UNIVERSIDADE FEDERAL DE JUIZ DE FORA
FACULDADE DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL

Técnicas de Aprendizado de Máquina Não Supervisionado para a Prevenção de Falhas em Máquinas de Chave

NIELSON SOARES

JUIZ DE FORA
2018

Técnicas de Aprendizado de Máquina Não Supervisionado para a Prevenção de Falhas em Máquinas de Chave

Dissertação apresentada ao Programa de Pós-graduação em Modelagem Computacional, da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do grau de Mestre em Modelagem Computacional.

Orientador: Prof. Dr. Leonardo Goliatt da Fonseca

Coorientador: Prof. Dr. Eduardo Pestana de Aguiar

JUIZ DE FORA

FACULDADE DE ENGENHARIA

2018

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Soares, Nielson.
Técnicas de Aprendizado de Máquina Não Supervisionado para a Prevenção de Falhas em Máquinas de Chave / Nielson Soares. -- 2018.
68 f. : il.

Orientador: Leonardo Goliatt da Fonseca
Coorientador: Eduardo Pestana de Aguiar
Dissertação (mestrado acadêmico) - Universidade Federal de Juiz de Fora, ICE/Engenharia. Programa de Pós-Graduação em Modelagem Computacional, 2018.

1. máquinas de chave. 2. inteligência computacional. 3. aprendizado de máquina. 4. previsão de falhas. I. Goliatt da Fonseca, Leonardo, orient. II. Pestana de Aguiar, Eduardo, coorient. III. Título.

Aplicação de Técnicas de Inteligência Computacional no Diagnóstico de Máquinas de Chave

NIELSON SOARES

Dissertação apresentada ao Programa de Pós-graduação em Modelagem Computacional, da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do grau de Mestre em Modelagem Computacional.

Aprovado em: 27/02/2018

Por:

Prof. D.Sc. Leonardo Goliatt da Fonseca - Orientador
Universidade Federal de Juiz de Fora

Prof. D.Sc. Eduardo Pestana de Aguiar - Coorientador
Universidade Federal de Juiz de Fora

Prof. D.Sc. Luciana Conceição Dias Campos
Universidade Federal de Juiz de Fora

Prof. D.Sc. Eduardo Krempser da Silva
Fundação Oswaldo Cruz

AGRADECIMENTOS

Agradeço primeiramente a Deus, que me proveu de todas as ferramentas necessárias para que eu alcançasse meus objetivos.

Aos meus pais, Nelson e Nilza, exemplos de vida, por serem meu porto seguro, sempre me apoiando nos momentos difíceis. Ao meu irmão Nedson, cujo carinho e amor foram fundamentais.

Aos meus tios, tias e primos que me abraçaram e que fizeram com que a distância para com meus pais não fosse tão sentida. Principalmente aos tios, Marco Aurélio e Gisele, que me receberam em seu apartamento.

À Mariana, minha namorada, e seus pais, João Batista e Alessandra, que me ajudaram muito nesses anos, e me fizeram parte da família.

À empresa MRS Logística S.A. pela sua parceria e que, sem ela, esta dissertação não seria possível.

Aos professores Leonardo Goliatt e Eduardo Aguiar, pela orientação, ajuda e, principalmente, paciência durante a elaboração deste trabalho.

Aos professores e funcionários do Programa de Pós-Graduação em Modelagem Computacional pelos ensinamentos proporcionados e suporte dado aos alunos, sempre buscando resolver os problemas da melhor forma.

Aos meus amigos de mestrado, que me deram o incentivo pra persistir apesar das adversidades e com quem compartilhei momentos maravilhosos ao longo desses anos. E à todos colegas e amigos que fiz durante o caminho percorrido e que estiveram ao meu lado.

"Computers themselves, and software yet to be developed, will revolutionize the way we learn".

Steve Jobs

RESUMO

As máquinas de chave são equipamentos eletromecânicos de grande importância em uma malha ferroviária. A ocorrência de falhas nesses equipamentos pode ocasionar interrupções das ferrovias e acarretar potenciais prejuízos econômicos. Assim, um diagnóstico precoce dessas falhas pode representar uma redução de custos e um aumento de produtividade. Essa dissertação tem como objetivo propor um modelo preditivo, baseado em técnicas de inteligência computacional, para a solução desse problema. A metodologia aplicada compreende o uso de técnicas de extração e seleção de características baseada em testes de hipóteses e modelos de aprendizado de máquina não supervisionado. O modelo proposto foi testado em uma base de dados disponibilizada por uma empresa ferroviária brasileira e se mostrou eficiente ao constatar como críticas as operações realizadas próximas à operação classificada como falha.

Palavras-chave: máquinas de chave; inteligência computacional; aprendizado de máquina; previsão de falhas

ABSTRACT

Railroad switch machines are important electromechanical equipment in a railway network, and the occurrence of failures in such equipment can cause railroad interruptions and lead to potential economic losses. Thus, an early diagnosis of these failures can represent a reduction of costs and an increase in productivity. This dissertation aims to propose a predictive model, based on computational intelligence techniques, to solve this problem. The applied methodology includes the use of features extraction and selection techniques based on hypothesis tests and unsupervised machine learning models. The proposed model was tested in a database made available by a Brazilian railway company and proved to be efficient when considering as critical the operations performed close to the operation classified as failure.

Key-words: railroad switch; computational intelligence; machine learning; failure prediction

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AMV	Aparelho de Mudança de Via
ANTF	Associação Nacional dos Transportes Ferroviários
ANTT	Agência Nacional de Transportes Terrestres
AWGN	*Additive White Gaussian Noise*
CBM	*Condition Based Maintenance*
CPS	*Cyber-Physical System*
C-SVM	*Classification Support Vector Machine*
DNIT	Departamento Nacional de Infraestrutura de Transportes
EFC	Estrada de Ferro Carajás
FDA	Função de Distribuição Acumulada
FDP	Função Densidade de Probabilidade
FDR	*False Discovery Rate*
FRESH	*FeatuRe Extraction based on Scalable Hypothesis tests*
HC	*Hierarchical Clustering*
ID	Indicador de Dano
KDE	*Kernel Density Estimation*
KS	*Kolmogorov–Smirnov*
MC	Máquinas de Chave
MRS	MRS Logística S.A.
PCA	*Principal Component Analysis*
RFFSA	Rede Ferroviária Federal Sociedade Anônima
SNR	*Signal to Noise Ratio*
ST	Série Temporal
TSC	*Time Series Classification*
TST	*Time Stopped Train*

SUMÁRIO

1 INTRODUÇÃO

1.1 CONTEXTO HISTÓRICO E RELEVÂNCIA DO TEMA

A primeira ferrovia do Brasil foi inaugurada em 1854 no Rio de Janeiro, entre o Porto de Estrela, situado ao fundo da Baía da Guanabara e a localidade de Raiz da Serra, em direção à cidade de Petrópolis e continha uma extensão de apenas 14,5 km. Posteriormente, várias outras foram surgindo em diversos pontos do país. Em 1884, o país contava com 6.116 km de extensão e em 1922 estimava-se que o país possuía um sistema ferroviário com, aproximadamente, 29.000 km de extensão. Na década de 50, as estradas de ferro no Brasil possuíam um total de 37.000 km de extensão de linhas, então, o Governo Federal decidiu por unir a administração dessas estradas, criando, em 1957, a sociedade anônima Rede Ferroviária Federal S.A. (RFFSA), que tinha como propósito administrar as estradas de ferro pertencentes à União. Entretanto, de 1980 a 1992, os sistemas ferroviários pertencentes à RFFSA sofreram uma queda substancial nos investimentos. Como consequência o Governo Federal colocou em prática um processo de desestatização, sendo a RFFSA leiloada, a partir de 1996, à iniciativa privada e sendo dissolvida no ano de 1999 (DNIT, 2018).

Figura 1 – Primeira estação ferroviária Brasil. Do acervo de Moisés Rodrigues Mano

Fonte: retirado de (Mayrink, 2016)

Após uma década, de 1996 à 2006, do processo de desestatização, foi concluído, em análise feita por Fleury (2006), que a privatização proporcionou, dentre diversos resultados,

um aumento no faturamento conjunto das empresas e nos investimentos realizados pelas concessionárias e uma queda no índice de acidentes. Apesar disso, houve uma queda na produtividade dos vagões, isto é, a quantidade de tonelada/quilômetros transportadas por ano, por cada vagão, foi reduzida em 14,9%.

Posteriormente, houve um crescimento de diversos setores da indústria brasileira, entre eles o setor ferroviário brasileiro, que, em dez anos, entre 2006 e 2016, aumentou o volume total de cargas transportadas no Brasil por trens em quase 30%, de acordo com o Anuário do Setor Ferroviário (ANTT, 2017), com destaque para as concessionárias EFC (Estrada de Ferro Carajás) e MRS Logística S.A., que aumentaram em quase 68% e 39% respectivamente, segundo a mesma publicação. A Tabela 1 apresenta uma informação mais detalhada sobre o transporte de cargas das concessionárias da Associação Nacional dos Transportadores Ferroviários (ANTF).

Tabela 1 – Produção de transporte ferroviário de cargas, em toneladas úteis (TU)

Produção de Transporte Ferroviário											
Concessionária	Toneladas Úteis (milhares de TU)										
	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
RMN	5.551	6.928	8.232	10.072	10.498	11.611	13.952	14.416	15.010	16.747	14.906
RMO	3.355	2.690	3.235	2.778	4.430	4.421	3.932	4.625	5.600	4.560	3.505
RMP	4.221	3.473	5.229	4.917	6.719	7.490	5.702	5.336	5.440	4.734	6.013
RMS	28.942	26.536	26.763	26.073	25.975	27.067	24.192	22.940	21.554	20.938	18.345
EFC	92.591	100.361	103.670	96.267	104.949	114.543	116.428	115.006	118.454	134.713	155.252
EFPO	1.511	862	996	646	471	400	306	210	356	369	440
EFVM	131.620	136.604	133.211	104.317	131.755	133.462	127.268	125.296	126.185	132.976	129.601
FCA	15.177	18.957	19.280	17.455	21.242	18.958	22.471	24.290	24.192	26.512	24.993
FNSTN	0	0	1.424	1.639	2.012	2.541	3.187	3.215	4.370	5.599	5.029
FTC	2.627	2.635	3.038	2.856	2.637	2.448	2.968	3.240	3.854	3.527	2.898
FTL	1.519	1.814	1.643	1.467	1.529	1.431	1.389	1.212	1.218	1.220	1.320
MRS	101.998	114.064	119.799	110.954	123.030	130.009	131.404	130.906	138.827	139.695	141.501
TOTAL	389.113	414.925	426.520	379.441	435.248	454.380	453.200	450.693	465.060	491.590	503.804

Fonte: (ANTT, 2017)

Nos próximos anos, esses números podem a aumentar devido à Lei nº 13.448 de 2017, que permite a antecipação da prorrogação dos contratos das concessionárias, possibilitando a antecipação de investimentos que só seriam realizados ao fim das concessões, tendo cada operadora um aporte médio de R$ 5 bilhões cada para investir na ampliação de capacidade de transporte, que, entre outras coisas, envolve a compra de equipamentos e softwares de última geração.

1.2 JUSTIFICATIVA

Com a ampliação das linhas férreas, ocorre um aumento no uso de equipamentos como os Aparelhos de Mudança de Via (AMV) e que possuem grande importância em uma malha ferroviária, pois possibilitam que locomotiva e composição sejam guiadas de uma ferrovia para outra. Como consequência, também há um aumento na probabilidade de falhas acontecerem. Essas falhas podem representar um grande custo para as empresas, seja ele econômico, devido à atraso nos transportes e acidentes, como colisões e descarrilhamentos, ou custo humano, quando esses acidentes causam vítimas.

A pesquisa feita por Dindar & Kaewunruen (2017) fez uma avaliação dos descarrilhamentos relacionados com AMVs no Reino Unido e o resultado pode ser visto na Figura 2, que compreende às causas imediatas que pode ocasionar um acidente.

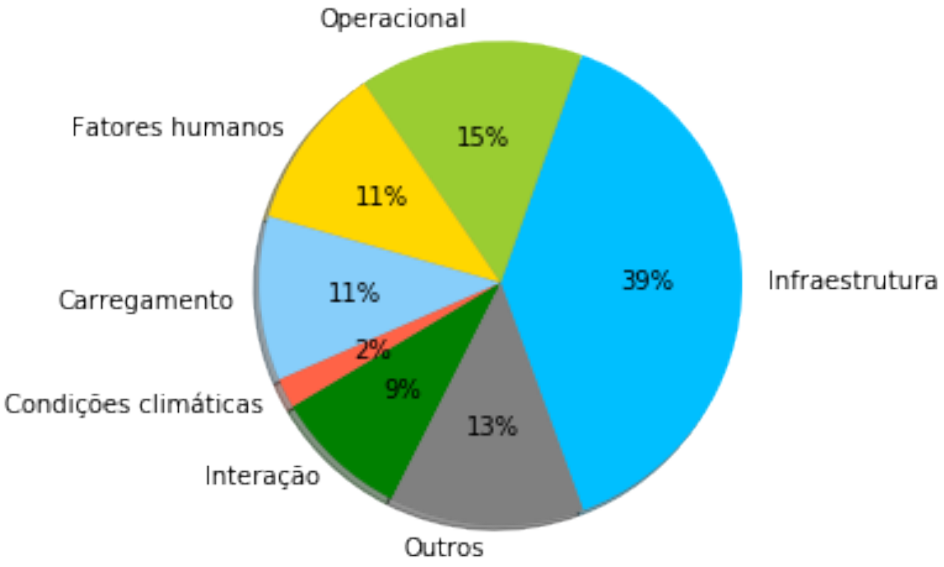


Figura 2 – Distribuição de causas de descarrilhamento de trens.

Nota-se que a maioria dos acidentes, 39%, estão relacionados com problemas de infraestrutura, como trilhos defeituosos ou falhas em diversos componentes que constituem um AMV, como agulhas ou máquinas de chave. Além das causas imediatas, os fatores causais são muito importantes na determinação da causa ou causas dos descarrilamentos

para evitar novos acidentes relacionados (Dindar; Kaewunruen, 2017). A Tabela 2 mostra esses fatores e a responsabilidade de cada um nas causas de descarrilhamento.

Tabela 2 – Resultados estatísticos dos fatores causais

Fatores Causais	Descrição	%
Manutenção	Problemas de instalação, inspeção inadequada, falhas não detectadas	32,6
Condições climáticas	Tempestades, gelo nas vias, altas temperaturas	16,6
Fatores humanos	Vandalismo, resposta lenta à sinalização ou comunicação	13,6
Geometria da via	Alargamento dos trilhos, falhas na bitola, torção dos trilhos	11,4
Componente do vagão	Rodas desgastadas	11,4
Problemas de sinalização	Sinalização bloqueada, Sinalização mal posicionada	4,5
Outros	Problemas de design, componentes vulneráveis	9,1

Fonte: (Dindar; Kaewunruen, 2017)

A partir da tabela, observa-se que uma baixa frequência de manutenção ou uma inspeção inadequada ou a falta de inspeção são os principais fatores relacionados com a ocorrência de falhas. Em razão disso, operações em tempo real de monitoramento e diagnóstico dos equipamentos são de vital importância, principalmente no que diz respeito à manutenção preditiva, visando evitar essas falhas, prevenindo acidentes, reduzindo o tempo de trem parado (TST - *Time Stopped Train*) e aumentando a produtividade da empresa (Aguiar *et al.*, 2014; Aguiar *et al.*, 2016).

Politicas de manutenção preditiva estão se tornando cada vez mais atrativas, aumentando o interesse de pesquisadores e profissionais do transporte em investigar a possibilidade de aplicação de técnicas de inteligência computacional objetivando a solução de problemas críticos, a fim de melhorar a eficiência e segurança dos sistemas de transporte (Li; Song; Cai, 2015; Soler *et al.*, 2015; Guan *et al.*, 2015). Uma delas é a Manutenção Baseada em Condição (CBM - *Condition Based Maintenance*), em que o equipamento é monitorado através de sensores e a ação é tomada baseando-se em análises preditivas realizadas em cima de dados do histórico de funcionamento desse equipamento, permitindo que a manutenção seja feita previamente à ocorrência da falha (Ellis; Byron, 2008).

1.3 OBJETIVOS

1.3.1 **Objetivo Geral**

O objetivo principal desta dissertação é propor um modelo de análise preditiva, através de uma medida de indicação de dano, baseada em técnicas de inteligência computacional, objetivando o diagnóstico precoce de falhas em máquinas de chave, equipamentos responsáveis por operar os AMVs. A metodologia aplicada compreende o uso de técnicas de extração e seleção de características baseada em testes de hipóteses e modelos de aprendizado de máquina não supervisionado

1.3.2 **Objetivos Específicos**

1. Estudar e implementar técnicas de extração de características para séries temporais, métodos de agrupamento e técnicas de visualização;

2. Estudar e implementar teste de hipóteses, um teste estatístico, a fim de se obter um conjunto de características apropriados para uma boa representação dos dados;

3. Integrar técnicas de extração de características com algoritmos de agrupamento e uma medida de indicador de dano para determinar a criticidade de manutenção de máquinas de chave.

1.4 ORGANIZAÇÃO DO TRABALHO

Este trabalho está dividido em cinco capítulos. No primeiro capítulo é feito uma apresentação do estudo realizado, realçando as justificativas e os objetivos do mesmo. No Capítulo 2 o problema de manutenção de equipamentos baseado em condição é levantando. É feita uma breve explicação sobre máquinas de chave e séries temporais, apresentando conceitos importantes para o desenvolvimento deste trabalho. No Capítulo 3 é feita uma descrição da base de dados utilizada e os conceitos das técnicas de inteligência computacional aplicadas são explicados. Os resultados obtidos são então mostrados no Capítulo 4, no qual também é apresentado as análises e discussões acerca desses resultados. O Capítulo 5 apresenta a conclusão com base nos resultados obtidos e as indicações para trabalhos futuros e no Apêndice A encontram uma lista com as técnicas utilizas para extrair características para testar o desempenho das técnicas de inteligência computacional.

2 REFERENCIAL TEÓRICO

2.1 APRESENTAÇÃO DO TEMA

O modal ferroviário é o sistema de transporte mais eficiente quando se trata de cargas de baixo valor agregado a grandes distâncias, como as commodities[1], principalmente devido à sua grande capacidade de carregamento. Diferentemente dos outros modais de transporte, no ferroviário não são os veículos que decidem as direções de movimento, sendo essa função pertinente aos trilhos da ferrovia, que determinam qual caminho seguir, não sendo possível que um veículo possa mudar de via por conta própria, principalmente em cruzamentos entre ferrovias como da Figura 3.

Figura 3 – Estação Ferroviária do Entroncamento, ponto de bifurcação.

Para solucionar esse problema, foram criados os AMVs, que permitiam que os veículos ferroviários pudessem se locomover de uma linha para outra, através da movimentação de suas agulha, descritas na Tabela 3. A Figura 4 mostra com mais detalhes um AMV instalado em um cruzamento entre linhas. Os AMVs são detalhados na Seção 2.1.1 a seguir.

