UNIVERSIDADE FEDERAL DE MINAS GERAIS

Escola de Engenharia

Programa de Pós Graduação em Engenharia de Produção

Rodrigo Barbosa De Santis

MODELOS DE APRENDIZADO DE MÁQUINA APLICADOS À MANUTENÇÃO PREDITIVA DE PEQUENAS CENTRAIS GERADORAS HIDRELÉTRICAS

Belo Horizonte

2023

Rodrigo Barbosa De Santis

MODELOS DE APRENDIZADO DE MÁQUINA APLICADOS À MANUTENÇÃO PREDITIVA DE PEQUENAS CENTRAIS GERADORAS HIDRELÉTRICAS

Versão final

Tese de doutorado apresentada ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Minas Gerais, na área de concentração em Modelagem Estocástica e Simulação, como requisito parcial para obtenção do título de Doutor em Engenharia de Produção.

Orientador: Prof. Dr. Marcelo Azevedo Costa

Belo Horizonte

2023

S236m

Santis, Rodrigo Barbosa de.
Modelos de aprendizado de máquina aplicados à manutenção preditiva de pequenas centrais geradoras hidrelétricas [recurso eletrônico] / Rodrigo Barbosa de Santis. - 2023.
1 recurso online (140 f. : il., color.) : pdf.

Orientador: Marcelo Azevedo Costa.

Tese (doutorado) - Universidade Federal de Minas Gerais, Escola de Engenharia.

Bibliografia: f. 123-140.

1. Engenharia de produção - Teses. 2. Usinas hidrelétricas - Teses. 3. Modelagem - Teses. 4. Falha de sistema (Engenharia) - Teses. I. Costa, Marcelo Azevedo. II. Universidade Federal de Minas Gerais. Escola de Engenharia. III. Título.

CDU: 658.5(043)

Ficha catalográfica elaborada pela Bibliotecária Letícia Alves Vieira - CRB-6/2337
Biblioteca Prof. Mário Werneck - Escola de Engenharia da UFMG

UNIVERSIDADE FEDERAL DE MINAS GERAIS

Escola de Engenharia

Programa de Pós-Graduação em Engenharia de Produção

FOLHA DE APROVAÇÃO

MODELOS DE SOFT-COMPUTING/MACHINE LEARNING PARA A PREVISÃO DE INDISPONIBILIDADE FORÇADA DE EQUIPAMENTOS EM PCHS/CGHS

RODRIGO BARBOSA DE SANTIS

Tese submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em ENGENHARIA DE PRODUÇÃO, como requisito para obtenção do grau de Doutor em ENGENHARIA DE PRODUÇÃO, área de concentração PESQUISA OPERACIONAL E INTERVENÇÃO EM SISTEMAS SOCIOTÉCNICOS, linha de pesquisa Modelagem Estocástica e Simulação.

Aprovada em 10 de março de 2023, pela banca constituída pelos membros:

Prof(a). Marcelo Azevedo Costa - Orientador

DEP/UFMG

Prof(a). Roberto da Costa Quinino

DEST/UFMG

Prof(a). Leonardo Goliatt da Fonseca

Universidade Federal de Juiz de Fora

Prof(a). Magno Silvério Campos

UFOP

Prof(a). Frederico Gualberto Ferreira Coelho

UFMG

Belo Horizonte, 10 de março de 2023.

Documento assinado eletronicamente por **Marcelo Azevedo Costa**, **Professor do Magistério Superior**, em 10/03/2023, às 15:31, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Leonardo Goliatt da Fonseca**, **Usuário Externo**, em 13/03/2023, às 20:36, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Frederico Gualberto Ferreira Coelho**, **Professor do Magistério Superior**, em 15/03/2023, às 10:15, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Magno Silvério Campos**, **Usuário Externo**, em 16/03/2023, às 17:15, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Roberto da Costa Quinino**, **Coordenador(a) de curso**, em 23/03/2023, às 21:56, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2132224** e o código CRC **C52A059C**.

AGRADECIMENTOS

Ao meu orientador Marcelo, pela autonomia e confiança depositada em mim para desenvolver este projeto, por me ensinar tanto e me inspirar a ser minha melhor versão.

