016/j.epsr.2016.02.002.
[28]	Purarjomandlangrudi, A., Ghapanchi, A. H. and Esmalifalak, M. (2014) "A data mining approach for fault diagnosis: An application of anomaly detection algorithm", <i>Measurement: journal of the International Measurement Confederation</i> , 55, pp. 343–352. doi: 10.1016/j.measurement.2014.05.029.
[29]	Souza, J. C. S. et al. (2001) "Fault location in electrical power systems using intelligent systems techniques", <i>IEEE transactions on power delivery</i> , 16(1), pp. 59–67. doi: 10.1109/61.905590.



[30]	Davidson, E. M. et al. (2006) "Applying multi-agent system technology in practice: Automated management and analysis of SCADA and digital fault recorder data," IEEE transactions on power systems: a publication of the Power Engineering Society, 21(2), pp. 559–567. doi: 10.1109/tpwrs.2006.873109.
[31]	Behkam, R. et al. (2022) "Generalized regression neural network application for fault type detection in distribution transformer windings considering statistical indices", COMPEL. The International Journal for Computation and Mathematics in Electrical and Electronic Engineering, 41(1), p. 381–409. doi: 10.1108/compel-06-2021-0199.
[32]	Zimnickas, T. et al. (2020) "A technique for frequency converter-fed asynchronous motor vibration monitoring and fault classification, applying continuous wavelet transform and convolutional neural networks", Energies, 13(14), p. 3690. doi: 10.3390/en13143690.
[33]	Klomjit, J. e Ngaopitakkul, A. (2020) "Comparison of artificial intelligence methods for fault classification of the 115-kV hybrid transmission system", Applied sciences (Basel, Switzerland), 10(11), p. 3967. doi: 10.3390/app10113967.
[34]	Skowron, M. et al. (2019) "Application of self-organizing neural networks to electrical fault classification in induction motors", Applied sciences (Basel, Switzerland), 9(4), p. 616. doi: 10.3390/app9040616.
[35]	Patel, D. et al. (2018) "Design and development of fault classification algorithm based on relevance vector machine for power transformer," IET electric power applications, 12(4), pp. 557–565. doi: 10.1049/ietepa.2017.0562.
[36]	Isermann R. (2005) <i>Fault-diagnosis systems an introduction from fault detection to fault tolerance</i> . Springer.
[37]	Bhasker, S. K. et al. (2015) "Differential protection of indirect symmetrical phase shift transformer and internal faults classification using wavelet and ANN", in TENCON 2015 - 2015 IEEE Region 10 Conference. IEEE.
[38]	Sadi, M.A.H.; AbuHussein, A.; Shoeb, M.A. (2021) "Transient Performance Improvement of Power Systems Using Fuzzy Logic Controlled Capacitive-Bridge Type Fault Current Limiter", IEEE Trans. Power Syst. 36, 323–335.
[39]	Rigatos, G.; Serpanos, D.; Zervos, N. (2017) "Detection of Attacks Against Power Grid Sensors Using Kalman Filter and Statistical Decision Making", IEEE Sens. J., 17, 7641–7648.
[40]	Madurai Elavarasan, R., Shafiullah, G. M., Raju, K., Mudgal, V., Arif, M. T., Jamal, T., Subramanian, S., Sriraja Balaguru, V. S., Reddy, K. S., & Subramaniam, U. (2020) COVID-19: Impact analysis and recommendations for power sector operation. Applied Energy, 279(115739), 115739. <a href="https://doi.org/10.1016/j.apenergy.2020.115739">https://doi.org/10.1016/j.apenergy.2020.115739</a> .
[41]	S. Leva, M. Mussetta, and E. Ogliari (2019) "PV module fault diagnosis based on microconverters and day-ahead forecast," IEEE Trans. Ind. Electron., vol. 66, no. 5, pp. 3928–3937.
[42]	S. Frizzo Stefenon et al. (2020) "Electrical insulator fault forecasting based on a Wavelet Neuro-Fuzzy System," Energies, vol. 13, no. 2, p. 484.
[43]	Y. Hu, J. Li, M. Hong, J. Ren, and Y. Man (2022) "Industrial artificial intelligence based energy management system: Integrated framework for electricity load forecasting and fault prediction," Energy (Oxf.), vol. 244, no. 123195, p. 123195.