2.1.1 **Aparelhos de Mudança de Via**

Os AMVs por serem de vital importância na variação da direção de veículos em uma ferrovia, são largamente empregados, principalmente em terminais de cargas, e devem ser constantemente monitorados. A Tabela 3 apresenta de forma mais detalhada a composição dos aparelhos de vias mais comuns.

[1] Termo usado para descrever produtos de baixo valor agregado, de base em estado bruto (matérias-primas), de qualidade quase uniforme e produzidos em escala.

Figura 4 – Cruzamento entre linhas.

Tabela 3 – Componentes mais comuns de um AMV

Componente	Descrição
Agulhas	Peças móveis e paralelas entre si. São ligadas por uma barra ao aparelho de manobra e são responsáveis por direcionar o veículo na direção desejada
Trilho de encosto	Peças adaptadas para servirem de batentes às agulhas
Jacaré	É a parte do AMV que possibilita o cruzamento entre duas direções de uma mesma linha de trilhos
Trilhos intermediários	São os trilhos que fazem a ligação das agulhas ao jacaré
Calços	São peças de ferro fundido, aparafusadas entre os trilhos e os contratrilhos, visando manter a distância entre eles estável.
Placas de deslizamento	São placas colocas sob as agulhas para facilitar a movimentação das mesmas durante a operação das máquinas de chave. Devem ser mantidas sempre lubrificadas.
Contratrilhos	São trilhos colocados na parte interna dos trilhos externos com finalidade de guiar o veículo durante sua passagem pelo jacaré, evitando as rodas de colidirem com o mesmo.
Aparelho de manobra	São as partes que permitem o movimento das agulhas, sendo elas: barra de conjugação, tirante e máquina de chave

Este estudo foi focado nas máquinas de chave (MC), pois, por serem responsáveis por toda movimentação dos AMVs, são fundamentais para que ocorra a mudança de via.

2.1.2 **Máquinas de Chave**

Hoje é possível encontrar três tipos de máquinas de chave no mercado, são divididas conforme sua forma de operação:

- Manuais;
- Pneumáticas;
- Elétricas;

2.1.2.1 *Manuais*

A máquinas de chave manuais são acionadas por meio de alavancas, molas ou cremalheiras. A Figura 5 mostra uma máquina de chave manual operada por alavanca.

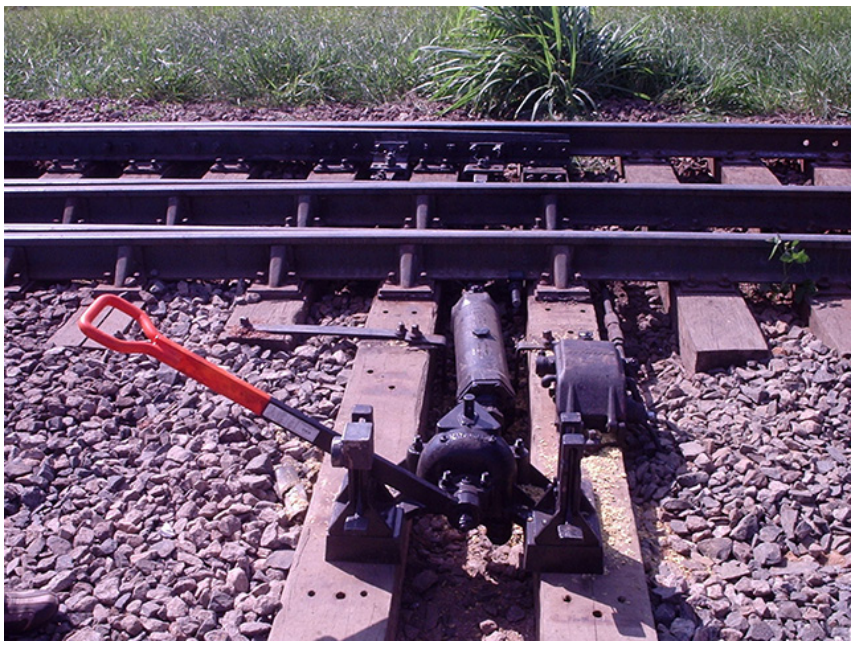


Figura 5 – Máquina de chave manual.

Fonte: retirado de (Cardoso, 2015)

2.1.2.2 *Pneumáticas*

São as máquinas que movimentam as agulhas por meio de dispositivos a ar comprimido. A Figura 6 apresenta um modelo de máquina de chave pneumática.

2.1.2.3 *Elétricas*

As máquinas de chave elétricas são movimentadas por meio de motores elétricos. Normalmente, essas máquinas possuem dispositivos elétricos que servem para detectar a

Figura 6 – Máquina de chave pneumática.

Fonte: retirado de (Mike, 2014)

posição da agulha. Podendo ser normal, quando as agulhas estão posicionadas de modo que o veículo ferroviário não altera de via, e reverso, quando as agulhas estão em posição oposta, alterando a direção do veículo. Um exemplo de máquina de chave elétrica é mostrada na Figura 7.

Figura 7 – Máquina de chave elétrica.

Fonte: retirado de (Mike, 2015)

A máquina de chave equipada com dispositivo indicador de posição deve obedecer à seguinte sequência de operação (DNIT, 2015):

- Abrir o circuito de indicação; e
- Destravar a máquina de chave:
  - Operar a chave;
  - Travar a máquina de chave; e
  - Restituir o circuito de indicação.

Devido à sua importância nas malhas ferroviária, as máquinas de chave elétrica têm sido o interesse de pesquisadores e profissionais da área, realizando diversos estudos relacionados. A maioria dos estudos disponíveis aplicam algoritmos de processamento de sinais e aprendizado de máquina para o diagnóstico de máquinas de chave.

Sistemas neurofuzzy para detecção e diagnóstico de falha foram abordados em Chen & Roberts (2006). Adachi, Kikuchi & Watanabe (2006) utilizaram técnicas de mineração de dados e métodos baseados em aprendizado estatístico, como testes de hipóteses, estimação de um intervalo de confiança e análise discriminante com regressão múltipla, objetivando a detecção de falhas em máquinas de chave elétricas. Márquez & Schmid (2007) aplicaram um Filtro de Kalman para suavizar o ruído no sinal de corrente de motor antes de realizar uma análise espectral para detecção de falhas. García, Pedregal & Roberts (2010) desenvolveram um sistema de detecção de falha baseado na comparação entre o sinal esperado e o sinal observado. No estudo feito por Asada & Roberts (2013), foi realizado uma extração de características utilizando transformada de wavelet de Haar em conjunto com o método de classificação C-SVM (*Classification Support Vector Machine*). Outra abordagem, proposta por Eker, Camci & Kumar (2010), foi a aplicação de análise de componentes principais para redução de características e um SVM com Kernel Gaussiano para classificação. Aguiar *et al.* (2017) e Aguiar *et al.* (2018) propuseram um sistema de inferência fuzzy do tipo 1 (Aguiar *et al.*, 2017), e do tipo 2 (Aguiar *et al.*, 2018) para a classificação das falhas mais comuns que podem ocorrer em máquinas de chave.

Esta dissertação propõe combinar as diversas técnicas de extração de características encontradas na literatura, selecionando as mais relevantes para encontrar uma solução de modo a tentar anteceder a ocorrência de falhas em máquinas de chave, permitindo a realização de uma manutenção preditiva.

A maioria dos estudos aqui citados, utilizam como base de dados a corrente de operação do motor elétrico das máquinas de chave, devido à facilidade de sua medição através de sensores não invasivos, não interferindo no processo de operação. Os dados de

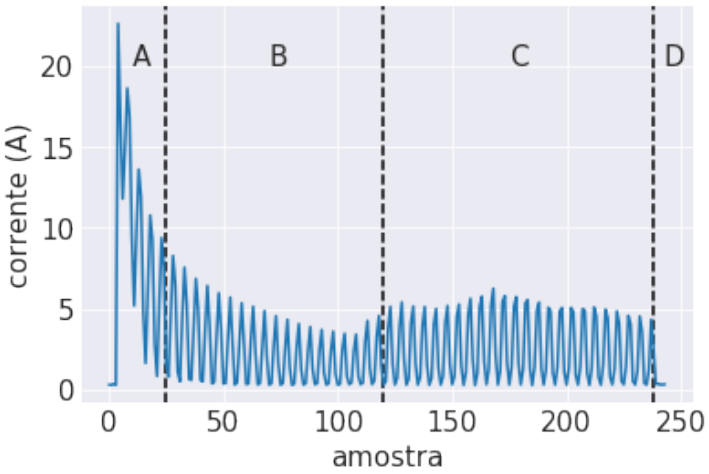


Figura 8 – Sinal de corrente observado.

corrente observados possuem uma curva característica, como pode ser vista na Figura 8. O valor da corrente é obtida em amostras de tempo discreto com intervalos de 10ms.

Nos primeiros instantes, na região A, há um pico de corrente em razão do torque necessário para tirar as agulhas da inércia e iniciar o movimento. Depois disso, na região B, a corrente cai devido ao deslizamento das agulhas, em seguida, na região C, nota-se um leve aumento da corrente, uma vez que é necessário mais torque para realização do travamento das agulhas na posição contrária e por fim, na região D, o desligamento do motor. Essas operações geram uma série de observações durante um período de tempo, criando o que é chamado de séries temporais. Essas séries temporais também são utilizadas na realização desta dissertação e o Capítulo 3 apresenta alguns conceitos sobre o assunto.

3 MATERIAIS E MÉTODOS

Séries temporais (ST) são meios utilizados para gerenciar coleções de dados que possuem valores observados em períodos regulares ou intervalos, e geralmente são representadas como uma sequência de eventos. Um evento é um par ordenado que consiste em um valor temporal e um valor de dados (Elmasri; Lee, 1998) como, por exemplo, medições de sensores em um equipamento, podendo ser representado como:

$$s_j(t) \rightarrow s_j(t + \Delta_t) \rightarrow s_j(t + 2\Delta_t) \rightarrow \ldots \rightarrow s_j(t + n\Delta_t) \tag{3.1}$$

em que, $s_j(t)$ é o estado de um equipamento em relação à uma medida específica do sensor j, repetida num intervalo de tempo igual a Δ_t. A Figura 9 apresenta um exemplo de séries temporais baseado nas medições de sensores de força e torque instalados em um robô, medidos em intervalos de tempo regulares (Lichman, 2013).

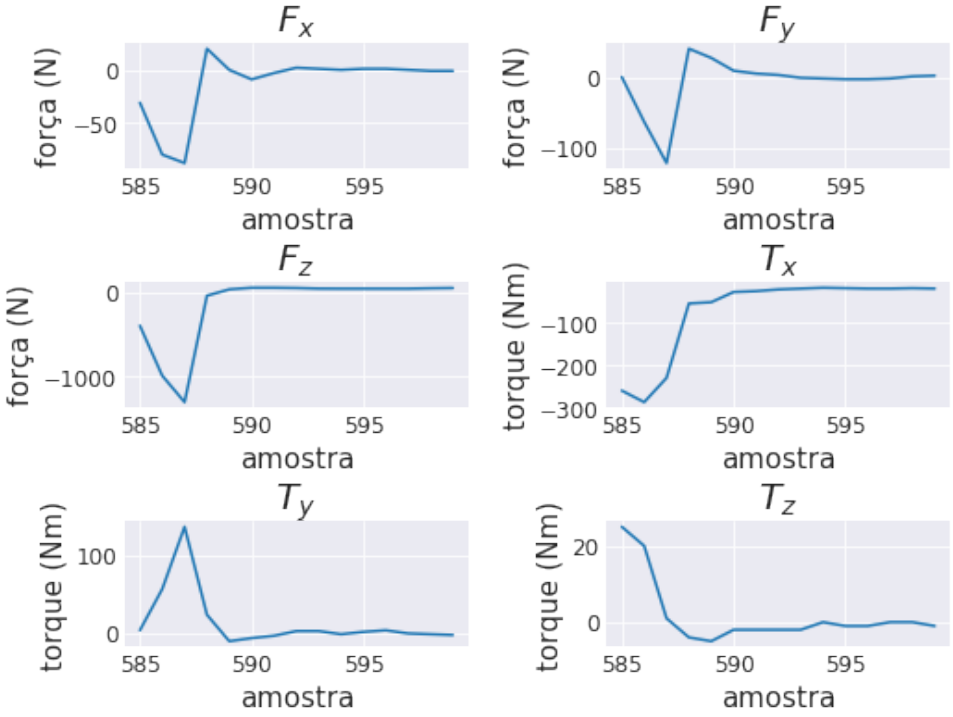


Figura 9 – Séries temporais resultante de sensores.

As STs podem apresentar diferentes comportamentos, então a extração de características se faz necessária, a fim de que se obtenha uma compreensão das forças subjacentes que originaram os dados observados. Uma investigação descritiva dos dados da série é muito importante para a escolha do modelo a ser aplicado, e determinado um modelo que se ajuste bem aos seus dados, análises podem ser feitas para fins de planejamento e controle. Assumindo um vetor $Y = (y_1, \ldots, y_m)$ em que y_i de Y compreende ao estado de

um equipamento no momento i, a predição de valores de Y baseado em diferentes STs, $S = (S_{i,j})_{i=1,...,m}$, implica em um problema de classificação de séries temporais (TSC - *Time Series Classification*). Existe na literatura diferentes propostas para atacar problemas de TSC. No trabalho de Bagnall *et al.* (2017) é possível encontrar uma visão geral sobre as propostas mais comuns na área de TSC.

Algumas propostas baseiam-se nas formas das séries, identificando as similaridades e classificando-as de acordo com a proximidade entre elas, como o estudo feito por Wang, Smith & Hyndman (2006), que aplica métodos como k-vizinhos mais próximos, que classifica um ST conforme a classe representada pela maioria de seus vizinhos. Outra abordagem é a aplicação de redes neuronais e aprendizado profundo, conhecido também como *deep learning*, que aprendem a representação dos dados, (Yang *et al.*, 2015; Hüsken; Stagge, 2003).

Nas seções a seguir, é apresentada a base de dados com as séries temporais utilizadas e também alguns conceitos sobre os métodos de inteligência computacional aplicados neste trabalho. Os métodos foram implementados utilizando a linguagem de programação de alto nível Pyhton, versão 2.7.0, em conjunto com as bibliotecas de código aberto *scikit-learn* (Pedregosa *et al.*, 2011), a qual contém ferramentas para o desenvolvimento de modelos de aprendizado de máquina, *NumPy* (Walt; Colbert; Varoquaux, 2011), pacote fundamental para a computação científica com o Python, e *pandas* (McKinney *et al.*, 2010), que oferece estruturas de dados de alto desempenho.

3.1 BASE DE DADOS

Os dados adquiridos para o desenvolvimento deste estudo foram disponibilizados pela MRS Logística S.A., uma empresa ferroviária brasileira, e foram obtidos através de quatro canais de uma placa de aquisição de dados industriais. Os mesmos consistem em séries temporais formadas por observações da corrente (A) de operação da máquina de chave medida em intervalos regulares de tempo de 10ms. Os sinais foram medidos a partir de seis máquinas de mesmo modelo em operação durante o período compreendido entre 23 de Junho e 4 de Julho de 2017.

As séries passaram por um processo de pré-tratamento, no qual foram retirados os períodos de ociosidade, em que as máquinas não foram acionadas, resultando em um total de 615 operações observadas, das quais, seis representam uma operação em que houve a ocorrência de alguma falha, sendo uma para cada máquina. As operações são divididas conforme abaixo:

- Máquina #1: 202 operações;

- Máquina #2: 66 operações;
- Máquina #3: 94 operações;
- Máquina #4: 76 operações;
- Máquina #5: 68 operações;
- Máquina #6: 109 operações;

Cada máquina possui dados de operação em sentido normal e em sentido reverso, os quais são compostos por corrente positiva e negativa, respectivamente. Foi feita uma estimativa de densidade Kernel (KDE - *Kernel Density Function*) (Elgammal *et al.*, 2002), usada para estimar a função de densidade de probabilidade da corrente e o resultado pode ser visto na Figura 10.

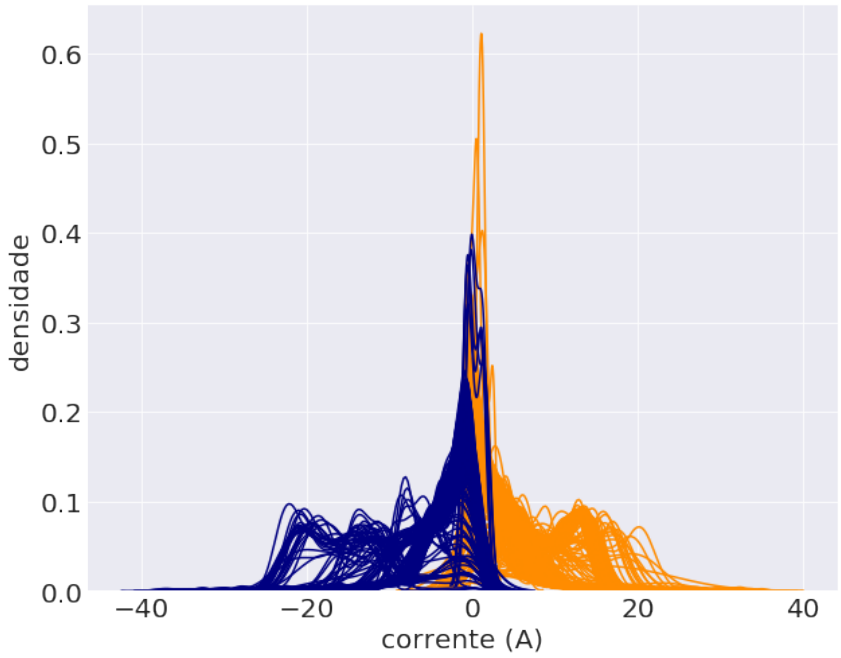


Figura 10 – Função densidade para os dados.

Os tipos de operação foram divididos por cor, sendo laranja para operação em sentido normal e azul para operação em sentido reverso. É possível ver na figura que as operações possuem distribuições distintas. Todavia, para realização deste trabalho, os sentidos não serão levados em consideração, visto que as falhas podem ocorrer em qualquer sentido.

3.2 MÉTODOS

O estudo foi conduzido de acordo com a metodologia esquematizada na Figura 11, e pode ser dividido nas seguintes etapas: tratamento de dados, experimentos computacionais e interpretação dos resultados.

