UNIVERSIDADE DO VALE DO RIO DOS SINOS - UNISINOS
UNIDADE ACADÊMICA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA
NÍVEL MESTRADO PROFISSIONAL

RICARDO DOS SANTOS COSTA

PCR: UM MODELO HÍBRIDO PARA PREVISÃO DE GRANDEZAS ELÉTRICAS APLICADO EM ESTUDO DE CASO DE UM REGULADOR DE TENSÃO EM OPERAÇÃO

São Leopoldo
2022

Ricardo dos Santos Costa

PCR: UM MODELO HÍBRIDO PARA PREVISÃO DE GRANDEZAS ELÉTRICAS APLICADO EM ESTUDO DE CASO DE UM REGULADOR DE TENSÃO EM OPERAÇÃO

Dissertação apresentada como requisito parcial para a obtenção do título de Mestre, pelo Programa de Pós-Graduação em Engenharia Elétrica da Universidade do Vale do Rio dos Sinos – UNISINOS

Orientador: Dr. Jorge Luis Victória Barbosa

Coorientador: Dr. Paulo Ricardo da Silva Pereira

São Leopoldo

2022

C838p Costa, Ricardo dos Santos.
PCR: um modelo híbrido para previsão de grandezas elétricas aplicado em estudo de caso de um regulador de tensão em operação / Ricardo dos Santos Costa. – 2022.
114 f. : il. ; 30 cm.

Dissertação (mestrado) – Universidade do Vale do Rio dos Sinos, Programa de Pós-Graduação em Engenharia Elétrica, 2022.
"Orientador: Dr. Jorge Luis Victória Barbosa
Coorientador: Dr. Paulo Ricardo da Silva Pereira."

1. Pré-processamento de dados. 2. Classificação. 3. Predição de grandezas elétricas. 4. Aprendizado de máquina. I. Título.

CDU 621.3

Dados Internacionais de Catalogação na Publicação (CIP)
(Bibliotecária: Amanda Schuster – CRB 10/2517)

Este trabalho é dedicado a memória de minha mãe, Julia Maris dos Santos Costa, que sempre se mostrou presente na minha vida acadêmica, o seu apoio foi o que possibilitou esse caminho, embora o tempo não tenha deixado ela participar dessa conquista, a sua vontade em me ver voando mais alto prevalece, e por isso eu continuo me esforçando.

“A tarefa não é tanto ver aquilo que ninguém viu, mas pensar o que ninguém ainda pensou sobre aquilo que todo mundo vê”.
Arthur Schopenhauer

AGRADECIMENTOS

Agradeço a todos que me apoiaram em minha trajetória acadêmica. A vida muitas vezes apresenta um o caminho difícil, mas as pessoas queridas nos dão força para vencer qualquer desafio. Agradeço a minha família pela presença e apoio. Agradeço ao orientador Dr. Jorge Luis Victória Barbosa por seu profissionalismo e incentivo. Agradeço ao meu colega e futuro doutor Jorge Arthur Schneider Aranda por todos os conselhos e encorajamentos. Agradeço em especial ao coorientador e amigo Paulo Ricardo da Silva Pereira por acreditar no meu potencial. Agradecimento especial a toda equipe da CEEE Equatorial que possibilitou e apoiou esse estudo, assim como os grandes amigos da Cooperativa Certaja Energia que sempre participaram das discussões e engrandeceram os resultados deste trabalho.

RESUMO

O crescente aumento da demanda por energia elétrica em conjunto com exigências impostas pelos órgãos reguladores têm levado o sistema de distribuição de energia elétrica convencional a evoluir para o contexto de redes inteligentes. A aquisição e análise de dados são questões centrais para a evolução do sistema de distribuição. As tecnologias de aprendizado de máquina estão ganhando espaço nos estudos aplicados, tornando esse um tema emergente, porém os estudos relacionados não abordam a preparação e análise dos dados antes da aplicação nos modelos de inteligência artificial. Nesse cenário esse trabalho responde à questão de pesquisa de como seria um modelo capaz de efetuar o pré-processamento de dados, classificá-los e realizar predições de grandezas elétricas. Em comparação com trabalhos relacionados há indicação que esse é o primeiro estudo que aborda o pré-processamento e agrupamento de dados por similaridade visando aumentar a eficácia dos modelos de inteligência artificial. Aplicando o método Fuzzy C-Means para classificação de dados são percebidos outliers que não eram encontrados antes da criação de clusters, também são percebidos pontos críticos de operação do sistema com o método de Grubbs. A etapa de regressão apresenta predições realizadas com redes neurais LSTM com até quatro passos de tempo a frente com erro médio absoluto percentual de 0,16% utilizando uma base de dados real de uma concessionária de distribuição de energia elétrica.

Palavras-Chave: Pré-processamento de dados, Classificação, Predição de grandezas elétricas, Aprendizado de máquina.

ABSTRACT

The growing demand for electricity and requirements imposed by regulatory agencies have led the conventional electricity distribution system to evolve into the context of smart grids. Data acquisition and analysis are central issues for the evolution of the distribution system. Machine learning technologies are gaining ground in applied studies, making this an emerging topic. The related studies do not address the preparation and analysis of data before application in artificial intelligence models. In this context, this work presents a model capable of performing the pre-processing, classification, and prediction of electrical quantities. Compared with related works, there is an indication that this is the first study that addresses the pre-processing and grouping of data by similarity to increase the effectiveness of artificial intelligence models. The Fuzzy C-Means method for data classification allows outliers to be found more assertively. The Grubbs method identifies critical operating points of the system. The regression stage presents predictions made with LSTM neural networks up to four-time steps ahead with a percentage absolute average error of 0.16% using a real database of an electric power distribution utility.

Keywords: Data pre-processing, Classification, Prediction of electrical quantities, Machine learning.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SIGLAS

AHP	*Analytic Hierarchy Process*
BDD	*Bad Data Detector*
BFS	*Backward-Forward Sweep*
BT	Baixa Tensão
CE	Critérios de Exclusão
CI	Critérios de Inclusão
CNN	*Convolutional Neural Network*
DBN	*Deep Belief Network*
DNN	*Deep Neural Network*
ELM	*Extreme Learning Machine*
EnKF	*Ensemble Kalman Filter*
EPSO	*Evolutionary Particle Swarm Optimization*
FDIA	Ataques de injeção de dados falsos
FPC	Fuzzy Partition Coefficient
GBM	*Gradient Boost Machines*
GCCM	*Gaussian Component Combination Method*
GD	Geração Distribuída
GM	*Generalized Maximum Likelihood*
GPQR	*Gaussian Process Quantile Regression*
HDLE	*Hybrid Deep Learning Ensemble*
LAV	*Least Absolute Value*
MM	Média Móvel
MMSE	*Minimum Mean Squares Estimation*
MOBRKGA	*Multiple Objective-Based Random Key Genetic Algorithm*
MT	Média Tensão
MTSL	*Multi-Time Scale Learning*
NTL	Perdas Não Técnicas
OLTCs	*On-Load Tap Changers*
PCA	Principal Components Analysis
PMU	*Phasor Measurement Unit*

PSOS-CGSA	*Particle Swarm Optimization with Sigmoid-based Acceleration Coefficients and Chaotic Gravitational Search Algorithm*
QE	Questões Específicas
QEST	Questão Estatística
QG	Questão Geral
REnKF	*Robust Ensemble Kalman Filter*
RL	Religador
RNA	Redes Neurais Artificiais
RT	Regulador de Tensão
RVMs	*Game-Theoretic Expansion of Relevance Vector Machines*
SC	*Spectral Clustering*
SCADA	*Supervisory Control And Data Acquisition*
SG	*Smart Grid*
SMs	*Smart Meters*
SVM	*Support Vector Machine*
TAP	Terminal de Ajuste de Potencial
TICs	Tecnologias de Informação e Comunicação
VVC	Volt-Var-Control
VVO	Volt-Var-Optimization

SUMÁRIO

1 INTRODUÇÃO

A eletricidade é um bem cada vez mais valioso para a sociedade, sendo fundamental em vários pontos da sociedade como na economia, saúde e lazer. A topologia do sistema elétrico permaneceu estável ao longo dos anos e opera com claras demarcações entre seus sistemas de geração, transmissão e distribuição, no entanto, cada setor evoluiu tecnologicamente de forma diferente de acordo com suas necessidades (BINDU; USHAKUMARI; SAVIER, 2021).

Nos últimos anos o conceito de *Smart Grid* (SG) tem ocupado maior espaço nos estudos sobre sistemas de distribuição de energia elétrica. Uma SG é formada por diferentes Tecnologias de Informação e Comunicação (TICs) e por diversos dispositivos sensoriais, que oferecem oportunidades de análise para concessionárias e consumidores (OURAHOU et al. 2020). Segundo Al-Badi et al. (2020) a SG integra relevante processamento de dados, comunicação de dados confiável e monitoramento, controle e supervisão do sistema.

A rede de distribuição de energia elétrica é geralmente uma rede radial, com diversos ramos e vários pontos de carga (YUEHAO et al. 2016). Com a integração da Geração Distribuída (GD), como a fotovoltaica e a eólica, os pontos de carga que optam por produzir energia renovável passam a ser ativos na rede. Por ser intermitente, a GD pode interferir na qualidade de energia fornecida pelas concessionárias gerando transgressões como sobre ou subtensão (MARUJO; ZANATTA; FLORÉZ, 2021). Nesse cenário, é necessário controlar a potência ativa e reativa presente no sistema. Segundo Zhang et al. (2021) as técnicas de Volt-Var-Control (VVC) e Volt-Var-Optimization (VVO) permitem regular o nível de tensão e reduzir as perdas. Ajustes inteligentes no Terminal de Ajuste de Potencial (TAP) dos Reguladores de Tensão (RTs) são utilizados para manter a tensão em um nível adequado (CANHA et al. 2017). Todas essas técnicas de ajustes nos níveis de tensão utilizam bases de dados obtidas nos sensores dos equipamentos, mostrando a importância que a correta aquisição e tratamento de dados possui na SG.

1.1 Definição do problema e questão de pesquisa

Um novo paradigma vem se tornando presente no dia a dia das concessionárias de energia, antigamente o controle dos sistemas de distribuição era

centralizado, onde um centro de operação era responsável por monitorar e controlar toda a rede de distribuição da concessionária. Com a expansão das redes de distribuição e a integração da GD esse modelo centralizado pode não ser capaz de atender a demanda de controle necessário para manter a estabilidade da rede. Com isso, cada vez mais equipamentos de medição inteligente e telecomandados estão sendo instalados na rede de distribuição para monitorar e auxiliar a conhecer o que acontece naquela determinada rede ao longo do tempo. Esses equipamentos inteligentes consideram fatores como dados cronológicos, tensão de fase, tensão de linha, potência ativa e reativa, corrente, entre outras informações.

Com o grande número de dados que esses equipamentos de medição geram, o novo desafio é encontrar formas eficientes de interpretar e utilizá-los de forma que possam auxiliar no controle da rede e acelerar o processo de tomada de decisão na operação do sistema.

Com a grande base de dados surgem também os problemas que envolvem as etapas de coleta, como perda de pacotes, repetições de leituras, aumento do armazenamento de dados, entre outros. Desta forma, a presente dissertação aborda o desenvolvimento de um modelo computacional que seja capaz de receber os dados coletados da rede de distribuição de energia elétrica, efetue o pré-processamento e entregue saídas que auxiliem nas tomadas de decisões no decorrer do dia a dia da operação do sistema.

Com isso a dissertação busca responder a seguinte questão de pesquisa: Como seria um modelo computacional que seja capaz de efetuar o pré-processamento de dados, considerando contextos locais dos dados como características de consumo de energia elétrica, e efetue a predição de grandezas elétricas que possam auxiliar na operação do sistema ou serem utilizadas em análises que melhorem o estado atual de operação de equipamentos especiais da rede de distribuição?

Com isso define-se a seguinte hipótese: “É possível criar um modelo que seja capaz de identificar padrões locais na base de dados de grandezas elétricas, efetuar o pré-processamento com base nesse contexto local e efetuar a predição da grandeza elétrica que for de interesse da concessionária de energia.”

1.2 Objetivos

O desenvolvimento da dissertação está pautado nos objetivos geral e específicos, como é apresentado nas seções a seguir.

1.2.1 Objetivo geral

Este trabalho, como objetivo geral, propõe desenvolver um modelo híbrido para efetuar o pré-processamento, classificação e predição de dados de grandezas elétricas, sejam elas medidas de tensão, corrente ou potência.

1.2.2 Objetivos específicos

De acordo com a questão de pesquisa e o objetivo geral apresentado, os objetivos específicos são os seguintes:

- realizar um estudo sobre o estado da arte das técnicas que estão sendo utilizadas para predição de grandezas elétricas;
- identificar os principais problemas encontrados na base de dados da concessionária de energia elétrica;
- reconhecer o funcionamento atual de algum equipamento especial instalado na rede de distribuição;
- especificar um modelo computacional para pré-processamento e predição de grandezas elétricas;
- identificar particularidades e padrões que possam existir nos dados de grandezas elétricas e classificá-los;
- avaliar diferentes técnicas de aprendizado de máquina aplicadas no contexto de grandezas elétricas;
- avaliar o modelo final com base em dados reais.

1.3 Método

A partir da definição do modelo computacional para pré-processamento e predição de grandezas elétricas, da implementação de um protótipo e de sua

aplicação em base de dados reais foi possível fazer avaliação da hipótese definida na seção 1.1. Para tal fim, as seguintes etapas ocorreram:

a) a primeira etapa envolveu estudos sobre os temas básicos: SG, RT, pré-processamento de dados, análise de séries temporais e aprendizado de máquina;
b) no próximo passo foi efetuada uma pesquisa sobre trabalhos que envolvam problemas de predição aplicados a rede de distribuição de energia elétrica, abordando temas como técnicas de estimação de estados, predição de grandezas elétricas, técnicas de controle de tensão, métodos de otimização aplicados a redes de distribuição e tipos de base de dados utilizadas nessas discussões;
c) a etapa consiste em um estudo preliminar sobre os principais problemas encontrados nas bases de dados das concessionárias de distribuição de energia elétrica e análise da operação de um RT;
d) com as etapas a, b e c concluídas foi possível especificar o modelo, incluindo testes de diferentes técnicas de regressão e classificação;
e) na quinta etapa o modelo foi avaliado em predições futuras, comparando o seu resultado com os dados reais obtidos do equipamento estudado;
f) na última etapa são discutidos os resultados obtidos e consolidados com as publicações que da dissertação. São destacadas conclusões, pontos fracos e contribuições, assim como os próximos passos a seguir após essa pesquisa.

1.4 Organização da dissertação

Esta dissertação está organizada em sete capítulos, sendo o primeiro a introdução, os demais são descritos a seguir:

a) Capítulo 2: Fundamentação teórica – apresenta os principais conceitos usados, como conceitos de SG, RT, e aprendizado de máquina.
b) Capítulo 3: Trabalhos relacionados – consiste no mapeamento sistemático na área de aprendizado de máquina aplicado a redes de distribuição, identificando modelos de predição, estimação de estados e otimização. É

apresentada ainda uma discussão sobre os métodos encontrados, uma taxonomia e comentários gerais.

c) Capítulo 4: Modelo PCR – contém o passo a passo da construção do modelo, inicia com a descrição de como o modelo é concebido, discute uma validação da metodologia com base no problema apresentado na seção 1.1 apresentando uma análise da base de dados discutindo problemas reais encontrados e valida a criação do modelo PCR, após a etapa de pré-processamento é discutida, seguindo para a classificação dos dados e correção de *outliers* com contexto, e por fim são discutidos métodos de regressão.
d) Capítulo 5: Estudo de caso – discute a assertividade das predições do modelo PCR com dados reais atuais e aprofunda discussões de aplicação do PCR para ajustes de performance do RT.
e) Capítulo 6: Contém as considerações finais da dissertação, as contribuições, a resposta a questão de pesquisa, a validação da hipótese e indica os próximos passos da pesquisa, bem como as publicações já realizadas.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta uma descrição dos conceitos e fundamentos utilizados para elaboração deste trabalho. Como destacado na introdução os dados utilizados nas análises são oriundos da rede de MT, essa escolha se deve ao fato de existirem equipamentos para coleta e transmissão dos dados instalados nos equipamentos especiais telecomandados da rede, o que facilita a aplicação de modelos que pretendem atuar em tempo real, um dos objetivos das SGs.

2.1 *Smart grids*

Não existe uma definição única para SG, cada autor define de maneira diferente. Em resumo, pode-se considerar que uma SG une diferentes tecnologias para monitoramento, controle, supervisão, automação e controle de redes de distribuição de energia elétrica. Uma rede nesse conceito deve ser confiável e avançada permitindo o fluxo de energia bidirecional com capacidade de autocura, adaptação e sustentabilidade (DILEEP, 2020).

De acordo com Huang et al. (2012) a crescente pressão pelo uso correto dos recursos do meio ambiente tem estimulado cada vez mais a utilização de novas fontes de energias renováveis, além disso o custo da energia elétrica e a crescente demanda tem levado mais pessoas a aderirem a microgeração.

As tecnologias para geração de energia elétrica mais utilizadas são as turbinas eólicas, microturbinas, sistemas fotovoltaicos, células de combustível, armazenamento de energia e aplicações de gerador síncrono (HIDAYATULLAH; STOJCEVSKI; KALAM, 2011).

A integração de GD afeta o controle de tensão em ambos os níveis de MT e Baixa Tensão (BT) (RUIZ-ROMERO et al. 2014). Além disso, segundo Vaziri et al. (2011) a alta penetração de GD pode gerar outros problemas como ilhamento, dificuldade de detecção de falhas, e problemas na proteção do sistema. Para contornar essas dificuldades uma opção é o tratamento e análise de dados obtidos SGs, como os oriundos dos RTs, descrito na seção 2.2.

A rede convencional é considerada cega, pois não é capaz de monitorar a si mesma, por outro lado uma SG possui diversos sensores conectados em toda a rede o que permite o automonitoramento (Hassan e Khan, 2021). A instalação de diferentes

sensores ao longo da rede permite que as funções do sistema *Supervisory Control And Data Acquisition* (SCADA) sejam aprimoradas pois permite que equipamentos sejam controlados remotamente e em tempo real, além dessa aplicabilidade a comunicação bilateral com medidores inteligentes instalados junto aos consumidores permite a criação da *Advanced Metering Infrastructure* (AMI), esses dados podem ser utilizados para gerenciamento de falhas, faturamento, previsão de demanda ente outros (Sartika, et al. 2021).

2.2 Reguladores de tensão

Os RTs são amplamente utilizados em redes de distribuição de energia elétrica. O princípio de funcionamento do RT é similar ao de um autotransformador, possuindo acoplamento magnético e elétrico, possui três terminais, F, C e FC este último conecta os dois enrolamentos através de um terminal comum (SIMONE, 2010).

Com isso o RT pode atuar como elevador ou rebaixador de tensão. Os RTs mais utilizados possuem 32 degraus de ajustes de tensão, chamados de TAPs, podendo fornecer ganhos de até ±10%. Deste modo cada TAP pode dar um ganho de ±0,625%. Nas redes de distribuição trifásicas são utilizados três RTs, um para cada fase. Sendo normalmente instalados em pontos estratégicos, que necessitem de ajustes nos níveis de tensão.

Os reguladores de tensão mais comuns fabricados hoje são monofásicos, eles podem fornecer flexibilidade e menor custo para operação e manutenção, resultando em maior eficiência operacional, qualidade de serviço e confiabilidade (Zhang, Hodge, e Attavvay, 2014).

O desempenho do RT é dependente do princípio operacional de seu controlador. O controlador é responsável por decidir se é necessário aumentar ou diminuir o TAP com base na amplitude de tensão medida pelo transformador de potencial. O controlador é programado de forma que um valor de referência e uma banda morta sejam definidos para ele. O valor de referência está no meio da banda morta. Por exemplo, se a tensão de referência é 1,02 p.u. com 0,02 p.u. como banda morta então a tensão local de referência deve estar entre 1,01 p.u. e 1,03 p.u., caso contrário é necessário o movimento do TAP. Se a tensão medida de PT exceder o limite superior da banda morta, então o comando de decremento do TAP é enviado para o comutador (Attar, Homaee, Repo, e Rekola,2018).

Em redes de distribuição que estão buscando o status de SGs estes equipamentos possuem sistemas acoplados para medição e transmissão de dados e até mesmo para ajuste de parâmetros. As medições coletadas possuem intervalo de coleta pequenos como a cada cinco, dez ou quinze minutos. Com o passar do tempo a quantidade de dados armazenada irá aumentar gradativamente, e assim como em toda base de dados, podem existir erros na fase de coleta, envio e armazenamento, sendo necessário utilizar técnicas para otimizar esses dados.

2.3 Pré-processamento de dados

Os conjuntos de dados estão se tornando cada vez maiores, o que torna o pré-processamento e a redução de dados técnicas essenciais nos cenários de descoberta de conhecimento (RAMÍREZ-GALLEGO et al. 2017).

Bancos de dados sem tratamento apresentam informações redundantes, ruidosas e por vezes não confiáveis. Desta forma a descoberta de conhecimento se tornará um problema muito difícil, uma vez que as etapas de preparação de dados exigem um tempo de processamento significativo em tarefas de aprendizado de máquina (ALEXANDROPOULOS et al. 2019).

Segundo García et al. (2016) as principais etapas do pré-processamento são balanceamento de dados, redução dos dados, dados imperfeitos. O balanceamento de dados busca fazer com que os dados não possuam uma classe majoritária que dificulte a análise. A redução dos dados aborda a geração e seleção de instâncias, discretização e seleção dos recursos. Esses métodos visam reduzir a complexidade dos conjuntos de dados, para que possam ser processados pelas soluções de mineração de dados (RAMÍREZ-GALLEGO et al. 2017). Estes dados tratados podem auxiliar na precisão de modelos de regressão e classificação como os propostos pelas tecnologias de aprendizado de máquina.

2.4 Aprendizado de máquina

O aprendizado de máquina é uma técnica emergente que tem como objetivo instruir computadores através de análise de dados para resolver um determinado problema (MUHAMMAD; YAN, 2015). Al-Sahaf et al. (2019) complementa afirmando que o aprendizado de máquina é um ramo da inteligência artificial baseado na ideia

de que os sistemas podem aprender a partir dos dados, identificar padrões ocultos e tomar decisões com pouca ou mínima intervenção humana. Segundo Adam e Smith (2008) o aprendizado de máquina é dividido em três subdomínios: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço.

O aprendizado supervisionado requer treinamento com dados rotulados que têm entradas e saídas desejadas, o aprendizado não supervisionado não requer dados de treinamento rotulados e o ambiente apenas fornece entradas sem os alvos desejados, já a aprendizagem por reforço possibilita a aprendizagem a partir do feedback recebido por meio de interações com um ambiente externo (QIU et al. 2016).

O aprendizado de máquina pode ser aplicado em diferentes análises, como abordagens que envolvam classificação, otimização ou regressão. A escolha sobre qual método utilizar está vinculado ao alvo que se deseja atingir e pela base de dados disponível. No âmbito dessa dissertação o alvo é a predição de valores de grandezas elétricas, sendo este um problema de regressão. Os dados disponíveis para análises são coletados ao longo do tempo por equipamentos instalados na rede de distribuição, como estes dados possuem uma amostragem de coleta fixa, e apresentam variações à medida que o tempo vai passando, podem ser classificados como uma série temporal.

2.5 Séries Temporais

Uma série temporal é uma sequência de pontos de dados indexados em uma ordem de tempo discreta, de forma geral é uma sequência tomada em sucessivos pontos distribuídos ao longo do tempo (Li, 2022).

Os dados analisados nesta dissertação representam uma série temporal, que possui intervalo de quinze minutos entre as amostras, Hyndman e Athanasopoulos (2021) citam que uma série temporal pode ser dividida em três componentes que são o ciclo de tendência, a sazonalidade e um componente restante, uma tendência existe quando a uma “mudança de direção” no gráfico de uma série temporal, quando ela pode passar de uma forma crescente para uma decrescente, já o fator sazonal aparece quando a série temporal é afetada por fatores sazonais como dia da semana ou época do ano.

2.6 Comentários sobre o capítulo

Neste capítulo foram apresentados conceitos estratégicos para o entendimento das técnicas que serão aplicadas nas etapas de pré-processamento, classificação e regressão dos dados. Esses conceitos serão discutidos e apresentados em aplicações práticas ao longo da dissertação.

As técnicas de aprendizagem de máquina estão ganhando espaço nos estudos relacionados a predição de grandezas elétricas, como será aprofundado no capítulo 3, essas técnicas contribuem para a evolução do sistema tradicional de distribuição de energia elétrica para o conceito de SG. Um equipamento essencial no conceito de SG é o RT, uma vez que pode resolver localmente problemas de afundamento de tensão devido a cargas elevadas instaladas e ajuda a manter a estabilidade do sistema frente as diversas flutuações de consumo de energia elétrica que ocorrem ao passar do tempo, essa importância fica ainda mais evidente quando se estuda a base de dados. Nos dados de energia elétrica as componentes de tendência e sazonalidade descritas são visivelmente expostas, o consumo de energia elétrica tende a ser maior em estações do ano que o calor seja extremo, por exemplo, bem como a interferência de dias úteis e não uteis com aumento ou baixa na produção do setor industrial.

3 TRABALHOS RELACIONADOS

Este capítulo explora a literatura relacionada a técnicas de suporte para melhorar a precisão de análises de dados aplicadas a SG. Usando a metodologia de estudo de mapeamento sistemático conforme apresentado por Rossei e Chren (2020), para a realização de uma revisão da literatura de trabalhos que aplicaram técnicas de análises de grandezas elétricas com o intuito de realizar previsões, estimar estados ou mesmo aprimorar a eficácia de modelos de análise de dados.

O mapeamento sistemático seguiu as diretrizes utilizadas por Aranda et al. (2019), definindo as seguintes etapas: definição das questões de pesquisa, elaboração do processo de busca em repositórios digitais e definição dos critérios para filtragem dos resultados. Este estudo encontrou inicialmente 15.148 artigos relacionados ao assunto, utilizando sete repositórios digitais que enfocam computação e eletricidade. Após um processo de filtragem, 37 artigos passaram por análise e discussão.

3.1 Questões de pesquisa

As questões de pesquisa direcionam a busca para estudos relacionados as técnicas para melhorar a estabilidade e precisão de técnicas de análises de dados. O quadro 1 apresenta uma Questão Geral (QG), seis Questões Específicas (QE) e uma Questão Estatística (QEST).

Quadro 1 – Questões de pesquisa

Questão Geral	
QG 1	Como os modelos de previsão e análise de dados foram usados para dar suporte aos estimadores de estado aplicados a redes inteligentes?
Questões Específicas	
QE 1	Quais são as técnicas usadas para alocação otimizada de medidores em redes inteligentes?
QE 2	Quais são as técnicas de análise de dados usadas para a previsão de grandezas elétricas em redes inteligentes?
QE 3	Quais são os tipos de bancos de dados usados para previsão de grandezas elétricas em redes inteligentes?
QE 4	Como as técnicas de previsão foram usadas para gerar pseudo-medições em redes inteligentes?
QE 5	Como as técnicas de previsão foram usadas para melhorar o controle de tensão em redes inteligentes?

QE 6	Quais são as técnicas usadas para tratar dados maliciosos em redes inteligentes?
Questão Estatística	
QEST 1	Qual é o número de publicações por ano e tipo?

Fonte: Elaborado pelo autor.

As questões buscaram destacar informações essenciais sobre o tema, delinear detalhes quantitativos dos artigos selecionados e definir dados cronológicos, respectivamente.

3.2 Processo de pesquisa

A busca de obras literárias consiste em três etapas: definir a *string* de pesquisa, selecionar bancos de dados e encontrar os resultados como discutido em Dalmina et al. (2019). A primeira etapa determina as palavras-chave destinadas a encontrar as respostas às perguntas da Tabela 1 e, em seguida, os sinônimos das palavras-chave são escolhidos para expandir o escopo. Finalmente, operadores booleanos como *"AND*" e *"OR"* são combinadas para formar a *string* de pesquisa. O quadro 2 apresenta os principais termos do tema e seus sinônimos selecionados.

Quadro 2 – Termos de pesquisa

Major terms	*Search Terms*
Energy	*(("Smart Grid" OR "Smart Energy") AND ("Power System" OR "Distribution System") AND*
Data and AI	*(“Data Analysis” OR “Big Data” OR “Statistical Analysis” OR “Machine Learning” OR “Deep Learning”) AND*
Electrical Quantities	*(“State Estimation” OR “State Estimator” OR “Prediction”) AND*
Smart Meters	*(“Load Model” OR “Load Modeling” OR “Measurements” OR "Metering" OR "Mensuration"))*

Fonte: Elaborado pelo autor.