[44]	N. W. Branco, M. S. M. Cavalca, S. F. Stefenon, and V. R. Q. Leithardt (2022) "Wavelet LSTM for fault forecasting in electrical power grids," <i>Sensors (Basel)</i> , vol. 22, no. 21, p. 8323.
[45]	Ribeiro, M.H.D.M.; da Silva, R.G.; Moreno, S.R.; Mariani, V.C.; Coelho, L.S. (2022) "Efficient bootstrap stacking ensemble learning model applied to wind power generation forecasting", <i>Int. J. Electr. Power Energy Syst.</i> , 136, 107712.
[46]	Liu H, Davidson RA, Rosowsky DV, Stedinger JR. (2005) "Negative binomial regression of electric power outages in hurricanes", <i>Journal of Infrastructure Systems</i> ; 11(4): 258–267.
[47]	Han SR, Guikema SD, Quiring SM. (2009) Improving the predictive accuracy of hurricane power outage forecasts using generalized additive models. <i>Risk Analysis</i> ; 29(10): 1443–1453.
[48]	Han SR, Guikema SD, Quiring SM, Lee K, Rosowsky D, Davidson RA. (2009) "Estimating the spatial distribution of power outages during hurricanes in the Gulf Coast region", <i>Reliability Engineering and System Safety</i> ; 94(2): 199–210.
[49]	Nateghi R, Guikema SD, Quiring, SM. (2014) "Power outage estimation for tropical cyclones: Improved accuracy with simpler models". <i>Risk Analysis</i> , 2014; 34(6): 1069–1078.
[50]	F. Yang, D. Cerrai, and E. N. Anagnostou (2021) "The effect of lead-time weather forecast uncertainty on outage prediction modeling," <i>Forecasting</i> , vol. 3, no. 3, pp. 501–516.
[51]	Wanik, D.; Anagnostou, E.; Hartman, B.; Frediani, M.; Astitha, M. (2015) "Storm outage modeling for an electric distribution net-work in northeastern", USA. <i>Nat. Hazards</i> , 79, 1359–1384.
[52]	Cerrai, D.; Wanik, D.W.; Bhuiyan, A.E.; Zhang, X.; Yang, J.; Frediani, M.E.B.; Anagnostou, E.N.; Bhuiyan, M.E. (2019) "Predicting storm outages through new representations of weather and vegetation", <i>IEEE Access</i> 2019, 7, 29639–29654.
[53]	Arif, A.; Wang, Z. (2018) "Distribution Network Outage Data Analysis and Repair Time Prediction Using Deep Learning", In <i>Proceedings of the 2018 IEEE International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)</i> , Boise, ID, USA, 23–28; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, USA; pp. 1–6.
[54]	Eskandarpour, R.; Khodaei, A. (2016) "Machine learning based power grid outage prediction in response to extreme events". <i>IEEE Trans. Power Syst.</i> 32, 3315–3316.
[55]	Eskandarpour, R.; Khodaei, A. (2017) "Leveraging accuracy-uncertainty tradeoff in SVM to achieve highly accurate outage pre-dictions", <i>IEEE Trans. Power Syst.</i> 2017, 33, 1139–1141.
[56]	He, J.; Cheng, M.X. (2021) "Machine learning methods for power line outage identification" <i>Electr. J.</i> , 34, 106885.
[57]	He, J.; Cheng, M.X.; Fang, Y.; Crow, M.L. (2019) "A machine learning approach for line outage identification in power systems" In <i>Proceedings of the 4th International Conference on Machine Learning, Optimization, and Data Science, LOD 2018, Volterra, Italy, 13–16 September 2018</i> ; Springer Science and Business Media LLC: Berlin, Germany; pp. 482–493.

[58]	Reilly, A.C.; Tonn, G.L.; Zhai, C.; Guikema, S.D (2017) "Hurricanes and Power System Reliability-The Effects of Individual Decisions and System-Level Hardening" Proc. IEEE 2017, 105, 1429–1442.
[59]	Owerko, D.; Gama, F.; Ribeiro, A. (2018) "Predicting Power Outages Using Graph Neural Networks". In Proceedings of the 2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Anaheim, CA, USA, 26–29 November; pp. 743–747.
[60]	Mendonça, M. J. C. et al. (2023) "Service quality performance indicators for electricity distribution in Brazil" Utilities policy, 80(101481), p. 101481. doi: 10.1016/j.jup.2022.101481.
[61]	Aneel (2021) "ren2021948", aneel.gov.br, <a href="https://www2.aneel.gov.br/cedoc/ren2021948.html">https://www2.aneel.gov.br/cedoc/ren2021948.html</a> , (accessed Nov. 16, 2023).
[62]	Aneel (2017) "Módulo 8 Revisão 8", aneel.gov.br, <a href="https://antigo.aneel.gov.br/documents/656827/14866914/M%C3%B3dulo8_Revisao_8/9c78cfab-a7d7-4066-b6ba-cfbda3058d19">https://antigo.aneel.gov.br/documents/656827/14866914/M%C3%B3dulo8_Revisao_8/9c78cfab-a7d7-4066-b6ba-cfbda3058d19</a> , (accessed Nov. 16, 2023).
[63]	E. Mariani, F. Mastroianni, and V. Romano (1984) "Field experiences in reenergization of electrical networks from thermal power plant," IEEE Trans. Power App. Syst., vol. PAS-103, pp. 1707–1713.
[64]	P. Gomes, A. C. S. de Lima and A. de Padua Guarini (2004) "Guidelines for power system restoration in the Brazilian system," in <i>IEEE Transactions on Power Systems</i> , vol. 19, no. 2, pp. 1159-1164, doi: 10.1109/TPWRS.2004.825862.
[65]	ANEEL (2023) "webappviewer", <a href="https://sigel.aneel.gov.br/">https://sigel.aneel.gov.br/</a> , <a href="https://sigel.aneel.gov.br/portal/apps/webappviewer/index.html?id=c1716d81d491426197bf497ace41db8d">https://sigel.aneel.gov.br/portal/apps/webappviewer/index.html?id=c1716d81d491426197bf497ace41db8d</a> (accessed Oct. 23, 2023).
[66]	SFE/ANEEL (2023) "Interrupcoes de Energia Eletrica nas Redes de Distribuicao", <a href="https://dadosabertos.aneel.gov.br">dadosabertos.aneel.gov.br</a> , Interrupções de Energia Elétrica nas Redes de Distribuição - Conjunto de dados - Dados Abertos - Agência Nacional de Energia Elétrica (aneel.gov.br) (accessed Oct. 21, 2023).
[67]	"ONS (2021) "Reevaluation of Restoration Procedures in the North and Northeast Region", (in Portuguese), ONS, Rio de Janeiro, Brazil, Tech. Rep. ons-3 073/2001.
[68]	Santos, G.M (2023) "History of Electric Power in Brazil: The "Electromemory" Project", <a href="https://eletromemoria.fflch.usp.br">eletromemoria.fflch.usp.br</a> , <a href="https://eletromemoria.fflch.usp.br/sites/eletromemoria.fflch.usp.br/files/projeto_de_pesquisa_eletromemoria_1_ingles.pdf">https://eletromemoria.fflch.usp.br/sites/eletromemoria.fflch.usp.br/files/projeto_de_pesquisa_eletromemoria_1_ingles.pdf</a> , (accessed Oct. 22, 2023).
[69]	Aneel (2021) "ren20211000", aneel.gov.br, <a href="https://www2.aneel.gov.br/cedoc/ren20211000.html">https://www2.aneel.gov.br/cedoc/ren20211000.html</a> , (accessed Nov. 10, 2023).
[70]	C. Liu, C. Chen, P. Li and Z. Chen (2023) "Optimal Scheduling of Distribution Network Maintenance Personnel Based on The Northern Goshawk Optimization Algorithm," 2023 IEEE 12th Data Driven Control and Learning Systems Conference (DDCLS), Xiangtan, China, pp. 506-510, doi: 10.1109/DDCLS58216.2023.10165975.
[71]	Perrier N., Agard B., Baptiste P., Frayret J.-M., Langevin A., Pellerin R., Riopel D., Trépanier M. (2013a) "A survey of models and algorithms for emergency