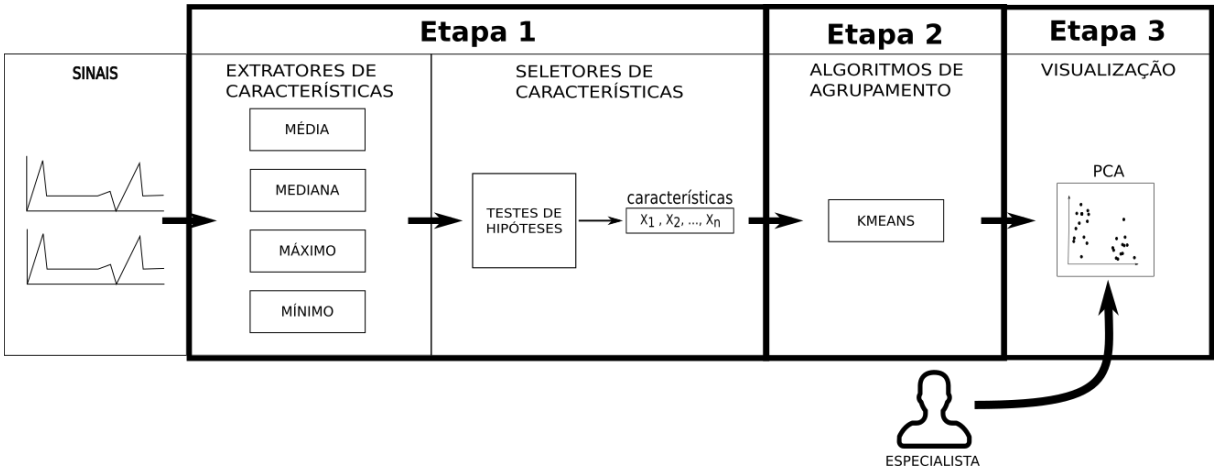


Figura 11 – Descrição da metodologia utilizada.

A primeira etapa consistiu na extração e seleção das características que mais representam as séries temporais formadas pela medição de sensores instalados nas máquinas de chave. A segunda etapa foi composta pelo processamento dos dados por meio dos modelos de aprendizado de máquina e a terceira compreendeu à análise dos resultados obtidos. As etapas e os métodos utilizados são descritos nas seções seguintes.

3.2.1 **Extração e Seleção de Características**

As séries temporais são naturalmente um tipo de dados de alta dimensionalidade e grande em tamanho (Rani; Sikka, 2012; Lin *et al.*, 2004), o que prejudica a análise desses dados por meio de diversos algoritmos, retardando o processo (Aghabozorgi; Shirkhorshidi; Wah, 2015). Em geral, os métodos de mineração de dados exigem altos custos computacionais ao serem aplicados em conjuntos muito grandes de dados (Krawczak; Szkatuła, 2014).

Para contornar esse problema, técnicas de redução de dimensionalidade de dados através de extração de características e seleção de características se fazem necessárias (Bolón-Canedo; Sánchez-Maroño; Alonso-Betanzos, 2015). Existem dois principais desafios ao executar análises em séries temporais, sendo eles: selecionar um conjunto de características que proporcione uma representação apropriada e selecionar uma medida adequada de dissimilaridade entre as mesmas (Wang *et al.*, 2013). Nesse cenário, a seleção de características após a extração se mostra eficiente, selecionando o conjunto de características

mais representativo, fazendo com que os algoritmos de aprendizado de máquina sejam mais rápidos e até mesmo produzam resultados melhores do que com os dados originais, especialmente algoritmos de agrupamento (Mörchen, 2003; Fulcher; Jones, 2014).

Este estudo fundamentou-se na técnica de extração e seleção de características proposta por Christ, Kempa-Liehr & Feindt (2016), chamada FRESH (*FeatuRe Extraction based on Scalable Hypothesis tests*), e consiste na extração de características baseada em testes de hipóteses escaláveis. Este método foi escolhido pois com ele é possível extrair diversos tipos de características e selecioná-las de acordo com uma determinada série temporal, sendo adequado quando não se sabe as propriedades que formam o comportamento da mesma. A seção a seguir é utilizada para descrever a proposta.

3.2.1.1 *FRESH*

O algoritmo caracteriza séries temporais com um mapeamento de características compreensivas e bem conhecidas. Em seguida, cada característica é individualmente avaliada considerando sua significância para prever o alvo em investigação. Como resultado, tem-se um vetor de p-valores que mensuram a significância de cada característica, que é então avaliado com base no procedimento de Benjamini & Yekutieli (2001) para decidir quais características devem ser mantidas. No Apêndice A é feito uma listagem das características levadas em consideração neste estudo.

Geralmente, séries temporais apresentam ruído e redundância, então, é importante manter o equilíbrio entre extrair características relevantes, porém sensíveis, e características robustas, mas que não são significativas (Christ; Kempa-Liehr; Feindt, 2016). Por exemplo, características como a mediana não são tão influenciadas por valores de *outlier*, já outras como o valor de máximo são particularmente sensíveis.

No trabalho de Radivojac *et al.* (2004), a relevância de uma característica X para um alvo Y é calculado como a diferença entre a distribuição de probabilidade condicional de X dado que $Y = y_1$, expressado por $f_{X|Y=y_1}$, e de X dado que $Y = y_2$, expressado por $f_{X|Y=y_2}$. Logo, a característica X será relevante para a predição de Y se, e somente se

$$\exists y_1, y_2 \quad \text{com} \quad f_Y(y_1) > 0, f_Y(y_2) > 0 : f_{X|Y=y_1} \neq f_{X|Y=y_2} \tag{3.2}$$

A Equação 3.2 é correspondente a dizer que X e Y são estatisticamente dependentes. A característica X é irrelevante quando contrário:

$$\exists y_1, y_2 \quad \text{com} \quad f_Y(y_1) > 0, f_Y(y_2) > 0 : f_{X|Y=y_1} = f_{X|Y=y_2} \tag{3.3}$$

A Equação 3.3 é correspondente a dizer que X e Y são estatisticamente independentes. Existem várias maneiras de se verificar se uma certa característica se encaixa nessas definições. Christ, Kempa-Liehr & Feindt (2016) investiga a relevância através de teste de hipóteses, um procedimento estatístico que permite tomar uma decisão entre duas hipóteses (Magalhães; Lima, 2000).

Para cada característica extraída $X_1, X_2, \ldots, X_n$ é aplicado um único teste de hipóteses, de forma independente, a fim de investigar as seguintes hipóteses:

$$\begin{aligned} H_0^i &= \{X_i \quad \text{não é relevante para} \quad Y\} \\ H_1^i &= \{X_i \quad \text{é relevante para} \quad Y\} \end{aligned} \tag{3.4}$$

O resultado de cada teste feito é o chamado p-valor, que é a probabilidade de se obter uma estatística de teste igual ou mais extrema que aquela observada em uma amostra, sob a hipótese nula. Neste trabalho, o p-valor é responsável por mensurar a probabilidade de uma certa característica de ser, ou não, relevante. Pequenos p-valores indicam maior relevância da característica testada.

Para este estudo, foi aplicado o teste de Kolmogorov–Smirnov (KS), que é um teste não paramétrico sobre a igualdade da função de distribuição acumulada (FDA) (Wilcox, 2005), resultando no teste das seguintes hipóteses:

$$\begin{aligned} H_0^i &= \left\{ f_{X_i|Y=y_1} = f_{X_i|Y=y_2} \right\} \\ H_1^i &= \left\{ f_{X_i|Y=y_1} \neq f_{X_i|Y=y_2} \right\} \end{aligned} \tag{3.5}$$

Em que $f_{X_i|Y=y_1}$ é a FDA da característica X_i para os dados de operação saudável e $f_{X_i|Y=y_2}$ é a FDA de X_i para os dados de operação de falha. Nota-se que as hipóteses da Equação 3.5 são equivalentes às da Equação 3.4. O teste KS considera a máxima diferença absoluta entre as FDAs das características, conforme descrito na Equação 3.6 e ilustrado na Figura 12.

$$D = \sup \left| f_{X_i|Y=y_1} - f_{X_i|Y=y_2} \right| \tag{3.6}$$

Na figura é possível ver a diferença considerada entre dois exemplos de FDA. Então, a hipótese nula H_0^i da Equação 3.5 é rejeitada se $D > D_{n,\alpha}$, em que $D_{n,\alpha}$ é um valor de ponto crítico que pode ser consultado na Tabela 5 do Apêndice B.

Entretanto, apesar dos testes feitos, existe a possibilidade de que características não relevantes sejam consideradas, ou seja, a hipótese H_0^i é rejeitada mesmo ela sendo verdadeira, conhecido como erro do tipo I, também chamado de falso positivo. Quando comparado múltiplas hipóteses e características, esses erros tendem a acumular (Curran-Everett, 2000). Benjamini & Hochberg (1995) então propôs controlar a FDR (*False*

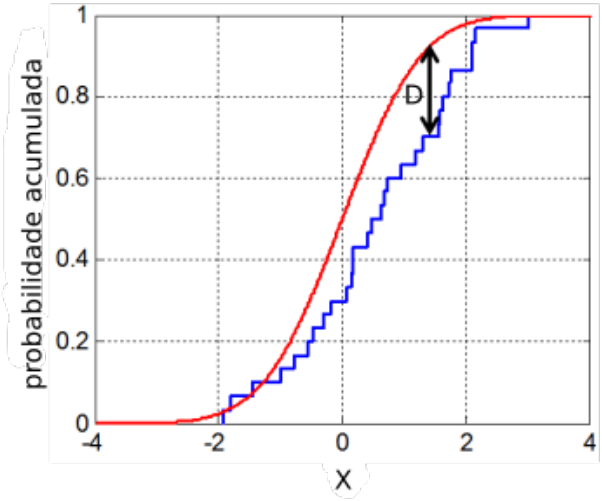


Figura 12 – Máxima diferença entre duas distribuições de probabilidade acumulada.

Discovery Rate), que é definida como a proporção de hipóteses nulas H_0 verdadeiras entre as hipóteses nulas rejeitadas e expressada pela Equação 3.7.

$$FDR = \frac{R}{V} \tag{3.7}$$

Onde, R é o número total de hipóteses rejeitadas e V é o número de hipóteses nulas verdadeiras rejeitadas.

Posteriormente o procedimento de Benjamini & Yekutieli (2001) foi proposto, em que as hipóteses eram rejeitadas baseando-se em seus p-valores enquanto a taxa FDR era controlada. O procedimento procura pela primeira interseção entre a sequência de p-valores $p_{(i)}$ ordenados e uma sequência linear expressa pela equação abaixo:

$$r_i = \frac{iq}{n \sum_{k=1}^{i} \frac{1}{k}} \tag{3.8}$$

Em que, n é o total de testes de hipótese feitos e q é valor de FDR que deseja-se controlar. A Figura 13 mostra como é aplicado o procedimento de Benjamini & Yekutieli (2001).

Foram simulados 216 p-valores, um para cada característica testada, e uma linha de rejeição foi traçada, levando em consideração manter um nível de $FDR = 10\%$. As características que possuem p-valores abaixo da interseção com a linha de rejeição são então selecionadas, rejeitando-se as demais, reduzindo de 216 para 31 características.

3.2.2 **Processamento dos Dados**

Na maioria das vezes, um conjunto de dados, devido a sua grande dimensionalidade, pode ser representado por características que possuem uma alta correlação entre si e

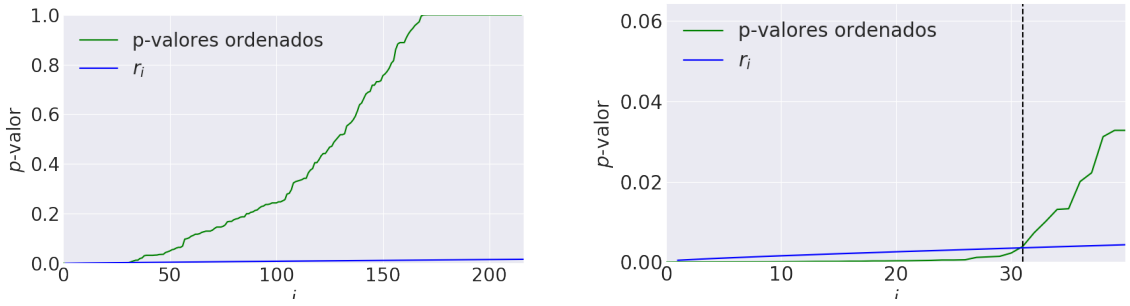


Figura 13 – Procedimento de Benjamini & Yekutieli para uma simulação de 216 p-valores e um nível FDR q de 10%.

Fonte: modificado de (Christ; Kempa-Liehr; Feindt, 2016)

também variáveis que não são significantes. Consequentemente, alguns dados possuem uma dificuldade de serem visualizados, dificultando a análise dos mesmos.

Entretanto, a aplicação de técnicas de visualização pode possibilitar ao usuário um entendimento mais robusto de algum evento. Possíveis correlações entre as variáveis podem ser evidenciadas, abrindo novas possibilidades de investigação. Neste trabalho empregou-se o método de Análise de Componentes Principais e de Indicador de Dano, possibilitando uma melhor visualização dos eventos, visando auxiliar na análise e interpretação dos resultados e são explicados a seguir.

3.2.2.1 *Análise de Componentes Principais*

A Análise de Componentes Principais (PCA - *Principal Component Analysis*) foi elaborada pela primeira vez por Pearson (1901). É um método muito utilizado para redução de dimensionalidade dos dados, eliminando redundâncias entres as características através de combinações lineares das variáveis originais.

Suponha um conjunto de variáveis $X = x_i$, $i = 1, 2, \ldots, n$. A não ser que X seja de baixa complexidade ou que n seja pequeno, na maioria das vezes, simplesmente visualizar esse conjunto de dados não é de grande ajuda, principalmente por causa da correlação entre as variáveis. Uma possibilidade então, seria investigar um conjunto menor de variáveis, derivado das variáveis originais, que preserve a maioria das informações dada pela variância dos dados. Esse novo conjunto de variáveis pode ser encontrado através de uma combinação linear $\boldsymbol{\alpha_k X}$ e que possua máxima variância. Essas novas variáveis são chamadas de componentes principais. A primeira componente principal, z_1, deve possuir a maior variância possível, ou seja, é responsável pela maior variabilidade dos dados, e é

definida pela seguinte combinação linear:

$$z_1 = \alpha_{11}x_1 + \alpha_{12}x_2 + \ldots + \alpha_{1n}x_n = \sum_{k=1}^{n} \alpha_{1k}x_k \tag{3.9}$$

Próximo passo é calcular a segunda componente principal, z_2, que possui a segunda maior variância e deve ser não correlacionada com a primeira componente. É definida pela seguinte equação:

$$z_2 = \alpha_{21}x_1 + \alpha_{22}x_2 + \ldots + \alpha_{2n}x_n = \sum_{k=1}^{n} \alpha_{2k}x_k \tag{3.10}$$

As seguintes componentes, então, são calculadas de forma análoga, de modo que, a j-ésima componente principal é dada pela combinação linear da Equação 3.11 e seja não correlacionada com as demais componentes (Jolliffe, 1986).

$$z_j = \alpha_{j1}x_1 + \alpha_{j2}x_2 + \ldots + \alpha_{jn}x_n = \sum_{k=1}^{n} \alpha_{jk}x_k \tag{3.11}$$

É um dos métodos estatísticos de múltiplas variáveis mais simples e conhecido, sendo muito utilizado como uma técnica de reconhecimento de padrões por algum tempo, obtendo excelentes resultados (Tibaduiza *et al.*, 2016).

PCA pode ser utilizada para visualização de dados complexos, analisando o conjunto de dados que contém as observações e que são normalmente descritas por variáveis dependentes e correlacionadas. A intenção é extrair os padrões a partir dos dados e exibir essa informação como um conjunto novo e ortogonal de variáveis, chamado de componentes principais, que podem ser mostradas como pontos (Abdi; Williams, 2010). Essas novas variáveis são combinações linear das variáveis originais e são independentes entre si. A Figura 14 apresenta um exemplo de aplicação de PCA.

3.2.2.2 *k-means*

O uso de modelos de aprendizado de máquina não supervisionado neste trabalho é devido a sua capacidade de fornecer uma classificação a partir de informações obtidas dos próprios dados. Assim, o algoritmo é capaz de fornecer uma classificação de forma automática, sem a necessidade de nenhuma pré-classificação existente, sem uma supervisão.

Os métodos mais utilizados em aprendizado de máquina não supervisionado são os métodos de agrupamento (*clustering*), que têm como objetivo a classificação de amostras em um espaço multidimensional de acordo com suas características intrínsecas, resultando em classes ou grupos (*clusters*) bem definidos (Jain; Dubes, 1988; Duda; Hart; Stork, 2000), de maneira a:

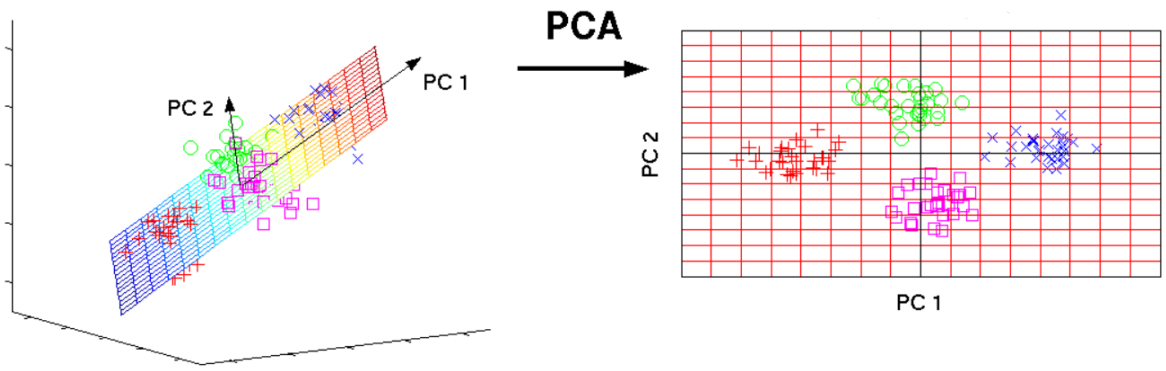


Figura 14 – Exemplo de PCA.

Fonte: modificado de (Scholz, 2006)

- Maximizar a similaridade entre as amostras de um mesmo grupo.
- Minimizar a similaridade entre as amostras de grupos distintos.

Na literatura é possível encontrar diversos modelos utilizados para agrupamento, como os algoritmos de agrupamento hierárquico (HC - *Hierarchical Clustering*), subdivididos em agrupamento por divisão (Kaufman; Rousseeuw, 2009) e agrupamento aglomerativo (Sneath, 1957). Algoritmos de agrupamento baseados em medidas de distância e similaridades, como o *Fuzzy c-means* (Hathaway; Bezdek; Hu, 2000) e o algoritmo *k-means* (MacQueen *et al.*, 1967). Agrupamento baseado na função de densidade de probabilidade (PDF - *Probability Density Function*), como o modelo de misturas de gaussianas (Zhuang *et al.*, 1996), entre outros. Escolheu-se o algoritmo de *k-means* por ser um dos mais conhecidos, muito simples, de fácil implementação e de baixa complexidade (Xu; Wunsch, 2005) e será explicado a seguir.

O algoritmo de agrupamento *k-means* é descrito em detalhes por Hartigan (1975). O objetivo é dividir M pontos com N dimensões em k diferentes grupos, minimizando a distância entre os pontos de um mesmo grupo, assim, maximizando a similaridade. O primeiro passo é definir os k diferentes centros, um para cada grupo, e então associar as amostras ao centro com o qual possua menor distância, consequentemente, ao seu respectivo grupo. Após, um novo centro é calculado para cada grupo e o primeiro passo é, então, repetido. Um laço então é formado, repetindo o procedimento até que os centros não sejam mais alterados. Uma síntese do algoritmo de *k-means* pode ser visto no Algoritmo 1.