A segunda etapa é definir os repositórios digitais nos quais aplicar a *string* de pesquisa. Este artigo tem como foco a distribuição de eletricidade e ciência da computação, por isso seguiu as orientações de Rossi e Chren (2020), estabelecendo: *ACM Digital Library, Google Scholar, IEEE Xplore, Science Direct, Scopus, Springer Link e Wiley Online Library*.

A pesquisa nos repositórios *ACM Digital Library, IEEE Xplore e Wiley Online Library* exigia o recurso de pesquisa avançada, alocando palavras-chave e operadores booleanos para os campos disponíveis. Além disso, o *Science Direct* limita o uso a oito operadores booleanos em uma única pesquisa, de modo que a busca completa foi aninhada em duas cadeias de caracteres: a primeira com os termos principais e a segunda com os sinônimos. Os repositórios *Google Scholar* e *Springer Link* permitiram que a *string* fosse totalmente escrita diretamente no campo de pesquisa.

3.3 Processo de filtragem

Após reunir as obras literárias por meio das palavras-chave, é necessário filtrar aquelas relacionadas ao tema e retirar aquelas que não sejam relevantes ao objetivo do trabalho.

Os seguintes Critérios de Inclusão (CI) foram usados para filtrar os artigos relacionados:

CI 1: um estudo publicado em uma conferência, *workshop* ou jornal;

CI 2: o estudo deve abordar assuntos relacionados, como SG, sistema de distribuição, estimadores de estado, medição de grandezas elétricas, análise de dados ou inteligência artificial;

CI 3: o estudo deve ser um artigo completo, implementado em sistemas reais ou simulados;

CI 4: artigos de conhecimento prévio dos autores que não apareceram na pesquisa inicial.

Para a remoção dos artigos foram usados os seguintes critérios de exclusão (CE):

CE 1: estudos publicados antes de 2010;

CE 2: estudos não escritos em inglês;

CE 3: estudos publicados na forma de dissertações ou teses;

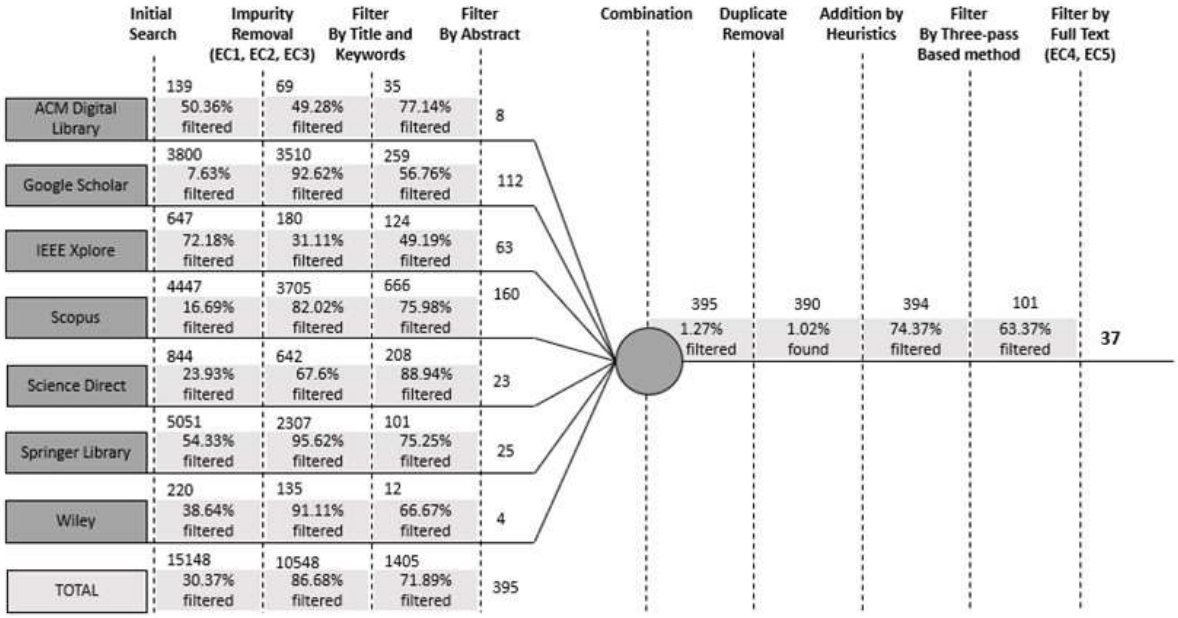
CE 4: estudos que não respondem a nenhuma das questões de pesquisa.

Esses critérios CI e CE permitiram a seleção dos estudos mais relevantes e a remoção de qualquer ruído. Inicialmente, os CE 1, CE 2 e CE 3 retiraram os papéis que não atendiam ao respectivo critério. Por fim, o CE 4 possibilitou a extração de

eventuais resíduos pela seguinte sequência: leitura dos títulos e palavras-chave, leitura do resumo, leitura do artigo pelo método das três etapas (KESHAV, 2007), e leitura do artigo na íntegra.

A Figura 1 mostra a quantidade de artigos selecionados de cada um dos repositórios digitais. As etapas do processo são destacadas, mostrando os valores absolutos e percentuais de cada etapa.

Figura 1 – Processo de filtragem



Fonte: Elaborado pelo autor.

A Figura 1 apresenta a direita os repositórios digitais selecionados para a busca, e a seguir todos os passos de filtragem apresentando a quantidade de artigos selecionados para o próximo passo da filtragem e a porcentagem de trabalhos excluídos do processo.

3.4 Resultados

O processo de filtragem descrito resultou em 37 de 15.148 trabalhos. A Tabela 1 apresenta os artigos selecionados no mapeamento sistemático, identificados por um número identificador (ID) que será usado para facilitar o entendimento das figuras. Também são apresentados o tipo de publicação, o *H-index,* e uma sinalização sobre quais questões a obra responde.

Tabela 1 – Artigos selecionados

Authors	*ID*	*Type*	*H-index*	*QE1*	*QE2*	*QE3*	*QE4*	*QE5*	*QE6*
Zhang et al.	*5*	*Journal*	*142*	*No*	*No*	*Yes*	*No*	*Yes*	*No*
Kabiri and Amjady	*30*	*Journal*	*142*	*No*	*No*	*Yes*	*No*	*No*	*No*
Dobbe et al.	*31*	*Journal*	*242*	*No*	*Yes*	*Yes*	*Yes*	*No*	*No*
Alzate et al.	*32*	*Journal*	*242*	*No*	*No*	*Yes*	*Yes*	*No*	*No*
Huang et al.	*33*	*Journal*	*111*	*No*	*No*	*Yes*	*Yes*	*No*	*No*
Huang et al.	*34*	*Journal*	*111*	*No*	*Yes*	*Yes*	*Yes*	*No*	*No*
Mestav et al.	*35*	*Conference*	*0*	*Yes*	*No*	*Yes*	*Yes*	*No*	*No*
Ullah et al.	*36*	*Journal*	*86*	*No*	*Yes*	*Yes*	*Yes*	*No*	*No*
Zhao et al.	*37*	*Conference*	*0*	*No*	*No*	*Yes*	*Yes*	*No*	*No*
Dehghanpour et al.	*38*	*Journal*	*142*	*No*	*No*	*Yes*	*Yes*	*No*	*No*
Yuan et al.	*39*	*Journal*	*242*	*No*	*No*	*Yes*	*Yes*	*No*	*No*
Yang et al.	*40*	*Journal*	*189*	*No*	*Yes*	*Yes*	*Yes*	*No*	*No*
Barbeiro et al.	*41*	*Conference*	*14*	*No*	*Yes*	*Yes*	*Yes*	*No*	*No*
Cao et al.	*42*	*Journal*	*242*	*No*	*Yes*	*Yes*	*Yes*	*No*	*No*
Pau et al.	*43*	*Journal*	*111*	*No*	*Yes*	*Yes*	*Yes*	*No*	*No*
Boroojeni et al.	*44*	*Journal*	*114*	*No*	*No*	*Yes*	*Yes*	*No*	*No*
Saviozzi et al.	*45*	*Journal*	*114*	*No*	*Yes*	*Yes*	*Yes*	*No*	*No*
Ahmad et al.	*46*	*Journal*	*116*	*No*	*Yes*	*Yes*	*Yes*	*No*	*No*
Manitsas et al.	*47*	*Journal*	*242*	*No*	*No*	*Yes*	*Yes*	*No*	*No*
Pegoraro et al.	*48*	*Journal*	*111*	*Yes*	*No*	*Yes*	*No*	*No*	*No*
Xygkis and Korres	*49*	*Conference*	*12*	*Yes*	*No*	*Yes*	*No*	*No*	*No*
Xiang et al.	*50*	*Journal*	*142*	*Yes*	*No*	*Yes*	*Yes*	*Yes*	*No*
Wang et al.	*51*	*Journal*	*142*	*Yes*	*No*	*Yes*	*Yes*	*No*	*No*
Milbradt et al.	*52*	*Conference*	*0*	*Yes*	*No*	*No*	*No*	*No*	*No*
Raposo et al.	*53*	*Journal*	*114*	*Yes*	*No*	*Yes*	*No*	*No*	*No*
Nguyen	*54*	*Journal*	*242*	*Yes*	*No*	*Yes*	*Yes*	*No*	*No*
Ye et al.	*55*	*Conference*	*14*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*No*
Liao et al.	*56*	*Journal*	*177*	*No*	*Yes*	*Yes*	*Yes*	*Yes*	*No*
Ye et al.	*57*	*Journal*	*116*	*No*	*Yes*	*Yes*	*No*	*Yes*	*No*
Zhang et al.	*58*	*Journal*	*142*	*No*	*No*	*Yes*	*No*	*No*	*No*
Wang et al.	*61*	*Journal*	*142*	*No*	*No*	*No*	*No*	*Yes*	*No*
Bagheri and Xu	*62*	*Journal*	*142*	*No*	*No*	*Yes*	*No*	*Yes*	*No*
Roberts et al.	*65*	*Journal*	*142*	*No*	*No*	*Yes*	*No*	*No*	*Yes*
Wang et al.	*66*	*Journal*	*67*	*No*	*No*	*Yes*	*No*	*No*	*Yes*
Zhang et al.	*67*	*Journal*	*142*	*No*	*No*	*Yes*	*No*	*No*	*Yes*
Raggi et al.	*68*	*Journal*	*177*	*No*	*No*	*Yes*	*No*	*No*	*Yes*
Dominguez et al.	*69*	*Conference*	*0*	*No*	*No*	*Yes*	*No*	*No*	*Yes*

Fonte: Elaborado pelo autor.

3.4.1 Como os modelos de previsão e análise de dados foram usados para dar suporte aos estimadores de estado aplicados a redes inteligentes?

Segundo Paruta et al. (2021) a injeção de energia renovável na rede de distribuição tem aumentado significativamente, tornando necessário o monitoramento e o controle da rede. No processo de estimação de estados, a redundância das medições melhora a precisão dos dados, atingindo assim a situação operacional do sistema (JI et al. 2021).

Kabiri e Amjady (2019) propõem um modelo híbrido usando dados obtidos do SCADA e *Phasor Measurement Unit* (PMU). Esses dados são separados e tratados em duas etapas diferentes. O *Weighted Least Squares* (WLS) usa dados SCADA, enquanto os dados PMU requerem o sistema de *Least Absolute Value* (LAV). Os autores destacam o excelente desempenho do estimador proposto, mas indicam a dificuldade de aplicação do método devido ao alto custo de implantação das PMU. Dobbe et al. (2020) apresentam um modelo baseado na estimativa dos *Minimum Mean Squares Estimation* (MMSE) para estimação de estado em tempo real. Essa abordagem se assemelha ao método WLS, mas se aplica aos princípios bayesianos. Segundo os autores, esse método permite a utilização de diversas fontes de dados, como SCADA, AMI e condições climáticas.

Alzate et al. (2019) implementam uma combinação de WLS e Levenberg-Marquardt (LM). Neste artigo, os autores usam dados de medidores inteligentes conhecidos como *Smart Meters* (SMs) em residências monitoradas para apoiar estratégias coordenadas de controle de tensão em tempo real. O estudo mostra resultados promissores quando aplicado em sistemas de BT com casas inteligentes. Huang et al. (2020) apresentam uma solução *Robust Ensemble Kalman Filter* (REnKF). Os autores utilizam dados de SMs no lado BT de subestações e modelos matemáticos no lado de MT para estimar o estado de nós não observados. Segundo os autores, este é o primeiro trabalho que utiliza essa técnica em redes de distribuição.

Huang et al. (2019) aplicam *Deep Belief Network* (DBN) para gerar pseudo-medidas, que funcionam como um banco de dados para estimação de estados baseado no *Gaussian Component Combination Method* (GCCM), que possui uma estrutura semelhante ao WLS. Mestav et al. (2018) destacam os benefícios do uso de técnicas de *Deep Neural Network* (DNN) para gerar pseudo-medições. Os autores também indicam que a combinação da inferência Bayesiana com DNNs traz

excelentes resultados para estimação de estados em tempo real. O método usa DNN para aproximar o problema MMSE. Ullah et al. (2020) apresentam um modelo capaz de estimar, com boa precisão, o estado de redes com alta penetração de energia renovável, algoritmo híbrido *Particle Swarm Optimization with Sigmoid-based Acceleration Coefficients and Chaotic Gravitational Search Algorithm* (PSOS-CGSA). O trabalho considera dados como perfis de carga, clima e perfil de consumo em dias diferentes.

Zhao et al. (2020) usam em seus trabalhos dados obtidos do SCADA e de SMs. Neste estudo, a técnica *Support Vector Machine* (SVM), uma técnica baseada em aprendizado de máquina, prevê a carga de curto prazo. A saída do modelo alimenta um estimador de estado com base na *Generalized Maximum Likelihood* (GM). Dehghanpour et al. (2019) propõem um método para gerar pseudo-medidas, com base em *Game-Theoretic Expansion of Relevance Vector Machines* (RVMs), a solução usa dados de tensão e consumo por hora coletados por meio de AMI para gerar dados para consumidores não monitorados.

Yuan et al. (2019) aplicam os dados de SMs no método *Spectral Clustering* (SC) baseado em *Multi-Time Scale Learning* (MTSL). Semelhante a Dehghanpour et al. (2019), os autores propõem um modelo que gera pseudo-medidas para consumidores não monitorados a partir de dados de consumo horário obtidos dos medidores. Yang et al. (2018) usam uma abordagem probabilística usando *Gaussian Process Quantile Regression* (GPQR) para previsão de carga de curto prazo usando dados históricos de demanda de energia e temperatura. O algoritmo tem a capacidade de estimar a carga com uma hora de antecedência.

Barbeiro et al. (2015) apresentam monitoramento em tempo real em redes BT mal caracterizadas, os autores aplicam uma técnica de aprendizado de máquina chamada *Evolutionary Particle Swarm Optimization* (EPSO), treinada pelo histórico de demanda de energia com uma amostragem de 15 minutos. Cao et al. (2020) usam DBN, uma técnica para prever cargas de BT. Pau et al. (2020) fazem uso de SMs, mas eles empregam os dados para monitorar redes MT e BT usando a teoria de propagação de incerteza. Após este estágio, um modelo WLS usa essas pseudo-medidas. Boroojeni et al. (2017) usam um modelo auto-regressivo e de média móvel para desenvolver estimativas de carga de curto e médio prazo, analisando dados sazonais e não sazonais separadamente. Saviozzi et al. (2019) propõem um sistema com Redes Neurais Artificiais (RNA) que usa dados de demanda de energia diários

com um intervalo de 15 minutos. O sistema também compila outros fatores cronológicos, como dia da semana, feriados, carga média diária e carga semanal.

Ahmad et al. (2019) usam o cálculo do fluxo de potência, realizado pela técnica *Backward-Forward Sweep* (BFS), para treinar um modelo baseado em RNA. Manitsas et al. (2012) usam RNA para gerar pseudo-medições para melhorar a estabilidade de um WLS, usando curvas de carga típicas e dados de fluxo de potência simulados. Outras abordagens melhoram estimação de estados, Pegoraro e Sulis (2013) apresentam um modelo robusto de alocação ótima de medidores em redes de distribuição. A técnica aplica a programação dinâmica através do princípio de Bellman. Após a alocação do medidor, a precisão do estimador de estados baseado em WLS é testada com a nova configuração dos medidores.

Xygkis e Korres (2016) usam a programação semi-definida formulado como um M-ótimo para alocação SMs visando melhorar a precisão de um estimador WLS. Xiang et al. (2014) usam simulação de Monte Carlo para alocação ótima de medidores. Este estudo sugere que a distribuição de medidores seja feita nos pontos de erros mais significativos para estimativa de grandezas elétricas. Alternativamente, Wang et al. (2016) usam simulação de Monte Carlo para geração de ruído de dados. O método WLS é a base do sistema proposto, que usa a magnitude e a fase da tensão como variáveis de entradas. Milbradt et al. (2013) oferecem uma abordagem multicritério para alocação de medidores. Com base no modelo WLS, utilizando o método *Analytic Hierarchy Process* (AHP), tratando como objetivo central aumentar a precisão do estimador. Este artigo também considera e detalha outros aspectos de engenharia, como custos de equipamentos, impacto na confiabilidade, redução de perdas e controle de tensão.

Raposo et al. (2020) apresentam um sistema de otimização de alocação de medidores capaz de funcionar mesmo que a topologia da rede mude. A implementação aplica a técnica nomeada como *Multiple Objective-Based Random Key Genetic Algorithm* (MOBRKGA). Nguyen (2015) avalia a carga de cada transformador na rede de distribuição usando o teorema do limite central. A incerteza da estimativa diminui à medida que mais medidores são instalados na rede, agregando dados de consumidores residenciais, comerciais e industriais. Ye et al. (2015) usam um método de estimador de estados baseado em inferência Bayesiana para verificar quedas de tensão. Usando os dados dos nós monitorados, o modelo estima as falhas onde não há sensores. A simulação de Monte Carlo verifica diferentes

cenários de falha. Liao et al. (2018) também abordam o problema da queda de tensão. Este trabalho apresenta uma *Convolutional Neural Network* (CNN) a para estimar a magnitude da queda de tensão em diferentes níveis de classificação, aplicando dados de fluxo de potência antes e após as falhas para treinamento.

Ye et al. (2019) aplicam o teorema de Bayes para analisar quedas de tensão. O estado da rede de distribuição antes e depois da falha é estudado para estimar problemas futuros. Zhang et al. (2020) usam um método de iteração numérica para aproximar a topologia da rede de distribuição, propondo um processo de dois estágios. Primeiro, um modelo de regressão linear baseado em medições históricas ou em tempo real de injeções de energia nodal e magnitude de tensão para estimar a topologia da rede. A seguir, a técnica de Newton-Raphson é aplicada para mitigar o erro. Mesmo sem usar medições de PMU, o modelo pode estimar fase e ângulos numericamente.

3.4.2 Quais são as técnicas usadas para alocação otimizada de medidores em redes inteligentes?

Segundo Bei Gou e Abur (2001) a alocação de medidores visa atingir a observabilidade do sistema através da obtenção de um número ótimo de medições em tempo real. A aplicação de técnicas de otimização de alocação de medidores melhora a estabilidade dos estimadores de estado, possibilitando investimentos com uma relação custo-benefício razoável, reduzindo o *payback*.

Pegoraro et al. (2013), Xygkis e Korres (2016), Xiang et al. (2014), Wang et al. (2016) e Milbradt et al. (2013) destacam os benefícios que a alocação de medidores traz para estimadores WLS. Os trabalhos de Pegoraro et al. (2013) e Xiang et al. (2014) apresentam métodos de alocação baseados na porcentagem de erro aceitável na predição de grandezas elétricas. Variando o número de medidores e a localização deles até atingir a meta. Um modelo de decisão binária para alocação de SMs é desenvolvido em Xygkis e Korres (2016), mesmo com um pequeno número de medidores, uma boa observabilidade da rede é atingida. Milbradt et al. (2013) apresenta uma abordagem multicritério. O método AHP analisa a precisão da estimativa de estado, custo do equipamento, impacto na confiabilidade e redução de perdas. Wang et al. (2016) apresentam um algoritmo para alocação ótima usando repetidas tentativas monitoradas localmente. Calculado pela cadeia de Markov e AHP,

o modelo pode ter boa observabilidade mesmo com uma mudança de topologia. Raposo et al. (2020) também consideram a reconfiguração da rede, o método de otimização visa reduzir as perdas técnicas anuais. O problema de estimativa de carga como um objetivo para alocação de medidores é tratado em Nguyen (2015). Usando um método analítico, o autor mostra o impacto sobre a média e a variância das estimativas em um alimentador de teste em vários cenários de posicionamento de medidores. Outros autores propuseram maneiras simples. Ullah et al. (2020) estimam que seja necessário monitorar 20% dos nós da rede de distribuição. Ye et al. (2015) considera que todos os nós de proteção devem ser monitorados, incluindo a saída do alimentador.

3.4.3 Quais são as técnicas de análise de dados usadas para a previsão de grandezas elétricas em redes inteligentes?

Alguns dos trabalhos mapeados neste estudo tratam da melhoria da estabilidade dos estimadores de estado, 27% dos artigos selecionados utilizam algum tipo de variação do método WLS. Conforme Primadianto e Lu (2017) os sistemas de distribuição de energia elétrica adaptaram e estão usando esta técnica, conhecida por seu uso em sistemas de transmissão.

No entanto, é frequente a tendência de usar novas tecnologias baseadas em aprendizado de máquina e aprendizado profundo, 24% dos artigos selecionados fazem esse tipo de abordagem. Soluções recentes também combinam diferentes técnicas, desenvolvendo sistemas híbridos no trabalho de Huang et al. (2020) e Pau et al. (2019).

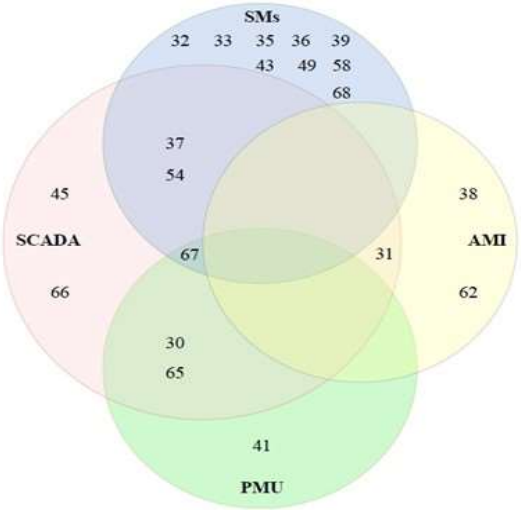
Entre as soluções baseadas em aprendizado de máquina, Saviozzi et al. (2019) e Ahmad et al. (2019) apresentam bons resultados com a técnica de RNA. Ambos os métodos mostram saídas estáveis mesmo ao agregar diferentes tipos de dados de entrada, provando que os modelos suportam a aplicação em redes complexas. Cao et al. (2020) apresentam um método adicional, usando uma variação do filtro de Kalman.

3.4.4 Quais são os tipos de bancos de dados usados para previsão de grandezas elétricas em redes inteligentes?

Dados de diferentes fontes estão disponíveis para a realização de estudos nas concessionárias de distribuição de energia elétrica, mas como utilizá-los ainda é um campo discutido. Dados históricos reais, e curvas de carga típicas são usadas em análises que nem sempre condizem com a realidade atual da rede de distribuição, além disso, o gerenciamento de desafios de big data e o custo computacional para processar todos os dados existentes também devem ser considerados (PRIMADIANTO; LU, 2017).

Entre os estudos selecionados, doze abordam o uso de medidores eletrônicos com comunicações e computação em tempo real chamados SMs. Os trabalhos de Zhao et al. (2020) e Nguyen (2015), combinam SMs e SCADA. Além desses dois métodos, Zhang et al. (2020) usa dados PMU. Wang et al. (2020) usam SCADA para detectar dados falsos, e Saviozzi et al. (2019) combinam esta tecnologia com a análise de dados históricos. Em Zhang et al. (2021), além dos dados históricos e do SCADA, são analisadas leituras AMI de consumo doméstico, dados de GD e dados meteorológicos públicos. Dehghanpour et al. (2019) usam AMI junto com dados históricos de magnitude de tensão e consumo horário. A Figura 2 mostra um resumo do uso dos medidores. A figura mostra uma tendência de uso de SMs, uma vez que doze artigos relatam seu uso. Sete estudos analisam dados SCADA. Com menos expressividade, aparecem PMU e AMI.

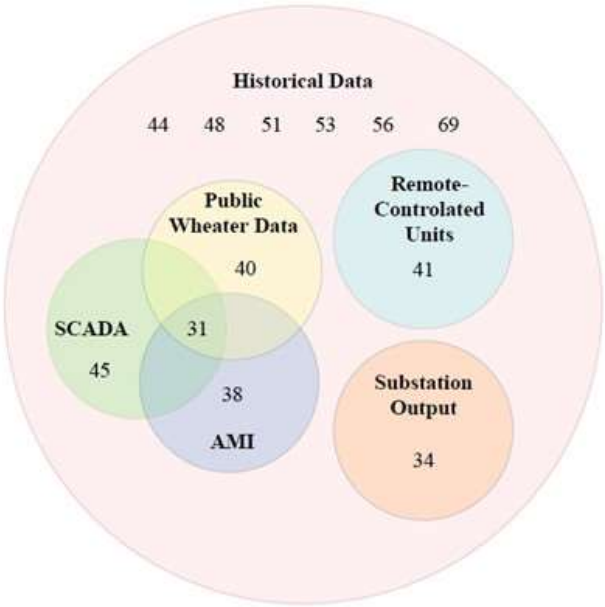
Figura 2 – Tipos de medidores para aquisição de dados



Fonte: Elaborado pelo autor.

Os desafios de big data estão presentes nesses artigos que usam dados históricos para estimar carga, tensão e outros parâmetros. Muitos dados são processados para tentar aproximar as estimativas da realidade. A figura 3 mostra os artigos que abordam esses desafios e combinam essa análise com outros tipos de medição.

Figura 3 – Tipos de dados



Fonte: Elaborado pelo autor.

A magnitude da tensão é a principal fonte de dados analisada no trabalho de Wang et al. (2016), Liao et al. (2018) e Ye et al. (2019). Ye et al. (2015) apresentam uma combinação entre magnitude de corrente, tensão e ângulo de fase. Medidas de potência ativa e reativa, dados de PMU são os outros parâmetros usados em Kabiri e Amjady (2019). Finalmente, Zhang et al. (2021), Mestav et al. (2018), Ahmad et al. (2019), Manitsas et al. (2012) e Pegoraro et al. (2013) usam dados de fluxo de potência. Manitsas et al. (2012) combinam esses dados com perfis de consumo típicos.

3.4.5 Como as técnicas de previsão foram usadas para gerar pseudo-medições em redes inteligentes?

Dehghanpour et al. (2019) considerou uma abordagem de três modelos de aprendizado de máquina. No primeiro estágio, o agrupamento espectral infere perfis

de carga diários típicos para clientes com SMs. Cada comportamento típico representa uma classe de perfil distinta. No segundo módulo, um modelo de aprendizagem multitemporal estima o consumo por hora a partir de dados históricos de consumo mensal. O terceiro estágio aplica um método de aprendizado bayesiano recursivo para aproximar os perfis de carga diária de consumidores não observados.

Huang et al. (2019) apresentam uma técnica de aprendizado profundo para predizer a carga de BT, realizando inicialmente uma abordagem determinística. O resultado alimenta o modelo probabilístico para gerar as pseudo medidas, por fim são usadas técnicas de reamostragem para melhorar a precisão. Mestav et al. (2018) treina dois modelos DBN. Eles estimam as saídas de injeção de potência ativa e reativa usando perfis de carga e medições reais.

Zhao et al. (2020) processam dados AMI. As medidas que apresentam valores maiores que cinco desvios-padrão do valor médio são desprezadas. Posteriormente, a correlação é verificada uma vez que os nós vizinhos possuem uma correlação significativa. Os dados dos SMs para monitoramento em tempo real estão disponíveis no trabalho de Yuan et al. (2019). Neste estudo, eles estimam o estado das redes BT e MT usando a teoria de propagação da incerteza. O histórico de dados obtidos dos SMs é utilizado no trabalho de Alzate et al. (2019) através da técnica SVM para carga de curto prazo e previsão de injeção de GD.

O método PSOS-CGSA proposto no trabalho de Pau et al. (2019) gera pseudo-medições de cargas com base em curvas de carga típicas para clientes sem GD, e para unidades GD não monitoradas faz uso de banco de dados histórico e condições climáticas.