	response logistics in electric distribution systems. Part I: Reliability planning with fault considerations" <i>Comput. Oper. Res.</i> , 40 (7) (2013), pp. 1895-1906, 10.1016/j.cor.2013.01.016.pplying Data Science: Business Case Studies Using SAS. SAS Institute.
[72]	Perrier, N., Agard, B., Baptiste, P., Frayret, J.-M., Langevin, A., Pellerin, R., Riopel, D., & Trépanier, M. (2013b) "A survey of models and algorithms for emergency response logistics in electric distribution systems. Part II: Contingency planning level" <i>Computers &amp; Operations Research</i> , 40(7), 1907–1922.
[73]	Wan, H., Chen, F. F., & Kuringer, G. W. (2011) "An intelligent decision support system for workforce forecast", Retrieved from University of Texas website: <a href="http://www.dtic.mil/dtic/tr/fulltext/u2/a537920.pdf">http://www.dtic.mil/dtic/tr/fulltext/u2/a537920.pdf</a> .
[74]	Ward, D. (1996) "Workforce demand forecasting techniques. <i>Human Resource Planning</i> ," 19(1), 54-55.
[75]	Berg, A. and Haubrich, H.-J. (2008) "Calculation of cost-minimal resources for the operation of electrical networks," in 2008 5th International Conference on the European Electricity Market. IEEE.
[76]	Safaei, N., Banjevic, D. and Jardine, A. K. S. (2012) "Workforce planning for power restoration: An integrated simulation-optimization approach," <i>IEEE transactions on power systems : a publication of the Power Engineering Society</i> , 27(1), pp. 442–449. doi: 10.1109/tpwrs.2011.2166090.
[77]	Yao, M.-J. and Min, K. J. (1998) "Repair-unit location models for power failures," <i>IEEE transactions on engineering management</i> , 45(1), pp. 57–65. doi: 10.1109/17.658661.
[78]	Wu, J. S. et al. (2005) "A fuzzy rule-based system for crew management of distribution systems in large-scale multiple outages," in 2004 International Conference on Power System Technology, 2004. PowerCon 2004. IEEE.
[79]	Langevin, A., Perrier, N., Agard, B., Baptiste, P., Frayret, J.-M., Pellerin, R., Riopel, D. and Trépanier, M. (2009) "Models and methods for emergency response logistics, in electric distribution systems," in <i>Proc. 2nd Int. Workshop Successful Strategies in Supply Chain Management</i> , Hong Kong, Jan. 8–9.
[80]	Duffuaa, S.O., Bendaya, M., Al-Sultan, K. S. and Andijani A. A. (2003) "Stochastic simulation model for maintenance manpower planning," in <i>Proc. 7th ISSAT Int. Conf. Reliability and Quality in Design</i> , pp. 230–236.
[81]	Choudhari, S. and Gajjar, H. (2018) "Simulation modeling for manpower planning in electrical maintenance service facility," <i>Business process management journal</i> , 24(1), pp. 89–104. doi: 10.1108/bpmj-06-2016-0125.
[82]	Shield, S. A. et al. (2021) "Major impacts of weather events on the electrical power delivery system in the United States," <i>Energy (Oxford, England)</i> , 218(119434), p. 119434. doi: 10.1016/j.energy.2020.119434.
[83]	Burillo, D. (2019) "Effects of climate change in electric power infrastructures," in <i>Power System Stability</i> IntechOpen.