A Figura 15 apresenta um exemplo de agrupamento utilizando *k-means*, em que um conjunto de dados foi separado em dois grupos, $k = 2$, representados pelas cores vermelho e azul.

Algoritmo 1: algoritmo *k-means*.

Entrada: (k) número de grupos, $X = (x_1, x_2, \ldots, x_n)$ dados a serem agrupados
Saída: $(G = (g_1, g_2, \ldots, g_k))$ Conjunto de grupos

```
1  início
2      para cada g_i em G faça
3          g_i ← x_j em X #seleção aleatória
4      fim para cada
5      para i = 1 até n faça
6          para j = 1 até k faça
7              distancia[x_i] ← min(x_i, g_j)
8          fim para
9      fim para
10     mudou ← false
11     repita
12         mudou ← false
13         para cada g_i em G faça
14             atualizaGrupo(g_i)
15         fim para cada
16         para i = 1 até n faça
17             para j = 1 até k faça
18                 distanciaMin ← min(x_i, g_j)
19                 se distanciaMin ≠ distancia[x_i] então
20                     distancia[x_i] ← distanciaMin
21                     mudou ← true
22                 fim se
23             fim para
24         fim para
25     até mudou = false;
26 fim
```

3.2.2.3 *Escore de Homogeneidade*

O escore de homogeneidade é uma métrica de avaliação utilizada para analisar o desempenho de um algoritmo de agrupamento, e foi aplicada nesta dissertação com o objetivo de comparar os resultados obtidos pelo *k-means* para diferentes conjuntos de características. Um resultado de agrupamento satisfaz a homogeneidade se a distribuição da classe dentro de cada grupo esta direcionada para uma única classe, ou seja, entropia zero. A entropia mede a incerteza de uma variável. Avalia-se a proximidade de um agrupamento do seu ideal examinando a entropia condicional da distribuição da classe dado o agrupamento proposto.

Seja n o número total de pontos de dados, o conjunto de classes dado por $C = \{c_i | i = 1, \ldots, n\}$, o conjunto de grupos, $G = \{g_i | i = 1, \ldots, m\}$ e $A = \{a_{ij}\}$, a tabela de contingência produzida pelo algoritmo de agrupamento que representa a solução do

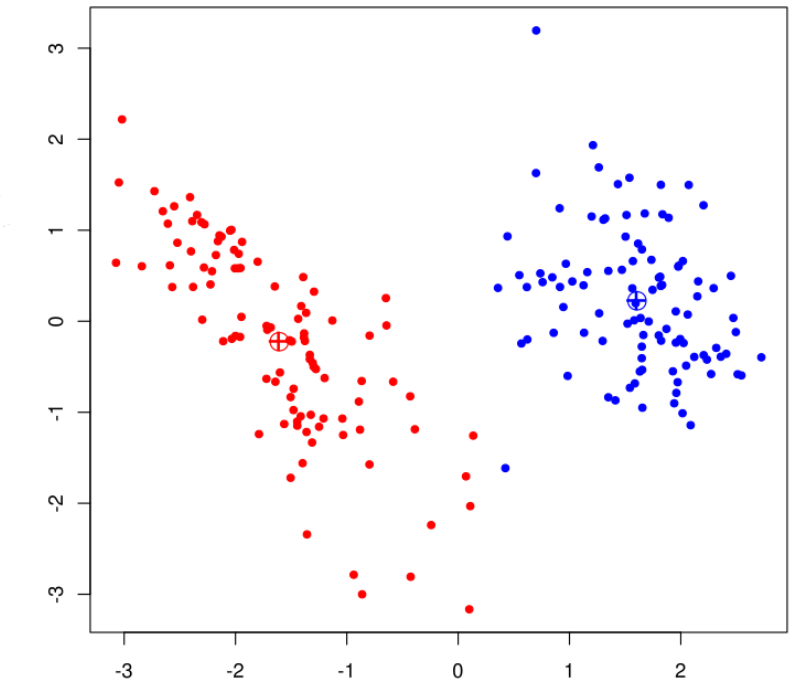


Figura 15 – Exemplo de *k-means* para $k = 2$.

agrupamento, de modo que a_{ij} é o número de pontos de dados que pertencem à classe c_i e elementos do grupo g_j. Um agrupamento é dito perfeitamente homogêneo quando sua entropia é $H(C|G) = 0$, pois nesse caso não há incerteza, visto que a probabilidade de que os elementos desse agrupamento pertença a uma única classe é igual a 1. Todavia, para uma situação diferente, esse valor depende do tamanho do conjunto de dados e da distribuição das classes, e para facilitar a análise da entropia condicional bruta, esse valor deve ser normalizado pela redução máxima na entropia que o agrupamento pode fornecer, $H(C)$. Em casos em que há somente uma classe, $H(C) = 0$, a homogeneidade é definida como 1. Para uma solução perfeitamente homogênea, esta normalização é $\frac{H(C|G)}{H(C)} = 0$, portanto, para se adaptar à convenção de que 1 é desejável e 0 é indesejável, a homogeneidade é definida conforme a Equação 3.12 (Rosenberg; Hirschberg, 2007):

$$h = \begin{cases} 1 & \text{se} \quad H(C,G) = 0 \\ 1 - \frac{H(C|G)}{H(C)} & \text{senão} \end{cases} \tag{3.12}$$

em que

$$H(C,G) = -\sum_{g=1}^{|G|}\sum_{c=1}^{|C|}\frac{a_{cg}}{n}\log\frac{a_{cg}}{\sum_{c=1}^{|C|}a_{cg}}$$

$$H(C) = -\sum_{c=1}^{|C|} \frac{\sum_{g=1}^{|G|} a_{cg}}{n} \log \frac{\sum_{g=1}^{|G|} a_{cg}}{n}$$

3.2.2.4 *Indicador de Dano*

Após encontrado os grupos resultantes dos algoritmos de agrupamento, o processo de identificação de defeito pode ser feito através de um indicador de dano (ID) calculado para cada observação. Esses indicadores são gerados baseando-se no método de detecção de *outliers* apresentado por Figueiredo *et al.* (2014).

Essencialmente, é calculada a distância Euclidiana entre um vetor de características X e o vetor C contendo os centroides de seus respectivos grupos. O ID é então a menor distância encontrada:

$$ID(\phi) = \min\left(\|x_\phi - c_1\|, \|x_\phi - c_2\|, \ldots, \|x_\phi - c_k\|\right) \tag{3.13}$$

4 RESULTADOS E DISCUSSÃO

Este capítulo apresenta os resultados obtidos através da aplicação dos modelos desenvolvidos nessa dissertação. Os resultados foram comparados com informações dos dados originais a fim de identificar os padrões encontrados. Após a etapa de extração e seleção de características, foi obtido o conjunto de características que mais representam os sinais de operação das máquinas de chave. As características foram extraídas levando em consideração cada operação das máquinas de chave de forma separada, ou seja, independente umas das outras. Esse conjunto foi então utilizado como entrada para o algoritmo de aprendizado de máquina não supervisionado aqui aplicado. Objetivando-se a verificar a generalização da proposta, os dados previamente classificados como operação de falha foram acrescidos de um ruído gaussiano, gerando novos dados nos quais a metodologia foi aplicada.

Os resultados são apresentados e discutidos na Seção 4.1 a seguir. Uma análise de componentes principais (PCA) foi realizada com o intuito de reduzir a dimensionalidade dos dados e encontrar padrões que facilitassem a sua visualização. Em seguida, foi empregado o algoritmo de agrupamento *k-means* buscando encontrar grupos que diferenciasse operações saudáveis e operações de falha. Posteriormente, dados sintéticos foram gerados, visando aumentar a quantidade de dados disponíveis e entender melhor o comportamento de um sinal de operação de falha. Então, de forma a quantificar o dano e identificar o comportamento dos dados, os IDs foram estimados. Por último, na Seção 4.2, foi feita uma análise em cima das características selecionadas pelo FRESH por meio do PCA, tendo como finalidade verificar a importância das mesmas para os resultados obtidos. O escore de homogeneidade foi utilizado como métrica de avaliação na comparação entre os conjuntos de características.

4.1 RESULTADOS COMPUTACIONAIS

Como primeiro passo após a extração e seleção de características, empregou-se o método de PCA para redução de dimensionalidade a fim de melhorar a visualização e investigar a disposição dos mesmos. O resultado pode ser visto na Figura 16, em que no eixo das abcissas tem-se a primeira componente principal e de maior variância e no eixo das ordenadas tem-se a segunda componente principal, de segunda maior variância e ortogonal à primeira componente. Por meio da figura é possível observar a qual máquina cada sinal pertence. Cada máquina é representada por uma cor diferente e os pontos com a borda em negrito são os dados que representam a ocorrência de falha na máquina de chave.

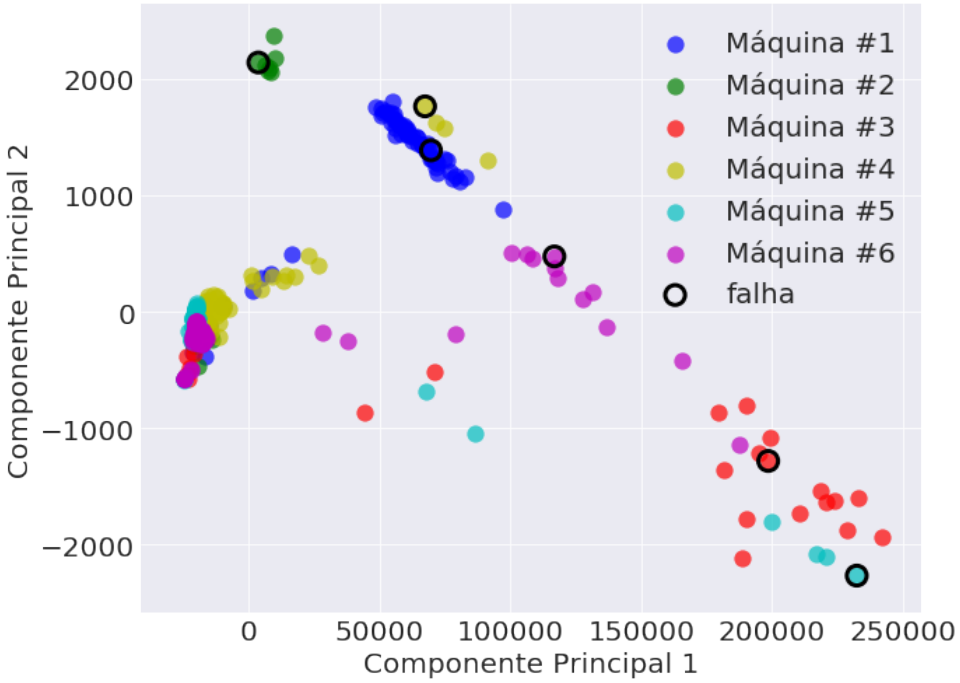
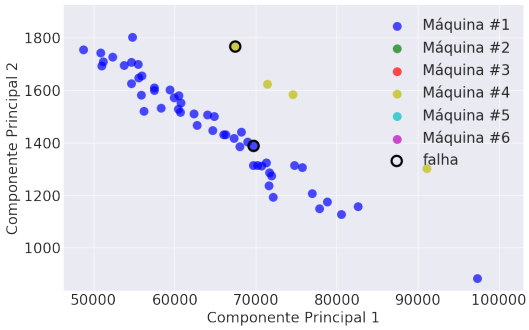


Figura 16 – Dados separados por máquinas de origem.

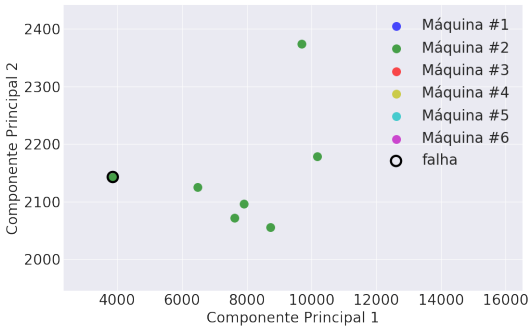
Nota-se que a maioria dos sinais se concentram do lado esquerdo central da figura, quase que sobrepostos entre si. Percebe-se também que nessa região estão somente os sinais pertinentes à uma operação saudável da máquina de chave. É possível ver que o sinais de pertinente à falha estão mais distantes da região citada anteriormente, e que também estão distantes entre si, com exceção das máquinas #1 e #4. Fato que pode ter ocorrido devido às máquinas estarem instaladas em diferentes localidades, ou também apresentarem falhas distintas. Percebe-se também, com a disposição dos dados na figura, que atributos os quais explicam e diferenciam os sentidos de operação entre normal e reverso não são levados em consideração durante a seleção de características, conforme esperado, visto que a detecção de falha deve ser independente do sentido em que a máquina está operando. Na Figura 17 pode ser visto uma aproximação da região em que não há sinal de falha e também das regiões em que há um sinal de falha para cada máquina.

Através das figuras foi possível verificar que a extração de características funcionou como esperado, selecionando aquelas que permitiram ver claramente uma diferença entre os sinais observados para operação saudável e de falha. Uma observação interessante é o fato de haver sinais que foram considerados saudáveis no entorno dos sinais de falha, e que apresentam data e hora próximos aos observados quando da falha. Esses sinais apresentam um comportamento análogo ao de falha, indicando que essa falha poderia ter sido identificada previamente.

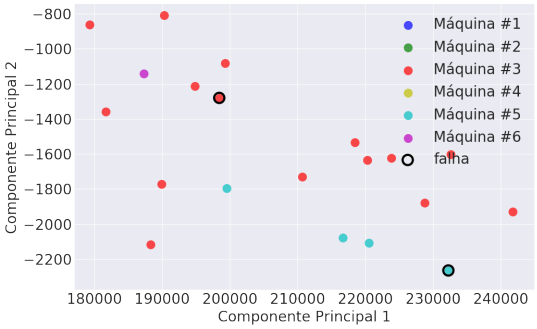
Em uma próxima etapa, foi empregado o algoritmo de agrupamento *k-means* sobre



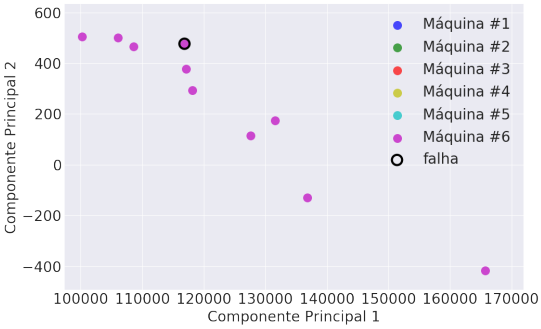
(a) Máquinas #1 e #4



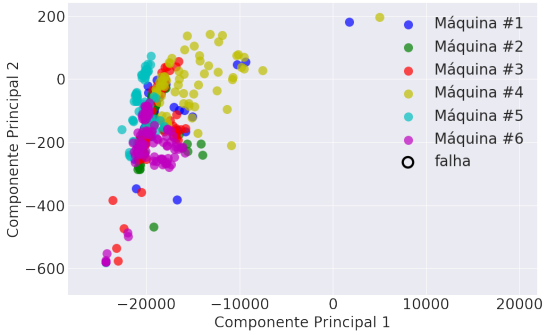
(b) Máquina #2



(c) Máquina #3 e #5



(d) Máquina #6



(e) Todas as máquinas

Figura 17 – Aproximação realizada na Figura 16.

as características selecionadas, tendo como objetivo encontrar grupos (*clusters*) com elementos que possuam maior similaridades entre si e verificar o comportamento mostrado na Figura 16. Novamente o PCA com duas componentes principais foi empregado com o intuito de facilitar a visualização dos dados. A Figura 18 mostra o resultado encontrado. Cada grupo encontrado é identificado por uma cor diferente e os pontos com borda em negrito são dados relacionados à ocorrência de uma falha.

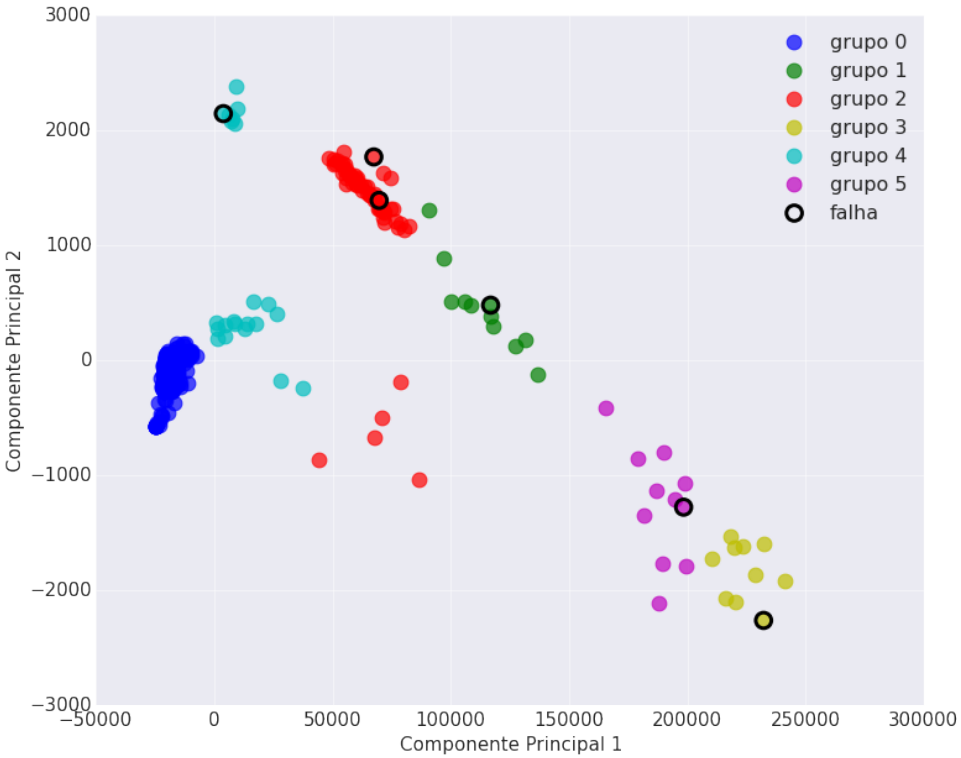


Figura 18 – Resultado da aplicação do *k-means* para $k = 6$.

Os grupos encontrados pelo *k-means* compreendem ao comportamento verificado anteriormente. O grupo 0 indica os sinais saudáveis das máquinas de chave, visto que não há nenhum ponto com borda em negrito em sua região. Nos demais grupos encontrados é possível ver a presença de sinais de falha, pontos com borda em negrito, podendo representar uma região crítica. Nessas regiões ocorrem uma falha ou há uma situação de falha iminente, devido a presença de alguns sinais previamente classificados como saudáveis estarem situados ali.