Manitsas et al. (2012) criam perfis de carga para todos os barramentos na rede. O fluxo de carga é simulado com esses dados e usado para treinar RNA. As simulações de Monte Carlo criam diferentes cenários. Os autores ressaltam que as simulações utilizadas para o treinamento dão mais força à técnica.

3.4.6 Como as técnicas de previsão foram usadas para melhorar o controle de tensão em redes inteligentes?

A adição de energias renováveis às redes de distribuição de energia está mudando o paradigma de uma rede passiva. Segundo Barker e De-Mello (2000) os componentes do GD podem desestabilizar o sistema de distribuição, afetando a

qualidade da energia e a regulação da tensão. Manter o nível de tensão dentro dos padrões aceitáveis deve ser um problema considerado ao realizar análises sobre as redes de distribuição.

O estimador de estados implementado no trabalho de Xiang et al. (2014) tem como objetivo estimar o estado em todos os pontos de conexão de uma rede de distribuição. Usando as estimativas, atinge-se o nível de tensão apropriado otimizando a posição de TAP dos transformadores. Os autores indicam que o principal problema de programação é a minimização do desvio da qualidade da tensão usando uma média ponderada.

Ye et al. (2019) discutem o problema da estimativa do estado do mergulho de tensão (VDSE). A partir das magnitudes das tensões ao longo do alimentador, o teorema de Bayes obtém a localização do curto-circuito. O objetivo deste modelo é estimar os valores das tensões residuais em barras não monitoradas com as magnitudes das tensões medidas em um número limitado de barras durante quedas de tensão. Para isso, é necessário conhecer a condição da rede quando ocorre a falha. Ye et al. (2015) também apresentou o uso do método baseado em inferência bayesiana. Este estudo usa medições simuladas para gerar pseudo medições de quedas de tensão para quantificar os indicadores de qualidade do índice SARFIx. Liao et al. (2018) a abordagem baseada na CNN é desenvolvida para realizar classificações de magnitudes de afundamento em barras não monitoradas em seis categorias possíveis com faixas de tensão especificadas definidas em p.u.

Wang et al. (2014) discutem a aplicação de um modelo exponencial para verificar a relação entre tensão e demanda - o modelo considera perfis de consumidores analisando os diferentes tipos de carga. Um Modelo de Controle Preditivo (MPC) para otimizar o uso de comutadores de derivação sob carga, termo do inglês *On-Load Tap Changers* (OLTCs) e dispositivos de compensação de reativos, como banco de capacitores, constituem o sistema VVO. O modelo atua no chaveamento dos dispositivos seguindo a saída GD esperada. Bagheri e Xu (2017) mostram um modelo de estimativa estatística para verificar os benefícios do uso de técnicas VVC baseadas em bancos de capacitores. Os autores investigam melhorias nas perdas totais de energia do sistema, níveis máximos e mínimos de tensão por meio de dados AMI. O modelo fornece feedback para melhorar as técnicas de VVC. Zhang et al. (2021) desenvolveram um aprendizado de reforço profundo multiagente, usando o método de varredura para trás e para a frente para gerar dados de fluxo de

energia precisos para treinamento. Os autores consideram o estado dos capacitores comutáveis, reguladores de tensão e inversores inteligentes, como as variáveis de ação para redução de perdas e controle de tensão. Este modelo tem resposta em milissegundos, possibilitando operação em tempo real.

3.4.7 Quais são as técnicas usadas para tratar dados maliciosos em redes inteligentes?

Os sistemas físico e cibernético estão profundamente interligados nas *SGs*. Consequentemente, os ataques cibernéticos podem trazer problemas econômicos e operacionais (NEJABATKHAH et al. 2020). Segundo Kaviani e Hedman (2021) ataques de injeção de dados falsos (FDIAs) podem alterar a entrada de dados do estimador de estados, afetando sua precisão.

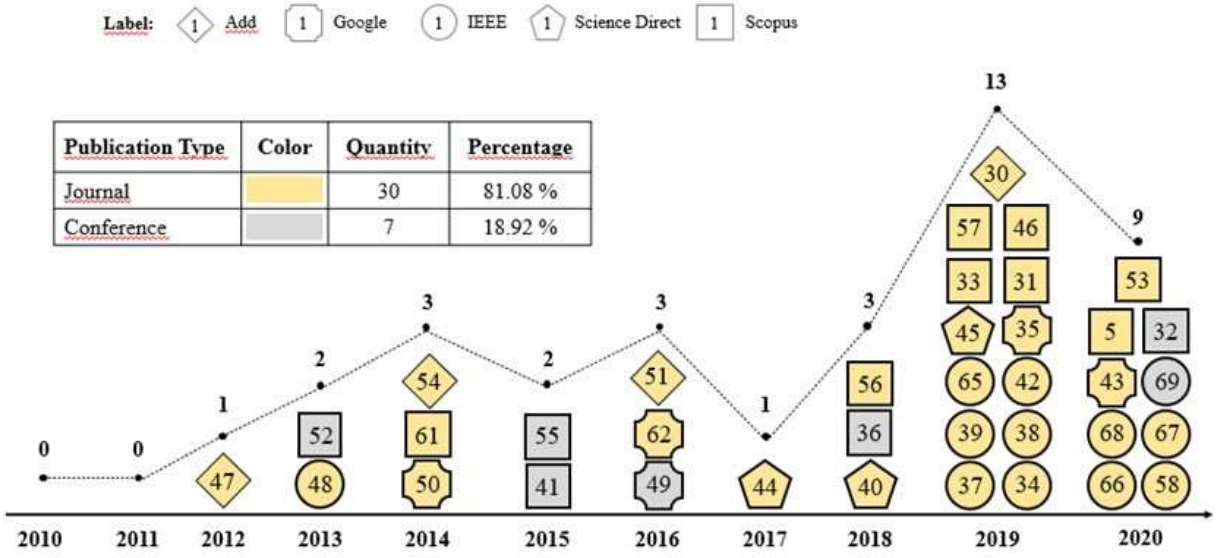
O trabalho de Roberts et al. (2020) tem como objetivo detectar a existência de um intruso antes que ocorra um ataque para desestabilizar os níveis de tensão. Ele apresenta um método semelhante ao bloco de recuperação de temporização *Early-Late Gate* no processamento de sinal. O objetivo é identificar o funcionamento padrão de reguladores de tensão e bancos de capacitores e então um sistema de aviso do operador é acionado se houver uma mudança no atraso de tempo estático ou na largura da banda morta. Wang et al. (2020) apresentam um método para detecção e localização de FDIAs. A técnica de aprendizado profundo utiliza RNA com um *Bad Data Detector* (BDD) padrão, obtendo os dados do sistema SCADA em tempo real. Inicialmente, o BDD remove dados inconsistentes, então o CNN identifica inconsistências no fluxo de potência, atuando como um classificador *multilabel*.

Zhang et al. (2021) mostram que os métodos BDD podem não ser eficazes contra FDIAs inteligentes. Portanto, é necessária a aplicação de diferentes técnicas para lidar com este problema. Os autores propõem o uso de autoencoders para extração de dados mais eficiente e para melhorar o treinamento da RNA. Outro ramo de dados maliciosos lida com perdas não técnicas (NTLs). Raggi et al. (2020) propõem o uso de dados de SMs aplicados ao método WLS para detectar e localizar NTLs, a técnica atua offline. Dominguez et al. (2020) comparam nove modelos de ML para detectar NTLs. Dados reais de consumo e sociodemográfico são usados para treinamento do modelo. Finalmente, os autores concluem que o modelo *Gradient Boost Machines* (GBM) possui a maior precisão.

3.4.8 Qual é o número de publicações por ano e tipo?

O mapeamento sistemático aqui apresentado ocorreu em janeiro de 2021, com a seleção dos artigos publicados a partir de 2010. A Figura 4 mostra os artigos selecionados, ordenados pelo ano de publicação, contando com uma legenda para o tipo de publicação e outra para indicar o repositório digital que devolveu o trabalho durante o processo de pesquisa. 60% dos artigos são publicações de 2019 e 2020, mostrando que a busca pela precisão e estabilidade nas análises de dados das SGs é uma questão atual.

Figura 4 – Publicações por ano, tipo e biblioteca digital



Fonte: Elaborado pelo autor.

Como pode ser visto na Figura 4, a maior parte dos trabalhos selecionados para análise e discussão foram publicados em *Journals* nos anos de 2019 e 2020.

3.5 Discussão

A pesquisa retornou diferentes modelos de estimadores de estado: WLS, LAV, MMSE, Filtro de Kalman, ML, GCCM, RNA, PSOS-CGSA e GM. Embora todas as técnicas apresentem resultados relevantes para o controle e operação dos sistemas de distribuição, não é fácil definir qual delas obtém os melhores resultados já que cada autor defende seu trabalho e os testes são realizados em diferentes condições, dificultando as comparações.

Zhao et al. (2020) comparam o desempenho do modelo GM proposto com o WLS. GM tem um bom desempenho com os estágios de injeção DG e injeção de dados, enquanto o modelo WLS precisa de mais robustez. Por outro lado, o tempo de resposta do modelo WLS ocorreu em 0,1s enquanto o GM precisou de 0,4s. Ullah et al. (2020) realizam dois tipos de testes, comparando inicialmente o modelo híbrido PSOS-CGSA com outros modelos metaheurísticos. O método apresenta erros menores em um sistema de 123 barramentos em comparação com o algoritmo WLS baseado em corrente de ramal.

Mestav et al. (2018) enfatizam que o método baseado em MMSE é complicado, mas uma abordagem baseada em DNN é possível. Os autores mostram que treinar RNA para estimar o estado diretamente é mais eficiente do que gerar pseudo-medidas durante o teste. O método mantém a estabilidade com injeção GD e dados inconsistentes. Huang et al. (2020) apresentam resultados positivos do modelo baseado em REnKF. A técnica robusta dos autores é testada com o modelo WLS e o modelo tradicional Ensemble Kalman Filter (EnKF). Como esperado, o método é menos sensível a dados ruins, apresentando mais estabilidade, porém com um custo computacional maior.

O método GCCM discutido no trabalho de Huang et al. (2019) apresenta maior robustez contra dados ruins. Ainda assim, deixa algumas questões em aberto, como a operação do modelo em cenários com interrupções na comunicação. Os autores destacam a necessidade de um número mais significativo de testes para validar o modelo.

A técnica de RNA vista no trabalho de Ahmad et al. (2019) mostra bom desempenho computacional na atualização dos estados da rede e boa estabilidade em comparação com WLS. Porém, o desempenho computacional não considerou o tempo de treinamento da RNA, indicando um campo promissor com carência de estudos. Barbeiro et al. (2015) mostram uma alternativa para contornar o tempo de treinamento, um autoencoder devidamente treinado com a técnica *Extreme Learning Machine* (ELM) traz ganhos de desempenho significativos.

Além disso, novas oportunidades surgiram, como o *Edge Computing*. Essa tecnologia permite a atuação diretamente em campo, de forma descentralizada, reduzindo o custo computacional - Huang et al. (2018) e Carvalho (2017) apresentam esta técnica. O modelo híbrido entre LAV e WLS de Kabiri e Amjady (2019) apresenta maior precisão e velocidade de resposta do que o método WLS tradicional. Devido às

medidas de PMU utilizadas, o custo computacional é maior, e exige um investimento financeiro mais considerável.

Todos os modelos para estimação de estados estudados apresentam vantagens e desvantagens, sendo essencial analisar cada cenário antes de implementar um modelo. Segundo Mestav (2018), ao atingir a observabilidade do sistema, com a aquisição de dados SCADA, as técnicas baseadas em WLS tendem a ter melhores resultados do que as baseadas em métodos Bayesianos, destacando a importância da análise.

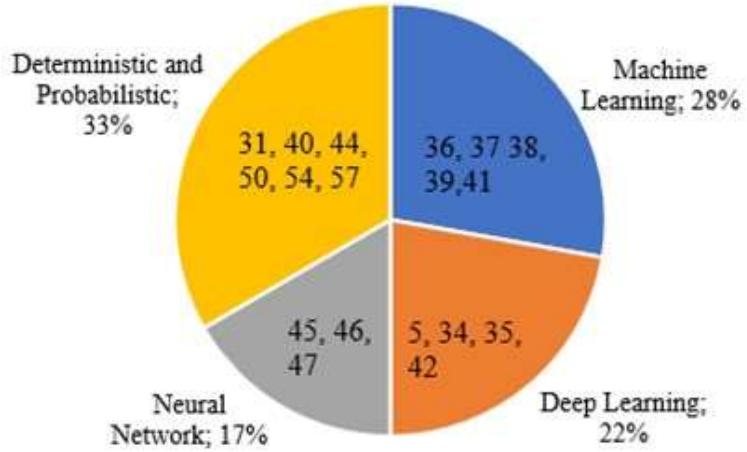
Além disso, é necessário reconhecer a necessidade de diferentes técnicas para melhorar os pontos fracos nas estimações. Nove artigos usam a alocação ideal do medidor como uma opção para aprimorar esse desempenho. Raposo et al. (2020) consideram a alocação de medidores para resolver o problema de reconfiguração da topologia da rede. Durante a operação, a topologia do sistema pode mudar em caso de falhas e manobras. Identificar essa topologia é uma tarefa crucial para o processo, como pode ser visto no trabalho de Zhang et al. (2020).

Xiang et al. (2014) ressaltam que aspectos qualitativos da operação, como estimativa de carga em tempo real, podem não ser alcançados usando técnicas tradicionais. Dessa forma, o uso de novos métodos de estimativa de carga pode ajudar a manter sua estabilidade diante de mudanças repentinas no consumo. Cao et al. (2020) compara o tempo de processamento para modelos de previsão RNA, DBN e *Hybrid Deep Learning Ensemble* (HEDL).

Enquanto as técnicas RNA e DBN apresentam um tempo computacional inferior a dez segundos, o modelo composto HEDL necessita de cem segundos. Mesmo que esses tempos sejam distantes, todos os modelos são aceitáveis para a operação, conseguindo boas previsões com 15 minutos de antecedência, com erros estimados em 1,31%. Massaoudi et al. (2021) apresentam uma comparação entre diferentes modelos de previsão de carga, este estudo citou quatro tipos de técnicas de previsão de carga. Modelos determinísticos e probabilísticos, aprendizado de máquina, aprendizado profundo e redes neurais, embora as RNAs sejam uma ramificação das técnicas de aprendizado de máquina, devido ao volume de trabalhos que aplicam essa técnica ela foi deixada em separado na análise.

A Figura 5 mostra quais artigos abordam cada modelo citado, apresentando o percentual de utilização de cada técnica.

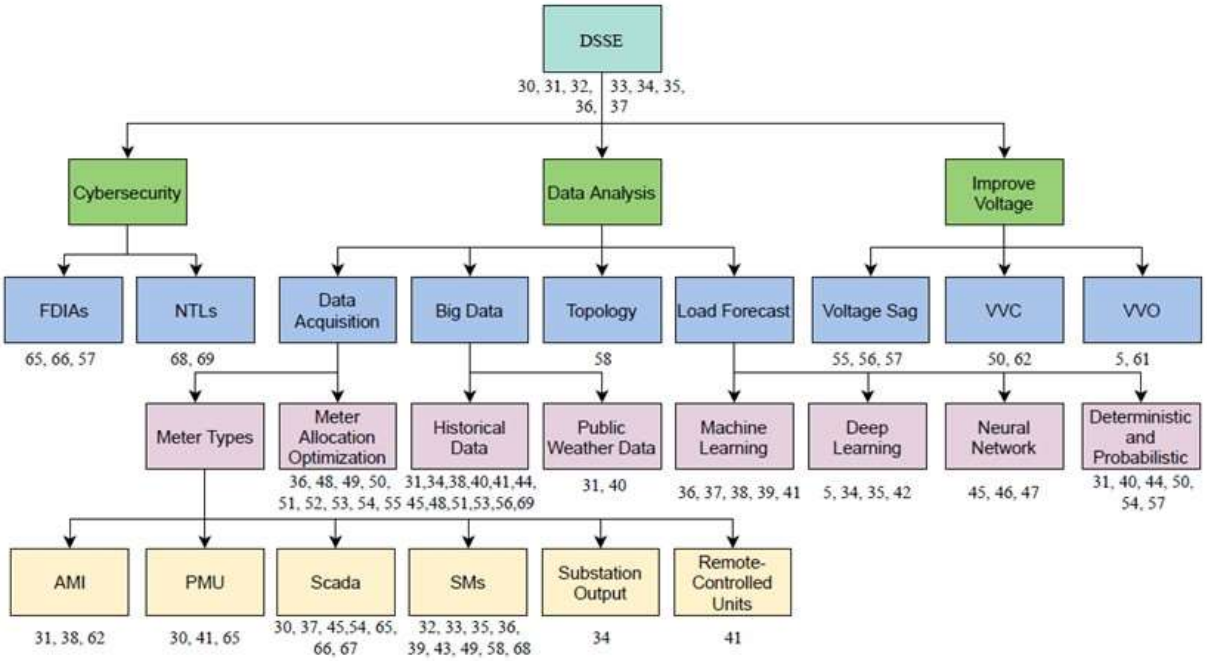
Figura 5 – Modelos utilizados para problemas de regressão



Fonte: Elaborado pelo autor.

O planejamento das operações dos equipamentos comutadores, como os reguladores de tensão abordados nesse trabalho, pode envolver essas previsões para melhorar assertividade na modelagem de funcionamento dos equipamentos. Alzate et al. (2019) apresentam um modelo de estimação com foco no controle de tensão. Segundo Bagheri e Xu (2017), os benefícios do VVC ficam evidenciados com uma resolução de vinte minutos. Uma compreensão mais ampla do equipamento comutável está disponível no trabalho de Queiroz et al. (2013). Roberts et al. (2020) usam um algoritmo de aprendizado passivo para entender o momento de operação do equipamento OLTC. A identificação do comportamento padrão de desempenho auxilia na detecção de possíveis invasores. Sete artigos mapeados abordam o problema do FDIA, mostrando que a segurança cibernética é uma questão crucial no contexto da SG. A Figura 6 apresenta uma taxonomia proposta para responder à questão geral, considerando todos os artigos analisados. Os conceitos essenciais encontrados no paradigma de estimação de estados foram explorados e estruturados.

Figura 6 – Taxonomia da estimação de estados



Fonte: Elaborado pelo autor.

A Figura 6 mostra os artigos que abordam cada tema proposto para discussão, os números apresentados abaixo de cada quadro estão vinculados ao ID do artigo apresentado na Tabela1.

3.6 Contribuições do estudo

Os trabalhos de Wang et al. (2020) e Zhang et al. (2021) apresentam uma proposta de BDD para pré-processamento de dados e eliminar dados não adequados, porém essas propostas abordam a segurança cibernéticas das SGs o tratamento dos dados não busca obter um melhor desempenho nas técnicas de aprendizado de máquina voltadas para otimização e regressão.

Dobbe et al. (2019) e Yang et al. (2018) mencionam o uso de dados climáticos além de dados históricos para efetuar a previsão de carga de curto prazo. Manitsas et al. (2012) utilizam os perfis dos consumidores como entrada para treinamento de uma RNA e assim gerar pseudo-medições. Esse trabalho apresenta boa assertividade e utiliza dados variados, mas não apresenta uma etapa de pré-processamento ou preparação dos dados para treinamento.

Dominguez et al. (2020) utilizam dados sociodemográficos para treinamento de diferentes modelos de aprendizado de máquina aplicados para detecção de perdas

não técnicas, embora o trabalho apresente bons resultados, não são apresentadas formas de pré-processamento dos dados oriundos de diferentes bases para culminar no objetivo de aprimorar a assertividade dos modelos testados.

Com essa lacuna observada, este estudo apresenta o desenvolvimento de um da etapa de pré-processamento de dados. Essa etapa efetua a limpeza dos dados, corrigi lacunas e padroniza os dados. Diferente dos trabalhos apresentados, nesta metodologia os dados serão correlacionados antes da etapa de treinamento das técnicas de aprendizado de máquina.

Outra contribuição deste trabalho é a classificação dos dados para a correção de *outliers*, tradicionalmente os dados de grandezas elétricas são analisados em três períodos distintos, manhã, tarde e noite. Nessa proposta os dados serão subdivididos em quantas partes forem necessárias para descrever o comportamento ideal das curvas de tensão e potência. De modo que cada subdivisão seja agrupada por dados que apresentam comportamento similar, relacionando os dados históricos de carga e tensão juntamente com a análise dos perfis dos clientes.

Desta forma, o diferencial deste trabalho é a preparação e subdivisão dos dados antes da aplicação na fase de treinamento das técnicas de aprendizado de máquina. Com isso espera-se obter ganhos em precisão e custo computacional.

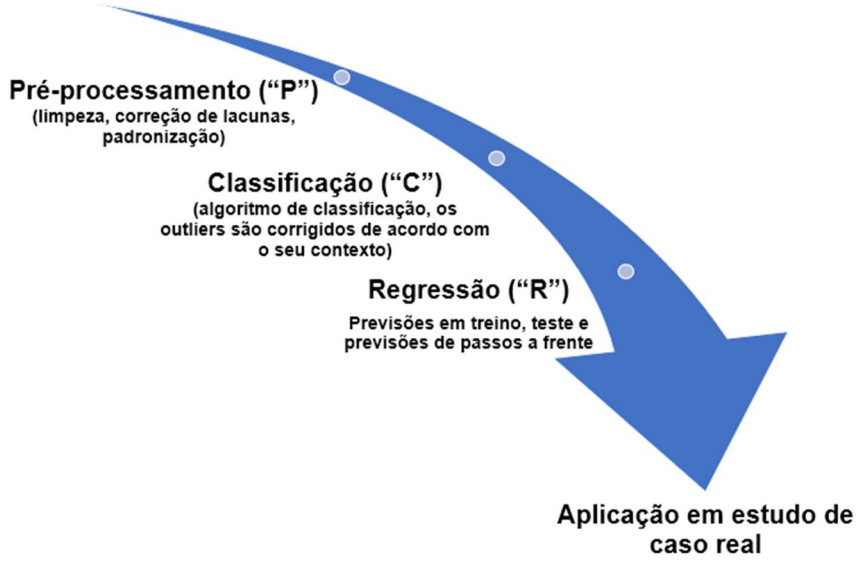
3.7 Comentários sobre o capítulo

Este capítulo teve dois objetivos principais, apresentar trabalhos relacionados com aplicação de predição de grandezas elétricas, e identificar os diferenciais que o modelo PCR possui frente as técnicas utilizadas. Conforme mencionado dois diferenciais são mencionados, o primeiro trata da descrição dos problemas encontrados nas bases de dados reais e como eles foram tratados, a segunda contribuição apresenta a aplicação de algoritmos de classificação para ampliar a capacidade de remoção de *outliers*. O capítulo 4 apresenta o modelo PCR com todos os passos e ganhos obtidos.

4 MODELO PCR

O objetivo principal deste trabalho é desenvolver um modelo híbrido para gerar previsões de grandezas elétricas. Para este objetivo o PCR inicia com a fase de pré-processamento de dados (“P”). Após os dados são classificados (“C”) de acordo com o comportamento da curva especificada, sendo ela de qualquer grandeza elétrica de interesse da concessionária, os *outliers* são corrigidos considerando o contexto dos dados dentro da classificação, por fim os dados são aplicados em técnicas de regressão (“R”) para gerar predições da grandeza elétrica selecionada. A Figura 7 apresenta o modelo “PCR”, o primeiro bloco contempla um *framework* para pré-processamento e contextualização cronológica dos dados, no bloco de classificação os dados tratados são utilizados por algoritmos de classificação buscando a melhor forma de agrupá-los, ao final os dados são aplicados em modelos de regressão para realizar predições.

Figura 7 – Modelo PCR



Fonte: Elaborado pelo autor.

Ao final o modelo será aplicado em um estudo de caso que busca sugerir ajustes nos TAPs de um RT utilizando os dados de saída do modelo, nessa etapa o objetivo é reduzir as comutações dos TAPs do equipamento, dando a ele uma sobrevida maior e reduzindo os custos de manutenção.

As seções a seguir apresentam todo o desenvolvimento do modelo, iniciando pelos passos iniciais que justificam o método, apresentando os erros encontrados em

uma base de dados real, no decorrer do capítulo as técnicas tradicionais de análise de séries temporais e de aprendizado de máquina são discutidas e aprofundadas buscando a melhor opção para o modelo PCR.

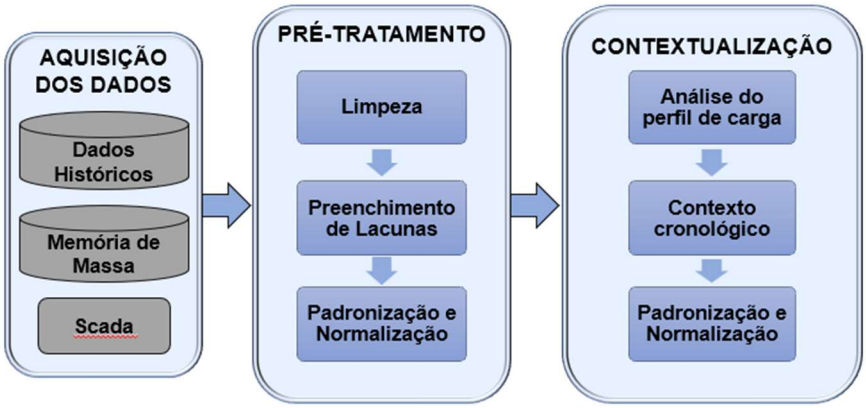
4.1 Pré-processamento

Os conjuntos de dados estão se tornando cada vez maiores, o que torna o pré-processamento e a redução de dados, técnicas essenciais nos cenários de descoberta de conhecimento (RAMÍREZ-GALLEGO et al. 2017). Conforme Alves, Cota e Castro (2022) para alcançar esse objetivo os dados precisam ser limpos, organizados e indexados.

Bancos de dados obtidos por rotinas automatizadas podem conter informações redundantes, ruidosas e por vezes não confiáveis. Desta forma a descoberta de conhecimento se tornará um problema muito difícil, uma vez que as etapas de preparação de dados exigem um tempo de processamento significativo em tarefas de aprendizado de máquina (ALEXANDROPOULOS et al. 2019), problema esse que pode ser atenuado pelo pré-processamento de dados.

Segundo García et al. (2016) as principais etapas do pré-processamento envolvem balanceamento, redução da base de dados de verificação de dados imperfeitos. O balanceamento de dados busca fazer com que os dados não possuam uma classe majoritária que dificulte a análise. A redução dos dados aborda a geração e seleção de instâncias, discretização e seleção dos recursos, já a análise de dados imperfeitos busca por falhas como dados repetidos e ausentes. Esses métodos visam reduzir a complexidade dos conjuntos de dados, para que possam ser processados pelas soluções de mineração de dados (RAMÍREZ-GALLEGO et al. 2017). Deste modo os passos para o pré-processamento da base de dados são exemplificados no decorrer deste capítulo, iniciando pela fase inicial que visa corrigir as falhas mais básicas, como dados zerados, repetidos e fora do padrão. Desta forma, a etapa de pré-processamento dos dados realizada nessa dissertação é organizada em três etapas, aquisição dos dados, pré-processamento e contextualização. Cada etapa possui diferentes aplicações como apresenta a Figura 8.

Figura 8 - Etapa inicial do modelo



Fonte: Elaborado pelo autor.

A primeira etapa consiste na aquisição dos dados dos equipamentos instalados na rede de MT. Inicialmente existem os dados históricos obtidos através do sistema SCADA, disponíveis nos diretórios da concessionaria de energia, esses dados são utilizados para efetuar as análises iniciais e para padronizar o formato dos relatórios. Os equipamentos possuem também memória de massa, onde dados mais atualizados podem ser adquiridos para tornar as análises mais próximas do estado atual da rede distribuição. Por fim o projeto no qual esse trabalho está inserido vai possibilitar o aperfeiçoamento do sistema SCADA, habilitando a coleta de dados em nível local e em tempo real.

No pré-processamento de dados ocorre o ajuste dos dados para eliminar problemas que ocorrem durante de coleta. Devido a problemas de comunicação ou falhas nos sensores alguns pacotes de dados podem ser perdidos ou mesmo duplicados, gerando erros nos dados armazenados. Assim é necessário que esses dados sejam tratados antes de usá-los em análises.

Inicialmente a base de dados será analisada buscando dados ausentes e duplicados. Os dados duplicados são aqueles em que a mesma informação é armazenada mais de uma vez, ou seja, o dado coletado no mesmo instante de tempo se repete mais de uma vez, essas duplicidades podem ser excluídas, porém os dados ausentes devem ser estudados e preenchidos de acordo com a curva de carga.