[84]	Mukherjee, S., Nateghi, R. and Hastak, M. (2018) "A multi-hazard approach to assess severe weather-induced major power outage risks in the U.S.," <i>Reliability Engineering &amp; System Safety</i> , 175, pp. 283–305. doi: 10.1016/j.res.2018.03.015.
[85]	Kezunovic, M., I. Dobson, and Y. Dong, (2008) "Impact of extreme weather on power system blackouts and forced outages: New challenges" <i>Seventh Balkan Power Conf.</i> , Šibenik, Croatia, Univerza v Ljubljani, 5 pp.
[86]	Wang, Y., Mo, Y., Wang, M., Zhou, X., Liang, L., & Zhang, P. (2018). Impact of conductor temperature time–space variation on the power system operational state. <i>Energies</i> , 11(4), 760. <a href="https://doi.org/10.3390/en11040760">https://doi.org/10.3390/en11040760</a>
[87]	Ward, D. M. (2013) "The effect of weather on grid systems and the reliability of electricity supply," <i>Climatic change</i> , 121(1), pp. 103–113. doi: 10.1007/s10584-013-0916-z.
[88]	Hines P, Apt J, Talukdar S (2009) "Large blackouts in North America: historical trends and policy implications" <i>Energy Policy</i> 37(12):5249–5259. doi:10.1016/j.enpol.2009.07.049.
[89]	EPRI (2006a) "Outline of guide for application of transmission line surge arresters - 42 to 765 kV: Extended Outline" Electric Power Research Institute, Palo Alto.
[90]	EPRI (2007) "AC flashovers on Henan Power 500-kV lines during rain. Electric Power Research Institute," Palo Alto.
[91]	Martikainen A, Pykälä M-L, Farin J (2007) "Recognizing climate change in electricity network design and construction. VTT Technical Research Centre of Finland" <a href="http://www.vtt.fi/inf/pdf/tiedotteet/2007/T2419.pdf">http://www.vtt.fi/inf/pdf/tiedotteet/2007/T2419.pdf</a> . Accessed 13 May 2013.
[92]	Mitchell JW (2009) "Power lines and catastrophic wildland fire in Southern California" In <i>Fire &amp; Materials 2009 Conference</i> , January 26 2009, San Francisco, California, USA.
[93]	Sunrise Powerlink Project (2008) "Attachment 1A: Effect of wildfires on transmission line reliability" California Public Utilities Commission. <a href="http://www.cpuc.ca.gov/Environment/info/asp/sunrise/deir/apps/a01/App%201%20ASR%20z_Attm%201A-Fire%20Report.pdf">http://www.cpuc.ca.gov/Environment/info/asp/sunrise/deir/apps/a01/App%201%20ASR%20z_Attm%201A-Fire%20Report.pdf</a> . Accessed 13 May 2013.
[94]	Hassan, M. ul <i>et al.</i> (2021) "Degradation assessment of in-service aerial bundled cables in coastal areas leading to prognosis using infrared thermography," <i>IET Generation, Transmission and Distribution</i> , 15(8), pp. 1348–1356. doi: 10.1049/gtd2.12106.
[95]	Shahid, M. A. <i>et al.</i> (2020) "Health diagnosis scheme for in-service low voltage Aerial Bundled Cables using super-heterodyned airborne Ultrasonic testing," <i>Electric power systems research</i> , 180(106162), p. 106162. doi: 10.1016/j.epsr.2019.106162.
[96]	Shahid, M. A. <i>et al.</i> (2021) "Novel health diagnostics schemes analyzing corona discharge of operational aerial bundled cables in coastal regions," <i>IEEE transactions on power delivery</i> , 36(1), pp. 84–91. doi: 10.1109/tpwrd.2020.2974751.

[97]	Goldstein, E. M. (1960) "The corrosion and oxidation of metals: Scientific principles and practical applications (Evans, ulick R.)," <i>Journal of chemical education</i> , 37(12), p. 662. doi: 10.1021/ed037p662.2.
[98]	Tomashov N.D. (1996) <i>Theory of Corrosion and Protection of Metals. The Science of Corrosion.</i> The MacMillan Co.; New York, NY, USA.
[99]	Rozenfel'd, I. L., and Greco, E. C. (1972) <i>Atmospheric Corrosion of Metals / I.L. Rozenfeld ; English Language Ed. Translated by Boris H. Tytell ; Edited by E.C. Greco.</i> Houston, Tex.: NACE, Print.
[100]	Barton K. (1976) <i>Protection against Atmospheric Corrosion.</i> John Wiley and Sons; New York, NY, USA.
[101]	Feliu S., Morcillo M. (1982) <i>Corrosión y Protección de los Metales en la Atmósfera.</i> Bellaterra; Barcelona, Spain. ISBN: 9788472900318.
[102]	Graedel T.E., Mc Gill R. (1986) Degradation of materials in the atmosphere. <i>Environ. Sci. Technol.</i> 20:1093–1100. doi: 10.1021/es00153a003.
[103]	Kucera V., Mattsson E. (1987) Atmospheric corrosion. In: Mansfeld F., editor. <i>Corrosion Mechanisms.</i> Marcel Dekker; New York, NY, USA. pp. 211–284.
[104]	Ambler, H. R. and Bain, A. A. J. (2007) "Corrosion of metals in the tropics," <i>Journal of applied chemistry</i> , 5(9), pp. 437–467. doi: 10.1002/jctb.5010050901.
[105]	Hansen, M. C. et al. (2003) "Global percent tree cover at a spatial resolution of 500 meters: First results of the MODIS vegetation continuous fields algorithm," <i>Earth interactions</i> , 7(10), pp. 1–15. doi: 10.1175/1087-3562(2003)007<0001:gptcaa>2.0.co;2.
[106]	Instituto Brasileiro de Geografia e Estatística (IBGE) (2017) <a href="http://web.archive.org/web/20170822164417/http://www.ibge.gov.br/home/geociencias/geografia/default_div_int.shtm">http://web.archive.org/web/20170822164417/http://www.ibge.gov.br/home/geociencias/geografia/default_div_int.shtm</a> , (copy archived Ago. 22. 2017.).
[107]	Instituto Brasileiro de Geografia e Estatística (IBGE) (1990) <a href="http://web.archive.org/web/20190120140831/http://biblioteca.ibge.gov.br/visualizacao/livros/liv2269_1.pdf">http://web.archive.org/web/20190120140831/http://biblioteca.ibge.gov.br/visualizacao/livros/liv2269_1.pdf</a> , (copy archived Jan. 20. 2019.).
[108]	Miles, S. B., Jagielo, N. and Gallagher, H. (2016) "Hurricane Isaac power outage impacts and restoration," <i>Journal of infrastructure systems</i> , 22(1), p. 05015005. doi: 10.1061/(asce)is.1943-555x.0000267.
[109]	Quinteros, F., Carrión, D. and Jaramillo, M. (2022) "Optimal power systems restoration based on energy quality and stability criteria," <i>Energies</i> , 15(6), p. 2062. doi: 10.3390/en15062062.
[110]	Adibi, M. M. and Fink, L. H. (1994) "Power system restoration planning," <i>IEEE transactions on power systems : a publication of the Power Engineering Society</i> , 9(1), pp. 22–28. doi: 10.1109/59.317561.
[111]	Liu, Y., Fan, R., & Terzija, V. (2016). Power system restoration: a literature review from 2006 to 2016. <i>Journal of Modern Power Systems and Clean Energy</i> , 4(3), 332–341. <a href="https://doi.org/10.1007/s40565-016-0219-2">https://doi.org/10.1007/s40565-016-0219-2</a> .