Aos meus colegas de doutorado Tiago, Álvaro, Leandro e Cassius, por me desafiarem a sonhar grande, e com nossa amizade e colaboração conseguir transformar ideias em resultados tangíveis.

À minha família: Carolina, Francisco, Regina, Danielle, Carolina e Rafael; por me darem a segurança e paz de espírito, para que eu pudesse me dedicar ao que gosto e acredito.

“If a machine is expected to be infallible, it cannot also be intelligent.”
- Alan Turing

RESUMO

A manutenção de pequenas centrais hidrelétricas é um tópico essencial para garantir a expansão de fontes de energias limpas e o fornecimento de energia necessária para as próximas décadas. No contexto industrial moderno, a manutenção preditiva guia intervenções e reparos de acordo com o estado de saúde da máquina, calculado a partir de técnicas estatísticas e computacionais. O trabalho atual tem como objetivo principal propor um modelo de manutenção específico para pequenas usinas hidrelétricas. A tese é apresentada em formato de coleção de artigos, sendo o primeiro uma revisão sistemática sobre manutenção preditiva no setor hidrelétrico, o segundo sobre o perfil de manutenção e operação das usinas e formulação do problema, e o terceiro sobre a aplicação do método de floresta de isolamento estendida para detecção de anomalias para diagnóstico inteligente de falhas. Como conclusão, apresentamos duas linhas de ação para trabalho para a tese final: a primeira na área de diagnóstico inteligente por tipo de falhas e a segunda em relação ao prognóstico de variáveis críticas para a operação.

Palavras-chave: Manutenção preditiva. Pequenas centrais hidrelétricas. Modelagem estatística e computacional. Diagnóstico inteligente de falhas. Prognóstico de falhas.

ABSTRACT

Maintenance in small hydroelectric plants is fundamental for guaranteeing the expansion of clean energy sources and supplying the energy estimated to be necessary for the coming decades. In the modern industrial context, predictive maintenance guides interventions and repairs based on the state of health of the machine, calculated from statistical and computational techniques. The current work has as main objective to propose a specific maintenance model for small hydroelectric plants. The thesis proposal is presented in the form of a collection of articles, the first being a systematic review on predictive maintenance in the hydroelectric sector, the second on the maintenance and operation profile of the plants and the formulation of the problem, and the third on the application of the method of extended isolation forest for anomaly detection for intelligent fault diagnosis. As a conclusion, we present two lines of action for work for the final thesis: the first in the area of intelligent diagnosis by type of failures and the second in relation to the prognosis of critical variables for the operation.

Keywords: Condition-based maintenance. Small hydroelectric plants. Computational and statistical modelling. Intelligent fault diagnosis. Fault prognosis.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AnEn	Analogs Ensemble
ACO	Ant Colony Optimization
ANN	Artificial Neural Network
ALIF	Adaptive Local Iterative Filtering
ARIMA	AutoRegressive Integrated Moving Average
ASCA	Adaptive Sine Cosine Algorithm
AVMD	Adaptive Variational Mode Decomposition
BEI	Brasil Energia Inteligente
BN	Bayesian Network
C-Index	Concordance Index Score
CBM	Condition Based Maintenance
CEEMD	Complementary Ensemble Empirical Mode Decomposition
CGH	Central Geradora Hidrelétrica
CN	CoxNet
CNN	Convolutional Neural Network
CNPQ	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CI	Confidence Interval
CPH	Cox Proportional Hazard
CS	Cuckoo Search
CSCA	Chaotic Sine Cosine Algorithm
CTT	Candidate to Target
CWT	Continuous Wavelet Transform
DTSF	Dynamic Time Scan Forecasting
EEMD	Ensemble Empirical Mode Decomposition
EIF	Extended Isolation Forest