Para finalizar a etapa de pré-processamento os dados são padronizados, o formato do relatório é diferente dependendo de como é extraído do repositório da concessionária, nessa etapa criasse um padrão a ser seguido para que todas as bases

sejam analisadas sob a mesma ótica. Nesse procedimento dados considerados desnecessários, que não influenciam diretamente na montagem das curvas das grandezas elétricas são removidos, deixando o *dataset* organizado e adequado para as etapas seguintes. A etapa de contextualização dos dados consiste na indicação cronológica dos dados, separando as informações referentes a data e hora de forma que possa existir um padrão de amostragem na base de dados, nessa fase os dados também são classificados quanto a dias úteis, feriados e finais de semana.

Conforme definido a etapa inicial o PCR irá procurar por lacunas, repetições e dados ausentes. Verificando a base de dados fornecida pela concessionária vários problemas como os mencionados foram encontrados. A Tabela 2 mostra diversos dados de corrente e tensão onde aparecem dados zerados, isso pode representar um defeito transitório na rede de distribuição, que levaria os dados de tensão e corrente para zero, ou uma falha na coleta de dados. O intervalo de aquisição de um novo dado é de quinze minutos.

Tabela 2 – Erros na base de dados

Data	Ia	Ib	Ic	Va	Vb	Vc
22/09/2021 11:00:01	79	86	79	7.485	7.491	7.613
22/09/2021 10:45:01	0	0	0	0,00	0,00	0,00
22/09/2021 10:30:01	0	0	0	0,00	0,00	0,00
22/09/2021 10:15:01	0	0	0	0,00	0,00	0,00
22/09/2021 10:00:01	0	0	0	0,00	0,00	0,00
22/09/2021 10:00:01	109	109	109	7.319	7.370	7.485
22/09/2021 09:45:02	0	0	0	0,00	0,00	0,00
22/09/2021 09:45:01	93	94	94	7.498	7.549	7.587
22/09/2021 09:30:01	99	99	99	7.472	7.504	7.555
22/09/2021 09:30:01	0	0	0	0,00	0,00	0,00
22/09/2021 09:15:01	88	84	86	7.504	7.632	7.632
22/09/2021 09:15:01	0	0	0	0,00	0,00	0,00

Fonte: Elaborado pelo autor.

Como pode ser visto na Tabela 2 existem diversos valores duplicados, ou seja, coletados no mesmo instante de tempo. Um ponto interessante da duplicidade encontrada é que em algumas vezes uma linha contém o dado correto e outra coletada no mesmo instante de tempo apresenta os dados zerados. Então nessa etapa ao invés de apenas excluir o dado repetido, é verificado se a repetição não contém o dado correto assim é excluído apenas o dado que comporta o erro. Outro erro encontrado

foi a ausência de dados por um longo período. A Tabela 3 mostra uma falha deste tipo, onde a aquisição de um dado novo ocorre quatro horas após a última coleta.

Tabela 3 – Dados ausentes

Data	Ia	Ib	Ic	Va	Vb	Vc
11/12/2020 15:30:00	60	66	63	7.421	7.498	7.748
11/12/2020 15:15:00	0	0	0	7.901	7.978	8.209
11/12/2020 15:00:00	0	0	0	8.702	8.977	8.977
11/12/2020 14:45:00	37	36	30	7.620	7.780	7.773
11/12/2020 14:30:00	0	0	0	7.671	7.837	7.722
11/12/2020 14:15:00	0	0	0	7.895	7.933	8.055
11/12/2020 10:15:00	95	105	98	7.459	7.331	7.447
11/12/2020 10:00:01	0	0	0	0,00	0,00	0,00
11/12/2020 10:00:00	100	108	99	7.434	7.319	7.415
11/12/2020 09:45:01	0	0	0	0,00	0,00	0,00
11/12/2020 09:45:00	96	100	97	7.466	7.383	7.427
11/12/2020 09:30:00	96	104	97	7.408	7.280	7.370

Fonte: Elaborado pelo autor.

Pode ser visto na Tabela 3 que o erro de dados duplicados aparece novamente, e após esse erro houve um período sem coleta de dados, voltando a normalidade na sequência. Esse problema será resolvido utilizando uma média baseada em dados históricos dentro do mesmo contexto do dado ausente, ou seja, dados que estão presentes no mesmo conjunto definido pela similaridade, possuindo o mesmo tipo de comportamento.

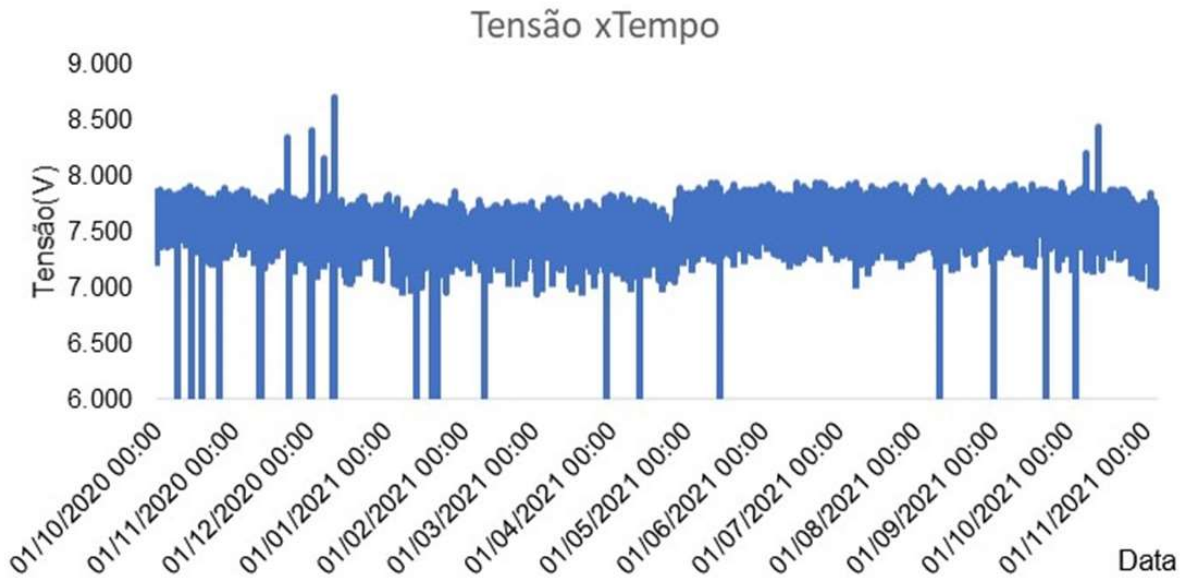
Para solucionar os problemas mencionados, foi desenvolvido um *framework* que efetua o pré-processamento dos dados e os armazena em um novo diretório para serem posteriormente utilizados em análises. Inicialmente o *framework* analisa o banco de dados e remove os dados duplicados, isso é feito analisando a duplicidade do registro que identifica a data e hora da coleta dos dados. Como foi desenvolvido em linguagem *python*, essa etapa consiste em apenas poucas linhas de códigos que excluem essa duplicidade.

Na segunda etapa são analisados os dados ausentes e zerados. A busca por dados zerados varreu todas as colunas da base de dados e onde encontrou o valor zerado ou nulo o substituiu pela média de cinco leituras anteriores e cinco posteriores ao erro encontrado, caso os dados posteriores também estejam zerados o algoritmo

busca pelo próximo dado válido e o insere no cálculo da média junto com o conjunto de dados anteriores.

A Figura 9 apresenta graficamente o efeito dos dados zerados na curva de tensão medida em volts. Como pode ser visto os dados zerados geram grandes curvas ao longo do tempo, se esses dados fossem utilizados diretamente em um modelo de regressão o custo computacional seria alto para conseguir reproduzir apenas as lacunas discrepantes da realidade.

Figura 9 – Dados zerados



Fonte: Elaborado pelo autor.

Na Figura 9 pode-se notar a presença de *outliers* na parte superior do gráfico, com valores de tensão superando 8500 volts, esses erros serão corrigidos na etapa de classificação.

Outra análise sobre os dados zerados é apresentada Figura 10, nessa análise pode-se observar que entre a variação dos dados zerados existe pontos que mantém a identidade da curva, o que válida a proposição feita anteriormente de utilizar a média das leituras encontradas antes e após o erro.

Figura 10 – Dados zerados particionados

Fonte: Elaborado pelo autor.

Aplicando a abordagem proposta obtém-se uma aproximação da curva real que deveria ter sido obtida nas leituras que continham erros, a Figura 11 apresenta essa correção dos dados, mostrando a nova curva de tensão.

Figura 11 – Correção dos dados zerados

Fonte: Elaborado pelo autor.

Outro erro da base de dados com grande interferência nas análises posteriores é a ausência de dados apresentado na Tabela 3, que mostra um intervalo de quatro horas sem a aquisição de nenhum registro. O intervalo de coleta de dados nos supervisórios de aquisição instalados é de quinze minutos, portanto para corrigir o problema de dados ausentes foi criada uma coluna que calcula o intervalo de aquisição de novos registros, quando esse intervalo é diferente de quinze minutos faz-se então uma etapa de análise, para observar se é apenas um erro no horário, por exemplo dado obtido no minuto quatorze ao invés de quinze, ou se ele é mesmo

ausente. Caso o algoritmo perceba que é mesmo um dado ausente é inserido uma nova linha entre os registros que inicialmente mantém os valores das medições iguais a zero, após a função de correção de dados zerados é chamada corrigindo esses pontos para a média local.

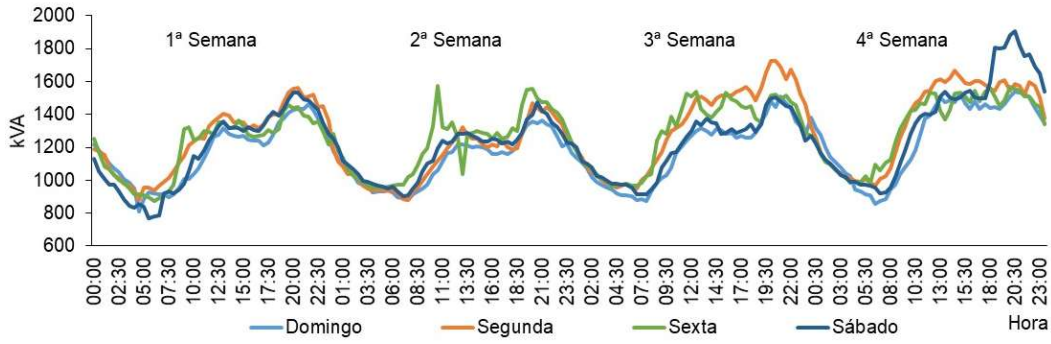
Ao final dessa etapa tem-se uma base de dados limpa de dados zerados e ausentes, garantindo a existência de 96 registros para cada dia, respeitando o intervalo de 15 minutos entre os registros. Com essa padronização realizada o próximo passo visa fazer a remoção dos *outliers* usando uma etapa de classificação para separar os dados em diferentes *clusters*, para que as particularidades das curvas sejam identificadas e consideradas, como apresentado a seguir.

4.2 Classificação dos dados

Além das leituras obtidas nos equipamentos, outros fatores podem influenciar no comportamento da rede de distribuição, como os tipos de consumidores conectados (indústrias, comércio ou residências), fatores climáticos como temperatura e condição do tempo, além de dados cronológicos que indicam a época do ano, dia da semana, feriados, ou seja, dados que identificam a sazonalidade existente no setor de distribuição.

A demanda por energia elétrica pode variar dependendo do dia semana. Sábado, domingos e feriados podem ter demandas diferentes dos dias úteis devido as atividades do comércio e indústrias. A Figura 12 apresenta uma análise para um equipamento predominantemente residencial. Os dados são relativos ao mês de outubro de 2020, mostrando todos os domingos, segundas, sextas e sábados do mês.

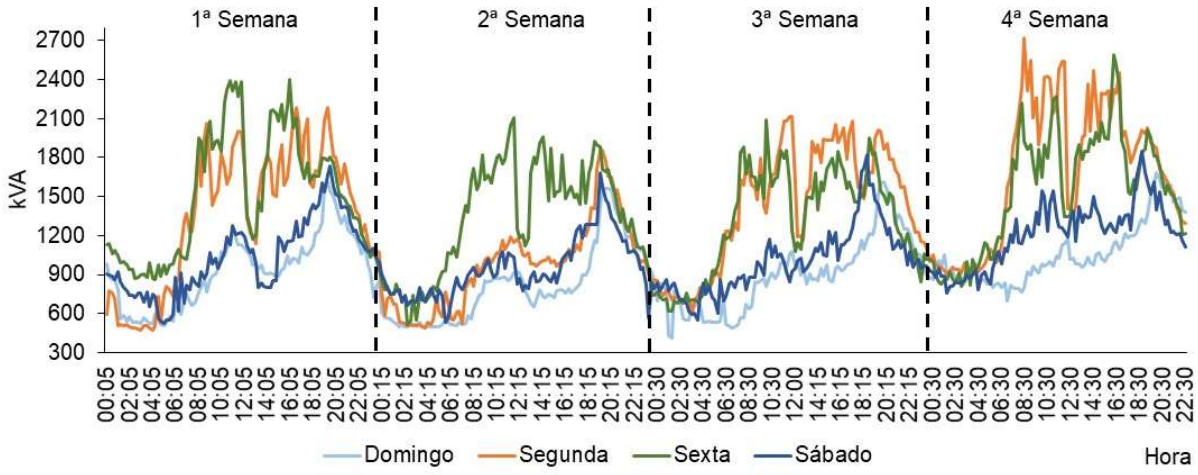
Figura 12 – Potência aparente em diferentes dias



Fonte: Elaborado pelo autor.

Como pode ser observado não existem diferenças significativas nas medições entre os dias úteis e não úteis nesse contexto, na verdade o maior pico de potência foi registrado em um sábado. Isso ocorre devido ao perfil dos consumidores, como a carga é predominantemente residencial, entende-se o maior consumo em um dia em que as pessoas estão em suas residências. Analisando outro equipamento, que possui um perfil de carga distinto, no mesmo período do ano, esse cenário muda, como apresenta a Figura 13.

Figura 13 - Potência aparente em diferentes dias



Fonte: Elaborado pelo autor.

Nessa análise a diferença entre os dias úteis e não úteis é mais clara, sábado e domingo possuem leituras menores no horário comercial. Também é possível verificar uma queda significativa da demanda nos dias úteis entre o horário de 11:30 e 13:30, horário comum de intervalo em muitas empresas.

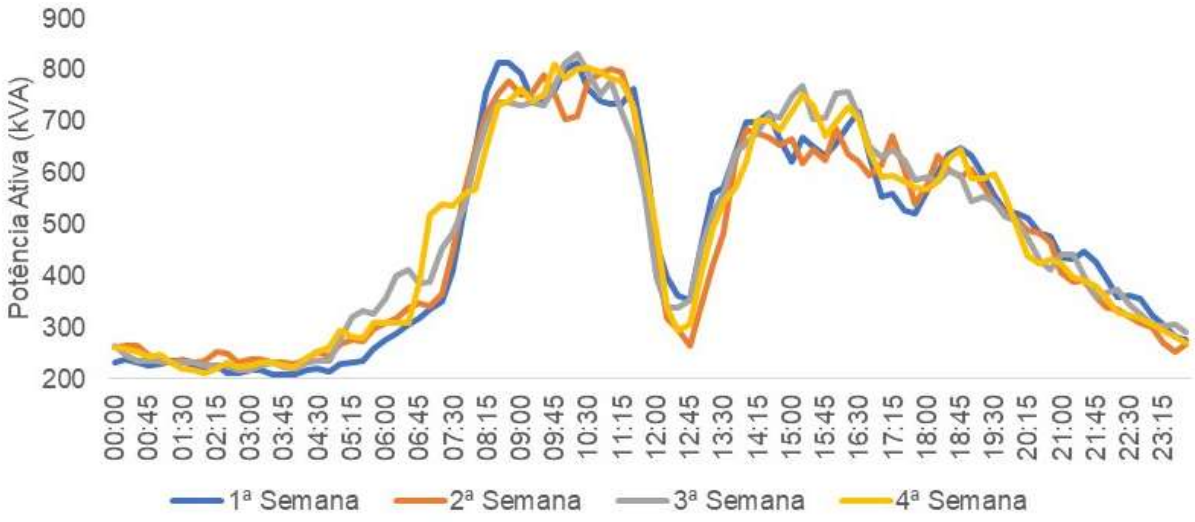
Outro fato interessante na Figura 13 é a demanda de uma segunda-feira presente na segunda semana do gráfico, apresentada na cor laranja. A demanda desse dia específico está bem abaixo dos outros, mostrando um evento atípico, uma vez que a curva de carga perdeu a sua assinatura. Como a análise está sendo feita sobre o mês de outubro de 2020, essa data trata do dia doze de outubro, feriado nacional no Brasil, desse modo pode-se observar que o comportamento da curva do feriado se assemelha ao de um dia não útil.

As análises preliminares, embora feitas sobre um conjunto pequeno de dados, apresentam como diversos fatores influenciam na demanda de energia elétrica. Deste modo uma aplicação de inteligência artificial em modelos que pretendam, por

exemplo, resolver problemas de regressão, como criar pseudo-medições, devem considerar todos esses aspectos. Tendo esses comportamentos analisados percebe-se a importância de que cada caso deve ser analisado de acordo com o seu contexto, considerando os fatores locais que interferem na modelagem das curvas de potência. Com isso nas próximas fases serão apresentadas as discussões referentes apenas aos dados coletados no RT alvo do estudo de caso do próximo capítulo.

Como foi mencionado anteriormente a correção dos *outliers* será feita baseada no contexto das curvas analisadas, para isso são criados *clusters* para subdividir os dados. Com base no histórico do equipamento pode-se notar pequenas sazonalidades mensais e subdivisões que se repetiam dentro de cada dia da semana, a Figura 14 apresenta um exemplo desta análise, são apresentados dados das quartas-feiras do mês de outubro de 2021, pode-se perceber que embora ajam pequenos pontos de pico as curvas apresentam o mesmo tipo de comportamento.

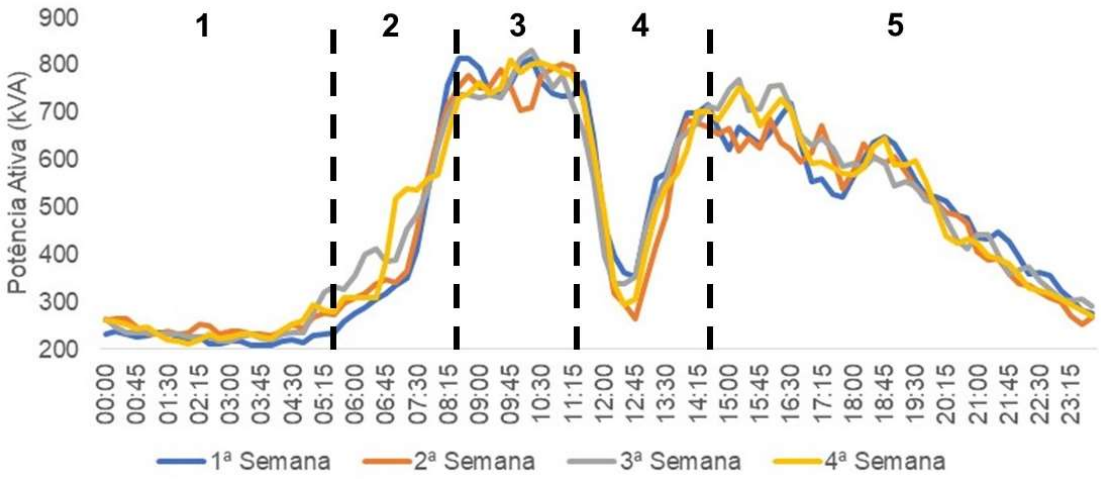
Figura 14 – Identidade das curvas de carga



Fonte: Elaborado pelo autor.

Com base na Figura 14 a divisão dos *clusters* deste exemplo pode ser feita em cinco estágios, a Figura 15 apresenta essa sugestão.

Figura 15 – Subdivisão em *clusters*



Fonte: Elaborado pelo autor.

Na Figura 15 a carga do dia foi subdividida em 5 clusters diferentes, de acordo com o comportamento da curva. Fazer essa subdivisão para um exemplo apenas não significa que as curvas se aplicam a toda a base de dados, para isso é necessário utilizar uma metodologia para classificar o quanto as curvas se assemelham ao padrão estabelecido.

Como discutido e apresentado na Figura 15, a subdivisão dos dias em *clusters* pode facilitar o reconhecimento de padrões, auxiliar na correção de *outliers* e ser fator central na criação de diferentes modelos de regressão para atuar em previsões de grandezas elétricas. A Figura 15 trouxe apenas uma ilustração da ideia de *clusters*, baseando-se no aspecto visual de uma curva, como este não é um método científico, é necessário usar ferramentas já reconhecidas e estabelecidas para este fim, duas metodologias que podem fazer essa classificação são os algoritmos de aprendizagem de máquina *Fuzzy C-Means* (FCM) e o *K-MEans* (KME). Os métodos são apresentados, discutidos e aplicados, ao final discutem-se os resultados e o melhor método é escolhido para incorporar o PCR

4.2.1 Método *k-means*

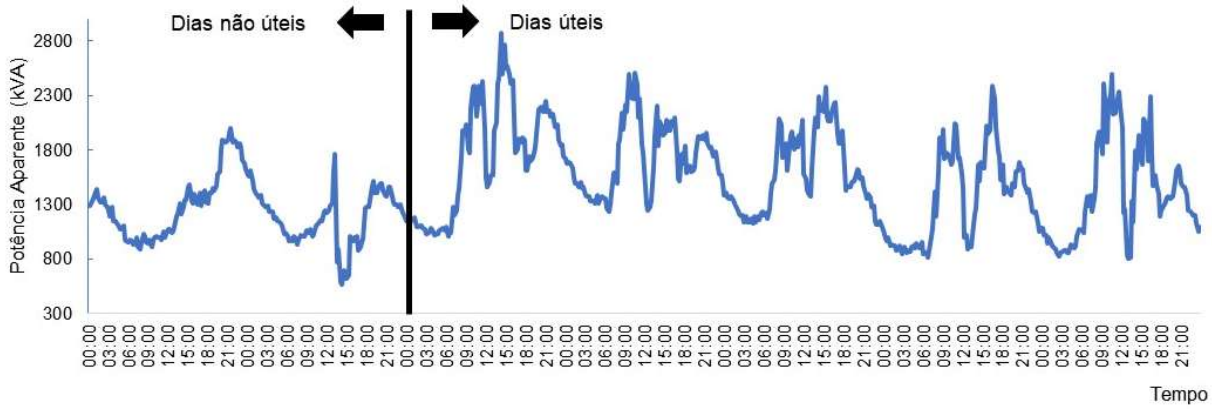
Esse algoritmo pertence ao grupo de aprendizado não supervisionado é utilizado para efetuar agrupamento de dados. A partir de uma base de dados o algoritmo mapeia os dados a fim de separá-los em diferentes grupos, assim o algoritmo KME divide os dados em “k” categorias após realizar múltiplas iterações.

Zhang et al. (2022) detalha que o algoritmo usa iterações para calcular a distância de cada ponto de dados ao centro do *cluster* mais próximo, após o método de atualização dos centros calcula o valor médio dos atributos correspondentes para atualizar o centro de cada cluster, na última etapa são comparadas as distâncias entre os centros, se essa distância respeitar um determinado limite o algoritmo encerra, caso contrário continua com as iterações.

Como o algoritmo busca agrupar os dados ao redor de um centroide, é razoável pensar que quanto mais similares forem os dados, melhor será a separação dos em grupos. Se os dados forem extremamente discrepantes necessitarão de um número grande de grupos para representá-los ou então a variância deles não será representada. Deste modo uma análise prévia para entender a curva de carga do equipamento é necessária para ajudar na performance do algoritmo.

O primeiro passo é reconhecer o perfil dos consumidores, a Figura 16 apresenta a curva de potência do RT estudado. Pode-se observar que existe uma diferença entre os dias úteis e não úteis, este fato se deve a existência de industriais e comércios, além dos clientes residenciais, por este motivo os picos de consumo são atingidos em horário comercial.

Figura 16 – Análise da curva de potência



Fonte: Elaborado pelo autor.

Além da diferença de consumo mencionada, é possível perceber que existe redução no consumo no período entre 12:00 horas e 13:00 horas, período em que as empresas do local devem fazer intervalo para refeição, com isso percebesse a importância de o algoritmo conseguir identificar a variância em um nível alto.

Para identificar o número de *clusters* mais indicado para a base de dados pode-se utilizar dois métodos em conjunto: o método *elbow* e o *silhouette score*. O método

elbow consiste em executar o algoritmo KME para diversos números de k, verificando qual o k mais indicado, pode-se pensar que esse método indica um valor em que o número de *clusters* não seja pequeno ao ponto de não reconhecer as variâncias e nem tão grande a ponto de deixar as diferenças intra-*clusters* pequenas demais a ponto de não serem observadas pela proximidade dos centroides. Com isso esse método apresenta uma curva onde o número ótimo para k fica onde temos a curva mais acentuada no gráfico, ou seja, onde a curva forma um cotovelo. Já o *silhouette score* é um coeficiente calculado usando a distância euclidiana média intra-*cluster*, quanto mais próximo um ponto está do seu centroide, menor a distância e, portanto, melhor agrupado está o *cluster*, a resposta deste cálculo é um valor entre -1 e 1, sendo 1 o ponto ótimo, valores próximos a zero indicam *clusters* sobrepostos, enquanto valores negativos indicam que amostras foram atribuídas ao *cluster* errado.

Com isso têm-se as ferramentas necessária para efetuar a classificação e verificação da assertividade do modelo, porém este não é o único método indicado para este fim, com objetivo de encontrar os melhores resultados, também foi estudado o método FCM.

4.2.3 Método *fuzzy c-means*

O método FCM é uma técnica de aprendizado não supervisionada utilizada para efetuar a classificação e agrupamento de dados em diferentes *clusters*. Esta técnica efetua a classificação dos dados de forma percentual, indicando a probabilidade de um dado pertencer a cada grupo previamente indicado, utilizando o mesmo princípio da técnica *fuzzy* tradicional (Xu, 2022).

Chen et al. (2022) salienta que a técnica é robusta quando aplicada para agrupar dados que podem pertencer a mais de um grupo, sendo adequada a utilização deste estudo que trata de curvas de corrente e tensão que variam de acordo com o tempo, mas possuem certa similaridade. Neste ponto esta técnica não tem uma indicação tão precisa para indicação do número de *clusters* ideal para cada problema como a destacada para o método KME, porém ela apresenta como retorno o coeficiente *Fuzzy Partition Coefficient* (FPC), este coeficiente é um valor entre 0 e 1 que é inversamente proporcional a probabilidade de um dado estar em mais de um cluster, então pode-se interpretá-lo da forma que quanto mais próximo de 1, maior será chance de que os dados não estejam em *clusters* sobrepostos, porém um valor

muito próximo de 1 indica que a variância dos dados representada pode ser pequena. Como descrito a interpretação deste coeficiente pode ser complicada, porém durante a aplicação que será descrita na seção 5.2 esse coeficiente ficará claro quanto a sua importância para a classificação do desempenho do método.

O método FCM foi utilizado com sucesso para classificação de perfis de carga como no trabalho de Anuar et al. (2021). Embora o método seja robusto ele apresenta a deficiência de que os cálculos da matriz de pertinência são de grau elevado, o que aumenta a complexidade computacional, fazendo com que esse algoritmo não seja indicado para uso em grandes bases de dados. Uma possível solução para este problema pode ser a redução da base de dados.

4.2.3 Método *Principal Components Analysis*

O método *Principal Components Analysis* (PCA) consiste em cálculos matemáticos que visam reduzir a dimensionalidade de uma base de dados, com a finalidade de que os dados principais mantenham a identidade dos dados, porém em uma base de dados menor.

Halstead et al. (2018) destaca que o método PCA é capaz de reduzir a dimensionalidade de um conjunto de dados preservando as características mais críticas. Rehman et al. (2020) apontam que o método consiste em transformar características possivelmente correlacionadas em um conjunto linear não correlacionado, após a transformação por PCA o primeiro componente possui a maior variância, o segundo componente possui a segunda maior variância, e assim subsequentemente.

Desta forma este método pode ser utilizado para agrupar dados que pertences a um contexto cronológico pré-estabelecido, reduzindo o custo computacional na execução dos algoritmos de classificação. Essa abordagem será destacada na próxima seção com a aplicação dos três métodos KME, FCM e PCA.