[112]	P. D. Allison, (1998) <i>Multiple Regression: A Primer</i> . Thousand Oaks, CA: Pine Forge. E Access 2019, 7, 29639–29654.
[113]	D. Maulud and A. M. Abdulazeez, (2020) “A Review on Linear Regression Comprehensive in Machine Learning”, JASTT, vol. 1, no. 2, pp. 140-147, Dec..
[114]	Stock, J. H.; Watson, M. W. <i>Introduction to Econometrics</i> . 3. ed. Addison-Wesley Series in Economics, v. 1. Addison-Wesley, 2010.
[115]	Jobson J. D. (1999) <i>Applied Multivariate Data Analysis</i> . Volume I: Regression and Experimental Design, Springer. DOI 10.1007/978-1-4612-0955-3.
[116]	Chai, T., & Draxler, R. R. (2014) "Root mean square error (RMSE) or mean absolute error (MAE)?" <a href="https://doi.org/10.5194/gmdd-7-1525-2014">https://doi.org/10.5194/gmdd-7-1525-2014</a> .
[117]	Lee, D. K., In, J., & Lee, S. (2015) "Standard deviation and standard error of the mean", Korean Journal of Anesthesiology, 68(3), 220–223. <a href="https://doi.org/10.4097/kjae.2015.68.3.220">https://doi.org/10.4097/kjae.2015.68.3.220</a> .
[118]	Svolba, G. (2019) <i>Applying Data Science: Business Case Studies Using SAS</i> . SAS Institute.
[119]	Chein F., (2019) <i>Introdução a modelos de Regressão Linear</i> , Enap. (accessed Nov. 23, 2023). ISBN: 978-85-256-0115-5.
[120]	Ren, J. et al. (2022) “Balanced MSE for imbalanced visual regression,” in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.
[121]	Chicco, D., Warrens, M. J. and Jurman, G. (2021) “The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation,” PeerJ. Computer science, 7(e623), p. e623. doi: 10.7717/peerj-cs.623.
[122]	Jobson, J. D. (2012) <i>Applied multivariate data analysis: Volume II: Categorical and multivariate methods</i> . 1992nd ed. New York, NY: Springer.
[123]	Allen, M. P. (1997) <i>Understanding Regression Analysis</i> . Netlibrary.
[124]	Marutho, D. et al. (2018) “The determination of cluster number at k-mean using elbow method and purity evaluation on headline news,” in 2018 International Seminar on Application for Technology of Information and <i>Communication</i> . IEEE.
[125]	Zhao W. (2018) "k-means: A revisit", Neurocomputing, Volume 291, 2018, Pages 195-206, ISSN 0925-2312, DOI: <a href="https://doi.org/10.1016/j.neucom.2018.02.072">https://doi.org/10.1016/j.neucom.2018.02.072</a> .
[126]	Block H. (2018) "Origins and extensions of the k-means algorithm" <i>Jornal Electronique d'Histoire des Probabilités et de la Statistique</i> . Vol 4, no2.
[127]	Diday, E. (1971): Une nouvelle methode de classification automatique et reconnaissance des formes: la méthode des nuées dynamiques. <i>Revue de Statistique Appliquée</i> XIX (2), 1970, 19-33.
[128]	Diday, E. (1973): The dynamic clusters method in nonhierarchical clustering. <i>Intern. Journal of Computer and Information Sciences</i> 2 (1), 61-88.

[129]	Diday, E., Schroeder, A. (1974): The dynamic clusters method in pattern recognition. In: J.L. Rosenfeld (ed.): Information Processing 74. Proc. IFIP Congress, Stockholm, August 1974. North Holland, Amsterdam, 691-697.
[130]	Diday, E. <i>et al.</i> (1979): Optimisation en classification automatique. Vol. I, II. Institut National der Recherche en Informatique et en Automatique (INRIA), Le Chesnay, France.
[131]	Vinod, H.D. (1969): Integer programming and the theory of grouping. J. Amer. Statist. Assoc. 64, 506-519.
[132]	Massart, D.L., Plastria, E., Kaufman, L. (1983): Non-hierarchical clustering with MASLOC. Pattern Recognition 16, 507-516.
[133]	Spath, H. (1980) <i>Cluster Analysis Algorithms for Data Reduction and Classification of Objects</i> , English translation: Cluster analysis algorithms for data reduction and classification of objects. Ellis Horwood Ltd., Chichester, UK.
[134]	Bholowalia P., Kumar A. (2014) "EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN" Internantional Journal of Computer Applications - Volume 105 - No. 9.
[135]	Lima F. M. (2023). Modelo Baseado em Dados para Previsão do Número Mensal de Interrupções em Conjuntos Elétricos Utilizando Dados Públicos - Repositório: v0 (Versão V0). Zenodo. <a href="https://doi.org/10.5281/zenodo.10033882">https://doi.org/10.5281/zenodo.10033882</a>
[136]	Balanda, K. P., & Macgillivray, H. L. (1988). Kurtosis: A critical review. The American Statistician, 42(2), 111–119. <a href="https://doi.org/10.1080/00031305.1988.10475539">https://doi.org/10.1080/00031305.1988.10475539</a> .
[137]	Hatem, Georges; Zeidan, Joe; Goossens, Mathijs; and Moreira, Carla (2022) "NORMALITY TESTING METHODS AND THE IMPORTANCE OF SKEWNESS AND KURTOSIS IN STATISTICAL ANALYSIS," BAU Journal - Science and Technology: Vol. 3: Iss. 2, Article 7.