Objetivando-se entender melhor o comportamento de um sinal de operação de falha, e devido ao fato de se ter um número pequeno de dados para este tipo de sinal, sendo um para cada máquina, novos dados foram gerados adicionando-se um ruído aos originais. O ruído branco aditivo gaussiano (AWGN - *Additive White Gaussian Noise*) é um ruído estatístico cuja função densidade de probabilidade (FDP) é igual a da distribuição normal, que é também conhecida como distribuição gaussiana (Barbu, 2013).

Para esta etapa, foi utilizado a relação sinal-ruído, também conhecida como SNR (*Signal-to-Noise Ratio*), que compara o nível de um sinal desejado com o nível do ruído de fundo e pode ser expressa pela Equação 4.1. Quanto maior o SNR, menor é o efeito do ruído sobre o sinal.

$$SNR = \frac{P_{sinal}}{P_{ruido}} \tag{4.1}$$

Por muitas razões a relação sinal-ruído possui uma faixa dinâmica, sendo os sinais geralmente expressos usando escala logarítmica de decibel, conforme mostra a Equação 4.2.

$$SNR_{dB} = 10\log_{10}\left(\frac{P_{sinal}}{P_{ruido}}\right) \tag{4.2}$$

Fazendo algumas substituições na Equação 4.2, o ruído, então, pode ser calculado de acordo com o SNR desejado utilizando-se a Equação 4.3.

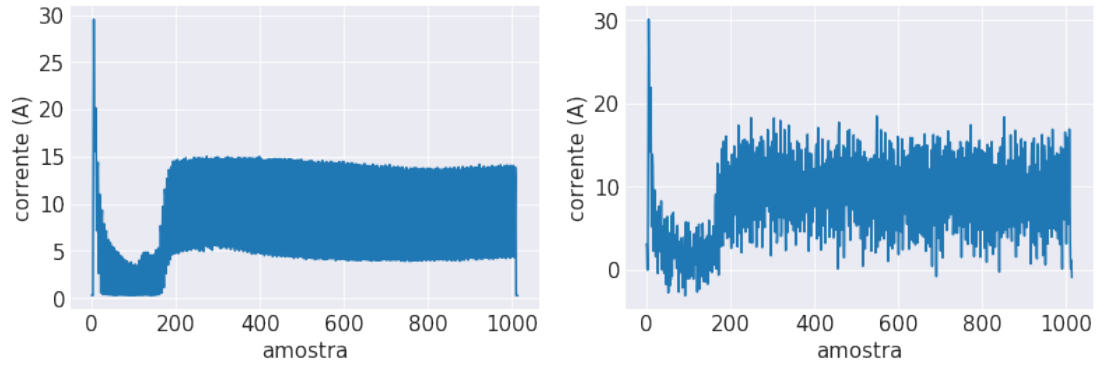
$$P_{ruido} = \frac{P_{sinal}}{10^{\left(\frac{SNR_{dB}}{10}\right)}} \tag{4.3}$$

A partir da Equação 4.3, os ruídos desejado foram caluculados e adicionados aos sinais originais para operação de falha. Foram gerados cinco novos sinais com três valores distintos de SNR para cada máquina, contabilizando um total de 90 novos sinais. A Figura 19 apresenta um exemplo de como é o sinal original e os sinais gerados.

Para analisar o comportamento dos sinais de operação de falha, os dados foram separados por máquina de origem, e os IDs foram utilizados como métrica de avaliação e estimados levando-se em consideração o centroide do grupo encontrado pelo *k-means* e que compreende os sinais que representam uma operação saudável. A partir dessa informação, foram gerados gráficos de barras para cada máquina e que podem ser vistos na Figura 20.

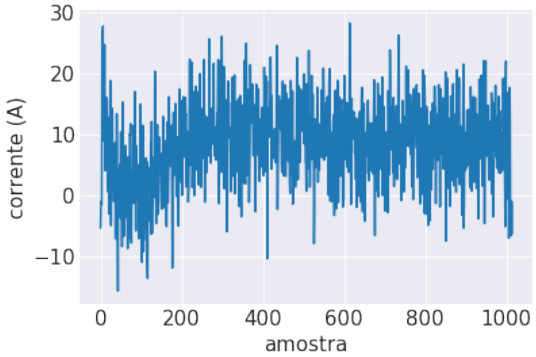
Os gráficos de barras são separados quanto a ocorrência de falha ou não e também pelo valor de SNR desejado para o ruído. Nota-se que todas as máquinas possuem comportamento parecido quanto aos valores dos IDs, apresentando valores pequenos para dados de operação saudável e valores altos para operação de falha. Também é possível notar que os sinais gerados com um ruído pequeno (SNR de 15dB) ficam no entorno do sinal de falha e possuem IDs similares. Quanto maior o ruído, e consequentemente menor SNR, maior é o valor do ID, conforme esperado, pois o sinal fica mais descaracterizado.

Como próximo passo, foram gerados gráficos com os IDs ordenados de maneira cronológica, na sequência em que as amostras foram coletadas, a fim de verificar o comportamento dos dados observados no entorno dos dados de operação de falha. Posteriormente, foi definido um limite linear correspondendo à um intervalo de confiança de 95% levando em consideração somente os dados de operação saudável, visando encontrar um limiar que definisse quando uma amostra tem comportamento análogo ao de uma operação saudável

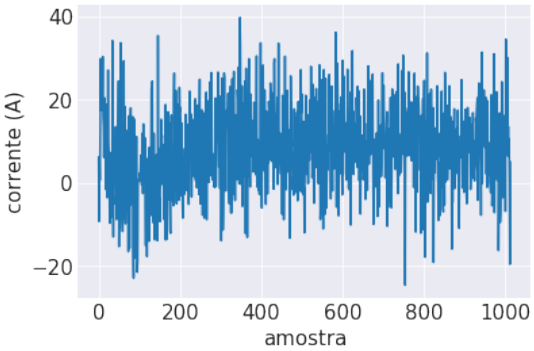


(a) Original

(b) SNR de 15dB



(c) SNR de 5dB

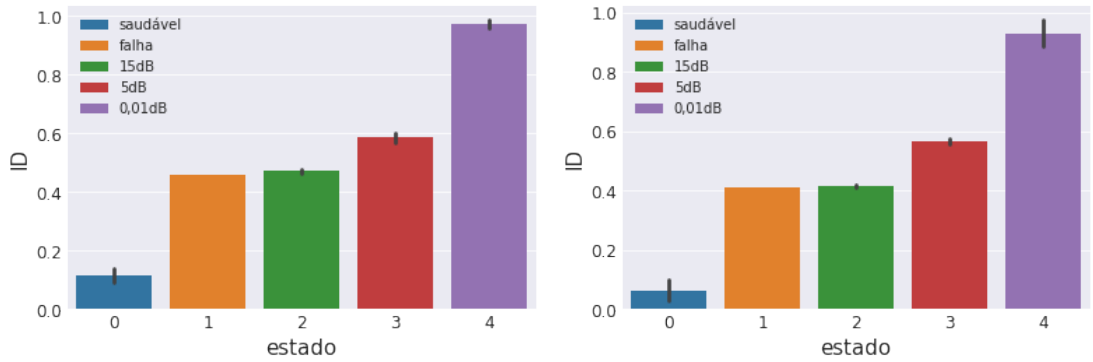


(d) SNR de 0,01dB

Figura 19 – Exemplo de sinal sem e com ruído.

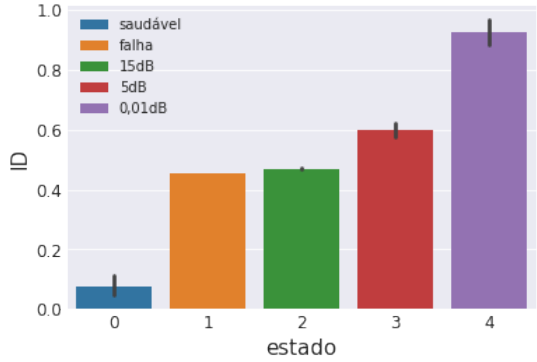
ou de uma operação de falha. Esses gráficos podem ser vistos na Figura 21, e os dados com a borda em negrito são os IDs pertinentes à ocorrência de uma falha.

É possível notar através da Figura 21 que a maioria dos IDs estão abaixo do limiar estimado, conforme esperado, visto que a maioria dos dados são de operações saudáveis. Os pontos com as bordas em negrito e que são as operações de falha estão posicionados acima desse limiar, também como era aguardado. Esse comportamento pode ser visto para todas as máquinas de chave. Outro comportamento verificado, são alguns pontos que estão acima do limiar em um instante, abaixo do limiar no próximo e voltando a estar acima no instante seguinte. Esse evento pode indicar que a falha pode estar em um sentido de operação da máquina, operando normalmente para o sentido oposto. Os pontos de operações saudáveis que estão acima do limiar são os mesmos pontos que estavam situados no entorno dos pontos de falha na Figura 16, e que também foram incluídos nos grupos relacionados com as falhas pelo *k-means*. Alguns desses sinais foram observados momentos antes à da falha e o alto valor do ID indica que a máquina de chave vinha operando em uma situação crítica, podendo alguma ação ter sido tomada para evitar a ocorrência da falha. Ainda aproveitando as informações disponíveis com os IDs, um mapa de calor foi criado e apresentado na Figura 22. Utilizou-se novamente as componentes principais para

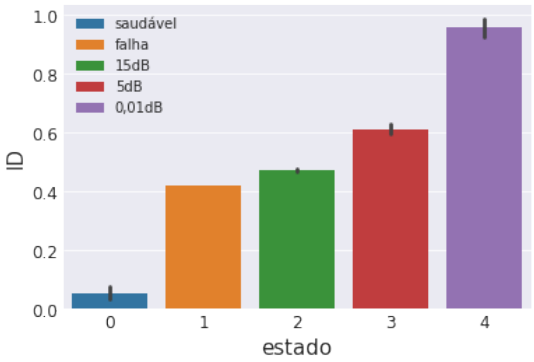


(a) Máquina #1

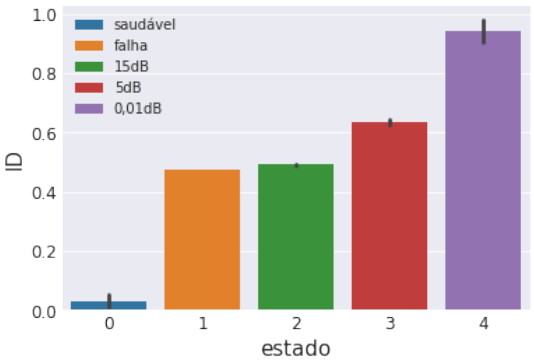
(b) Máquina #2



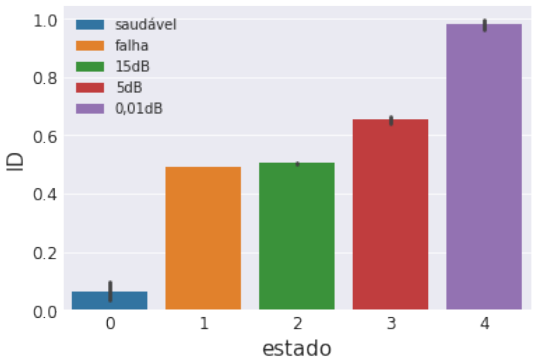
(c) Máquina #3



(d) Máquina #4



(e) Máquina #5

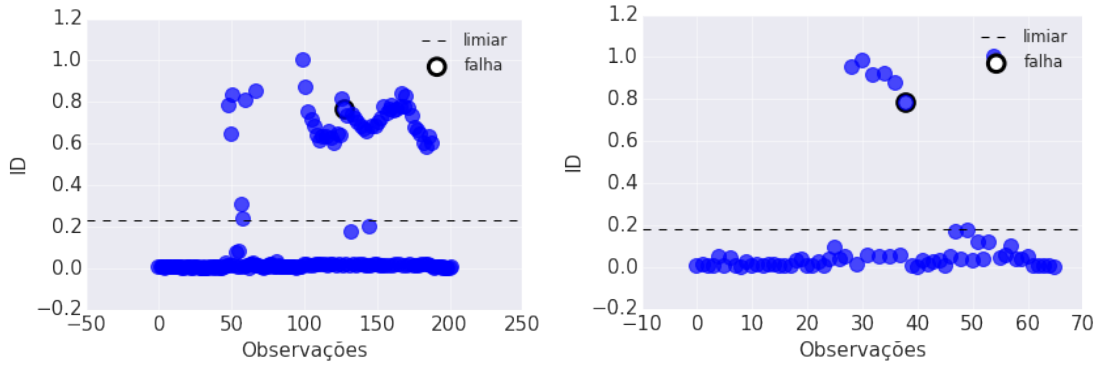


(f) Máquina #6

Figura 20 – Gráfico de caixas com os IDs para cada máquina.

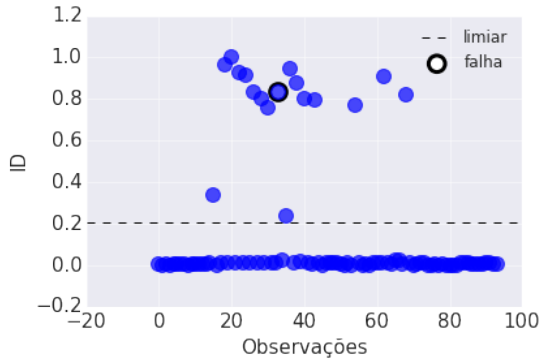
redução de dimensionalidade e exibição do resultado. Os dados de operação saudáveis, definidos pelos IDs mais baixos, são representados pelos pontos em azul e de acordo com que os IDs vão aumentando, o que representa um comportamento análogo à uma operação de falha, esses pontos vão sendo representados por cores mais quentes, como amarelo e vermelho.

É possível ver na figura, uma região que compreende à operações saudáveis para as máquinas de chave, caracterizada pelos pontos em tom de azul, indicando um baixo valor de ID. Conforme os pontos vão se afastando dessa região, o risco de falha vai aumentando.

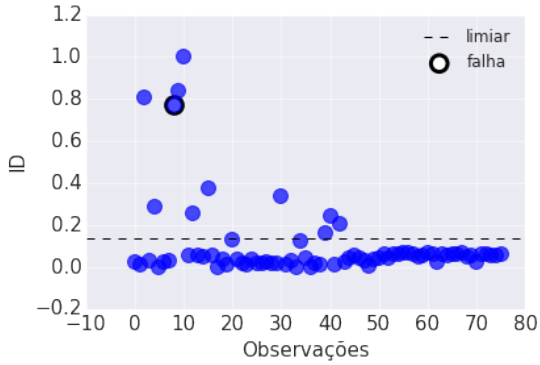


(a) Máquina #1

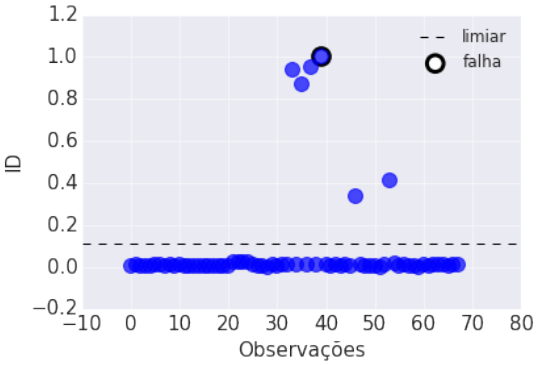
(b) Máquina #2



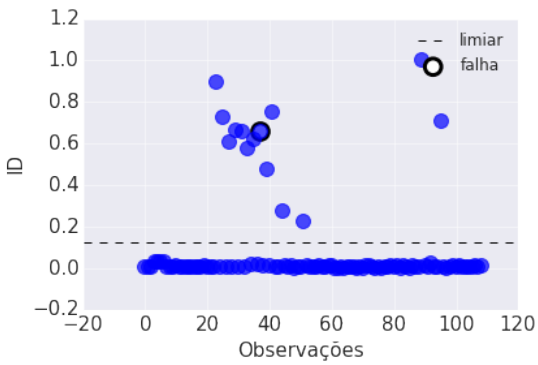
(c) Máquina #3



(d) Máquina #4



(e) Máquina #5



(f) Máquina #6

Figura 21 – IDs estimados a partir do algoritmo de agrupamento *k-means*

Essa análise pode auxiliar um profissional da área a monitorar o estado de operação da máquina, permitindo-o tomar decisões de maneira antecipada referente à manutenção dos equipamentos.

4.2 ANÁLISE DAS CARACTERÍSTICAS

Por intermédio do PCA, foi possível verificar a relevância das características selecionadas para analisar as séries temporais no cálculo das componentes principais. Essa relevância é dada pelas constantes α_{jk} da Equação 3.11, e que atuam como pesos indicando

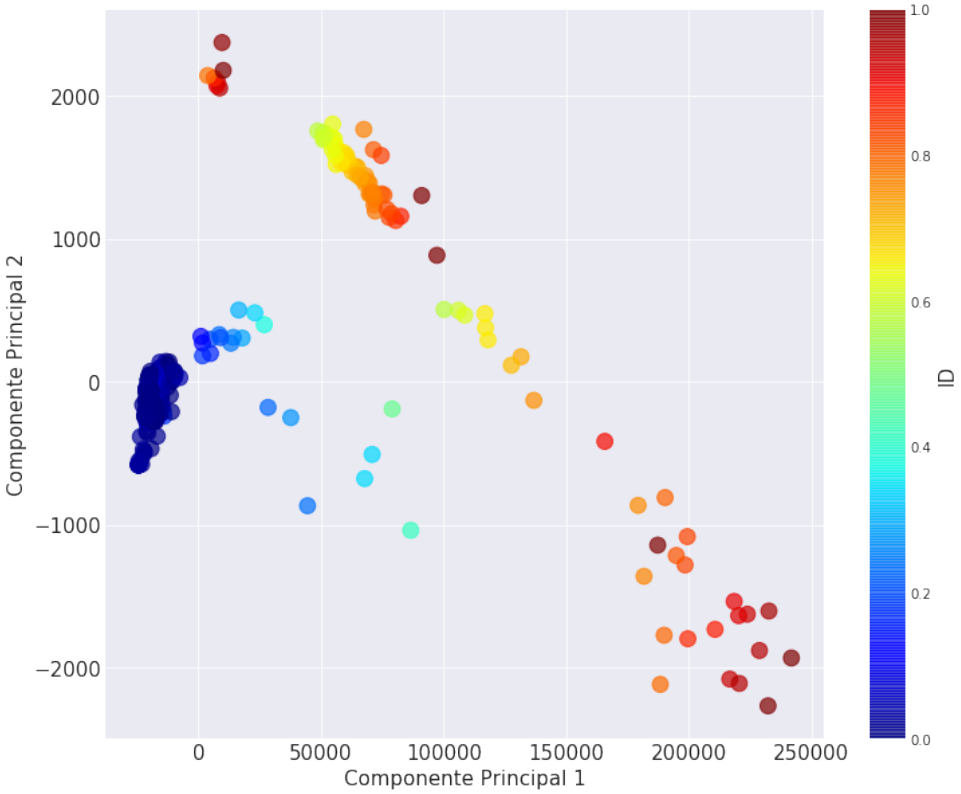


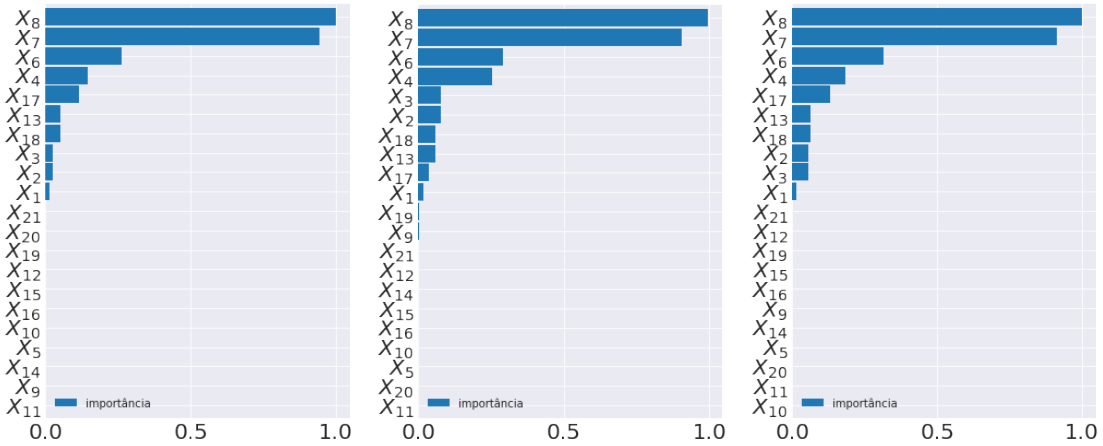
Figura 22 – Mapa de calor para os IDs estimados.

a importância de cada característica no cálculo das componentes principais. A Figura 23 foi criada, de modo que, quanto maior a barra, maior a relevância da respectiva característica. A descrição de cada característica pode ser vista na Tabela 4.