4.2.4 Subdivisão dos dados em clusters

Como mencionado nas seções anteriores o objetivo desta etapa é agrupar os dados em *clusters*, relacionando os dados por seu contexto cronológico e pela identidade da curva de cada grandeza elétrica mensurada, nomeadas tensão,

corrente e potência. Além disso, busca-se uma forma de melhorar a acurácia de modelos de regressão que serão aplicados nas próximas seções.

Com isso, o primeiro passo foi classificar a base de dados de acordo com o contexto cronológico, indicando o dia, mês, ano, hora e dia da semana. Como os dados possuem amostragem de quinze minutos uma classificação de 1 a 96 foi aplicada para subdividir cada dia. Essas informações foram adicionadas a base de dados que havia sido previamente tratada para a eliminação de dados ausentes, duplicados e inválidos.

O método KME e FCM possuem diferença quanto a entrada de dados, por isso algumas simulações foram feitas de forma diferente, porém preservando o mesmo objetivo final para poder determinar o melhor algoritmo para esse estudo de caso.

No primeiro cenário de avaliação são utilizados os dados de todas as medições de um mês completo, e em um segundo cenário é avaliado apenas um dia da semana específico deste mesmo mês, utilizando como exemplo a segunda-feira e o mês de janeiro de 2021.

A Tabela 4 apresenta a entrada de dados para o modelo FCM, a primeira coluna mostra o número da amostra, subdividida de 1 a 96, as próximas 4 colunas trazem os dados de cada segunda-feira do mês, e a coluna tratada pelo PCA. Desta forma a entrada real do FCM para o segundo cenário é a primeira e a última coluna apresentada na Tabela 4, as colunas intermediárias farão parte do experimento do primeiro cenário apenas. De forma sucinta um experimento utiliza os dados da forma como são coletados, ou seja, sequencialmente, no segundo é considerado apenas um dia da semana agrupado pelo método PCA.

Tabela 4 – Entrada de dados FCM

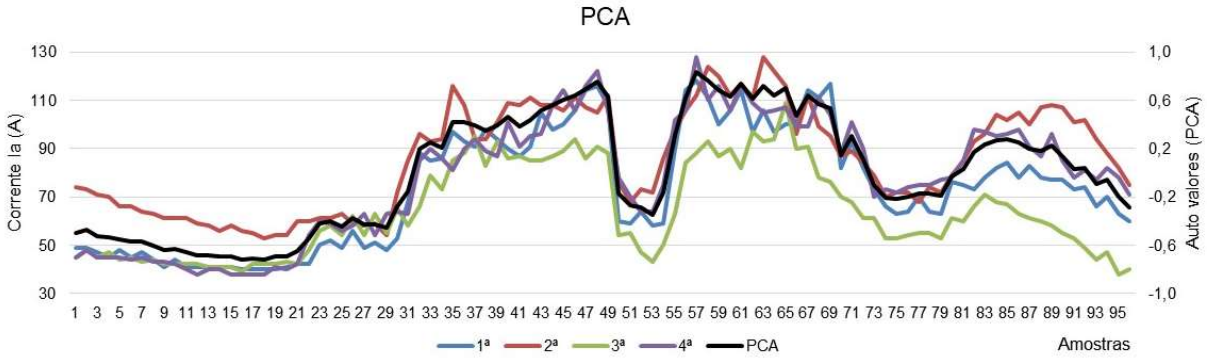
Entrada 1	Leituras intermediárias de corrente (A)				Entrada 2
Amostras	1ª Semana	2ª Semana	3ª Semana	4ª Semana	PCA
1	49	74	45	45	-0,5008
2	49	73	48	48	-0,4739
3	47	71	45	45	-0,5286
4	45	70	47	45	-0,5373
5	48	66	44	45	-0,5529

Fonte: Elaborado pelo autor.

Para o método KME a entrada de dados não necessita estar no formato bidimensional (x, y) deste modo as informações das amostras não precisam necessariamente estar presentes. Deste modo a primeira coluna da Tabela 4 pode ser descartada.

Como apresentado na Tabela 4 o método PCA transformou os dados das quatro semanas em uma representação em uma única coluna que contém os dados principais, para verificar a validade do método, a Figura 17 apresenta graficamente as quatro semanas plotadas no eixo y principal e o retorno do PCA no eixo y secundário, como pode ser observado a identidade da curva é mantida pelo PCA e os componentes principais que a representam estão contidos no gráfico.

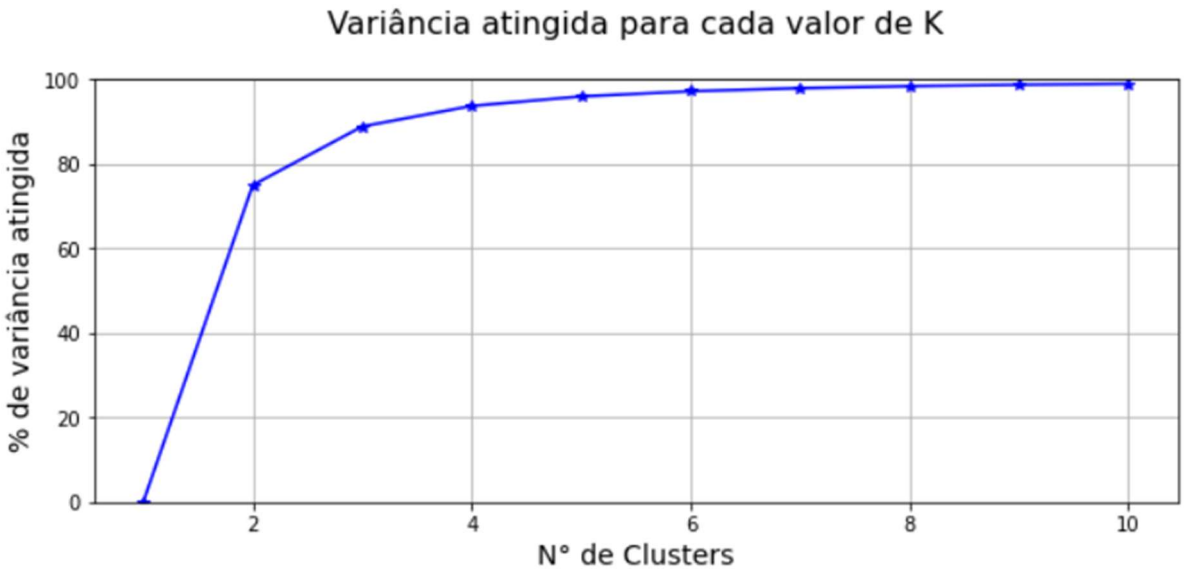
Figura 17 – Método PCA



Fonte: Elaborado pelo autor.

Para definir a quantidade correta de *cluster* a serem utilizados em cada método foi utilizada a curva de *Elbow* para o método KME, e o indicador FPC para o FCM. A Figura 18 apresenta a curva para o cenário de análise que envolve as segundas-feiras do mês de janeiro de 2021, como pode ser observado os melhores resultados ocorrem para 3,4 e 5 *clusters* após o gráfico entra na fase de platô e não existe ganho real na variância atingida com o aumento do número de *clusters*.

Figura 18 – Curva de *elbow* para análise do número de *clusters*



Fonte: Elaborado pelo autor.

Já o método FCM retorna o indicador FPC de forma numérica, a Tabela 5 apresenta o FPC para cada número de clusters.

Tabela 5 – FPC para análise de quantidade de *clusters*

Clusters	FPC
2	0.85485
3	0.80175
4	0.77341
5	0.75553
6	0,74326
7	0,73413
8	0,72723
9	0,72425
10	0,71060

Fonte: Elaborado pelo autor.

Para comparar os resultados do método KME foi utilizado o *silhouette score* mencionado na seção 5.2.1, como apontado esse indicador retorna um coeficiente entre -1 e 1, sendo 1 o melhor resultado. A Tabela apresenta esse coeficiente para os cenários destacados.

Tabela 6 - *Silhouette score* diferentes cenários para o método KME

Clusters	Todos os dados	Segundas PCA
3	0,5893	0,4967
4	0,5705	0,4808
5	0,5588	0,4679
6	0,5504	0,4240

Fonte: Elaborado pelo autor.

Como pode ser observado a maior pontuação ocorreu com o uso de todos os dados de forma sequencial, esse resultado acaba descartando o uso do algoritmo PCA para reduzir a base de dados durante a classificação, uma vez que ao diminuir a variância dos dados se limita também a probabilidade de o algoritmo a alcançar. A Tabela 7 apresenta o mesmo experimento para o método FCM.

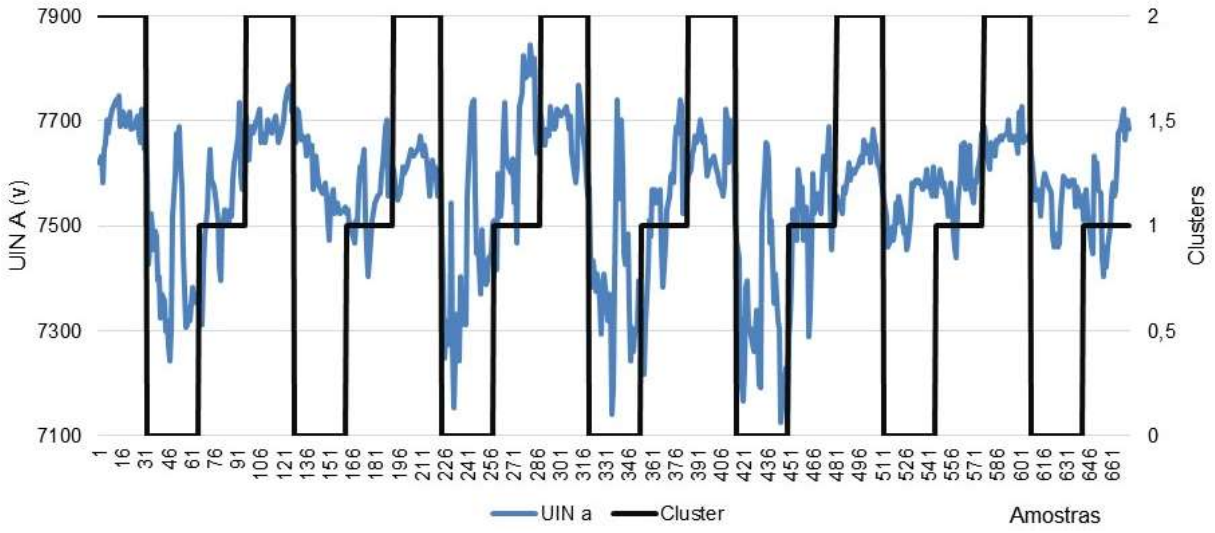
Tabela 7 - *Silhouette score* diferentes cenários para o método FCM

Clusters	Todos os dados	Segundas PCA
2	0,6269	0,6188
3	0,5897	0,5768
4	0,5713	0,5528
5	0,5601	0,5361
6	0,5531	0,5240

Fonte: Elaborado pelo autor.

Como visto nas Tabelas 6 e 7 o método FCM com uso de dois *clusters* teve a melhor performance em agrupar os dados, porém como visto na Figura 17 a variância é abaixo de 80% para este número, portanto o uso de três a cinco *clusters* é mais recomendada considerando os dois fatores estudados, curva de *elbow* e *silhoutte score*. A Figura 19 apresenta a subdivisão dos dados em três *clusters*, mostrando no eixo y principal a tensão e no eixo y secundário os *clusters*, são apresentados dados referentes a sete dias.

Figura 19 – Classificação dos dados



Fonte: Elaborado pelo autor.

Pode-se perceber na Figura 19 que a subdivisão em *clusters* representa a tendência da curva de tensão em cada fase do dia, mantendo-se estável mesmo sob a variação da carga em dias úteis e dias não úteis, que aparecem mais à direita do gráfico. Com essa etapa concluída o próximo passo é a verificação dos outliers considerando todos os passos realizados até aqui.

4.2.5 Detecção e correção de *outliers*

Após a classificação dos dados o próximo passo é a correção dos *outliers*. Inicialmente será feita a verificação da existência dos mesmos e após a demonstração do efeito dessa correção.

A detecção de *outliers* será feita utilizando o método de Grubbs, assim como apresentado no trabalho de Meng et al. (2021). Essa etapa foi realizada filtrando os dados pelo *cluster* e pelo mês do registro, por exemplo, separando os dados referentes a um ano (doze meses) com subdivisão em dois *clusters* são verificados os dados separados em 24 grupos, ou seja, dois vetores são criados para cada mês, este ensaio citado como exemplo retornou à existência de *outlier* em apenas um caso. O retorno do *framework* é a referência da ocorrência, indicado o mês e o *cluster* da ocorrência, como apresentado na Figura 20, a partir desta informação o especialista da área pode visualizar através da análise gráfica se existe mesmo um outlier ou se

apenas se trata de uma ocorrência normal que pode voltar a ocorrer com o passar do tempo.

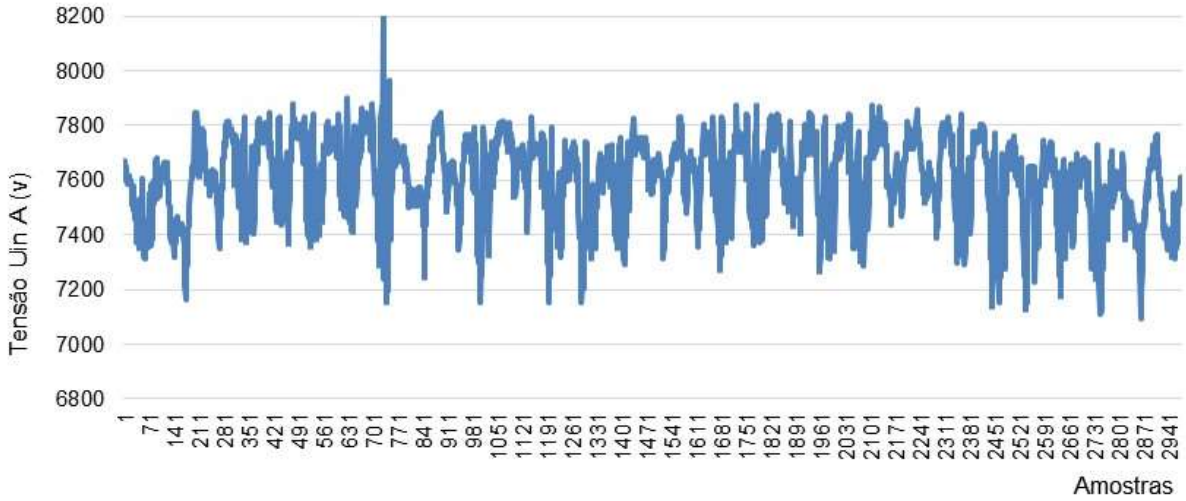
Figura 20 – Indicação da ocorrência de *outliers*

```
Grupos de dados sem outliers: 23
Grupos de dados com outliers: 1
Identificação do mês e cluster das ocorrências: [10, 1]
```

Fonte: Elaborado pelo autor.

Através da referência fornecida é possível plotar o gráfico do mês em questão para verificar se existe mesmo um *outlier*, a Figura 21 apresenta esse gráfico de tensão ao longo do tempo para a referência citada na Figura 20.

Figura 21 – Identificação do *outlier*



Fonte: Elaborado pelo autor.

Como pode ser observado na Figura 21 existe um ponto em que tensão está acima de 8200 v, sendo um *outlier* visível. Alterando o número de *clusters* é possível detectar a presença de mais *outliers*. Isso é explicado pelo fato de que ao aumentar o número de *clusters* a variância dos dados atingida na classificação é maior, como apresentado na curva de *Elbow*, assim espera-se que mais *outliers* locais sejam observados, antes que o número de *clusters* alcance o estado de platô da curva de *Elbow*. A Tabela 8 apresenta a quantidade de *outliers* que são detectadas para cada faixa.

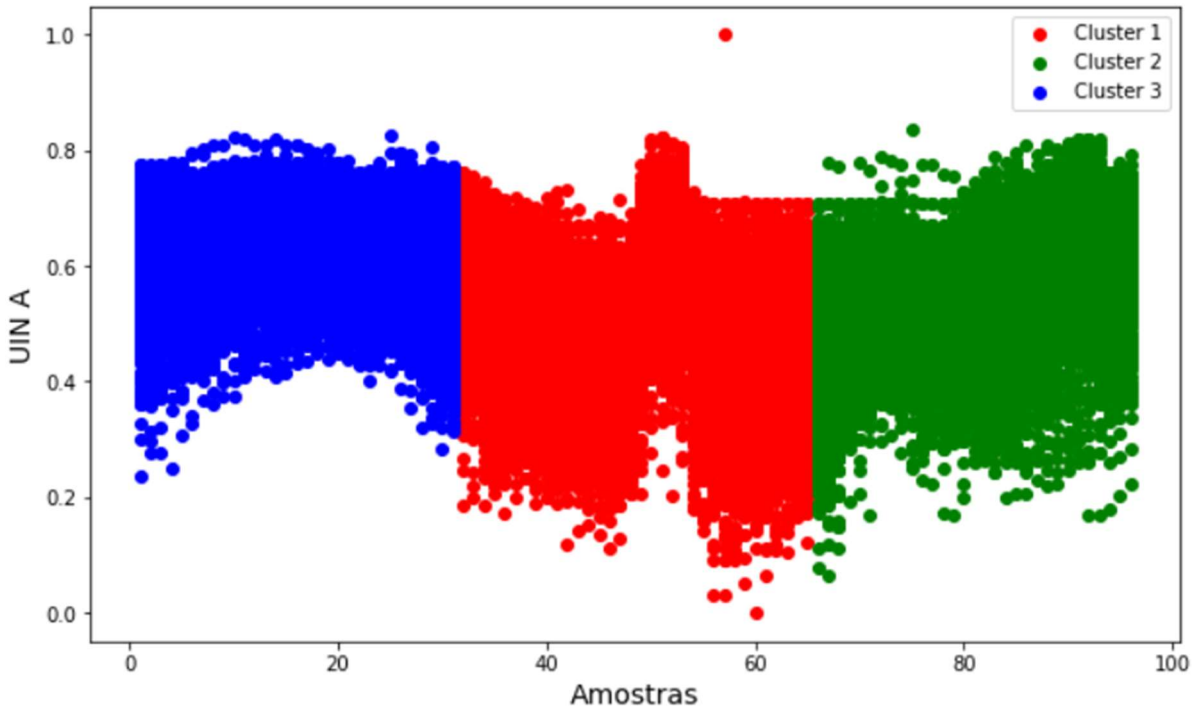
Tabela 8 – *Outliers* verificados em cada grupo pelo método Grubbs

Clusters	Grupos verificados	*Outliers*
2	24	1
3	36	8
4	48	2
5	60	3
6	72	6

Fonte: Elaborado pelo autor.

Como visto utilizando 3 *clusters* o método Grubbs encontrou ocorrência de *outliers* em 8 casos, ou seja, em 33,3% dos grupos. Também é possível verificar a existência dos *outliers* utilizando um gráfico de dispersão, separando os dados em faixa percentual do *clusters* no eixo x e a tensão medida no eixo y, como mostra a Figura 22.

Figura 22 – Gráfico de dispersão com 3 *clusters*

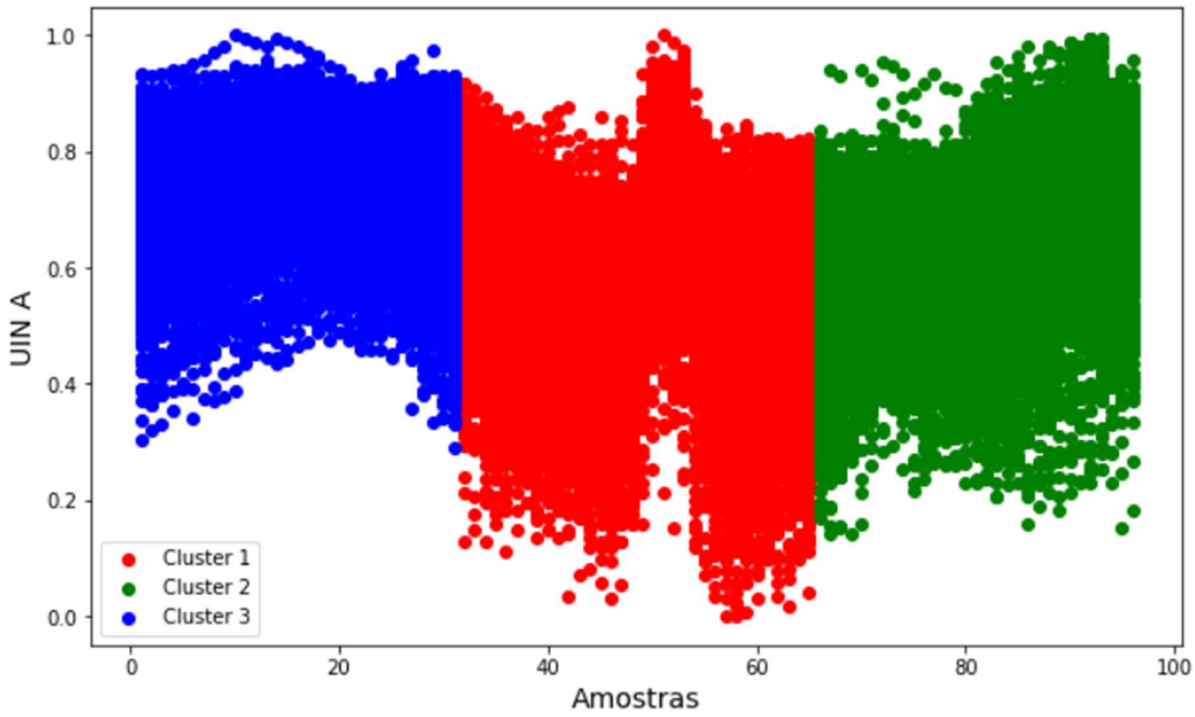


Fonte: Elaborado pelo autor.

A correção dos *outliers* foi realizada utilizando a média local das leituras coletadas antes e depois do dado fora da curva. Analisando novamente o gráfico de

dispersão, nota-se a remoção destes dados, sem que a identidade verificada anteriormente seja perdida, como mostra a Figura 23.

Figura 23 - Gráfico de dispersão com *outliers* corrigidos



Fonte: Elaborado pelo autor.

Nesse ponto as etapas de pré-processamento, contextualização e classificação dos dados por similaridade estão prontas, deixando a base de dados pronta para ser utilizada no algoritmo de regressão

4.3 Regressão

O aprendizado de máquina é uma técnica emergente que tem como objetivo instruir computadores através de análise de dados para resolver um determinado problema (MUHAMMAD; YAN, 2015). Al-Sahaf et al. (2019) complementam afirmando que o aprendizado de máquina é um ramo da inteligência artificial baseado na ideia de que os sistemas podem aprender a partir dos dados, identificar padrões ocultos e tomar decisões com pouca ou mínima intervenção humana. Com isso nessa etapa do desenvolvimento do PCR buscou-se definir qual melhor método de regressão para efetuar a predição de grandezas elétricas de curto prazo, nessa linha, inicialmente é

realizado um estudo sobre a série temporal formada pelos dados históricos e após são aplicados em redes neurais artificiais.

4.3.1 Análise da série temporal e métricas de desempenho

Uma série temporal apresenta o comportamento de uma variável ao longo do tempo, e pode ser dividida em três componentes, sazonalidade, ciclo de tendencia e outra componente, Hyndman e Athanasopoulos (2021). A sazonalidade é um componente de grande importância quando se faz análises sobre o sistema de distribuição de energia elétrica, pois a carga é diferente dependendo da estação do ano, portanto essa componente deve aparecer na subdivisão da série temporal. A componente restante mencionada pode conter qualquer outro comportamento que descreva o comportamento da série temporal, neste estudo podemos pensar no efeito do tipo de cliente que está conectado à rede, que tem a sua peculiaridade no impacto de consumo e deve ser analisado. A componente ciclo de tendência traz o comportamento da curva, descrevendo a tendência dos acontecimentos sejam crescentes ou decrescentes.

Morettin e Toloi (2006) destacam que as séries temporais podem ser utilizadas para descrever o comportamento da série através de gráficos que separam as componentes da série e para fazer previsões de valores futuros da série. Com isso no primeiro momento são extraídos dados relevantes para entendimento da série temporal através da análise gráfica e após a aplicação dos algoritmos de regressão.

A primeira análise da série temporal visa comprovar que se trata de uma série estacionária, de forma geral ao pensar sobre o sistema de distribuição de energia elétrica ele é um sistema controlado, contando com equipamentos que mantém a tensão e a corrente controladas como RTs e religadores. Para confirmar essa hipótese pode-se utilizar o teste de Dickey e Fuller (1979), valores negativos para o teste indicam que a série é estacionária com certo grau de certeza, aplicando o teste na base de dados encontrou-se um valor abaixo do componente crítico 0,01, indicando com 99% de certeza que a série é estacionária, como apresenta a Figura 24.

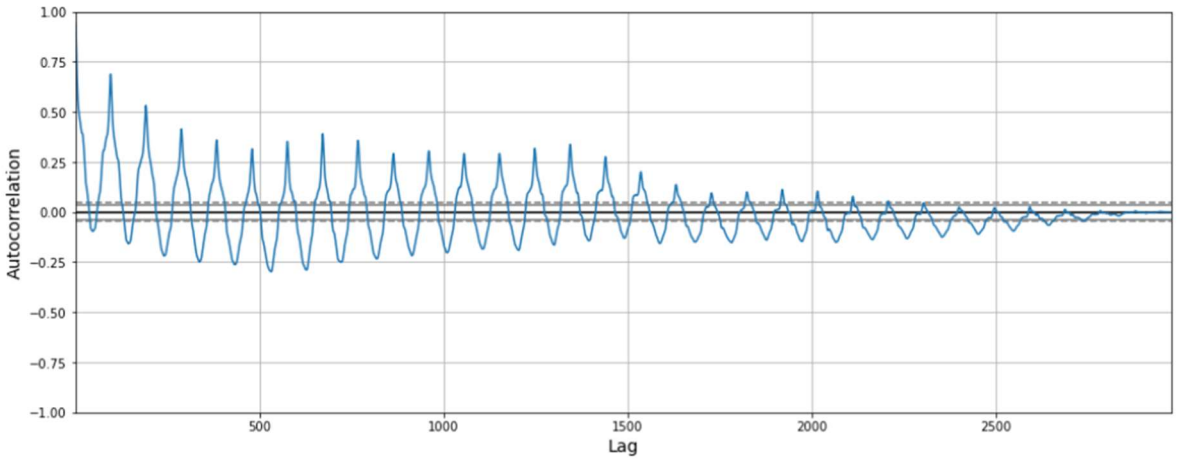
Figura 24 – Teste ADF

```
Teste adf: -8.739279997534338
p-valor: 3.039823041470027e-14
Critical values:
        1%: -3.431
        5%: -2.862
        10%: -2.567
```

Fonte: Elaborado pelo autor.

A biblioteca *statsmodel* (Seabold e Perktold, 2010) apresenta além do método ARIMA, algumas funções importantes que podem ser utilizadas como base para utilização de modelos mais recentes como as redes neurais e o método do gradiente. Utilizando essa biblioteca é possível verificar se existe autocorrelação na série temporal. A existência de autocorrelação implica que um determinado dado da série está correlacionado com dados anteriores, o grau da autocorrelação irá determinar quantos dados anteriores estão correlacionados. Plotando o gráfico de autocorrelação é possível verificar que ela está presente na série temporal como mostra a Figura 25. O eixo “y” apresenta o grau da autocorrelação e o eixo “x” o número de lags, ou seja, o número de dados anteriores que possuem autocorrelação. As linhas tracejadas apresentam a faixa de significância, valores que estão dentro dessa faixa geram uma série aleatória, portanto sem autocorrelação.

Figura 25 – Gráfico de autocorrelação

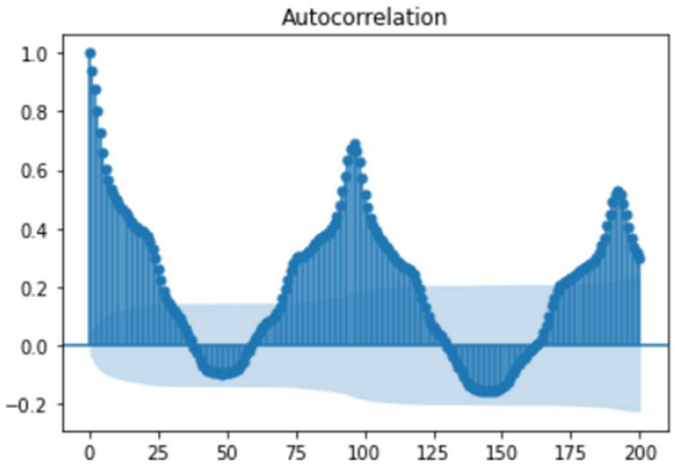


Fonte: Elaborado pelo autor.