## Anexo A – Correlação entre os atributos e a quantidade de interrupções.

A 1 –Valores de correlação entre os atributos do modelo e a variável de quantidade de interrupções por conjunto de unidades consumidoras.

Atributos do Modelo	Quantidade de Interrupções
MAX_PRECIPITACAO TOTAL, DIARIO (AUT)(mm)	0,170
MAX_PRESSAO ATMOSFERICA MEDIA DIARIA (AUT)(mB)	0,059
MAX_TEMPERATURA DO PONTO DE ORVALHO MEDIA DIARIA (AUT)(Â°C)	0,164
MAX_TEMPERATURA MAXIMA, DIARIA (AUT)(Â°C)	0,074
MAX_TEMPERATURA MEDIA, DIARIA (AUT)(Â°C)	0,113
MAX_TEMPERATURA MINIMA, DIARIA (AUT)(Â°C)	0,142
MAX_UMIDADE RELATIVA DO AR, MEDIA DIARIA (AUT)(%)	0,098
MAX_UMIDADE RELATIVA DO AR, MINIMA DIARIA (AUT)(%)	0,094
MAX_VENTO, RAJADA MAXIMA DIARIA (AUT)(m/s)	0,067
MAX_VENTO, VELOCIDADE MEDIA DIARIA (AUT)(m/s)	-0,034
MED_PRECIPITACAO TOTAL, DIARIO (AUT)(mm)	0,184
MED_PRESSAO ATMOSFERICA MEDIA DIARIA (AUT)(mB)	0,056
MED_TEMPERATURA DO PONTO DE ORVALHO MEDIA DIARIA (AUT)(Â°C)	0,166
MED_TEMPERATURA MAXIMA, DIARIA (AUT)(Â°C)	0,073
MED_TEMPERATURA MEDIA, DIARIA (AUT)(Â°C)	0,108
MED_TEMPERATURA MINIMA, DIARIA (AUT)(Â°C)	0,143
MED_UMIDADE RELATIVA DO AR, MEDIA DIARIA (AUT)(%)	0,112
MED_UMIDADE RELATIVA DO AR, MINIMA DIARIA (AUT)(%)	0,115
MED_VENTO, RAJADA MAXIMA DIARIA (AUT)(m/s)	0,014
MED_VENTO, VELOCIDADE MEDIA DIARIA (AUT)(m/s)	-0,042
TOTAL_PRECIPITACAO TOTAL, DIARIO (AUT)(mm)	0,214
Pop estimada 2021	-0,087
Produto Interno Bruto, a preços correntes (R\$ 1.000)	-0,084
Produto Interno Bruto per capita, a preços correntes (R\$ 1,00)	-0,062

FLAG_LIT	0,013
MUN_MINIMUM_BOUND_RADIO	0,406
AREA	0,337
NUMERO_CONSUMIDORES_DO_CONJUNTO	0,499
NUMERO_UNIDADES_CONSUMIDORAS	0,679
FLAG_INTERRUPCOES	1,000
percent_tree_cover	0,140
FLAG_INTERRUPCOES_lognorm	0,796

## Anexo B – Atributos considerados após a aplicação da estatística T.

**B 1 - Variáveis selecionadas após a aplicação do teste com a estatística T no caso avaliado da distribuidoras Energisa Sergipe.**

1	MAX_PRECIPITACAO TOTAL, DIARIO (AUT)(mm)
2	TOTAL_PRECIPITACAO TOTAL, DIARIO (AUT)(mm)
3	MED_PRECIPITACAO TOTAL, DIARIO (AUT)(mm)
4	MAX_TEMPERATURA DO PONTO DE ORVALHO MEDIA DIARIA (AUT)(Â°C)
5	MAX_TEMPERATURA MAXIMA, DIARIA (AUT)(Â°C)
6	MAX_UMIDADE RELATIVA DO AR, MINIMA DIARIA (AUT)(%)
7	MAX_VENTO, RAJADA MAXIMA DIARIA (AUT)(m/s)
8	MED_TEMPERATURA MAXIMA, DIARIA (AUT)(Â°C)
9	MED_UMIDADE RELATIVA DO AR, MEDIA DIARIA (AUT)(%)
10	MED_UMIDADE RELATIVA DO AR, MINIMA DIARIA (AUT)(%)
11	MED_VENTO, VELOCIDADE MEDIA DIARIA (AUT)(m/s)
12	MUN_MINIMUM_BOUND_RADIO
13	NUMERO_CONSUMIDORES_DO_CONJUNTO
14	percent_tree_cover
15	mes_12
16	mes_2
17	mes_3