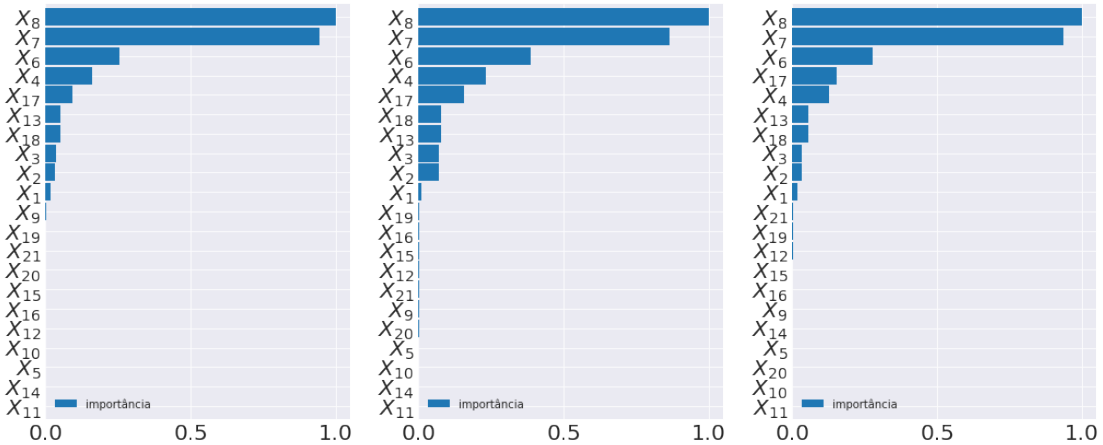
De acordo com a Figura 23, a maioria das características extraídas possuem nenhuma relevância para as componentes principais calculadas, e consequentemente para identificação do estado do sinal. Quanto as características que possuem alguma relevância, nota-se que são as mesmas para todas as máquinas, porém são diferentes em ordem de importância, com exceção para as três primeiras. Como as máquinas são de mesmo modelo, essa diferença pode ser devido alguma diferença na instalação das mesmas, como o local por exemplo. As três características mais importantes são as mesmas para todas as máquinas. A primeira é a energia absoluta, que é expressada pela Equação 4.4, sendo a soma dos valores absolutos ao quadrado da série temporal. A segunda é soma absoluta das mudanças, expressada pela Equação 4.5 e é o valor absoluto de mudanças consecutivas na série temporal. A terceira é o comprimento da série temporal.

$$E = \sum_{i=1,\ldots,n} |x_i|^2 \tag{4.4}$$

$$SM = \sum_{i=1,\ldots,n-1} |x_{i+1} - x_i| \tag{4.5}$$



(a) Máquina #1 (b) Máquina #2 (c) Máquina #3



(d) Máquina #4 (e) Máquina #5 (f) Máquina #6

Figura 23 – Gráficos de barras com a relevância das característica para cada MC.

A fim de verificar se a partir de somente essas três características é possível identificar um comportamento de falha para os dados, a Figura 24 foi gerada. A análise foi feita para cada máquina, os valores são as médias dos valores encontrados para cada característica e eles são separados pela ocorrência ou não de falha.

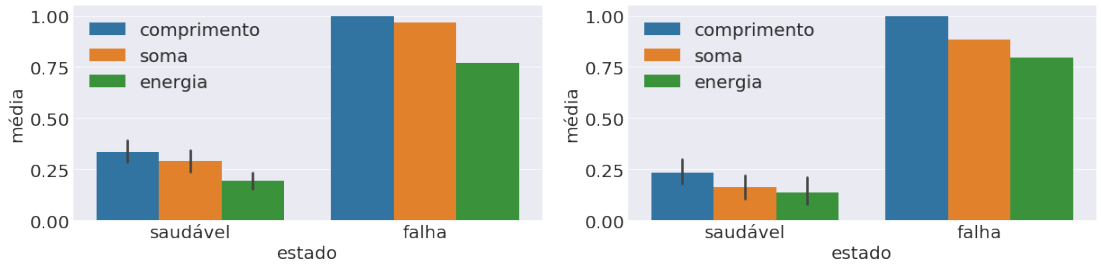
Foi verificado um comportamento semelhante para todas as máquinas, em que houve um aumento nos valores medidos para as três características analisadas. Um valor maior para a energia absoluta pode ser devido um aumento na corrente de operação ou um aumento no tempo de operação da máquina, ambos causados por alguma falha, o que também poderia explicar o aumento no comprimento do sinal observado. Já para o somatório dos valores absolutos da mudança, o aumento pode estar relacionado com o fato de o sinal apresentar uma maior variação da corrente, ou seja, os sinais de uma operação de falha são menos uniformes que os sinais de operação saudável, devido a uma irregularidade na movimentação das agulhas pela máquina de chave. Por acreditar que as características de energia e comprimento estão altamente relacionadas entre si, excluiu-se

Tabela 4 – Conjunto de características selecionadas pelo FRESH

Características	Descrição
X_1	número de picos para 5 vizinhos de cada lado.
X_2	número de picos após suavização pela transformada wavelet de nível 1.
X_3	número de picos após suavização pela transformada wavelet de nível 5.
X_4	número de valores abaixo da média.
X_5	porcentagem de valores duplicados.
X_6	comprimento.
X_7	soma absoluta das mudanças consecutivas.
X_8	energia.
X_9	média da soma das mudanças consecutivas entre os quantis $0,0$ e $0,8$.
X_{10}	razão de valores duplicados.
X_{11}	coeficientes da equação de Friedrich.
X_{12}	média da soma das mudanças consecutivas entre os quantis $0,2$ e $0,8$.
X_{13}	número de picos para 3 vizinhos de cada lado.
X_{14}	autocorrelação.
X_{15}	média da soma das mudanças consecutivas entre os quantis $0,0$ e $1,0$.
X_{16}	média da soma das mudanças consecutivas.
X_{17}	número de valores acima da média.
X_{18}	número de picos para 1 vizinho de cada lado.
X_{19}	média da soma das mudanças consecutivas entre os quantis $0,2$ e $1,0$.
X_{20}	média da soma das mudanças consecutivas entre os quantis $0,2$ e $0,6$.
X_{21}	média da soma das mudanças consecutivas entre os quantis $0,4$ e $0,8$.

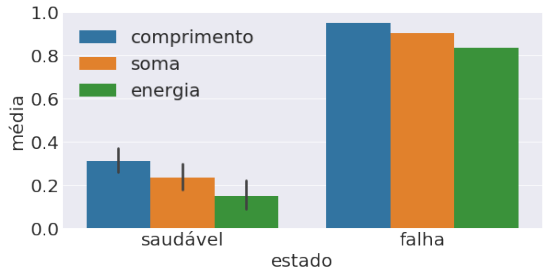
a última, e as duas restantes foram utilizadas como entrada do algoritmo de *k-means*. O *k-means* foi aplicado novamente com a intenção de investigar se, somente com essas duas características, é possível encontrar os mesmos grupos encontrados na Figura 18, e o resultado pode ser visto na Figura 25, na qual cada cor diferente representa um grupo encontrado, e os pontos com a borda em negrito são dados pertinentes à falha.

Verifica-se que as figuras são muito parecidas e os resultados encontrados são similares qualitativamente. De forma a avaliar essa semelhança quantitativamente, comparou-se as classificações resultantes do agrupamento para 21 e 2 características, respectivamente, através do escore de homogeneidade, que foi estimado através de 10 iterações para garantir

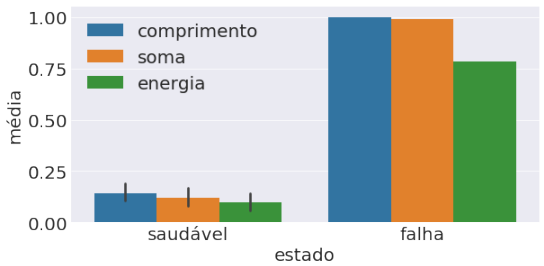


(a) Máquina #1

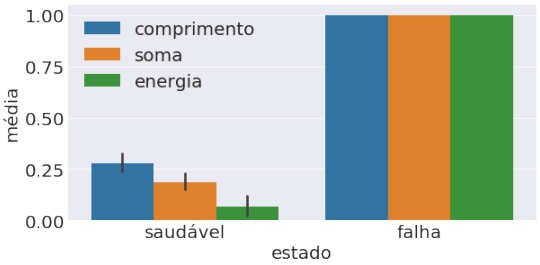
(b) Máquina #2



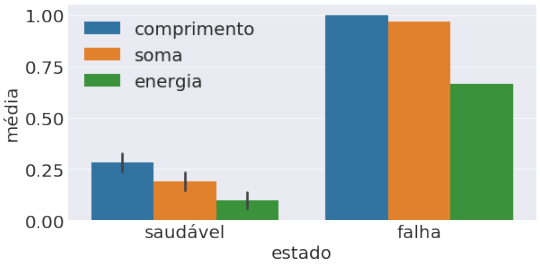
(c) Máquina #3



(d) Máquina #4



(e) Máquina #5



(f) Máquina #6

Figura 24 – Gráficos de barras que mostram a diferença no comportamento das características entre operações saudáveis e de falha.

a robustez do resultado. O escore encontrado foi de 1, que corrobora que utilizar somente as duas características selecionadas através do PCA apresenta o mesmo resultado de quando são utilizadas todas as características selecionadas pelo FRESH. Como próximo passo, utilizou somente as características de energia e soma da variação absoluta da corrente para investigar o comportamento dos dados. Extraiu-se a raiz quadrada da energia, visto que a mesma é uma função quadrática. Novamente, foi criado um mapa de calor, utilizando os IDs, visando auxiliar nas discussões. O resultado é apresentado na Figura 26.

É visível uma separação da figura em nove sub-regiões. Nota-se que no primeiro quadrante, localizado da esquerda pra direita e de baixo pra cima, encontra-se os dados de operação saudável, isto é, de baixo ID. É possível observar uma relação linear entre as variáveis e que os valores baixos indicam que, em uma operação saudável, as máquinas de chave usam menos energia e possuem uma menor variação da corrente. No quadrante do meio, essa relação linear começa a se perder, e um aumento na energia e na variação

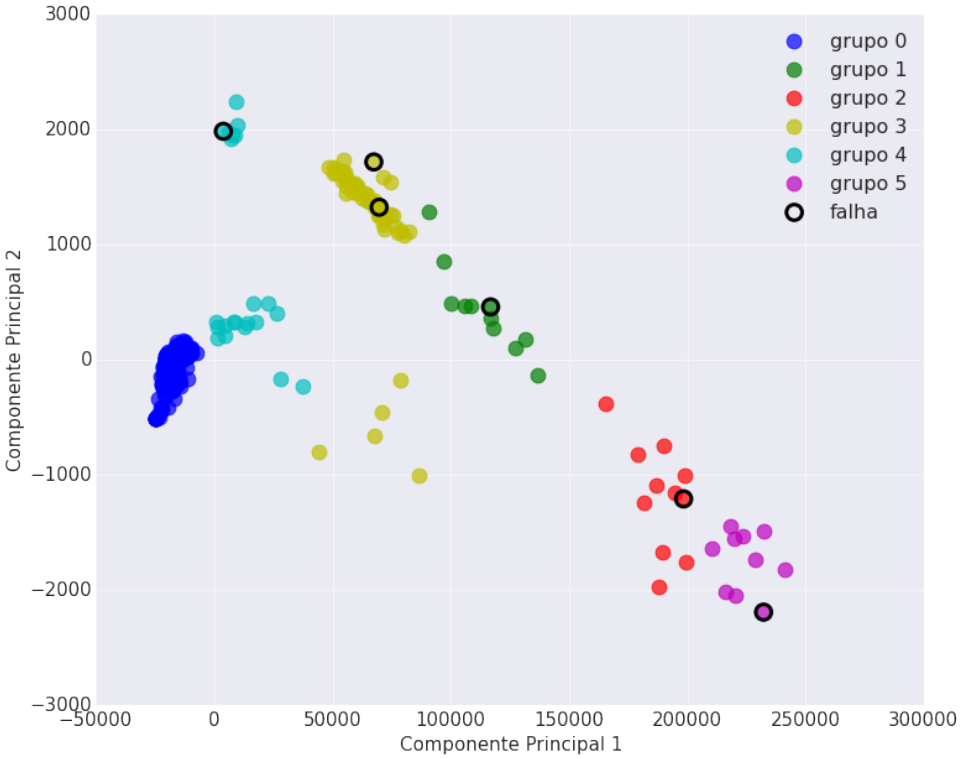


Figura 25 – Resultado da aplicação do *k-means* para $k = 6$.

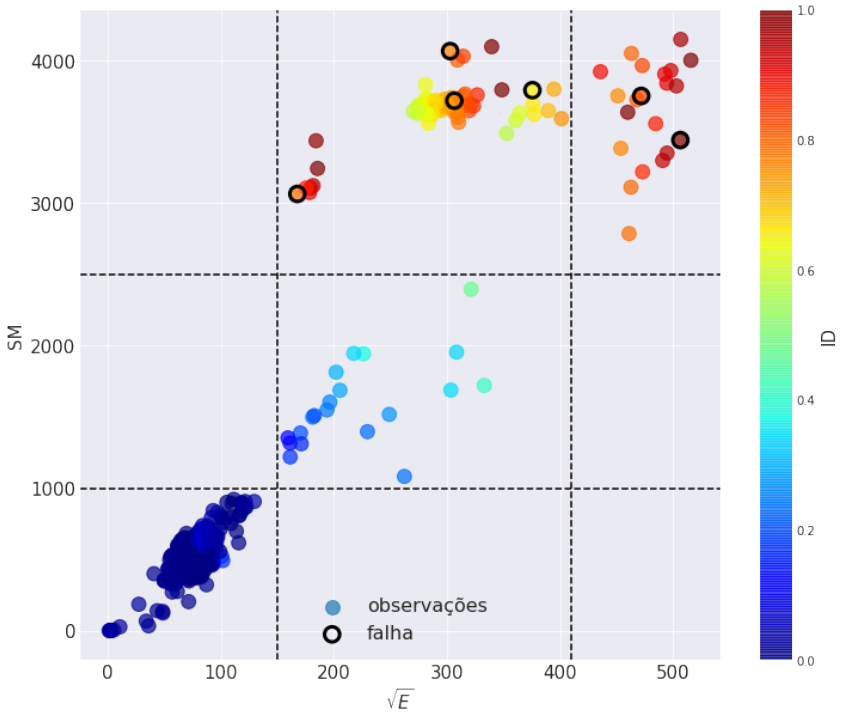


Figura 26 – Mapa de calor com o comportamento das operações baseado nas duas características mais relevantes.

de corrente pode indicar um processo de danificação do equipamento. Nos quadrantes localizados mais acima, foi verificado, por meio do comportamento dos dados, uma quebra

dessa relação linear. Nota-se uma variação maior da corrente, novamente indicando que os sinais de operação de falha são menos uniformes.

A energia varia consideravelmente na região onde encontram-se os dados de falha, o que pode indicar uma diferença entre as falhas observadas. Por exemplo, as falhas em que há uma maior energia podem ter ocorrido devido a falta de lubrificação de algum componente ou o impedimento das agulhas por meio de algum objeto, levando ao equipamento exercer uma força maior quando operado. E as falhas com baixo valor de energia podem ter ocorrido devido a um mal funcionamento ou regulagem de algum componente.

Com base nessa análise, é possível delimitar uma região segura de operação, localizada no primeiro quadrante, em que as mudanças são menores que 1000 e a raiz quadrada da energia é menor que 140, uma região de alerta, localizada no quadrante central, e outra crítica, quando as mudanças possuem valores maiores que 2500.

Resumindo:

$$\begin{cases} \begin{matrix} \sqrt{E} \leq 140 \\ SM \leq 1000 \end{matrix} & \text{para região segura} \\ \begin{matrix} \sqrt{E} > 140 \\ 1000 < SM \leq 1000 \end{matrix} & \text{para região de alerta} \\ SM > 2500 & \text{para região crítica} \end{cases} \tag{4.6}$$

5 CONCLUSÃO E TRABALHOS FUTUROS

A metodologia proposta neste trabalho apresentou uma abordagem de extração e seleção de característica para séries temporais, combinada com técnicas de inteligência computacional para a previsão de falhas em máquinas de chave. Utilizou-se o algoritmo FRESH proposto por Christ, Kempa-Liehr & Feindt (2016) para extração devido à sua capacidade em encontrar um conjunto de características bem representativas quando as propriedades que caracterizam o comportamento da série temporal utilizada são desconhecidas. Na etapa que compreendeu a análise dos sinais, empregou-se o PCA e algoritmo de agrupamento *k-means*, uma técnica de aprendizado de máquina não supervisionado. Pelo fato de os dados serem provenientes de máquinas de chave que estão operando em campo, e não em um ambiente controlado, algumas informações referentes aos mesmos não são acuradas, confiáveis. Então, optou-se pelo algoritmo de *k-means*, pois o mesmo é conhecido por extrair informações a partir dos próprios dados, não sendo necessário uma pré-classificação.