É possível verificar no gráfico que existe uma autocorrelação grande com o número de *lags* próximo a 0, e posteriormente existe outro ponto de forte autocorrelação com o número de *lags* próximo de 100. Este valor deve indicar o

número 96, que representa os dados de um dia anterior. Utilizando a Função de AutoCorrelação “ACF” e definindo o número máximo de *lags* igual a 200 é possível verificar essa hipótese, como mostra a Figura 26. Desta vez a faixa de significância é representada pela área demarcada ao redor do eixo “x”.

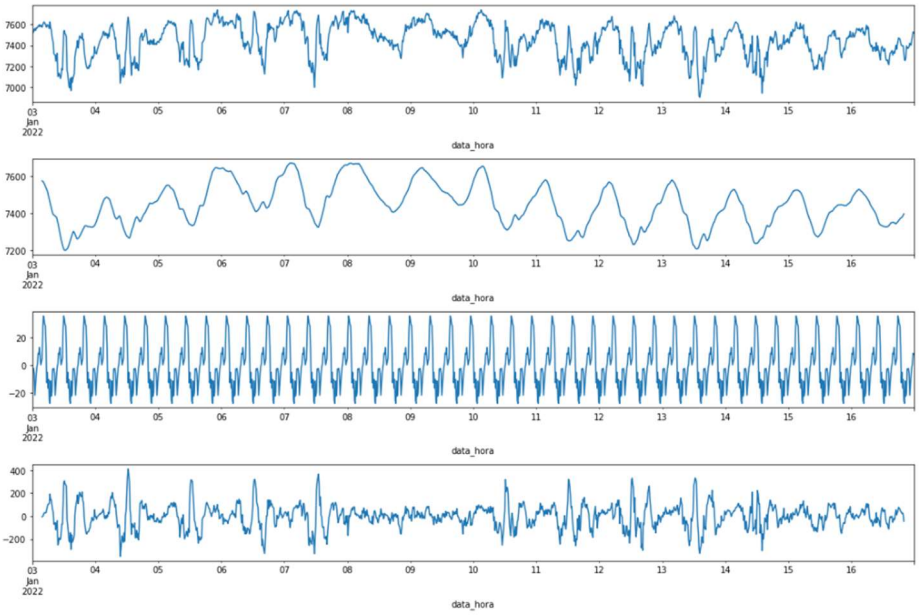
Figura 26 - ACF



Fonte: Elaborado pelo autor.

Outra análise importante é a verificação das componentes da série temporal, como mencionada no início da seção, a função *seasonal decompose* permite separar a série em componente de tendência, componente sazonal e resíduo. A Figura 27 apresenta essa análise.

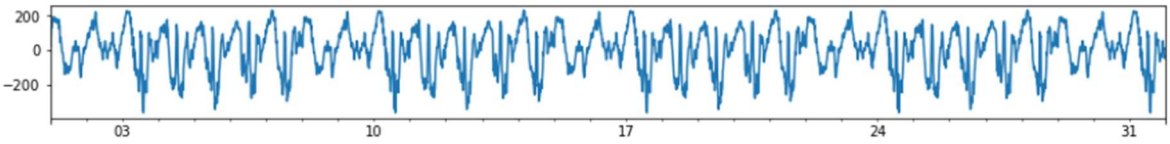
Figura 27 – Componentes da série temporal



Fonte: Elaborado pelo autor.

Como visto na Figura 27 a série possui todas as componentes mencionadas anteriormente, analisando a componente sazonal do gráfico e utilizando como período 672, relativo a uma semana de dados, ou seja, amostras a cada quinze minutos, resultam em 96 amostras por dia (4*24*7=672) é possível perceber que a componente sazonal tende a se repetir no ciclo semanal, como mostra a Figura 28.

Figura 28 - Sazonalidade



Fonte: Elaborado pelo autor.

Anteriormente a Figura 13 apresentou o comportamento da curva típica de carga do RT estudado, foi possível perceber que o consumo de energia elétrica era maior nos dias úteis e reduzido aos finais de semana e feriados, essa análise corrobora a indicação de uma sazonalidade com ciclos semanais como apresentado. Com essas análises conclui-se a análise preliminar da série temporal onde verificou-se que a autocorrelação está presente nos dados é pode ser utilizada para criar uma janela de dados buscando melhorar a assertividade dos modelos de aprendizado de máquina. Na próxima seção são testados diferentes algoritmos de regressão, avaliados com dados de treino e teste e ao final do capítulo os dados são usados para gerar previsões futuras para definir qual mais adequado para uso com essa base de dados.

Para avaliar a precisão dos modelos foi utilizada a raiz do erro quadrático médio (RMSE), obtido através da equação:

$$R\,MSE = \sqrt{\sum_{i=1}^{n}\left(Yi - \widehat{Yi}\right)^2 * \frac{1}{n}} \quad (1)$$

Também foram apurados o erro médio absoluto (MAE) que indica a diferença média do erro entre o dado real e previsto, sendo apresentado na mesma unidade do dado de entrada, e o erro percentual médio absoluto (MAPE) que apresenta a média do erro de forma percentual. Esses indicadores são dados pelas seguintes equações:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Yi - \widehat{Yi}\right| \quad (2)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|Yi - \widehat{Yi}|}{Yi} * 100 \quad (3)$$

4.3.2 Redes neurais artificiais

As redes neurais artificiais são métodos de previsão baseados em modelos matemáticos simples do cérebro. Eles permitem relacionamentos não lineares complexos entre a variável de resposta e seus preditores (Hyndman e Athanasopoulos, 2021). Existem diversos tipos de redes neurais, como as redes diretas, redes recorrentes e redes competitivas. Uma rede neural tradicional considera que as entradas e saídas são independentes, para a previsão de séries temporais esse método pode não ser o mais indicado, Masum e Chiverton (2018). Medsker e Jain (2001) ressaltam que modelos de redes neurais recorrentes são os mais indicados para realizar previsões em séries temporais, utilizando no estudo um modelo do tipo *Long Short Term Memory* (LSTM), trata-se de uma rede neural recorrente capaz de reter informações em intervalos arbitrários, sendo então adequada para análises que envolvem séries temporais.

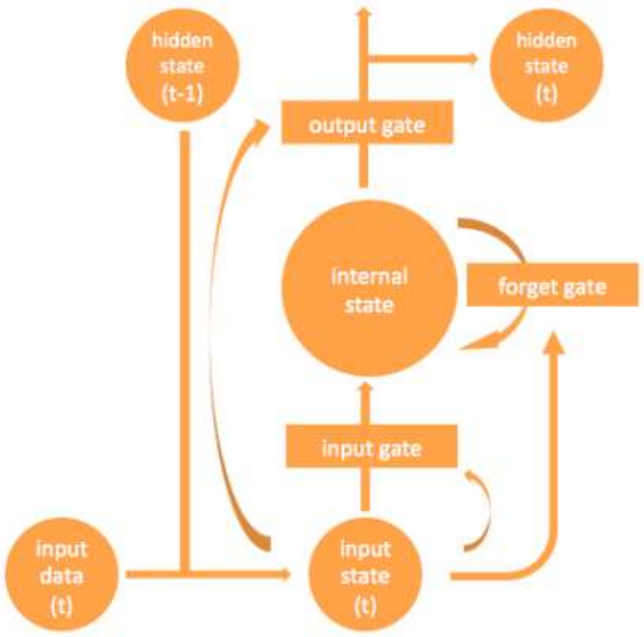
Os modelos LSTM vêm ganhando espaço para previsões de séries temporais e apresentando bons resultados como nos trabalhos de Tokgoz e Unal (2018), Choi e Lee (2018) e Abbasimehr e Paki (2021). As redes LSTM possuem blocos de memória conectados em suas camadas o que a torna mais eficiente para análise de sequências.

Segundo Wang e Raj (2017) a rede LSTM possui os seguintes componentes críticos:

- Estado oculto: determina o que esquecer, a entrada e a saída do passo seguinte;
- Estado de entrada: combinação entre entrada atual e estado oculto;
- Estado interno: valores utilizados como memória;
- Porta de entrada: decide se o estado de entrada passa para o estado interno;
- Porta de esquecimento: determina se o estado interno deve esquecer o estado anterior;
- Porta de saída: decide se o estado interno passa seu valor para a saída e ao estado oculto do próximo passo.

A Figura 29 apresenta o esquema de funcionamento das redes LSTM, interligando as portas e estados mencionados.

Figura 29 – Unidade LSTM



Fonte: Wang e Raj (2017).

Os dados em um modelo LSTM precisam ser transformados para o formato de vetor, como descrito em Koparanov, Georgiev e Shterev (2020). Com isso esse trabalho irá abordar diferentes modelos LSTM, inicialmente foi testado um modelo básico com uma camada de entrada LSTM com quatro neurônios. Outro método testado, é chamado de *Time Step*, e pode ser incorporado ao método anterior, nesse método os dados são transformados adicionando os passos de tempo como atributo de entrada da rede neural.

Outra opção testada é o método *Stateful*, ele adiciona mais um recurso as redes LSTM, nessa configuração o estado obtido ao analisar um lote de entrada será armazenado e atualizado no lote posterior, ou seja, o estado da saída de um lote é reutilizado como estado inicial para as amostras do próximo lote. Outra vantagem da utilização das redes LSTM é a possibilidade de empilhamento das redes, gerando as chamadas redes neurais profundas, nessa configuração, nomeada de *Stacked*, uma camada LSTM secundária é adicionada ficando duas camadas empilhadas. Como a intenção do trabalho é utilizar o modelo com melhor desempenho, estes modelos serão testados sob o ponto de vista dos mesmos parâmetros, utilizando 70% dos dados para treino e 30% para teste. Após é utilizado a análise de correlação feita anteriormente para criar uma janela de dados para gerar a saída, ou seja, para gerar

uma saída podem ser observados *n* valores anteriores. A Tabela 9 apresenta os resultados destes ensaios.

Tabela 9 – Análise com diferentes janelas de dados

Método	RMSE (v)		MAE (v)		MAPE (%)		Tempo de treinamento (s)
	Treino	Teste	Treino	Teste	Treino	Teste	
Básico	61,26	58,79	42,70	41,23	0,577	0,549	32,25
Time Step	61,13	59,22	41,07	39,28	0,556	0,524	36,34
Stateful	65,83	62,30	45,82	42,92	0,622	0,573	28,19
Básico n=2	60,62	57,06	40,97	38,24	0,555	0,511	32,63
Básico n=4	60,05	59,22	41,38	42,71	0,560	0,568	34,63
Time Step n=2	60,26	57,37	41,02	38,64	0,556	0,516	38,89
Stateful n=2	63,99	59,99	44,27	41,27	0,601	0,551	35,19

Fonte: Elaborado pelo autor.

Como apresentado na Tabela 9, o melhor resultado foi obtido com o método básico utilizando dois dados anteriores para predizer o próximo valor. Adicionando uma segunda camada ao método com melhor desempenho obteve-se um resultado similar ao anterior, o ganho na acurácia foi pequeno se comparado ao aumento do custo computacional, apresentado da coluna tempo de treinamento, como mostra a Tabela 10.

Tabela 10 – Ensaio com duas camadas empilhadas

Método	RMSE		MAE		MAPE		Tempo de treinamento (s)
	Treino	Teste	Treino	Teste	Treino	Teste	
Deep LSTM n = 2	60,62	56,79	40,90	38,36	0,554	0,512	56,35

Fonte: Elaborado pelo autor.

Com esse experimento conclui-se que o método utilizando a entrada de dados com dois valores anteriores apresenta o melhor desempenho quanto a acurácia e ao tempo de retreinamento.

Os modelos de regressão aplicados buscam generalizar a série de forma que ela possa ser reproduzida ou estimada no futuro, deste modo ao analisar as

componentes da série temporal citada anteriormente a componente ciclo de tendência pode ser melhor aproximada pelos algoritmos de regressão, segundo Hyndman e Athanasopoulos (2021) observações que estão próximas no tempo tendem a ter um valor próximo, portanto parte da aleatoriedade dos dados pode ser removida com a utilização da Média Móvel (MM), deixando a componente de ciclo de tendência mais suave. Alguns modelos aplicam a MM obtendo bons resultados nas predições, como os trabalhos de ArunKumar et al. (2021) e Salman e Kanigoro (2021), os modelos híbridos apresentados mostram ganhos de acurácia quando combinados com o uso de MM. Utilizando a MM sendo calculada com um dado anterior obteve-se uma melhora na acurácia do método como apresenta a Tabela 11.

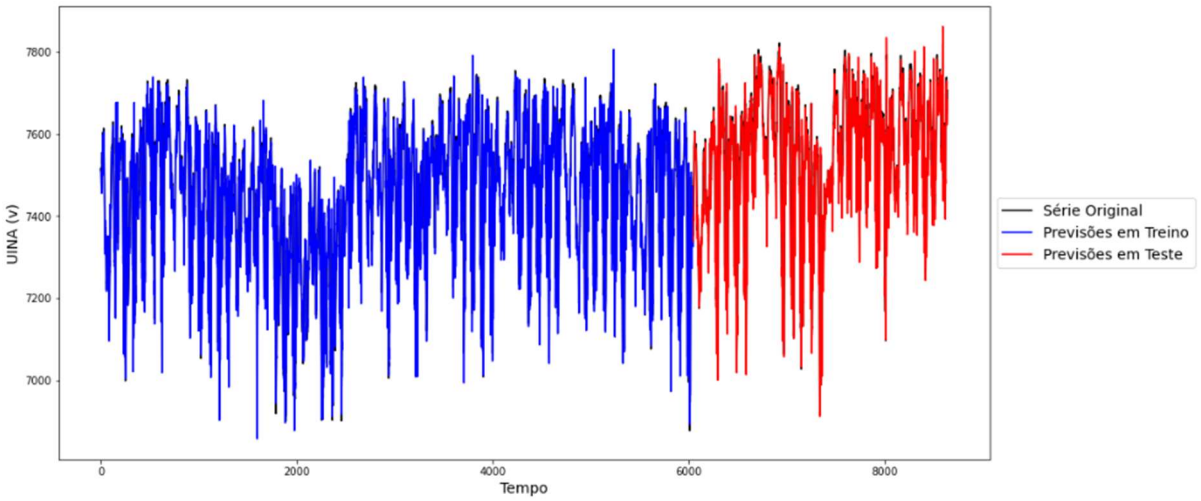
Tabela 11 – Ensaio com duas camadas empilhadas

Método	RMSE		MAE		MAPE		Tempo de treinamento (s)
	Treino	Teste	Treino	Teste	Treino	Teste	
Básico n = 2 MM(2)	35,92	34,63	25,94	25,94	0,350	0,345	33,59

Fonte: Elaborado pelo autor.

A Figura 30 apresenta o gráfico obtido no último ensaio com o uso da MM.

Figura 30 – Previsões em treino e teste método LSTM



Fonte: Elaborado pelo autor.

Com isso fica finalizado o teste utilizando redes neurais artificiais LSTM, apresentando um bom resultado em treino e teste. Como o objetivo do modelo é efetuar predições para dados futuros apenas a validação em treino e teste não é suficiente, embora já indiquem que o caminho seguido é assertivo, com isso a próxima seção apresenta a determinação da janela de dados necessária para treinar o modelo e aplicá-lo em predições futuras.

4.5 Parcimônia e definição da janela de dados

A parcimônia ou economia de parâmetros é um critério desejável na construção de modelos estatísticos (Box et al. 2016). Cada parâmetro desconhecido que é estimado está sujeito a variações, por isso escolher representação de modelos com o menor número de parâmetros possível é adequado (Ledolter e Abraham, 1981). Com isso escolher uma janela de tempo adequada para treinamento do PCR é adequado, uma vez que utilizar mais dados que o necessário pode resultar na inserção de variações desnecessárias ao modelo.

Em muitos campos há fortes evidências que a memória longa tem implicações na habilidade de previsão dos modelos, isso implica que há dependência não desprezível entre o presente e os pontos no passado (Graves, Watkins e Franze, 2017). Yaziz, Zakaria e Ahmad (2017) apresentam um estudo sobre o efeito da utilização de uma base histórica de dados relativo ao preço do ouro de 1971 a 2013, os autores concluem que a utilização de apenas 12,25% dos dados, relativo a 5 anos, é suficiente para garantir a assertividade do modelo de predição, este estudo ressalta a importância de utilizar a janela de dados correta, reduzir o tamanho da base de dados pode trazer ganhos significativos no custo computacional, dando ganho para modelos que busquem a aplicação em tempo real.

A Tabela 12 apresenta um ensaio com a variação da janela de dados usada para treinamento do modelo PCR, foi mantido fixo o mesmo ponto para ser o final da série de dados, e os dados usados para treino foram aumentando semanalmente até utilizar todos os dados disponíveis, que são de janeiro de 2021 a março de 2022.

Tabela 12 – Definição da janela de dados usada para treinamento

Semana de dados	RMSE (v)		MAE (v)		MAPE (%)		Tempo de treinamento (s)
	Treino	Teste	Treino	Teste	Treino	Teste	
1	29,84	42,00	20,19	31,94	0,264	0,425	3
2	33,58	35,60	22,95	25,58	0,303	0,339	4
3	33,88	31,38	23,29	21,88	0,308	0,289	6
4	33,73	30,49	24,14	22,26	0,319	0,293	7
5	34,55	31,32	23,89	21,75	0,318	0,286	8
6	32,96	35,09	22,78	25,81	0,303	0,338	10
7	33,88	34,46	23,72	25,14	0,317	0,329	11
8	34,22	31,60	23,81	22,19	0,318	0,292	12
9	34,13	34,43	23,58	24,49	0,316	0,322	14
10	33,96	34,99	23,53	24,52	0,315	0,324	15
20	34,33	34,32	23,94	24,76	0,322	0,327	29
30	34,42	33,28	23,90	23,19	0,320	0,308	43
40	33,88	33,78	23,66	23,58	0,315	0,314	58
50	34,24	34,95	23,84	24,28	0,316	0,324	71
Todas	34,25	34,48	24,13	24,16	0,320	0,324	97

Fonte: Elaborado pelo autor.

Como pode ser observado o melhor resultado segundo as métricas utilizadas ocorreu com a utilização de 5 semanas para retreino do modelo PCR. O ensaio utilizando 30 semanas também apresentou um resultado positivo, como é de conhecimento existe uma sazonalidade semanal na série temporal, porém não se descarta a sazonalidade com intervalo maior de tempo, uma vez que o consumo de energia elétrica possui peculiaridades quanto a climas frio e quentes, por isso esses dois valores de janela de dados para treino serão considerados na etapa de validação.

4.6 Comentários sobre o capítulo

O decorrer deste capítulo apresentou passo a passo a construção do modelo PCR, discutindo algoritmos de aprendizado de máquina que por hipótese tenderiam a atender as necessidades da questão de pesquisa apresentada na introdução. No

decorrer do estudo foi evidenciado que alguns algoritmos atendem melhor que outros como o caso do método KME e do FCM, onde experimentalmente foi escolhido o que obteve melhor resultado frente ao objetivo do estudo.

O método PCA embora não tenho sido inserido na composição final do PCR, mostrou uma boa resposta em apresentar os itens principais das curvas típicas do alimentador estudado, como apresentado na Figura 17. Esse método mostra-se eficiente para a generalização dos dados e pode ser utilizado para, por exemplo, reduzir a análise das curvas típicas de carga ao reduzir a quantidade de dados a serem analisados, podendo ser utilizado para agrupar as curvas de dias úteis e não úteis.

No decorrer dos ensaios para realizar as predições verificou-se que alguns métodos como a adição de mais camadas ocultas no método LSTM pode apenas trazer um maior custo computacional sem necessariamente trazer ganhos de acurácia ao modelo final, mostrando a importância de ajustar todos os hiper parâmetros do modelo antes de colocar em uso.

Já no final do capítulo foi discutida a precisão do modelo com diferentes dados usados para treinamento, nessa análise percebe-se a redução do custo computacional ao reduzir o tamanho da base de dados. Com isso, mesmo sendo de conhecimento que as redes neurais se adaptam bem a base de dados mais longas, com conhecimento sobre o comportamento da curva do equipamento estudado pode-se vislumbrar a utilização de base de dados reduzidas e manter uma boa acurácia no modelo.

5 ESTUDO DE CASO

Neste capítulo são apresentadas as validações do PCR, mostrando a sua aplicação completa em uma base de dados nova fornecida pela concessionária de energia elétrica.

5.1 Aplicando o PCR em uma nova base de dados

Os dados apresentados até esta etapa da dissertação são de janeiro de 2021 a março de 2022. Para validação do PCR, ele foi usado na predição de dados futuros, ou seja, o modelo PCR é utilizado em dados de abril a junho de 2022.

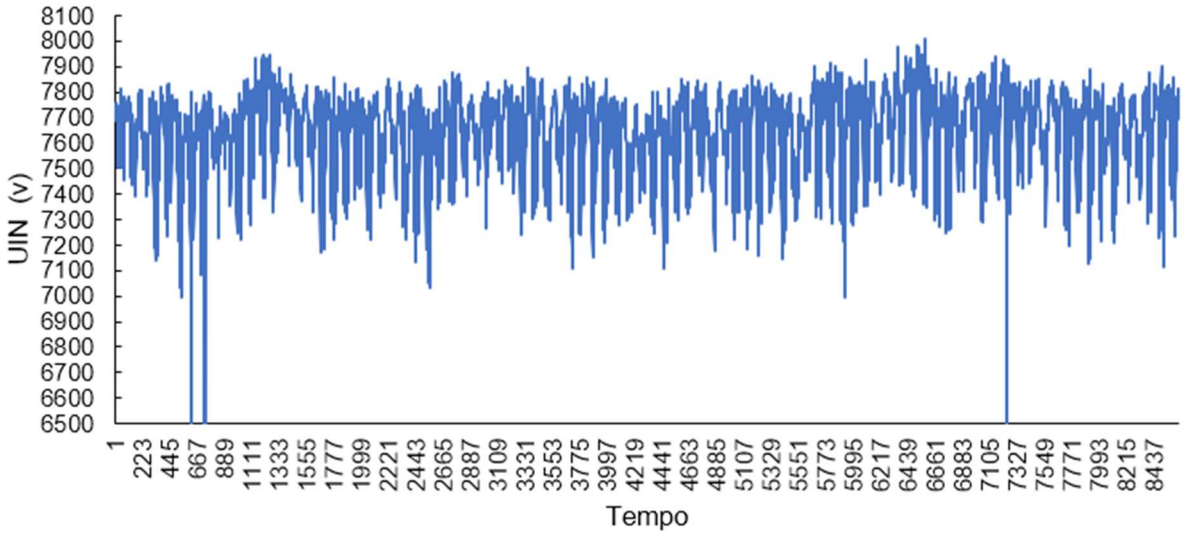
Estes dados não estiveram envolvidos em nenhum momento durante a construção do modelo. Inicialmente será apresentada as fases de pré-processamento e classificação dos dados novos, e posteriormente o modelo treinado com a base de dados inicial é empregado para gerar predições de etapas futuras, ou seja, são usados os dados anteriores a março de 2022 para treinar o modelo e gerar as predições para abril. Os resultados são então comparados com os dados reais adquiridos.

5.1.1 Validando o pré-processamento

A nova base de dados apresenta três meses de dados, sendo eles abril, maio e junho de 2022. Para validar as etapas de pré-processamento e classificação PCR, foram executadas as funções criadas para este objetivo, sem acionar a fase de regressão, ou seja, o algoritmo foi executado até a fase de tratamento dos *outliers*.

A Figura 31 apresenta os dados originais, da forma como foram coletados, sem qualquer intervenção. Como pode ser observado mesmo os dados sendo atuais, coletados nos meses em que essa dissertação estava sendo produzida, os erros descritos na análise inicial continuam sendo percebidos, esse fato valida a utilização de um *framework* que atue em tempo real, tratando o novo dado coletado antes de aplicar na fase de predição.

Figura 31 – Dados originais, nova base



Fonte: Elaborado pelo autor.

Como pode ser observado na Figura 31 existem pontos em que as medições estão zeradas e possíveis pontos de *outliers*. Com a aplicação do PCR inicialmente ele padroniza os dados e remove os zeros, como já discutido no capítulo 4 e após classifica e indica os *outliers*. Com isso, a Figura 32 apresenta a indicação dos possíveis *outliers,* a verificação é feita para três meses divididos em três *clusters*, implicando em 9 grupos de verificação.

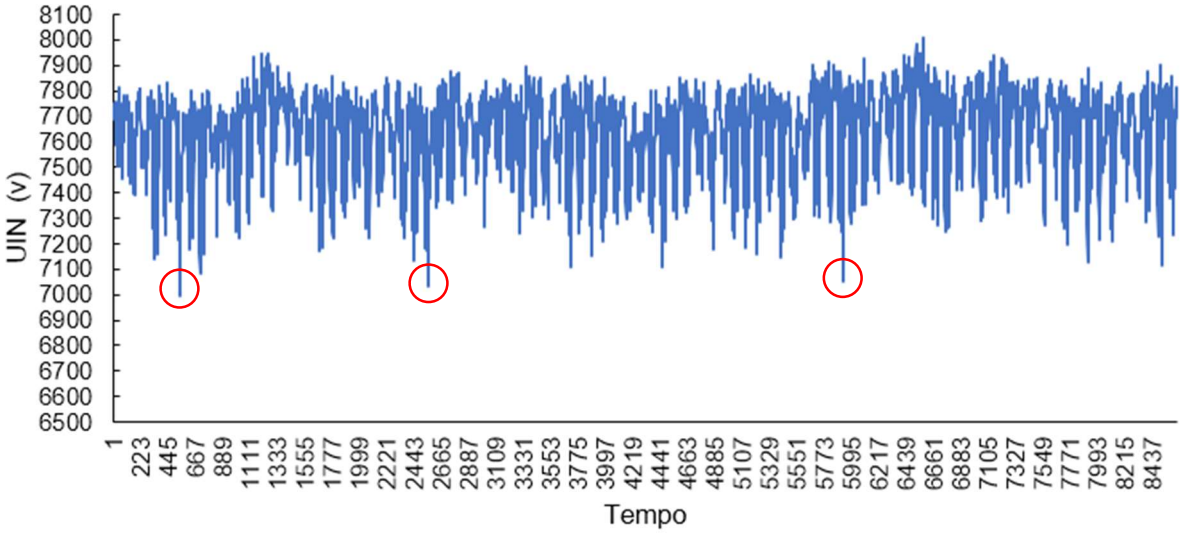
Figura 32 – Verificação de *outliers*

```
Grupos de dados sem outliers: 6
Grupos de dados com outliers: 3
Identificação do mês e cluster das ocorrências: [4, 2, 5, 0, 6, 2]
```

Fonte: Elaborado pelo autor.

Analisando localmente os candidatos a *outliers* apenas o grupo “5, 0” foi corrigido. Os outros dois grupos apresentam valores baixos de tensão, próximo a 7000 v, a opção por não alterar esses valores se deu porque nestes instantes de tempo a corrente medida estava alta e o TAP do RT estava no 16, ou seja, o sistema estava sofrendo estresse alto ao operar sob carga alta, visualizando essa informação optou-se por deixar esse dado no relatório para que outras análises possam considerá-lo. Sob o ponto de vista da regressão, esses valores serão suavizados pela aplicação da MM, diminuindo sua interferência em um novo de treinamento do PCR. A Figura 33 apresenta os dados após as fases iniciais do PCR, mostrando os outliers mantidos.

Figura 33 – Etapa de validação “PC”



Com isso as primeiras etapas do PCR ficam validadas com aplicação em dados reais. No próximo capítulo é apresentada a validação da etapa de regressão com a predição de eventos futuros.

5.1.2 Predições para janelas de tempo futuras

Neste ensaio foram utilizando os dados de fevereiro e março de 2022, equivalentes a 3360 amostras, ou seja, cinco semanas (5*7*96) para treinamento, as duas últimas amostras de março são utilizadas para prever a primeira de abril de 2022. Após a cada novo dado recebido de abril o PCR irá predizer o próximo, portanto considerando o instante atual como “t” o PCR irá predizer “t+1”. A Figura 34 apresenta os valores para RMSE, MAE e MAPE para o treinamento.