18	mes_6
19	mes_7
20	mes_8
21	mes_9
22	UF_AL
23	UF_AM
24	UF_BA
25	UF_DF
26	UF_GO
27	UF_MA
28	UF_MS
29	UF_MT
30	UF_PA
31	UF_PB
32	UF_PE
33	UF_PR
34	UF_RJ
35	UF_RN
36	UF_RR
37	UF_SC
38	UF_SP
39	UF_TO
40	Amazônia Legal_Sim
41	Semiárido_Sim
42	veg_class_Campinarana Florestada
43	veg_class_Estepe Gramineo-Lenhosa
44	veg_class_Floresta Ombrófila/Floresta Estacional
45	veg_class_Floresta Estacional Semidecidual Montana
46	veg_class_Floresta Estacional Semidecidual Submontana
47	veg_class_Floresta Ombrófila Densa Aluvial
48	veg_class_Floresta Ombrófila Densa Montana

49	veg_class_Floresta Ombrófila Densa Submontana
50	veg_class_Floresta Ombrófila Densa Terras Baixas
51	veg_class_Massa Dagua Continental
52	veg_class_Savana Arborizada
53	veg_class_Savana Estépica/Floresta Estacional
54	veg_class_Savana Florestada
55	veg_class_Savana Gramineo-Lenhosa
56	veg_class_Savana/Floresta Ombrófila
57	veg_class_Vegetação Ombrófila Aberta Submontana
58	veg_class_Vegetação Ombrófila Aberta Terras Baixas
59	veg_class_Vegetação Secundária e Atividades Agrárias
60	veg_class_áreas das Formações Pioneiras Vegetação com Influência Fluvial e/ou Lacustre
61	veg_class_áreas das Formações Pioneiras Vegetação com Influência Fluvio-marinha
62	veg_class_áreas das Formações Pioneiras Vegetação com Influência Marinha
63	labels_1
64	labels_3
65	VERAO
66	OUTONO
67	INVERNO
68	PRIMAVERA
69	FLAG_INTERRUPCOES_SET_LAST_YEAR

**B 2 - Variáveis selecionadas após a aplicação do teste com a estatística T no caso avaliado da distribuidoras Energisa Cemig.**

1	TOTAL_PRECIPITACAO TOTAL, DIARIO (AUT)(mm)
2	MED_PRECIPITACAO TOTAL, DIARIO (AUT)(mm)
3	MAX_TEMPERATURA DO PONTO DE ORVALHO MEDIA DIARIA (AUT)(Â°C)
4	MAX_TEMPERATURA MAXIMA, DIARIA (AUT)(Â°C)
5	MAX_UMIDADE RELATIVA DO AR, MINIMA DIARIA (AUT)(%)

6	MAX_VENTO, RAJADA MAXIMA DIARIA (AUT)(m/s)
7	MED_TEMPERATURA MAXIMA, DIARIA (AUT)(Â°C)
8	MED_UMIDADE RELATIVA DO AR, MEDIA DIARIA (AUT)(%)
9	MED_UMIDADE RELATIVA DO AR, MINIMA DIARIA (AUT)(%)
10	MED_VENTO, RAJADA MAXIMA DIARIA (AUT)(m/s)
11	MUN_MINIMUM_BOUND_RADIO
12	NUMERO_CONSUMIDORES_DO_CONJUNTO
13	percent_tree_cover
14	mes_11
15	mes_12
16	mes_2
17	mes_3
18	mes_6
19	mes_7
20	UF_AL
21	UF_AM
22	UF_BA
23	UF_DF
24	UF_GO
25	UF_MA
26	UF_MS
27	UF_MT
28	UF_PA
29	UF_PB
30	UF_PE
31	UF_PR
32	UF_RJ
33	UF_RN
34	UF_RR
35	UF_SC
36	UF_SP

37	UF_TO
38	Amazônia Legal_Sim
39	Semiárido_Sim
40	veg_class_Campinarana Florestada
41	veg_class_Estepe Gramineo-Lenhosa
42	veg_class_Floresta Ombrófila/Floresta Estacional
43	veg_class_Floresta Estacional Semidecidual Montana
44	veg_class_Floresta Estacional Semidecidual Submontana
45	veg_class_Floresta Ombrófila Densa Montana
46	veg_class_Floresta Ombrófila Densa Submontana
47	veg_class_Floresta Ombrófila Densa Terras Baixas
48	veg_class_Massa Dagua Continental
49	veg_class_Savana Arborizada
50	veg_class_Savana Florestada
51	veg_class_Savana Gramineo-Lenhosa
52	veg_class_Savana/Floresta Ombrófila
53	veg_class_Vegetação Ombrófila Aberta Submontana
54	veg_class_Vegetação Ombrófila Aberta Terras Baixas
55	veg_class_Vegetação Secundária e Atividades Agrárias
56	veg_class_áreas das Formações Pioneiras Vegetação com Influência Fluvial e/ou Lacustre
57	veg_class_áreas das Formações Pioneiras Vegetação com Influência Fluvio-marinha
58	veg_class_áreas das Formações Pioneiras Vegetação com Influência Marinha
59	labels_1
60	labels_3
61	VERAO
62	OUTONO
63	INVERNO
64	PRIMAVERA