Após a execução da metodologia proposta, os resultados apresentados mostraram a eficiência do FRESH, que encontrou um subconjunto de características capazes de identificar os padrões que definem e separam uma operação saudável e uma operação de falha, como visto na Figura 16. Com a aplicação do *k-means* foi possível identificar as similaridades existentes em um mesmo tipo de sinal e evidenciar as dissimilaridades entre os tipos diferentes, permitindo encontrar grupos que caracterizassem cada tipo de operação, apresentados na Figura 18. A partir da definição desses grupos, foi possível então aplicar o conceito de Identificador de Dano (ID) como métrica de avaliação, visando examinar o comportamento do sinal observado. Com base nos IDs e na Figura 21, foi possível definir um limiar que indicasse quando uma máquina está operando em estado saudável ou com alguma anomalia, permitindo que uma ação, uma decisão, possa ser tomada por um profissional da área, antecipando a ocorrência de uma falha, evitando assim a paralisação do equipamento e consequentemente, aumentando a produtividade da empresa.

O PCA foi utilizando com o intuito de reduzir a dimensionalidade dos dados a fim de melhorar a visualização dos mesmos. No entanto, a aplicação do mesmo permitiu executar uma análise em cima das características selecionadas pelo FRESH. Como consequência, concluiu-se que, mesmo que o subconjunto de características selecionado tenha apresentado bons resultados, uma quantidade menor de características poderia ter sido utilizada, uma vez que, com somente duas, foi possível obter os mesmos resultados. A utilização dessas duas características permitiu fazer uma análise sobre o comportamento das máquinas de chave quando uma operação é saudável ou de falha, em que foi possível verificar um

comportamento distinto entre as falhas, principalmente quanto à energia, e que pode ser o indicativo de ocorrência de diferentes tipos de falhas.

A Alemanha, no ano de 2011, através de um conferência realizada pelo governo em Hanôver, introduziu um novo conceito chamado de Indústria 4.0 e que se refere à quarta revolução industrial (Drath; Horch, 2014), e que concentra-se na implantação de produtos e processos inteligentes dentro das indústrias (Brettel *et al.*, 2014). Essas indústrias inteligentes devem lidar com a necessidade de desenvolvimento rápido de produtos, alta produtividade, produção flexível e ambientes complexos (Vyatkin *et al.*, 2007), através da aplicação de sistemas ciber-físicos (CPS - *Cyber-Physical Systems*), que se referem a uma nova geração de sistemas com integração entre capacidades computacional e física e que possibilitam a comunicação entre humanos, máquinas e produtos, ou seja, comunicação direta entre sistemas, máquinas, produtos e pessoas (Baheti; Gill, 2011; Einsiedler, 2013; Damm *et al.*, 2010).

Em razão disso, há um crescente interesse na busca de soluções para problemas relacionados à indústrias, como políticas de manutenção, e que se enquadram dentro da definição de CPS e, por esse motivo, propõe-se como trabalho futuro ampliar o escopo deste trabalho, objetivando o estudo de uma metodologia capaz de analisar séries temporais de diversos tipos, por meio de métodos de análise escaláveis e que se adaptam à série temporal utilizada, fornecendo informações com agilidade e eficácia, prevendo a ocorrência de falhas e identificando-as, e assim, auxiliar na tomada de decisões e formação de estratégias dentro das indústrias.

REFERÊNCIAS

ABDI, Hervé; WILLIAMS, Lynne J. Principal component analysis. **Wiley interdisciplinary reviews: computational statistics**, Wiley Online Library, v. 2, n. 4, p. 433–459, 2010.

ADACHI, Hisanobu; KIKUCHI, Makoto; WATANABE, Yoshihiro. Electric switch machine failure detection using data-mining technique. **Quarterly Report of RTRI**, Railway Technical Research Institute, v. 47, n. 4, p. 182–186, 2006.

AGHABOZORGI, Saeed; SHIRKHORSHIDI, Ali Seyed; WAH, Teh Ying. Time-series clustering–a decade review. **Information Systems**, Elsevier, v. 53, p. 16–38, 2015.

AGUIAR, Eduardo; NOGUEIRA, Fernando; AMARAL, Renan; FABRI, Diego; ROSSIGNOLI, Sérgio; FERREIRA, José Geraldo; VELLASCO, Marley; TANSCHEIT, Ricardo; RIBEIRO, Moisés; VELLASCO, Pedro. Classification of events in switch machines using bayes, fuzzy logic system and neural network. In: SPRINGER. **International Conference on Engineering Applications of Neural Networks**. [S.l.], 2014. p. 81–91.

AGUIAR, Eduardo P; FERNANDO, M de A; VELLASCO, Marley MBR; RIBEIRO, Moisés V. Set-membership type-1 fuzzy logic system applied to fault classification in a switch machine. **IEEE Transactions on Intelligent Transportation Systems**, IEEE, 2017.

AGUIAR, Eduardo P de; AMARAL, Renan PF; VELLASCO, Marley MBR; RIBEIRO, Moisés V. An enhanced singleton type-2 fuzzy logic system for fault classification in a railroad switch machine. **Electric Power Systems Research**, Elsevier, v. 158, p. 195–206, 2018.

AGUIAR, Eduardo P de; FERNANDO, M de A; AMARAL, Renan PF; FABRI, Diego F; SÉRGIO, C de A; FERREIRA, José G; VELLASCO, Marley MBR; TANSCHEIT, Ricardo; VELLASCO, Pedro CG da S; RIBEIRO, Moisés V. Eann 2014: a fuzzy logic system trained by conjugate gradient methods for fault classification in a switch machine. **Neural Computing and Applications**, Springer, v. 27, n. 5, p. 1175–1189, 2016.

ANTT, Agência Nacional de Transportes Terrestres. **Anuário Estatístico**. 2017.

ASADA, Tomotsugu; ROBERTS, Clive. Improving the dependability of dc point machines with a novel condition monitoring system. **Proceedings of the Institution of Mechanical Engineers, Part F: Journal of rail and rapid transit**, Sage Publications Sage UK: London, England, v. 227, n. 4, p. 322–332, 2013.

BAGNALL, Anthony; LINES, Jason; BOSTROM, Aaron; LARGE, James; KEOGH, Eamonn. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. **Data Mining and Knowledge Discovery**, Springer, v. 31, n. 3, p. 606–660, 2017.

BAHETI, Radhakisan; GILL, Helen. Cyber-physical systems. **The impact of control technology**, IEEE Control Systems Society, v. 12, p. 161–166, 2011.

BARBU, Tudor. Variational image denoising approach with diffusion porous media flow. In: HINDAWI PUBLISHING CORPORATION. **Abstract and Applied Analysis**. [S.l.], 2013. v. 2013.

BENJAMINI, Yoav; HOCHBERG, Yosef. Controlling the false discovery rate: a practical and powerful approach to multiple testing. **Journal of the royal statistical society. Series B (Methodological)**, JSTOR, p. 289–300, 1995.

BENJAMINI, Yoav; YEKUTIELI, Daniel. The control of the false discovery rate in multiple testing under dependency. **Annals of statistics**, JSTOR, p. 1165–1188, 2001.

BOLÓN-CANEDO, Verónica; SÁNCHEZ-MAROÑO, Noelia; ALONSO-BETANZOS, Amparo. **Feature selection for high-dimensional data**. [S.l.]: Springer, 2015.

BOX, George EP; JENKINS, Gwilym M; REINSEL, Gregory C; LJUNG, Greta M. **Time series analysis: forecasting and control**. [S.l.]: John Wiley & Sons, 2015.

BRETTEL, Malte; FRIEDERICHSEN, Niklas; KELLER, Michael; ROSENBERG, Marius. How virtualization, decentralization and network building change the manufacturing landscape: An industry 4.0 perspective. **International Journal of Mechanical, Industrial Science and Engineering**, v. 8, n. 1, p. 37–44, 2014.

CARDOSO, Elcio. **Substituindo os Acionadores de Desvios AMV eletromagnéticos por manuais**. 2015. `http://redeferroviariarioseco.blogspot.com.br/2015/10/substituindo-os-acionadores-de-desvios.html`. Acessado em: 18-01-2018.

CHEN, J; ROBERTS, C. Effective condition monitoring of line side assets. IET, 2006.

CHRIST, Maximilian; KEMPA-LIEHR, Andreas W; FEINDT, Michael. Distributed and parallel time series feature extraction for industrial big data applications. **arXiv preprint arXiv:1610.07717**, 2016.

CURRAN-EVERETT, Douglas. Multiple comparisons: philosophies and illustrations. **American Journal of Physiology-Regulatory, Integrative and Comparative Physiology**, Am Physiological Soc, v. 279, n. 1, p. R1–R8, 2000.

DAMM, Werner; ACHATZ, Reinhold; BEETZ, Klaus; BROY, Manfred; DAEMBKES, Heinrich; GRIMM, Klaus; LIGGESMEYER, Peter. Nationale roadmap embedded systems. In: **Cyber-Physical Systems**. [S.l.]: Springer, 2010. p. 67–136.

DINDAR, Serdar; KAEWUNRUEN, Sakdirat. Assessment of turnout-related derailments by various causes. In: SPRINGER. **International Congress and Exhibition"Sustainable Civil Infrastructures: Innovative Infrastructure Geotechnology"**. [S.l.], 2017. p. 27–39.

DNIT, Departamento Nacional de Infra-Estrutura. **AMV – Equipamentos de manobra**. 2015. `http://www.dnit.gov.br/download/consultas-publicas/ferroviario/pim/`. Acessado em: 18-01-2018.

DNIT, Departamento Nacional de Infra-Estrutura. **O Histórico da ferrovia**. 2018. `http://www1.dnit.gov.br/ferrovias/historico.asp`. Acessado em: 17-01-2018.

DRATH, Rainer; HORCH, Alexander. Industrie 4.0: Hit or hype?[industry forum]. **IEEE industrial electronics magazine**, IEEE, v. 8, n. 2, p. 56–58, 2014.

DUDA, Richard O; HART, Peter E; STORK, David G. Pattern classification and scene analysis part 1: Pattern classification. **Wiley, Chichester**, 2000.

EHLERS, RS. Análise de séries temporais, 2003. **Departamento de Estatística, Universidade Federal do Paraná**, 2012.

EINSIEDLER, Ingrid. Embedded systeme für industrie 4.0. **Product. Manag**, v. 18, p. 26–28, 2013.

EKER, Omer Faruk; CAMCI, Fatih; KUMAR, Uday. Failure diagnostics on railway turnout systems using support vector machines. In: LULEÅ TEKNISKA UNIVERSITET. **International Workshop and Congress on eMaintenance: 22/06/2010-24/06/2010**. [S.l.], 2010. p. 248–251.

ELGAMMAL, Ahmed; DURAISWAMI, Ramani; HARWOOD, David; DAVIS, Larry S. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. **Proceedings of the IEEE**, IEEE, v. 90, n. 7, p. 1151–1163, 2002.

ELLIS, Byron A; BYRON, A. Condition based maintenance. **The Jethro Project**, v. 10, p. 1–5, 2008.

ELMASRI, Ramez; LEE, Jae Young. Implementation options for time-series data. In: **Temporal Databases: Research and Practice**. [S.l.]: Springer, 1998. p. 115–128.

FIGUEIREDO, Eloi; RADU, Lucian; WORDEN, Keith; FARRAR, Charles R. A bayesian approach based on a markov-chain monte carlo method for damage detection under unknown sources of variability. **Engineering Structures**, Elsevier, v. 80, p. 1–10, 2014.

FLEURY, Paulo Fernando. Pontos fortes e fracos da fase pós-privatização. **Valor Setorial Ferrovias, Setembro**, p. 54–55, 2006.

FRIEDRICH, Rudolf; SIEGERT, Silke; PEINKE, Joachim; SIEFERT, M; LINDEMANN, M; RAETHJEN, J; DEUSCHL, G; PFISTER, G *et al.* Extracting model equations from experimental data. **Physics Letters A**, Elsevier, v. 271, n. 3, p. 217–222, 2000.

FULCHER, Ben D; JONES, Nick S. Highly comparative feature-based time-series classification. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 26, n. 12, p. 3026–3037, 2014.

GARCÍA, Fausto P; PEDREGAL, Diego J; ROBERTS, Clive. Time series methods applied to failure prediction and detection. **Reliability Engineering & System Safety**, Elsevier, v. 95, n. 6, p. 698–703, 2010.

GUAN, Ke; AI, Bo; ZHONG, Zhangdui; LÓPEZ, Carlos F; ZHANG, Lei; BRISO-RODRÍGUEZ, Cesar; HROVAT, Andrej; ZHANG, Bei; HE, Ruisi; TANG, Tao. Measurements and analysis of large-scale fading characteristics in curved subway tunnels at 920 mhz, 2400 mhz, and 5705 mhz. **IEEE Transactions on Intelligent Transportation Systems**, IEEE, v. 16, n. 5, p. 2393–2405, 2015.

HARTIGAN, John A. **Clustering Algorithms**. 99th. ed. New York, NY, USA: John Wiley & Sons, Inc., 1975. ISBN 047135645X.

HATHAWAY, Richard J; BEZDEK, James C; HU, Yingkang. Generalized fuzzy c-means clustering strategies using l/sub p/norm distances. **IEEE transactions on Fuzzy Systems**, IEEE, v. 8, n. 5, p. 576–582, 2000.

HÜSKEN, Michael; STAGGE, Peter. Recurrent neural networks for time series classification. **Neurocomputing**, Elsevier, v. 50, p. 223–235, 2003.

JAIN, Anil K; DUBES, Richard C. **Algorithms for clustering data**. [S.l.]: Prentice-Hall, Inc., 1988.

JOLLIFFE, Ian T. Principal component analysis and factor analysis. In: **Principal component analysis**. [S.l.]: Springer, 1986. p. 115–128.

KAUFMAN, Leonard; ROUSSEEUW, Peter J. **Finding groups in data: an introduction to cluster analysis**. [S.l.]: John Wiley & Sons, 2009.

KRAWCZAK, Maciej; SZKATUŁA, Grażyna. An approach to dimensionality reduction in time series. **Information Sciences**, Elsevier, v. 260, p. 15–36, 2014.

LI, Dan-Yong; SONG, Yong-Duan; CAI, Wen-Chuan. Neuro-adaptive fault-tolerant approach for active suspension control of high-speed trains. **IEEE Transactions on Intelligent Transportation Systems**, IEEE, v. 16, n. 5, p. 2446–2456, 2015.

LICHMAN, M. **UCI Machine Learning Repository**. 2013. Disponível em: <http://archive.ics.uci.edu/ml>.

LIN, Jessica; VLACHOS, Michail; KEOGH, Eamonn; GUNOPULOS, Dimitrios. Iterative incremental clustering of time series. In: SPRINGER. **International Conference on Extending Database Technology**. [S.l.], 2004. p. 106–122.

MACQUEEN, James *et al.* Some methods for classification and analysis of multivariate observations. In: OAKLAND, CA, USA. **Proceedings of the fifth Berkeley symposium on mathematical statistics and probability**. [S.l.], 1967. v. 1, n. 14, p. 281–297.

MAGALHÃES, Marcos Nascimento; LIMA, Antonio Carlos Pedroso de. **Noções de probabilidade e estatística**. [S.l.]: IME-USP São Paulo:, 2000.

MÁRQUEZ, Fausto Pedro García; SCHMID, Felix. A digital filter-based approach to the remote condition monitoring of railway turnouts. **Reliability Engineering & System Safety**, Elsevier, v. 92, n. 6, p. 830–840, 2007.

MAYRINK, Amarildo José. **Estação Guia de Pacobaíba - O Porto da Estrela, no "fundo"da Baía de Guanabara.** 2016. `http://otremexpresso.blogspot.com.br/2016/06/estacao-guia-de-pacobaiba-o-porto-da.html`. Acessado em: 18-01-2018.

MCKINNEY, Wes *et al.* Data structures for statistical computing in python. In: SCIPY AUSTIN, TX. **Proceedings of the 9th Python in Science Conference**. [S.l.], 2010. v. 445, p. 51–56.

MIKE, Jersey. **PHOTOS: Port Road Trips - COLA Interlocking and Tower**. 2014. `http://position-light.blogspot.com.br/2014/05/photos-port-road-trips-cola.html`. Acessado em: 18-01-2018.

MIKE, Jersey. **PHOTOS: Port Road Trips - PILOT to WEST ROCK**. 2015. http://position-light.blogspot.com.br/2015/05/photos-port-road-trips-pilot-to-west.html. Acessado em: 18-01-2018.

MÖRCHEN, Fabian. **Time series feature extraction for data mining using DWT and DFT**. [S.l.]: Univ., 2003.

MUSHTAQ, Rizwan. Augmented dickey fuller test. 2011.

PEARSON, Karl. Liii. on lines and planes of closest fit to systems of points in space. **The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science**, Taylor & Francis, v. 2, n. 11, p. 559–572, 1901.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PINCUS, Steven M; GLADSTONE, Igor M; EHRENKRANZ, Richard A. A regularity statistic for medical data analysis. **Journal of clinical monitoring**, Springer, v. 7, n. 4, p. 335–345, 1991.

RADIVOJAC, Predrag; OBRADOVIC, Zoran; DUNKER, A Keith; VUCETIC, Slobodan. Feature selection filters based on the permutation test. In: SPRINGER. **ECML**. [S.l.], 2004. p. 334–346.

RANI, Sangeeta; SIKKA, Geeta. Recent techniques of clustering of time series data: a survey. **International Journal of Computer Applications**, Foundation of Computer Science, v. 52, n. 15, 2012.

RICHMAN, Joshua S; MOORMAN, J Randall. Physiological time-series analysis using approximate entropy and sample entropy. **American Journal of Physiology-Heart and Circulatory Physiology**, Am Physiological Soc, v. 278, n. 6, p. H2039–H2049, 2000.

ROSENBERG, Andrew; HIRSCHBERG, Julia. V-measure: A conditional entropy-based external cluster evaluation measure. In: **Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)**. [S.l.: s.n.], 2007.

SCHOLZ, Matthias. Approaches to analyse and interpret biological profile data. 2006.

SNEATH, Peter HA. The application of computers to taxonomy. **Microbiology**, Microbiology Society, v. 17, n. 1, p. 201–226, 1957.

SOLER, Manuel; LÓPEZ, Jesús; PEDRO, José Manuel Mera Sánchez de; MAROTO, Joaquin. Methodology for multiobjective optimization of the ac railway power supply system. **IEEE Transactions on Intelligent Transportation Systems**, IEEE, v. 16, n. 5, p. 2531–2542, 2015.

TIBADUIZA, Diego A; MUJICA, Luis E; RODELLAR, José; GüEMES, Alfredo. Structural damage detection using principal component analysis and damage indices. **Journal of Intelligent Material Systems and Structures**, v. 27, n. 2, p. 233–248, 2016.

VYATKIN, Valeriy; SALCIC, Zoran; ROOP, Partha S; FITZGERALD, John. Now that's smart! **IEEE Industrial Electronics Magazine**, IEEE, v. 1, n. 4, p. 17–29, 2007.

WALT, Stéfan van der; COLBERT, S Chris; VAROQUAUX, Gael. The numpy array: a structure for efficient numerical computation. **Computing in Science & Engineering**, IEEE, v. 13, n. 2, p. 22–30, 2011.