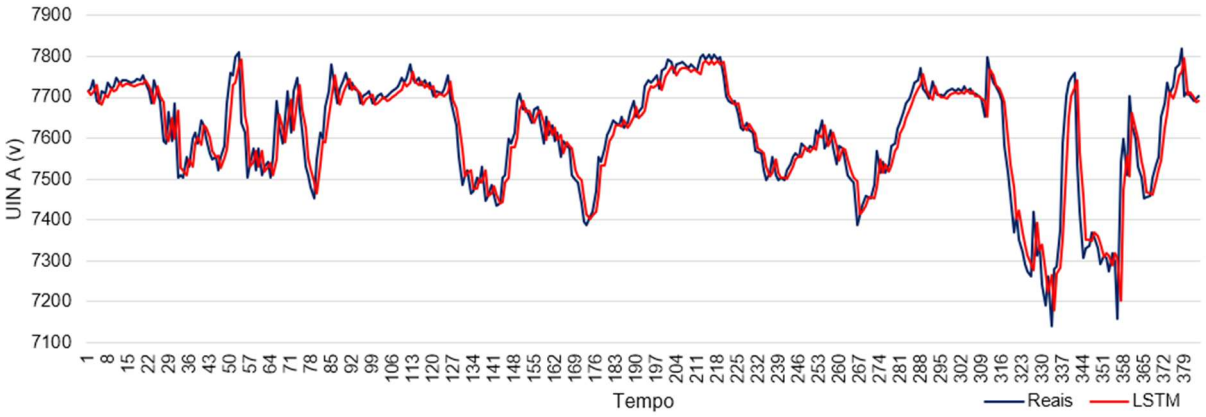
Figura 34 – Treinamento com dados novos

```
RMSE em Treino: 34.78
MAE em Treino: 24.19
MAPE em Treino: 0.326%
```

Fonte: Elaborado pelo autor.

Aplicando o PCR, a primeira predição do modelo foi 7714 v enquanto o valor real medido foi de 7716 v, o primeiro dado gerado pelo modelo PCR ficou próximo do real, seguindo o experimento para 384 amostras referentes a quatro dias, obteve-se o resultado apresentado pela Figura 35.

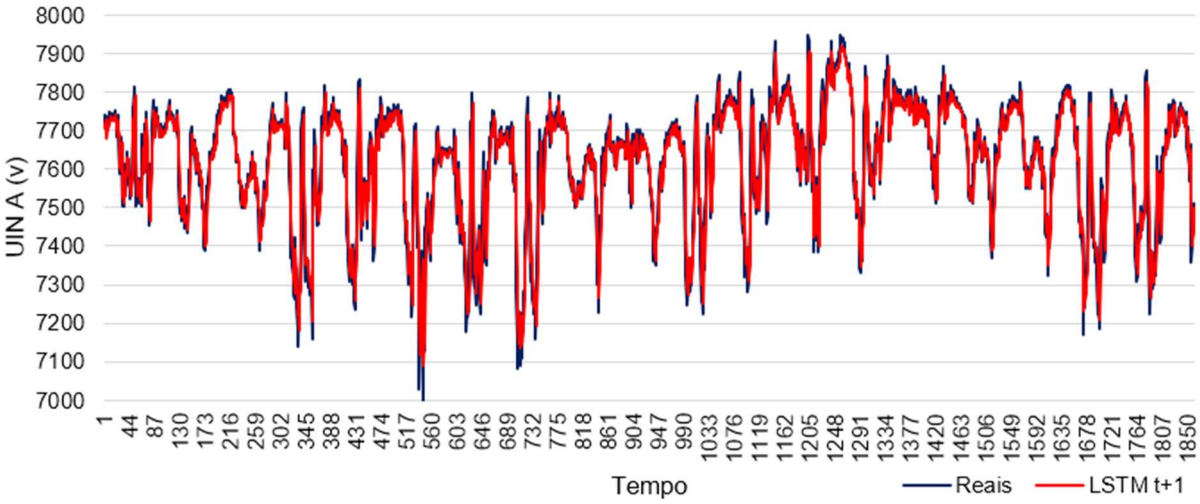
Figura 35 – Predições t+1 para 4 dias



Fonte: Elaborado pelo autor.

Como visto na Figura 35 o PCR conseguiu generalizar de maneira satisfatória a tensão mensurada no RT. Expandindo a plotagem para um gráfico de 20 dias é possível verificar a continuidade da generalização da curva como mostra a Figura 36.

Figura 36 - Predições t+1 para 20 dias



Fonte: Elaborado pelo autor.

Utilizando o PCR para alcançar janelas maiores, como predições de 2, 3 e 4 intervalos de tempo a frente o resultado vai decaindo conforme a janela de tempo vai aumentando, como esperado a capacidade de generalização do PCR diminui com a distância no tempo da série temporal aumentando. A Tabela 13 apresenta o ensaio de treino para a predição de 2, 3 e 4 amostras a frente.

Tabela 13 - Avaliação do PCR para janelas de predições futuras em treino

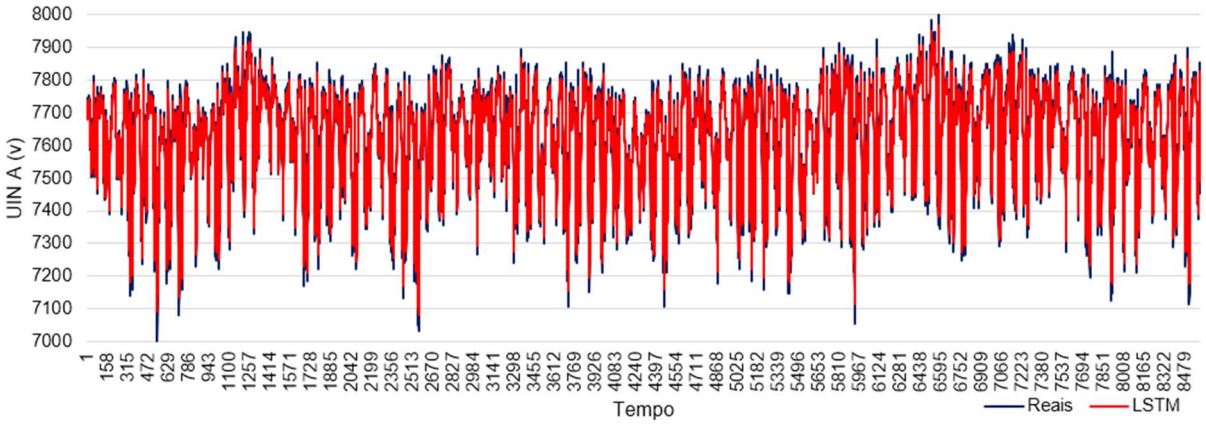
Semana de dados	RMSE (v)		MAE (v)		MAPE (%)	
	Treino	Teste	Treino	Teste	Treino	Teste
t+2	50,16	45,50	34,04	30,99	0,453	0,407
t+3	60,65	57,45	42,21	40,81	0,761	0,734
t+4	68,15	63,86	47,95	44,48	0,937	0,983

Fonte: Elaborado pelo autor.

Analisando as métricas de avaliação a janela de tempo t+2 possui acurácia relativamente boa, porém as janelas com maior alcance apresentam erros maiores. Desta forma conclui-se que a forma como o modelo foi treinada não o deixa capaz de generalizar a série a ponto de predizer dados mais distante na linha do tempo. Deste modo esse retreino será ajustado na seção 5.1.3.

Para finalizar os ensaios foi simulado um modelo adaptativo, qual seja, o PCR é aplicado para efetuar a predição de um dia completo e então é retreinado com dados de 5 semanas, assim a cada 96 dados novos adquiridos o modelo é retreinado e gera as predições com um passo de tempo a frente, a Figura 37 apresenta as predições feitas com quinze minutos de antecedência, para os dados de abril a junho de 2022, apresentando resultados promissores.

Figura 37 - Predições t+1 para abril, maio e junho de 2022



Fonte: Elaborado pelo autor.

A Tabela 14 apresenta a avaliação de desempenho para as predições da Figura 37, são apresentados o RMSE, MAE, MAPE e o maior erro absoluto encontrado. Comparado com o modelo estático anterior o modelo adaptativo mostrou um ganho alto na acurácia, sendo o mais indicado para realizar as predições.

Tabela 14 - Avaliação do PCR

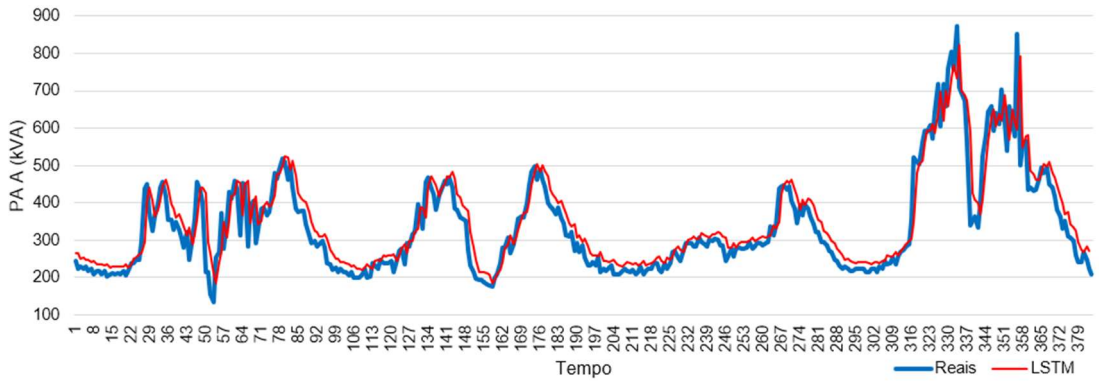
Janela de predição	RMSE (v) Predições	MAE (v) Predições	MAPE (%) Predições	Maior erro absoluto (v) Predições
t+1	15,97	12,26	0,161	100

Fonte: Elaborado pelo autor.

Como visto os resultados para a janela t+1 são expressivos, pois o maior erro é de 100 v, com as medições variando de 7000 v a 8000 v. Outro fato interessante para a janela t+1 é que 4453 registros preditos tiveram erro absoluto menor que 10 v, dentro de um total de 8600 registros gerados, ou seja, mais da metade das predições tiveram erro inferior a 10 v.

A Figura 38 apresenta o ensaio para a janela t+2, e desta vez para mostrar a versatilidade do PCR o ensaio apresenta a grandeza elétrica potência aparente, essa alteração de grandeza elétrica é feita selecionado a *flag* de interesse no modelo.

Figura 38 – Desempenho do PCR para janela t+2 e potência aparente



Fonte: Elaborado pelo autor.

Com essa base de dados o modelo PCR não apresentou erros significativos, de maneira geral o desempenho foi melhor que o esperado. Com esses experimentos fica validado a acurácia do modelo PCR utilizando os parâmetros apresentados na Tabela 9, com destaque para o parâmetro *window size* igual a dois. O desempenho foi adequado com diferentes grandezas elétricas, utilizando dados reais e futuros para essa verificação. Na próxima seção discute-se a nova janela de treinamento do PCR, uma vez que a predição para janela de t+3 e t+4 não apresentou bom resultado.

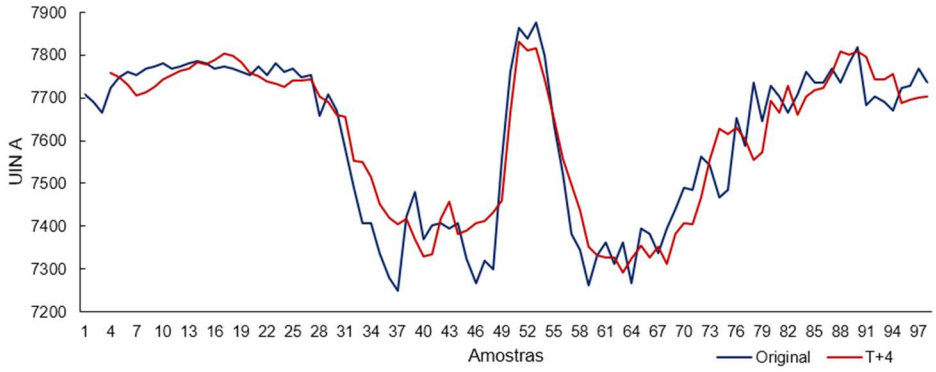
5.1.3 Ajustes, retreino e nova validação

Nas seções anteriores foram apresentadas as validações do PCR e sua acurácia ao prever dados futuros com até 30 minutos de antecedência. Esta seção busca visualizar um ajuste para que o modelo tenha maior acurácia nas janelas com 45 minutos e uma hora de antecedência.

Como visto na Figura 26, existe forte autocorrelação dos dados com *lag* de 96, esse parâmetro será definido como o valor relativo a *window size* do PCR. A Tabela 12 mostrou o desempenho do modelo com variação da quantidade de dados usada para treinamento, com isso para aumentar o alcance do PCR esses dados serão ajustados para 30 semanas. Com esses ajustes o desempenho do PCR melhorou para as janelas de dados de 45 minutos e uma hora.

A Figura 39 apresenta as predições quatro passos de tempo a frente, pode-se observar que o PCR consegue acompanhar a curva de referência, errando de forma mais acentuada nos pontos críticos onde a tensão caí ou aumenta de forma abrupta.

Figura 39 – Predições com uma hora de antecedência



Fonte: Elaborado pelo autor.

Este modelo foi definido com quatro neurônios de saída, sendo cada um responsável pela saída de um passo de tempo, ou seja, t+1, t+2, t+3 e t+4. A Tabela 15 apresenta os indicadores de desempenho para essa variação do PCR.

Tabela 15 – Avaliação do PCR ajustado

Janela de predição	RMSE (v) Predições	MAE (v) Predições	MAPE (%) Predições
t+1	51,70	40,28	0,533
t+2	56,56	42,85	0,568
t+3	60,72	46,75	0,620
t+4	66,70	52,00	0,690

Fonte: Elaborado pelo autor.

Como pode ser observado o resultado melhorou para as janelas de 45 minutos e uma hora a frente, porém foi pior para as janelas de 15 e 30 minutos. Com isso sugere-se a criação de dois modelos distintos para realizar a predição de forma mais assertiva para cada intervalo de interesse de acordo com estratégia da concessionária de energia elétrica.

5.2 Discussão sobre aplicações do PCR

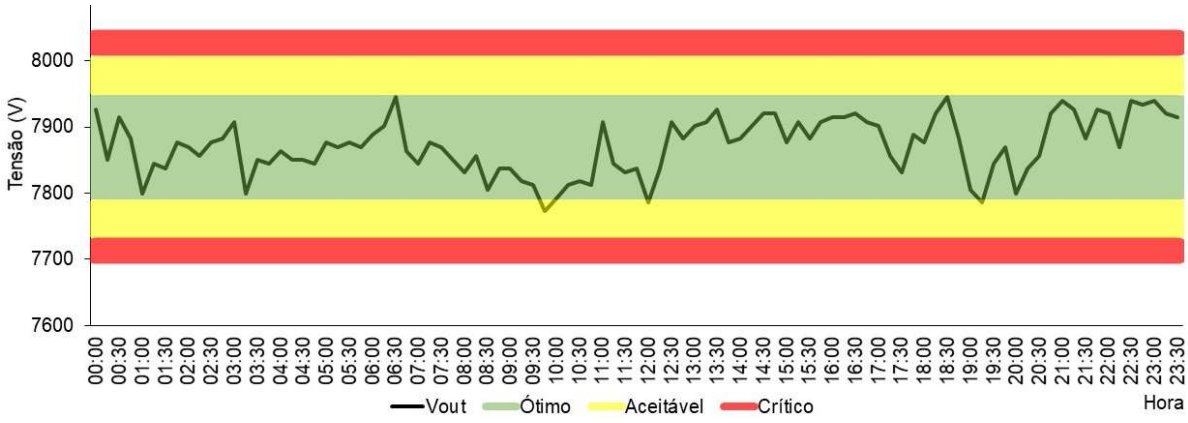
Os dados entregues pelo PCR podem usados para indicar ajustes nos RTs da rede de distribuição. Na seção 5.1.1, foram apresentados valores considerados como *outliers* pelo método Grubbs, porém, embora fossem dados fora da curva, não se tratava de outliers e sim de momentos de estresse máximo do RT, onde a tensão caiu por não existir mais uma faixa de ajuste para aumentar o nível de tensão.

As predições geradas com o PCR com até uma hora de antecedência podem ser utilizadas para planejar manobras para reduzir o carregamento do alimentador evitando que o RT fique sobrecarregado e sem possibilidade de ajuste de tensão. Esta é uma análise crítica para a concessionária, uma vez que níveis baixos de tensão podem gerar compensações financeiras a serem pagas aos consumidores.

A predição da tensão de entrada do RT e da potência podem ajudar a sugerir o nível de TAP adequado para a operação equipamento. Pode ser considerado a faixa de insensibilidade configurada no equipamento e um valor pré-determinado como aceitável fora da faixa de insensibilidade. Na prática isso implica determinar uma zona ótima do nível de tensão de saída do RT (insensibilidade), uma zona aceitável e uma zona crítica. Dessa forma, a faixa de TAPs pode ser definida de forma a seguir os níveis de tensão esperados na saída do equipamento. A Figura 40 apresenta essa alternativa, apresentado a tensão de saída do RT e as zonas descritas. A tensão de

ajuste do equipamento é 7870 V, a zona ótima utilizada nesse exemplo é de 1%, a zona aceitável varia de 1% a 1,75% enquanto a zona crítica fica no intervalo de 1,75% a 2,25%. Essa configuração pode ser definida contando com a opinião de especialistas da concessionária, considerando os fatores locais, níveis de transgressões e penalidades que possam ocorrer.

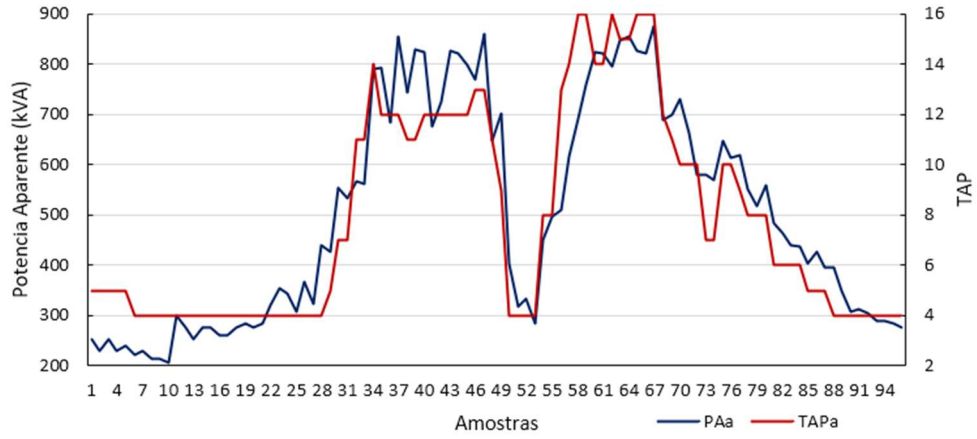
Figura 40 – Zonas definidas para otimização dos TAPs



Fonte: Elaborado pelo autor.

Com a aplicação das predições pode-se variar a faixa de insensibilidade e visualizar a variação dos níveis da tensão de saída do equipamento. Como analisado no decorrer desta dissertação o comportamento das curvas típicas deste equipamento tem diferenças entre dias úteis e não úteis, bem como existem grandes variações no consumo de energia elétrica ao longo do dia, oscilando entre momentos de pico e de estabilidade transitória. A Figura 41 apresenta a potência aparente e o TAP do RT ao longo de um dia.

Figura 41 – Potência aparente e TAP do RT



Fonte: Elaborado pelo autor.

Como pode ser visto na Figura 41 inicialmente a potência está em um valor baixo assim como no final do gráfico, mesmo assim existe uma comutação no TAP do RT. Após uma rampa de subida o consumo entra em uma faixa de estabilidade transitória entre as amostras 34 e 48, nesse instante existem algumas comutações no TAP, assim como após a metade do gráfico (meio-dia). Como pode ser observado existem várias comutações de TAP que talvez possam ser eliminadas com a visualização das predições, que indicam as alterações no nível de carregamento que tendem a acontecer nas janelas de tempo posteriores.

Também pode ser observado que existem pontos críticos de operação do equipamento, onde o TAP se encontra no valor máximo (16). Essa visualização de momentos críticos pode ser indicada com uma flag pelo modelo PCR bem como ter sua frequência mensurada ao longo do dia, com análises mais profundas pode-se sugerir ajustes da tensão de referência e faixa de insensibilidade utilizando ainda a compensação de tensão resistiva $\mathrm{U_R}$ e a compensação de tensão reativa $\mathrm{U_X}$, definindo os parâmetros através da análise da tensão desejável em carga máxima e mínima (PEREIRA et al. 2011). Essas análises serão observadas novamente na seção sobre os próximos passos deste estudo.

5.3 Comentários sobre o capítulo

Neste capítulo foi apresentada a validação do modelo PCR, inicialmente a fase de validação mostrou o efeito do pré-processamento de dados, como pode ser visto na Figura 31, mesmo sendo dados recentes, datados de abril a junho de 2022, existem valores iguais a zero, essa constatação mostra além da validação do modelo que corrigiu adequadamente os erros na base de dados, ele deve ser executado em tempo real para a geração das predições, uma vez que a cada amostra recebida pelo modelo pode trazer em si um erro que pode prejudicar as predições.

Após a correção da base a análise dos *outliers* trouxe considerações importantes como: O que seria realmente um *outlier*? Até onde a ferramenta pode remover de forma automatizada um possível *outlier*?

Um dos objetivos da ferramenta é alcançar a padronização dos dados utilizados para outras análises pela concessionária de distribuição de energia elétrica. Manter um padrão pode trazer grandes ganhos computacionais, pois não se perderia tempo ajustando as bases de dados uma vez que elas já se encontram devidamente

ajustadas, economizando o retrabalho a cada análise nova. Porém como apresentado na Figura 33, os pontos identificados como *outlier* pelo método de Grubbs, tratavam-se na verdade de pontos em que o equipamento que faria os ajustes na rede para que aquele ponto de tensão baixa não ocorresse, já estava no seu limite e não havia possibilidade de ajuste.

Essa análise pontual mostra que a detecção e correção de *outliers* não é uma tarefa trivial quando se trata de sistemas de distribuição de energia elétrica, uma vez que remover essas amostras prejudicaria outras análises que buscassem visualizar a qualidade do sistema e impediria que o modelo de regressão reproduzisse através das predições esse possível sinistro, que aconteceu e pode voltar a ocorrer.

No decorrer da etapa de validação verificou-se que o modelo que apresentou a melhor acurácia na fase de construção do PCR não foi capaz de generalizar a série para quatro passos de tempo a frente, sendo necessário novo ajuste no modelo para então apresentar um resultado que pode ser utilizado no dia a dia da operação. Essas considerações mostram que a série temporal produzida com as amostras do sistema de distribuição de energia possui suas particularidades e deve ser analisada com cuidado, generalizações que atendem um alimentador podem não atender qualquer outro alimentador da rede de distribuição, uma vez que a sazonalidade percebida em diferentes regiões será possivelmente diferente.

6 CONSIDERAÇÕES FINAIS

Esta dissertação levantou a hipótese que era possível criar um modelo capaz de identificar padrões locais na base de dados de grandezas elétricas, efetuar o pré-processamento com base nesse contexto local e efetuar a predição da grandeza elétrica de interesse da concessionária de energia, através da metodologia adotada foi criado o modelo denominado PCR.

Como visto nos trabalhos relatados grande parte dos estudos analisados usaram algum tipo de aprendizado de máquina para realizar as predições, demonstrando que esse era um tema emergente no setor de distribuição de energia elétrica, ainda sobre os trabalhos relatados apenas dois trabalhos utilizavam o contexto climático como entrada para gerar as predições, o que é um indicativo relevante quantas as afirmações sobre a parcimônia relatada pelos autores Box et al. 2016 e Ledolter e Abraham, 1981.

Buscando reduzir a quantidade de dados usados esta dissertação apresentou um ensaio onde obteve-se boa resposta do modelo utilizando uma base de dados menor para o treinamento, pode-se concluir nessas análises que as predições de horizontes menores podem utilizar menos dados para treinamento enquanto as mais longas precisam de base de dados maiores. A definição da janela de dados utilizada para treinamento passa também pela análise da sazonalidade contida na série de dados, por esse motivo é importante conhecer as particularidades de cada alimentador ou de cada setor da rede de distribuição de energia elétrica antes de fazer a determinação dos parâmetros de treinamento do modelo.

Com todas essas considerações o modelo PCR se mostrou capaz de efetuar o pré-processamento dos dados, eliminando os problemas como dados ausentes, zerados e repetidos, armazenando os dados em um novo diretório para ser então utilizado pelas outras fases do modelo ou por análises externas a execução do PCR.

A etapa de classificação dos dados auxilia na detecção de *outliers*, como demonstrado, a utilização dos *clusters* encontrou o*utliers* que não eram percebidos ao analisar a base de dados completa. Essa etapa também contribuiu para que pontos de estresse máximo do sistema fossem reconhecidos, como os pontos em que o TAP do RT estava no limite e a tensão continuava a decair com o aumento da carga.

Por fim as contribuições com a etapa de regressão trouxeram resultados promissores para a utilização em tempo real, tendo em vista que o PCR tem um tempo

de retreino baixo e o modelo adaptativo apresentou boa acurácia com os dados novos utilizados na validação.

6. 1 Contribuições

Este trabalho trouxe algumas contribuições para a área, como visto os trabalhos relatados não abordam todo o problema que envolve a análise da base de dados antes de chegar à etapa de predição e geração de pseudomedições. Este trabalho apresentou todos os passos de desenvolvimento podendo ser utilizado como guia para desenvolvimento de análises similares.

Durante a pesquisa não foram encontrados trabalhos que utilizem a criação de *clusters* para efetuar a correção de *outliers,* embora essa abordagem traga a limitação de mostrar ao especialista da área onde se encontra o *outlier* para que ele seja analisado antes de ser corrigido este é um diferencial da dissertação.

As predições geradas possuem uma acurácia alta não apenas em treino e teste, mas também aplicadas a dados reais e realmente produzindo predições de até quatro passos de tempo a frente com assertividade embasadas para serem utilizadas na programação de manobras e intervenções na rede de distribuição de energia elétrica.

6.2 Trabalhos futuros

Com base no estudo desenvolvido foram encontradas algumas questões para serem discutidas em trabalhos futuros. A amostragem dos dados coletados é de quinze minutos o que não permite ver a quantidade real de comutações do RT, as análises feitas na seção 5.2 podem ser expandidas com uma frequência maior na aquisição dos dados verificando desta forma quantas comutações podem ser evitadas com o uso das predições da tensão de entrada do equipamento.

O desenvolvimento da dissertação foi realizado com a base de dados fornecida pelo RT instalado em uma região próxima a área central do alimentador, como os resultados das predições são promissores, apresentando boa acurácia para até 4 passos de tempo futuros, abre-se espaço para verificar como o uso destas predições podem auxiliar na estimação de estados de todo o alimentador. Com base nessa ideia podem-se instalar outros medidores ao longo do alimentador utilizando as premissas

vistas na seção 3.4.2 sobre alocação de medidores e evoluir para a estimação de estados da rede distribuição baseada nas predições geradas nas bordas.

6.3 Produção científica

Durante o desenvolvimento desta dissertação foram produzidos artigos de autoria do próprio autor e participação como coautor.

Um deles publicado no XXIII Congresso Brasileiro de Automática e selecionado para submissão de uma versão estendida, contando com a participação do professor Dr. Paulo Ricardo da Silva Pereira e do Dr. Ederson Pereira Madruga, sendo o artigo:

- Dos S. Costa, R., Da S. Pereira, P. R., & P. Madruga, E. (2020). Análise Multivariável para Priorização de Obras em Redes de Distribuição de Energia Elétrica com Foco nos Indicadores de Qualidade de Energia. Anais Do Congresso Brasileiro de Automática 2020. Congresso Brasileiro de Automática - 2020. https://doi.org/10.48011/asba.v2i1.1221

A versão estendida deste artigo foi produzida em inglês para o periódico *Journal of Control, Automation and Electrical Systems,* intitulada: *A Hybrid Model for Investment Prioritization and Performance Analysis in Electrical Power Distribution Systems.* No ato da entrega desta dissertação este artigo encontra-se aceito para publicação aguardando próxima edição do periódico.

Da mesma forma, foi aceito no CLAGTEE 2022, artigo que comtempla parte desta dissertação intitulado:

- Costa, R. S., Aranda, J. A. S., Firmino, C., Pereira, P. R. S., Barbosa, J. L. V., Silva, E. L. M., Vianna, M. P. (2022). Pré-tratamento de Dados e Previsão de Grandezas Elétricas Utilizando Modelos de Redes Neurais LSTM. CLAGTEE 2022 XIV Latin-American Congress on Electricity Generation and Transmission, Rio de Janeiro, Brazil, 2022.