**B 3 - Variáveis selecionadas após a aplicação do teste com a estatística T no caso avaliado da distribuidoras Copel.**

1	MAX_PRECIPITACAO TOTAL, DIARIO (AUT)(mm)
2	TOTAL_PRECIPITACAO TOTAL, DIARIO (AUT)(mm)
3	MED_PRECIPITACAO TOTAL, DIARIO (AUT)(mm)
4	MAX_TEMPERATURA MAXIMA, DIARIA (AUT)( $^{\circ}$ C)
5	MAX_UMIDADE RELATIVA DO AR, MINIMA DIARIA (AUT)(%)
6	MAX_VENTO, RAJADA MAXIMA DIARIA (AUT)(m/s)
7	MED_TEMPERATURA MAXIMA, DIARIA (AUT)( $^{\circ}$ C)
8	MED_UMIDADE RELATIVA DO AR, MINIMA DIARIA (AUT)(%)
9	MED_VENTO, VELOCIDADE MEDIA DIARIA (AUT)(m/s)
10	MUN_MINIMUM_BOUND_RADIO
11	AREA
12	NUMERO_CONSUMIDORES_DO_CONJUNTO
13	percent_tree_cover
14	mes_12
15	mes_2
16	mes_6
17	mes_7
18	UF_AL
19	UF_AM
20	UF_BA
21	UF_DF
22	UF_GO
23	UF_MA
24	UF_MS
25	UF_MT
26	UF_PA
27	UF_PB
28	UF_PE
29	UF_PR

30	UF_RJ
31	UF_RN
32	UF_RO
33	UF_RR
34	UF_SC
35	UF_SP
36	UF_TO
37	Amazônia Legal_Sim
38	Semiárido_Sim
39	veg_class_Campinarana Florestada
40	veg_class_Estepe Gramineo-Lenhosa
41	veg_class_Floresta Ombrófila/Floresta Estacional
42	veg_class_Floresta Estacional Semidecidual Montana
43	veg_class_Floresta Estacional Semidecidual Submontana
44	veg_class_Floresta Ombrófila Densa Montana
45	veg_class_Floresta Ombrófila Densa Submontana
46	veg_class_Floresta Ombrófila Densa Terras Baixas
47	veg_class_Massa Dagua Continental
48	veg_class_Savana Arborizada
49	veg_class_Savana Estépica/Floresta Estacional
50	veg_class_Savana Florestada
51	veg_class_Savana Gramineo-Lenhosa
52	veg_class_Savana/Floresta Ombrófila
53	veg_class_Vegetação Ombrófila Aberta Submontana
54	veg_class_Vegetação Ombrófila Aberta Terras Baixas
55	veg_class_Vegetação Secundária e Atividades Agrárias
56	veg_class_áreas das Formações Pioneiras Vegetação com Influência Fluvial e/ou Lacustre
57	veg_class_áreas das Formações Pioneiras Vegetação com Influência Fluvio-marinha
58	veg_class_áreas das Formações Pioneiras Vegetação com Influência Marinha
59	labels_3



60	VERAO
61	OUTONO
62	INVERNO
63	PRIMAVERA

**B 2 - Variáveis selecionadas após a aplicação do teste com a estatística T no caso avaliado da distribuidoras CPFL.**

1	MAX_PRECIPITACAO TOTAL, DIARIO (AUT)(mm)
2	TOTAL_PRECIPITACAO TOTAL, DIARIO (AUT)(mm)
3	MED_PRECIPITACAO TOTAL, DIARIO (AUT)(mm)
4	MAX_TEMPERATURA MAXIMA, DIARIA (AUT)(Â°C)
5	MAX_UMIDADE RELATIVA DO AR, MINIMA DIARIA (AUT)(%)
6	MAX_VENTO, RAJADA MAXIMA DIARIA (AUT)(m/s)
7	MED_TEMPERATURA MAXIMA, DIARIA (AUT)(Â°C)
8	MED_UMIDADE RELATIVA DO AR, MEDIA DIARIA (AUT)(%)
9	MED_UMIDADE RELATIVA DO AR, MINIMA DIARIA (AUT)(%)
10	MED_VENTO, VELOCIDADE MEDIA DIARIA (AUT)(m/s)
11	MUN_MINIMUM_BOUND_RADIO
12	NUMERO_CONSUMIDORES_DO_CONJUNTO
13	percent_tree_cover
14	mes_12
15	mes_2
16	mes_3
17	mes_6
18	mes_7
19	UF_AL
20	UF_AM
21	UF_BA
22	UF_DF
23	UF_GO
24	UF_MA
25	UF_MS
26	UF_MT
27	UF_PA
28	UF_PB
29	UF_PE
30	UF_PR
31	UF_RJ
32	UF_RN
33	UF_RR
34	UF_SC
35	UF_SP
36	UF_TO
37	Amazônia Legal_Sim
38	Semiárido_Sim
39	veg_class_Campinarana Florestada
40	veg_class_Estepe Gramineo-Lenhosa
41	veg_class_Floresta Ombrófila/Floresta Estacional
42	veg_class_Floresta Estacional Semidecidual Montana
43	veg_class_Floresta Estacional Semidecidual Submontana

44	veg_class_Floresta Ombrófila Densa Montana
45	veg_class_Floresta Ombrófila Densa Submontana
46	veg_class_Floresta Ombrófila Densa Terras Baixas
47	veg_class_Massa Dagua Continental
48	veg_class_Savana Arborizada
49	veg_class_Savana Estépica/Floresta Estacional
50	veg_class_Savana Florestada
51	veg_class_Savana Gramineo-Lenhosa
52	veg_class_Savana/Floresta Ombrófila
53	veg_class_Vegetação Ombrófila Aberta Submontana
54	veg_class_Vegetação Ombrófila Aberta Terras Baixas
55	veg_class_Vegetação Secundária e Atividades Agrárias
56	veg_class_áreas das Formações Pioneiras Vegetação com Influência Fluvial e/ou Lacustre
57	veg_class_áreas das Formações Pioneiras Vegetação com Influência Fluvio-marinha
58	veg_class_áreas das Formações Pioneiras Vegetação com Influência Marinha
59	labels_1
60	labels_3
61	VERAO
62	OUTONO
63	INVERNO
64	PRIMAVERA