WANG, Xiaoyue; MUEEN, Abdullah; DING, Hui; TRAJCEVSKI, Goce; SCHEUERMANN, Peter; KEOGH, Eamonn. Experimental comparison of representation methods and distance measures for time series data. **Data Mining and Knowledge Discovery**, Springer, v. 26, n. 2, p. 275–309, 2013.

WANG, Xiaozhe; SMITH, Kate; HYNDMAN, Rob. Characteristic-based clustering for time series data. **Data mining and knowledge Discovery**, Springer, v. 13, n. 3, p. 335–364, 2006.

WEISSTEIN, Eric W. **CRC concise encyclopedia of mathematics**. [S.l.]: CRC press, 2002.

WILCOX, R. Kolmogorov–smirnov test. **Encyclopedia of biostatistics**, Wiley Online Library, 2005.

XU, Rui; WUNSCH, Donald. Survey of clustering algorithms. **IEEE Transactions on neural networks**, Ieee, v. 16, n. 3, p. 645–678, 2005.

YANG, Jianbo; NGUYEN, Minh Nhut; SAN, Phyo Phyo; LI, Xiaoli; KRISHNASWAMY, Shonali. Deep convolutional neural networks on multichannel time series for human activity recognition. In: **IJCAI**. [S.l.: s.n.], 2015. p. 3995–4001.

ZHUANG, Xinhua; HUANG, Yan; PALANIAPPAN, Kannappan; ZHAO, Yunxin. Gaussian mixture density modeling, decomposition, and applications. **IEEE Transactions on Image Processing**, IEEE, v. 5, n. 9, p. 1293–1302, 1996.

A EXTRATORES DE CARACTERÍSTICAS UTILIZADOS

Para obter um conjunto de características relevantes de uma série temporal, a técnica de FRESH depende de um mapeamento de características $\theta_k : \mathbb{R}^n \to \mathbb{R}$. Cada um desses extratores de características $\theta(S)$ leva em conta uma única série temporal como argumento e retornam uma característica ou um vetor de características, sendo que, no último caso, cada elemento do vetor retornado é tratado como uma característica separada (Christ; Kempa-Liehr; Feindt, 2016). Abaixo é apresentado uma lista com algumas funções para extração de características mais utilizadas pelo FRESH.

- **Energia:** Retorna a energia absoluta da série temporal, que é a soma dos valores absolutos ao quadrado:

$$E = \sum_{i=1,\ldots,n} |s_i|^2 \tag{A.1}$$

- **Soma das mudanças consecutivas:** Retorna a soma sobre o valor absoluto de mudanças consecutivas na série temporal S

$$\sum_{i=1,\ldots,n-1} |s_{i+1} - s_i| \tag{A.2}$$

- **Autocorrelação:** Calcula o valor de uma função de agregação (por exemplo, variância ou média) da autocorrelação, usada para encontrar padrões de repetição em uma série, assumindo diferentes defasamentos possíveis (de 1 ao comprimento de S).

$$\frac{1}{n-1} \sum_{l=1,\ldots,n} \frac{1}{(n-1)\sigma^2} \sum_{t=1}^{n-1} (s_t - \mu)(s_{t+l} - \mu) \tag{A.3}$$

  em que n é o comprimento da série temporal S, σ^2 é sua variância e μ sua média.

- **Regressão linear:** Calcula uma regressão linear de mínimos quadrados para os valores da série temporal. Este recurso pressupõe que o sinal seja amostrado uniformemente.

- **Entropia:** Implementa um algoritmo de entropia aproximado vetorizado. Em Estatística, uma entropia aproximada é uma técnica utilizada para quantificar a quantidade de regularidade e a imprevisibilidade das flutuações em relação aos dados da série temporal (Pincus; Gladstone; Ehrenkranz, 1991).

- **Processo auto-regressivo:** Faz o ajusto na máxima probabilidade incondicional de um processo auto-regressivo de ordem k, ou $AR(k)$ (Ehlers, 2012). O parâmetro

k é o defasamento máximo do processo.

$$S_t = \varphi_0 + \sum_{i=1}^{k} \varphi_i S_{t-i} + \varepsilon_t \tag{A.4}$$

em que os coeficientes φ são retornados.

- **Teste de Dickey-Fuller aumentado:** O Teste de Dickey-Fuller aumentado ou Teste ADF (*Augmented Dickey-Fuller*) é um teste de raiz unitária em séries temporais. Retorna o valor do respectivo teste estatístico. Mais detalhes pode ser visto em Mushtaq (2011).

- **Autocorrelação com um defasamento:** Calcula a autocorrelação de um defasamento (lag) especificado, de acordo com a fórmula:

$$\frac{1}{(n-1)\sigma^2} \sum_{t=1}^{n-1} (s_t - \mu)(s_{t+l} - \mu) \tag{A.5}$$

em que n é o comprimento da série temporal S, σ^2 sua variância e μ é a média. l denota o defasamento.

- **Entropia em colunas:** Inicializa os valores de S em colunas (bins) equidistantes de max_bins. Em seguida, calcula o valor de

$$\sum_{k=0}^{min(max\_bins, len(S))} p_k \log(p_k) \cdot \mathbf{1}_{(p_k>0)} \tag{A.6}$$

em que p_k é a porcentagem de amostras na coluna k.

- **Valor médio em um dado intervalo de quantis:** Primeiro ajusta um corredor dado pelos quantis ql (menor quantil do corredor) e qh (maior quantil do corredor) da distribuição de S. Em seguida, calcula o valor médio e absoluto das mudanças consecutivas da série S dentro desse corredor.

- **Valores acima da média:** Retorna o número de valores em S que são superiores à média de S.

- **Valores abaixo da média:** Retorna o número de valores em S que são inferiores à média de S.

- **Transformada de wavelet:** Calcula uma transformada de wavelet contínua para a wavelet de Ricker, também conhecida como "chapéu mexicano", que é definida por

$$\frac{2}{\sqrt{3a}\pi^{\frac{1}{4}}}(1 - \frac{S^2}{a^2}) \exp(-\frac{S^2}{2a^2}) \tag{A.7}$$

em que a é o parâmetro de largura da função wavelet.

- **Energia dividida em segmentos:** Calcula a soma dos quadrados do segmento i de N segmentos expressos como uma razão entre soma dos quadrados de toda a série temporal

  Toma como parâmetros de entrada o número de segmentos para dividir a série e o número do segmento (começando em zero) para retornar uma característica.

- **Transformada de Fourier:** Calcula os coeficientes da Transformada de Fourier discreta Unidimensional para entrada real, através de um algoritmo de transformada rápida de Fourier

$$A_k = \sum_{m=0}^{n-1} a_m \exp\left\{-2\pi i \frac{mk}{n}\right\}, k = 0, \ldots, n-1 \tag{A.8}$$

  Os coeficientes resultantes serão complexos, podendo retornar a parte real, a parte imaginária, o valor absoluto e o ângulo em graus.

- **Primeira localização do máximo:** Retorna a primeira localização do valor máximo de S. A posição é calculada em relação ao comprimento de S.

- **Primeira localização do mínimo:** Retorna a primeira localização do valor mínimo de S. A posição é calculada em relação ao comprimento de S.

- **Coeficientes de Friedrich:** Coeficientes de polinomio $h(S)$, que foi ajustado à dinâmica determinística do modelo de Langevin

$$\dot{S}(t) = h(S(t)) + N(0, R) \tag{A.9}$$

  Como descrito por Friedrich *et al.* (2000). Para séries temporais curtas, este método é altamente dependente dos parâmetros.

- **Duplicidade:** Verifica se algum valor em S ocorre mais de uma vez.

- **Duplicidade do valor máximo:** Verifica se o valor máximo de S é observado mais de uma vez.

- **Duplicidade do valor mínimo:** Verifica se o valor mínimo de S é observado mais de uma vez.

- **Índice relativo à massa para um quantil:** Calcula o índice relativo i em que q% da massa da série temporal S fica de i. Por exemplo, para $q = 50\%$ o centro de massa das séries temporais é retornado.

- **Curtose:** Retorna a curtose de S (calculada com o coeficiente ajustado de Fisher-Pearson).

$$kurtosis(S) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{s_i - \bar{S}}{std(S)}\right)^4 - 3 \tag{A.10}$$

como definido por Weisstein (2002).

- **Desvio padrão maior que um intervalo:** Variável booleana indicando se o desvio padrão de S é maior do que r vezes o intervalo = diferença entre max e min de S. Por isso, verifica se

$$std(S) > r * (max(S) - min(S)) \tag{A.11}$$

- **Última localização do máximo:** Retorna a última localização relativa do valor máximo de S. A posição é calculada em relação ao comprimento de S.

- **Última localização do mínimo:** Retorna a última localização relativa do valor mínimo de S. A posição é calculada em relação ao comprimento de S.

- **Comprimento:** Retorna o comprimento de S

$$length(S) = n \tag{A.12}$$

- **Sequência mais longa acima da média:** Retorna o comprimento da subsequência consecutiva mais longa em S maior que a média de S.

- **Sequência mais longa abaixo da média:** Retorna o comprimento da subsequência consecutiva mais longa em S menor que a média de S.

- **Máximo:** Calcula o maior valor da série temporal S.

$$max(S) = max(s_1, s_2, \ldots, s_n) \tag{A.13}$$

- **Mínimo:** Calcule o menor valor da série temporal S.

$$min(S) = min(s_1, s_2, \ldots, s_n) \tag{A.14}$$

- **Média:** Retorna a média de S.

$$mean(S) = \bar{S} = \frac{1}{n}\sum_{i=1}^{n} s_i \tag{A.15}$$

- **Mediana:** Retorna a mediana de S.

$$median(S) = \begin{cases} s_{(n+1)/2} & : \text{se n ímpar} \\ \frac{1}{2}\left(s_{n/2} + s_{(n/2+1)}\right) & : \text{se n par} \end{cases} \tag{A.16}$$

- **Média das diferenças:** Retorna a média sobre as diferenças entre os valores das séries temporais subsequentes que são:

$$\frac{1}{n}\sum_{i=1,\ldots,n-1} s_{i+1} - s_i \tag{A.17}$$

- **Média das diferenças absolutas:** Retorna a média sobre as diferenças absolutas entre os valores das séries temporais subsequentes que são:

$$\frac{1}{n}\sum_{i=1,\ldots,n-1} |s_{i+1} - s_i| \tag{A.18}$$

- **Valor médio da derivada segunda:** Retorna o valor médio de uma aproximação central da derivada segunda.

$$\frac{1}{n}\sum_{i=1,\ldots,n-1} \frac{1}{2}(s_{i+2} - 2 \cdot s_{i+1} + s_i) \tag{A.19}$$

- **Número de cruzamentos:** Calcula o número de cruzamentos de S em m. Um cruzamento é definido como dois valores sequenciais onde o primeiro valor é inferior a m e o próximo é maior ou vice-versa. Se você definir m para zero, você receberá o número de zero cruzamentos.

- **Número de picos:** Calcula o número de picos de pelo menos o suporte n na série temporal S. Um pico de suporte n é definido como uma subsequência de S onde ocorre um valor, que é maior que os seus vizinhos a esquerda e à direita.

- **Picos de uma transformada de wavelet:** Procura por diferentes picos em S. Para fazer isso, S é suavizado por uma wavelet de Ricker e para larguras variando de 1 a n. O número de picos que ocorrem em escalas de largura suficientes e com Relação Sinal-Ruído (SNR) suficientemente alta é retornado.

- **Autocorrelação parcial:** Calcula o valor da função de autocorrelação parcial em um dado defasamento (*lag*). A autocorrelação parcial de *lag* k de uma série temporal $\{s_t, t = 1 \ldots n\}$ é igual à correlação parcial de s_t e s_{tk}, ajustada para as variáveis intermediárias $\{s_{t-1}, \ldots, s_{t-k+1}\}$ (Box *et al.*, 2015). Pode ser definido como:

$$\alpha_k = \frac{Cov(s_t, s_{t-k}|s_{t-1}, \ldots, s_{t-k+1})}{\sqrt{Var(s_t|s_{t-1}, \ldots, s_{t-k+1})Var(s_{t-k}|s_{t-1}, \ldots, s_{t-k+1})}} \tag{A.20}$$

- **Porcentagem de valores únicos:** Retorna a porcentagem de valores únicos, que estão presentes na série temporal S mais de uma vez.

$$\frac{\#\text{diferentes valores que ocorrem mais de uma vez}}{\#\text{diferentes valores}} \tag{A.21}$$

Isso significa que a porcentagem é normalizada para o número de valores únicos.

- **Proporção de valores únicos:** Retorna a proporção de valores únicos, que estão presentes na série temporal S mais de uma vez.

$$\frac{\#\text{de pontos de dados que ocorrem mais de uma vez}}{\#\text{de todos os pontos de dados}} \tag{A.22}$$

Isso significa que a relação é normalizada para o número de pontos de dados na série temporal

- **Quantil:** Calcula o q quantil de S. Este é o valor de S superior a $q\%$ dos valores ordenados de S.

- **Contagem de valores dado um intervalo:** Contagem de valores observados dentro do intervalo [min, max].

- **Proporção de valores maiores que o desvio padrão:** A proporção de valores que são maiores que $r * std(S)$ $(r\sigma)$ longe da média de S.

- **Razão de valores únicos** Retorna um fator que é um se todos os valores na série temporal S ocorrerem apenas uma vez e abaixo de um, se esse não for o caso. Em princípio, ele simplesmente retorna:

$$\frac{\#\text{de valores únicos}}{\#\text{total de valores}} \tag{A.23}$$

- **Entropia amostral:** Calcula e retorne entropia amostral de S (Richman; Moorman, 2000).

- **Assimetria (skewness):** Retorna a assimetria (*skewness*) da amostra de S (calculada com o coeficiente ajustado de Fisher-Pearson).

$$skewness(S) = \frac{n^2}{(n-1)(n-2)} \frac{\frac{1}{n}\sum_{i=1}^{n}(s_i - \bar{S})^3}{\left(\frac{1}{n-1}\sum_{i=1}^{n}(s_i - \bar{S})^2\right)^{\frac{3}{2}}} \tag{A.24}$$

- **Densidade espectral:** Calcula a densidade espectral de potência cruzada da série temporal S em diferentes frequências. Para fazer isso, a série temporal é deslocada para o domínio da frequência.

- **Variância:** Retorna a variância de S.

$$var(S) = \frac{1}{n}\sum_{i=1}^{n}(s_i - \bar{S})^2 \tag{A.25}$$

- **Desvio padrão:** Retorna o desvio padrão de S.

$$std(S) = \sqrt{var(S)} \tag{A.26}$$

- **Soma de pontos de dados repetidos:** Retorna a soma de todos os pontos de dados, que estão presentes na série temporal S mais de uma vez.

- **Soma dos valores repetidos:** Retorna a soma de todos os valores, que estão presentes na série temporal S mais de uma vez.

- **Soma dos valores:** Calcula a soma dos valores da série temporal S.

- **Simetria:** Variável booleana que indica se a distribuição de S parece simétrica. Este é o caso se:

$$|mean(S) - median(S)| < r * (max(S) - min(S)) \tag{A.27}$$

- **Número de vezes de um valor:** Contagem de ocorrências de um certo valor na série temporal S.

- **Variância maior que desvio padrão:** Variável booleana que indica se a variância de S é maior que seu desvio padrão.

B TABELA DE VALROES CRÍTICOS PARA O TESTE DE KOMOLGOROV-SMIRNOV

Tabela 5 – Tabela de valores críticos $D_{n,\alpha}$ do teste de KS para $\alpha = 5\%$ e 10%.

n	0,05	0,01	n	0,05	0,01	n	0,05	0,01	n	0,05	0,01	n	0,05	0,01
1	0,9750	0,9950	**21**	0,2872	0,3443	**41**	0,2076	0,2490	**61**	0,1709	0,2051	**81**	0,1487	0,1784
2	0,8419	0,9293	**22**	0,2809	0,3367	**42**	0,2052	0,2461	**62**	0,1696	0,2034	**82**	0,1478	0,1773
3	0,7076	0,8290	**23**	0,2749	0,3295	**43**	0,2028	0,2433	**63**	0,1682	0,2018	**83**	0,1469	0,1763
4	0,6239	0,7342	**24**	0,2693	0,3229	**44**	0,2006	0,2406	**64**	0,1669	0,2003	**84**	0,1460	0,1752
5	0,5633	6685	**25**	0,2640	0,3166	**45**	0,1984	0,2380	**65**	0,1657	0,1988	**85**	0,1452	0,1742
6	0,5193	0,6166	**26**	0,2591	0,3106	**46**	0,1963	0,2354	**66**	0,1644	0,1973	**86**	0,1444	0,1732
7	0,4834	0,5758	**27**	0,2544	0,3050	**47**	0,1942	0,2330	**67**	0,1632	0,1958	**87**	0,1435	0,1722
8	0,4543	0,5418	**28**	0,2499	0,2997	**48**	0,1922	0,2306	**68**	0,1620	0,1944	**88**	0,1427	0,1713
9	4300	0,5133	**29**	0,2457	0,2947	**49**	0,1903	0,2283	**69**	0,1609	0,1930	**89**	0,1419	0,1703
10	0,4092	0,4889	**30**	0,2417	0,2899	**50**	0,1884	0,2260	**70**	0,1597	0,1917	**90**	0,1412	0,1694
11	0,3912	0,4677	**31**	0,2379	0,2853	**51**	0,1866	0,2239	**71**	0,1586	0,1903	**91**	0,1404	0,1685
12	0,3754	0,4490	**32**	0,2342	0,2809	**52**	0,1848	0,2217	**72**	0,1576	0,1890	**92**	0,1396	0,1676
13	0,3614	0,4325	**33**	0,2308	0,2768	**53**	0,1831	0,2197	**73**	0,1565	0,1878	**93**	0,1389	0,1667
14	0,3489	0,4176	**34**	0,2274	0,2728	**54**	0,1814	0,2177	**74**	0,1554	0,1865	**94**	0,1382	0,1658
15	0,3376	0,4042	**35**	0,2242	0,2690	**55**	0,1798	0,2157	**75**	0,1544	0,1853	**95**	0,1375	0,1649
16	0,3273	0,3920	**36**	0,2212	0,2653	**56**	0,1782	0,2138	**76**	0,1534	0,1841	**96**	0,1368	0,1641
17	0,3180	0,3809	**37**	0,2183	0,2618	**57**	0,1767	0,2120	**77**	0,1524	0,1829	**97**	0,1361	0,1632
18	0,3094	0,3706	**38**	0,2154	0,2584	**58**	0,1752	0,2102	**78**	0,1515	0,1817	**98**	0,1354	0,1624
19	0,3014	0,3612	**39**	0,2127	0,2552	**59**	0,1737	0,2084	**79**	0,1505	0,1806	**99**	0,1347	0,1616
20	0,2941	0,3524	**40**	0,2101	0,2521	**60**	0,1723	0,2067	**80**	0,1496	0,1795	**100**	0,1340	0,1608