Além disso houve a participação como coautor no artigo publicado no periódico *Computers and Electrical Engineering*, nomeado:

- Aranda, J. A. S., dos Santos Costa, R., de Vargas, V. W., da Silva Pereira, P. R., Barbosa, J. L. V., & Vianna, M. P. (2022). *Context-aware Edge Computing and Internet of Things in Smart Grids: A systematic mapping study. Computers and Electrical Engineering*, 99, 107826. https://doi.org/10.1016/j.compeleceng.2022.107826

REFERÊNCIAS

Adam, B., & Smith, I. F. (2008). Reinforcement Learning for Structural Control. Journal of Computing in Civil Engineering, 22(2), 133–139. https://doi.org/10.1061/(asce)0887-3801(2008)22:2(133)

Ahmad, F., Tariq, M., & Farooq, A. (2019). A novel ANN-based distribution network state estimator. International Journal of Electrical Power & Energy Systems, 107, 200-212. https://doi.org/10.1016/j.ijepes.2018.11.019

Al-Badi, A. H., Ahshan, R., Hosseinzadeh, N., Ghorbani, R., & Hossain, E. (2020). Survey of Smart Grid Concepts and Technological Demonstrations Worldwide Emphasizing on the Oman Perspective. Applied System Innovation, 3(1), 5. https://doi.org/10.3390/asi3010005

Alexandropoulos, S.-A. N., Kotsiantis, S. B., & Vrahatis, M. N. (2019). Data preprocessing in predictive data mining. The Knowledge Engineering Review, 34. https://doi.org/10.1017/s026988891800036x

Al-Sahaf, H., Bi, Y., Chen, Q., Lensen, A., Mei, Y., Sun, Y., Tran, B., Xue, B., & Zhang, M. (2019). A survey on evolutionary machine learning. Journal of the Royal Society of New Zealand, 49(2), 205–228. https://doi.org/10.1080/03036758.2019.1609052

Alzate, E. B., Bueno-Lopez, M., Xie, J., & Strunz, K. (2019). Distribution System State Estimation to Support Coordinated Voltage-Control Strategies by Using Smart Meters. IEEE Transactions on Power Systems, 34(6), 5198-5207. https://doi.org/10.1109/tpwrs.2019.2902184

Anuar, N., Baharin, N. K. K., Nizam, N. H. M., Fadzilah, A. N., Nazri, S. E. M., & Lip, N. M. (2021). Determination of Typical Electricity Load Profile by Using Double Clustering of Fuzzy C-Means and Hierarchical Method. In 2021 IEEE 12th Control and System Graduate Research Colloquium (ICSGRC) (p. 277–280). Apresentado em 2021 IEEE 12th Control and System Graduate Research Colloquium (ICSGRC), Shah Alam, Malaysia: IEEE. https://doi.org/10.1109/ICSGRC53186.2021.9515295

Aranda, J. A. S., Dias, L. P. S., Barbosa, J. L. V., de Carvalho, J. V., Tavares, J. E. da R., & Tavares, M. C. (2019). Collection and analysis of physiological data in smart environments: a systematic mapping. Journal of Ambient Intelligence and Humanized Computing, 11(7), 2883-2897. https://doi.org/10.1007/s12652-019-01409-9

ArunKumar, K. E., Kalaga, D. V., Sai Kumar, Ch. M., Chilkoor, G., Kawaji, M., & Brenza, T. M. (2021). Forecasting the dynamics of cumulative COVID-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: Auto-Regressive Integrated Moving Average (ARIMA) and Seasonal Auto-Regressive Integrated Moving Average (SARIMA). Applied Soft Computing, 103, 107161. https://doi.org/10.1016/j.asoc.2021.107161

Attar, M., Homaee, O., Repo, S., & Rekola, J. (2018). Importance Investigation of Load Models Consideration in Stand-Alone Voltage Regulators Placement in Distribution Systems. In 2018 8th International Conference on Power and Energy

Systems (ICPES) (p. 146–150). Colombo, Sri Lanka: IEEE. https://doi.org/10.1109/ICPESYS.2018.8626919

Bagheri, P., & Xu, W. (2017). Assessing Benefits of Volt-Var Control Schemes Using AMI Data Analytics. IEEE Transactions on Smart Grid, 8(3), 1295-1304. https://doi.org/10.1109/tsg.2016.2603421

Barker, P. P., & De Mello, R. W. Determining the impact of distributed generation on power systems. I. Radial distribution systems. 2000 Power Engineering Society Summer Meeting (Cat. No.00CH37134). 2000 Power Engineering Society Summer Meeting. https://doi.org/10.1109/pess.2000.868775

Bei Gou, & Abur, A. (2001). An improved measurement placement algorithm for network observability. IEEE Transactions on Power Systems, 16(4), 819-824. https://doi.org/10.1109/59.962432

Bindu, S., Ushakumari, S., & Savier, J. S. (2021). Linear Distribution System State Estimation with Integration of DG. Technology and Economics of Smart Grids and Sustainable Energy, 6(1). https://doi.org/10.1007/s40866-020-00101-8

Borges Rodrigues, R., & Sanches Mantovani, J. R. (2020). Proposta de uma Metodologia para Alocação de Dispositivos de Manobra e Proteção em Redes de Distribuição para Melhoria dos Índices de Continuidade. Anais Do Congresso Brasileiro de Automática 2020. Congresso Brasileiro de Automática - 2020. https://doi.org/10.48011/asba.v2i1.1487

Boroojeni, K. G., Amini, M. H., Bahrami, S., Iyengar, S. S., Sarwat, A. I., & Karabasoglu, O. (2017). A novel multi-time-scale modeling for electric power demand forecasting: From short-term to medium-term horizon. Electric Power Systems Research, 142, 58-73. https://doi.org/10.1016/j.epsr.2016.08.031

Box, G. E. P., Jenkins G. M., Reinsel G. C., and Ljung G. M. Time Series Analysis: Forecasting and Control, Fifth Edition. 2016 John Wiley & Sons. Inc. Published 2016 by John Wiley & Sons. Inc.

C. e Silva, F., H. de Sousa, R., G. S. Chaves, A., H. G. Barbosa, B., & D. Ferreira, D. (2020). Classificador Fuzzy-genético aplicado ao processamento de linguagem natural. Anais Do Congresso Brasileiro de Automática 2020. Congresso Brasileiro de Automática - 2020. https://doi.org/10.48011/asba.v2i1.1202

Cao, Z., Wan, C., Zhang, Z., Li, F., & Song, Y. (2020). Hybrid Ensemble Deep Learning for Deterministic and Probabilistic Low-Voltage Load Forecasting. IEEE Transactions on Power Systems, 35(3), 1881-1897. https://doi.org/10.1109/tpwrs.2019.2946701

Carvalho, O., Roloff, E., & Navaux, P. O. A. (2017). A Distributed Stream Processing based Architecture for IoT Smart Grids Monitoring. Companion Proceedings of The10th International Conference on Utility and Cloud Computing. UCC '17: 10th International Conference on Utility and Cloud Computing. https://doi.org/10.1145/3147234.3148105

Cota, M. P., Alves, C. M. O., & Castro, M. R. G. (2022). MLV-Viewer: Universal Decision Support System. In J. Mejia, M. Muñoz, Á. Rocha, H. Avila-George, & G. M. Martínez-Aguilar (Orgs.), New Perspectives in Software Engineering (Vol. 1416, p. 123–133). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-89909-7_10

Chen, H., Das, S., Morgan, J. M., & Maharatna, K. (2022). Prediction and classification of ventricular arrhythmia based on phase-space reconstruction and fuzzy c-means clustering. Computers in Biology and Medicine, 142, 105180. https://doi.org/10.1016/j.compbiomed.2021.105180

Dalmina, L., Barbosa, J. L. V., & Vianna, H. D. (2019). A systematic mapping study of gamification models oriented to motivational characteristics. Behaviour & Information Technology, 38(11), 1167-1184. https://doi.org/10.1080/0144929x.2019.1576768

Dehghanpour, K., Yuan, Y., Wang, Z., & Bu, F. (2019). A Game-Theoretic Data-Driven Approach for Pseudo-Measurement Generation in Distribution System State Estimation. IEEE Transactions on Smart Grid, 10(6), 5942-5951. https://doi.org

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the Estimators for Autoregressive Time Series With a Unit Root. Journal of the American Statistical Association, 74(366), 427. https://doi.org/10.2307/2286348

Dileep, G. (2020). A survey on smart grid technologies and applications. Renewable Energy, 146, 2589–2625. https://doi.org/10.1016/j.renene.2019.08.092

Dobbe, R., van Westering, W., Liu, S., Arnold, D., Callaway, D., & Tomlin, C. (2020). Linear Single- and Three-Phase Voltage Forecasting and Bayesian State Estimation With Limited Sensing. IEEE Transactions on Power Systems, 35(3), 1674-1683. https://doi.org/10.1109/tpwrs.2019.2955893

Dominguez, J. A., Rivera, A., Botina, K., Perdomo, G. A., Montoya, O., Campillo, J., & Delahoz, E. (2020). Data-driven framework for the detection of non-technical losses in distribution grids. 2020 IX International Congress of Mechatronics Engineering and Automation (CIIMA). 2020 IX International Congress of Mechatronics Engineering and Automation (CIIMA). https://doi.org/10.1109/ciima50553.2020.9290186

García, S., Luengo, J., & Herrera, F. (2016). Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. Knowledge-Based Systems, 98, 1–29. https://doi.org/10.1016/j.knosys.2015.12.006

Graves, T., Gramacy, R., Watkins, N., & Franzke, C. (2017). A Brief History of Long Memory: Hurst, Mandelbrot and the Road to ARFIMA, 1951–1980. Entropy, 19(9), 437. https://doi.org/10.3390/e19090437

Halstead, J. E., Smith, J. A., Carter, E. A., Lay, P. A., & Johnston, E. L. (2018). Assessment tools for microplastics and natural fibres ingested by fish in an urbanised estuary. Environmental Pollution, 234, 552–561. https://doi.org/10.1016/j.envpol.2017.11.085

Han, Y., Huang, G., Song, S., Yang, L., Wang, H., & Wang, Y. (2021). Dynamic Neural Networks: A Survey. https://doi.org/10.48550/ARXIV.2102.04906

Hassan, S., & Khan, S. (2021). Review of Advances in Smart Grids, Blackout Mitigation, and Applications in Bangladesh. In 2021 International Conference on Electromechanical and Energy Systems (SIELMEN) (p. 207–212). Iasi, Romania: IEEE. https://doi.org/10.1109/SIELMEN53755.2021.9600283

Hidayatullah, N. A., Stojcevski, B., & Kalam, A. (2011). Analysis of Distributed Generation Systems, Smart Grid Technologies and Future Motivators Influencing Change in the Electricity Sector. Smart Grid and Renewable Energy, 02(03), 216–229. https://doi.org/10.4236/sgre.2011.23025

Huang, B. B., Xie, G. H., Kong, W. Z., & Li, Q. H. (2012). Study on smart grid and key technology system to promote the development of distributed generation. IEEE PES Innovative Smart Grid Technologies. 2012 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia). https://doi.org/10.1109/isgt-asia.2012.6303265

Huang, M., Wei, Z., Zhao, J., Jabr, R. A., Pau, M., & Sun, G. (2020). Robust Ensemble Kalman Filter for Medium-Voltage Distribution System State Estimation. IEEE Transactions on Instrumentation and Measurement, 69(7), 4114-4124. https://doi.org/10.1109/tim.2019.2945743

Huang, Y., Lu, Y., Wang, F., Fan, X., Liu, J., & Leung, V. C. M. (2018). An Edge Computing Framework for Real-Time Monitoring in Smart Grid. 2018 IEEE International Conference on Industrial Internet (ICII). 2018 IEEE International Conference on Industrial Internet (ICII). https://doi.org/10.1109/icii.2018.00019

Huang, Y., Xu, Q., Hu, C., Sun, Y., & Lin, G. (2019). Probabilistic State Estimation Approach for AC/MTDC Distribution System Using Deep Belief Network with Non-Gaussian Uncertainties. IEEE Sensors Journal, 1-1. https://doi.org/10.1109/jsen.2019.2926089

Hyndman, R.J., & Athanasopoulos, G. (2021) Forecasting: principles and practice, 3rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp3. Accessed on 30/06/2022.

Ji, X., Yin, Z., Zhang, Y., Wang, M., Zhang, X., Zhang, C., & Wang, D. (2021). Real-time robust forecasting-aided state estimation of power system based on data-driven models. International Journal of Electrical Power & Energy Systems, 125, 106412. https://doi.org/10.1016/j.ijepes.2020.106412

Kabiri, M., & Amjady, N. (2019). A New Hybrid State Estimation Considering Different Accuracy Levels of PMU and SCADA Measurements. IEEE Transactions on Instrumentation and Measurement, 68(9), 3078-3089. https://doi.org/10.1109/tim.2018.2872446

Kaviani, R., & Hedman, K. W. (2021). A Detection Mechanism Against Load-Redistribution Attacks in Smart Grids. IEEE Transactions on Smart Grid, 12(1), 704-714. https://doi.org/10.1109/tsg.2020.3017562

Keshav, S. (2007). How to read a paper. ACM SIGCOMM Computer Communication Review, 37(3), 83-84. https://doi.org/10.1145/1273445.1273458

Ledolter, J., & Abraham, B. (1981). Parsimony and Its Importance in Time Series Forecasting. Technometrics, 23(4), 411–414. https://doi.org/10.1080/00401706.1981.10487687

L. N. Canha, P. R. Pereira, R. Milbradt, A. da Rosa Abaide, K. E. Kork Schmitt and M. de Abreu Antunes, "Intelligent voltage regulator to distributed voltage control in smart grids," 2017 52nd International Universities Power Engineering Conference (UPEC), 2017, pp. 1-6, doi: 10.1109/UPEC.2017.8231977.

Li, PhD, H. (2022). Time-Series Analysis. In Numerical Methods Using Java (p. 979–1172). Berkeley, CA: Apress. https://doi.org/10.1007/978-1-4842-6797-4_15

Liao, H., Milanovic, J. V., Rodrigues, M., & Shenfield, A. (2018). Voltage Sag Estimation in Sparsely Monitored Power Systems Based on Deep Learning and System Area Mapping. IEEE Transactions on Power Delivery, 33(6), 3162-3172. https://doi.org/10.1109/tpwrd.2018.2865906

Manitsas, E., Singh, R., Pal, B. C., & Strbac, G. (2012). Distribution System State Estimation Using an Artificial Neural Network Approach for Pseudo Measurement Modeling. IEEE Transactions on Power Systems, 27(4), 1888-1896. https://doi.org/10.1109/tpwrs.2012.2187804

Marujo, D., Zanatta, G. L., & Floréz, H. A. R. (2021). Optimal management of electrical power systems for losses reduction in the presence of active distribution networks. Electrical Engineering. https://doi.org/10.1007/s00202-020-01182-5

Massaoudi, M., Refaat, S. S., Chihi, I., Trabelsi, M., Oueslati, F. S., & Abu-Rub, H. (2021). A novel stacked generalization ensemble-based hybrid LGBM-XGB-MLP model for Short-Term Load Forecasting. Energy, 214, 118874. https://doi.org/10.1016/j.energy.2020.118874

Meng, J., Wang, S., Li, G., Jiang, L., Zhang, X., Liu, C., & Xie, Y. (2021). Iterative-learning error compensation for autonomous parking of mobile manipulator in harsh industrial environment. Robotics and Computer-Integrated Manufacturing, 68, 102077. https://doi.org/10.1016/j.rcim.2020.102077

Mestav, K. R., Luengo-Rozas, J., & Tong, L. (2018). State Estimation for Unobservable Distribution Systems via Deep Neural Networks. 2018 IEEE Power & Energy Society General Meeting (PESGM). 2018 IEEE Power & Energy Society General Meeting (PESGM). https://doi.org/10.1109/pesgm.2018.8586649

Milbradt, R. G., Canha, L. N., Zorrilla, P. B., Abaide, A. R., Pereira, P. R., & Schmaedecke, S. M. (2013). A multicriteria approach for meter placement in monitoring of smart distribution systems. 2013 48th International Universities' Power Engineering Conference (UPEC). 2013 48th International Universities' Power Engineering Conference (UPEC). https://doi.org/10.1109/upec.2013.6714874

MORETTIN, P. A., TOLOI, C. M. C. Análise de Séries Temporais. 2ª edição. São Paulo: Egard Blucher, 2006.

Muhammad, I., & Yan, Z. (2015). SUPERVISED MACHINE LEARNING APPROACHES: A SURVEY. ICTACT Journal on Soft Computing, 05(03), 946–952. https://doi.org/10.21917/ijsc.2015.0133

Nejabatkhah, F., Li, Y. W., Liang, H., & Reza Ahrabi, R. (2020). Cyber-Security of Smart Microgrids: A Survey. Energies, 14(1), 27. https://doi.org/10.3390/en14010027

Nguyen, D. T. (2015). Modeling Load Uncertainty in Distribution Network Monitoring. IEEE Transactions on Power Systems, 30(5), 2321-2328. https://doi.org/10.1109/tpwrs.2014.2364819

Ourahou, M., Ayrir, W., EL Hassouni, B., & Haddi, A. (2020). Review on smart grid control and reliability in presence of renewable energies: Challenges and prospects. Mathematics and Computers in Simulation, 167, 19-31. https://doi.org/10.1016/j.matcom.2018.11.009

Paruta, P., Pidancier, T., Bozorg, M., & Carpita, M. (2021). Greedy placement of measurement devices on distribution grids based on enhanced distflow state estimation. Sustainable Energy, Grids and Networks, 26, 100433. https://doi.org/10.1016/j.segan.2021.100433

Pau, M., Patti, E., Barbierato, L., Estebsari, A., Pons, E., Ponci, F., & Monti, A. (2019). Design and Accuracy Analysis of Multilevel State Estimation Based on Smart Metering Infrastructure. IEEE Transactions on Instrumentation and Measurement, 68(11), 4300-4312. https://doi.org/10.1109/tim.2018.2890399

Pegoraro, P. A., & Sulis, S. (2013). Robustness-Oriented Meter Placement for Distribution System State Estimation in Presence of Network Parameter Uncertainty. IEEE Transactions on Instrumentation and Measurement, 62(5), 954-962. https://doi.org/10.1109/tim.2013.2243502

Pereira Barbeiro, P. N., Teixeira, H., Pereira, J., & Bessa, R. (2015). An ELM-AE State Estimator for real-time monitoring in poorly characterized distribution networks. 2015 IEEE Eindhoven PowerTech. 2015 IEEE Eindhoven PowerTech. https://doi.org/10.1109/ptc.2015.7232679

Pereira, P., Canha, L., Milbradt, R., Abaide, A., Schmaedecke, S., Arend, G., & Madruga, E. (2011, May). Optimization of voltage regulators settings and transformer tap zones in distribution systems with great load variation using distribution automation and the smart grids initiatives. 2011 8th International Conference on the European Energy Market (EEM). 2011 European Energy Market (EEM). https://doi.org/10.1109/eem.2011.5953038

Primadianto, A., & Lu, C.-N. (2017). A Review on Distribution System State Estimation. IEEE Transactions on Power Systems, 32(5), 3875-3883. https://doi.org/10.1109/tpwrs.2016.2632156

Qiu, J., Wu, Q., Ding, G., Xu, Y., & Feng, S. (2016). A survey of machine learning for big data processing. EURASIP Journal on Advances in Signal Processing, 2016(1). https://doi.org/10.1186/s13634-016-0355-x

Quiroz, J. E., Reno, M. J., & Broderick, R. J. (2013). Time series simulation of voltage regulation device control modes. 2013 IEEE 39th Photovoltaic Specialists Conference (PVSC). 2013 IEEE 39th Photovoltaic Specialists Conference (PVSC). https://doi.org/10.1109/pvsc.2013.6744472

Raggi, L., Trindade, F., Carnelossi da Cunha, V., & Freitas, W. (2020). Non-Technical Loss Identification by Using Data Analytics and Customer Smart Meters. IEEE Transactions on Power Delivery, 1-1. https://doi.org/10.1109/tpwrd.2020.2974132

Ramírez-Gallego, S., Krawczyk, B., García, S., Wozniak, M., & Herrera, F. (2017). A survey on data preprocessing for data stream mining: Current status and future directions. Neurocomputing, 239, 39–57. https://doi.org/10.1016/j.neucom.2017.01.078

Raposo, A. A. M., Rodrigues, A. B., & da Guia da Silva, M. (2020). Robust meter placement for state estimation considering Distribution Network Reconfiguration for annual energy loss reduction. Electric Power Systems Research, 182, 106233. https://doi.org/10.1016/j.epsr.2020.106233

Rehman, A., Khan, A., Ali, M. A., Khan, M. U., Khan, S. U., & Ali, L. (2020). Performance Analysis of PCA, Sparse PCA, Kernel PCA and Incremental PCA Algorithms for Heart Failure Prediction. In 2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE) (p. 1–5). Apresentado em 2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), Istanbul, Turkey: IEEE. https://doi.org/10.1109/ICECCE49384.2020.9179199

Roberts, C., Scaglione, A., Jamei, M., Gentz, R., Peisert, S., Stewart, E. M., McParland, C., McEachern, A., & Arnold, D. (2020). Learning Behavior of Distribution System Discrete Control Devices for Cyber-Physical Security. IEEE Transactions on Smart Grid, 11(1), 749-761. https://doi.org/10.1109/tsg.2019.2936016

Rossi, B., & Chren, S. (2020). Smart Grids Data Analysis: A Systematic Mapping Study. IEEE Transactions on Industrial Informatics, 16(6), 3619-3639. https://doi.org/10.1109/tii.2019.2954098

Ruiz-Romero, S., Colmenar-Santos, A., Mur-Pérez, F., & López-Rey, Á. (2014). Integration of distributed generation in the power distribution network: The need for smart grid control systems, communication and equipment for a smart city — Use cases. Renewable and Sustainable Energy Reviews, 38, 223–234. https://doi.org/10.1016/j.rser.2014.05.082

Sartika, N., Sukmana, Y., Effendi, M. R., Rusliana, I., Khomisah, K., & Yuningsih, Y. (2021). A Systematic Literature Review on IoT-based Smart Grid. In 2021 7th International Conference on Wireless and Telematics (ICWT) (p. 1–5). Bandung, Indonesia: IEEE. https://doi.org/10.1109/ICWT52862.2021.9678415

Saviozzi, M., Massucco, S., & Silvestro, F. (2019). Implementation of advanced functionalities for Distribution Management Systems: Load forecasting and modeling

through Artificial Neural Networks ensembles. Electric Power Systems Research, 167, 230-239. https://doi.org/10.1016/j.epsr.2018.10.036

Salman, A. G., & Kanigoro, B. (2021). Visibility Forecasting Using Autoregressive Integrated Moving Average (ARIMA) Models. Procedia Computer Science, 179, 252–259. https://doi.org/10.1016/j.procs.2021.01.004

Seabold, S., and Perktold J. “statsmodels: Econometric and statistical modeling with python.” Proceedings of the 9th Python in Science Conference. 2010.

Simone, G. A. Transformadores: teoria e exercícios. São Paulo: editora Érica, 2010.

Ullah, Z., Elkadeem, M. R., Wang, S., & Radosavljevic, J. (2020). A Novel PSOS-CGSA Method for State Estimation in Unbalanced DG-Integrated Distribution Systems. IEEE Access, 8, 113219-113229. https://doi.org/10.1109/access.2020.3003521

Vaziri, M., Vadhva, S., Oneal, T., & Johnson, M. (2011, July). Smart grid, Distributed Generation, and standards. 2011 IEEE Power and Energy Society General Meeting. 2011 IEEE Power & Energy Society General Meeting. https://doi.org/10.1109/pes.2011.6039277

Wang, H., Zhang, W., & Liu, Y. (2016). A Robust Measurement Placement Method for Active Distribution System State Estimation Considering Network Reconfiguration. IEEE Transactions on Smart Grid, 1-1. https://doi.org/10.1109/tsg.2016.2606700

Wang, S., Bi, S., & Zhang, Y.-J. A. (2020). Locational Detection of the False Data Injection Attack in a Smart Grid: A Multilabel Classification Approach. IEEE Internet of Things Journal, 7(9), 8218-8227. https://doi.org/10.1109/jiot.2020.2983911

Wang, Z., Wang, J., Chen, B., Begovic, M. M., & He, Y. (2014). MPC-Based Voltage/Var Optimization for Distribution Circuits With Distributed Generators and Exponential Load Models. IEEE Transactions on Smart Grid, 5(5), 2412-2420. https://doi.org/10.1109/tsg.2014.2329842

Xiang, Y., Ribeiro, P. F., & Cobben, J. F. G. (2014). Optimization of State-Estimator-Based Operation Framework Including Measurement Placement for Medium Voltage Distribution Grid. IEEE Transactions on Smart Grid, 5(6), 2929-2937. https://doi.org/10.1109/tsg.2014.2343672

Xu, R. (2022). Fuzzy C-means Clustering Image Segmentation Algorithm Based on Hidden Markov Model. Mobile Networks and Applications. https://doi.org/10.1007/s11036-022-01917-7

Xygkis, T. C., & Korres, G. N. (2016). Optimal allocation of smart metering systems for enhanced distribution system state estimation. 2016 Power Systems Computation Conference (PSCC). 2016 Power Systems Computation Conference (PSCC). https://doi.org/10.1109/pscc.2016.7540966

Yang, Y., Li, S., Li, W., & Qu, M. (2018). Power load probability density forecasting using Gaussian process quantile regression. Applied Energy, 213, 499-509. https://doi.org/10.1016/j.apenergy.2017.11.035

Yaziz, S. R., Zakaria, R., & Ahmad, M. H. (2017). Determination of sample size for higher volatile data using new framework of Box-Jenkins model with GARCH: A case study on gold price. Journal of Physics: Conference Series, 890, 012161. https://doi.org/10.1088/1742-6596/890/1/012161

Ye, G., Nijhuis, M., Cuk, V., & Cobben, J. F. G. (2019). Incorporating network uncertainties in voltage dip state estimation. International Journal of Electrical Power & Energy Systems, 113, 888-896. https://doi.org/10.1016/j.ijepes.2019.06.005

Ye, G., Xiang, Y., Cuk, V., & Cobben, J. F. G. (2015). Voltage dip state estimation in distribution networks by applying Bayesian inference. IECON 2015 - 41st Annual Conference of the IEEE Industrial Electronics Society. https://doi.org/10.1109/iecon.2015.7392216

Yuan, Y., Dehghanpour, K., Bu, F., & Wang, Z. (2019). A Multi-Timescale Data-Driven Approach to Enhance Distribution System Observability. IEEE Transactions on Power Systems, 34(4), 3168-3177. https://doi.org/10.1109/tpwrs.2019.2893821

Yuehao, Y., Hui, L., Wei, B., Zhaohui, L., Hao, Z., & Yaoheng, D. (2016, August). A distribution network state estimation method based on power user electric energy data acquisition system. 2016 China International Conference on Electricity Distribution (CICED). 2016 China International Conference on Electricity Distribution (CICED). https://doi.org/10.1109/ciced.2016.7576317

Zhang, J., Wang, Y., Weng, Y., & Zhang, N. (2020). Topology Identification and Line Parameter Estimation for Non-PMU Distribution Network: A Numerical Method. IEEE Transactions on Smart Grid, 11(5), 4440-4453. https://doi.org/10.1109/tsg.2020.2979368

Zhang, Y., Wang, J., & Chen, B. (2021). Detecting False Data Injection Attacks in Smart Grids: A Semi-Supervised Deep Learning Approach. IEEE Transactions on Smart Grid, 12(1), 623-634. https://doi.org/10.1109/tsg.2020.3010510

Zhang, Y., Wang, X., Wang, J., & Zhang, Y. (2021). Deep Reinforcement Learning Based Volt-VAR Optimization in Smart Distribution Systems. IEEE Transactions on Smart Grid, 12(1), 361-371. https://doi.org/10.1109/tsg.2020.3010130

Zhang, E., Li, H., Huang, Y., Hong, S., Zhao, L., & Ji, C. (2022). Practical multi-party private collaborative k-means clustering. Neurocomputing, 467, 256–265. https://doi.org/10.1016/j.neucom.2021.09.050

Zhang X., Hodge, J., & Attavvay, M. (2014). Intelligent Voltage Regulator in Smart Grid Distribution System. In 2014 China International Conference on Electricity Distribution (CICED) (p. 1716–1720). Shenzhen, China: IEEE. https://doi.org/10.1109/CICED.2014.6991996

Zhao, J., Huang, C., Mili, L., Zhang, Y., & Min, L. (2020). Robust Medium-Voltage Distribution System State Estimation using Multi-Source Data. 2020 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT). 2020 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT). https://doi.org/10.1109/isgt45199.2020.9087787