EMD	Empirical Mode Decomposition
EWT	Ensemble Wavelet Transform
ETS	ExponenTial Smoothing state space mode
FMSA	Failure Mechanism and Symptoms Analysis
FEM	Finite Element Method
FFT	Fast Fourier Transformation
FRDFE	Fuzzy Recursive Decision Feedback Extension
FT	Fault Tree
FI	Fuzzy Inference
GBS	Gradient Boosting Survival Analysis
GSO	Gram-Schmidt orthogonal
GU	Generating Unit
HD	Hamiltonian dynamic
HGU	Hydro-Generator Unit
HI	Health Index
HFS	Hours of Forced Shutdown
HLU	Hydraulic Lubrication Unit
HOEC	Hours turned Off due to External Conditions
HOS	Higher-Order Statistics
HSS	Hours of Scheduled Shutdown
ICA	Independent Component Analysis
iForest	Isolation Forest
IoT	Internet of Things
ITD	Intrinsic Time-Scale Decomposition
iTree	Isolation Tree
KELM	Kernel Extreme Learning Machine

KICA-PCA	Kernel Independent Component and Principal Component Analysis
KZC	Kutta-Zhoukowski conditions
LS-SVM	Least Square and Support Vector Machine
MASE	Mean Absolute Scaled Error
MOSSA	Multi-Objective Salp Swarm Algorithm
NESO	National Electric System Operator
ON	Hours In Service
OPM	Operation Procedure Manual
OSEEMD	Over-Sampling Ensemble Empirical Mode Decomposition
OWA	Overall Weighted Average
PCA	Principal Component Analysis
PCH	Pequena Central Hidrelétrica
PN	Petri Net
RHD	Reserve Hours Disconnected
RSF	Random Survival Forest
SES	Simple Exponential Smoothing
SHP	Small Hydroelectric Plant
SLR	Systematic Literature Review
sMAPE	Symmetric Mean Absolute Percentage Errror
SSA	Spectral Signal Analysis
SVM	Support Vector Machine
TD	Temporal Distance
TSF	TsFresh algorithm
TTC	Target to Candidate
VMD	Variational Mode Decomposition
WT	Wavelet Transform

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização

Vivemos a era da informação.

Se você, leitor, nasceu no milênio passado, provavelmente teve a oportunidade de acompanhar diversas disrupções se tornarem obsoletas. Tomemos o exemplo dos tocadores de música: a primeira versão do Walkman foi lançada em 1979, capaz de reproduzir músicas direto de uma fita cassete do seu bolso para seus ouvidos. Com sua versão atualizada para CD, este tipo de sistema dominou o mercado 32 anos, até o primeiro iPod ser lançado, em 2001. E hoje, em 2020, o iPod entrou oficialmente na lista de produtos vintage. Hoje a música está presente em todos os lugares em plataformas de *streaming*: você pode ouvir desde as clássicas composições de Vivaldi, até o último lançamento de uma banda *pop* koreana. E onde quiser: celular, computador, televisão. A maioria destes dispositivos nem existiam há 1 século atrás.

O ritmo de transformação da sociedade em que vivemos é incomparável ao de qualquer outra época. São 7.494 bilhões de pessoas habitando o mesmo globo, conectados, criando e compartilhando conhecimento. Barreiras estão sendo derrubadas: comerciais, linguísticas, culturais. Somos uma colmeia do tamanho do mundo. Este ritmo se torna mais evidente quando olhamos para nossos antepassados. O homo sapiens desenvolveu suas primeiras ferramentas de perfuração e corte há cerca de 100 mil anos, com o objetivo de extrair tutano dos ossos de animais. A maioria das ferramentas e culturas agrícolas se desenvolveram durante 5-10 mil A.C., e até hoje se fazem presente em nossa sociedade. Da mesma forma, o que iremos deixar de legado para as próximas gerações, são os frutos da revolução digital.

Neste cenário caótico e dinâmico, este trabalho se faz presente. Buscamos aplicar as mais recentes técnicas computacionais a um setor vital para a continuidade de nossas ações, o energético. A eletricidade iluminou o mundo, e hoje a maioria dos utensílios que usamos dependem dela. Muito brevemente, carros serão inteiramente movidos pela energia elétrica. Diversos países como Reino Unido e França já estabeleceram planos para que carros movidos a combustão sejam eliminados até 2040. Para atendermos toda esta demanda, precisamos continuar investindo em pesquisa e desenvolvimento de fontes de energia limpas.

Dentre estas, destacamos a importância das pequenas centrais hidrelétricas (PCHs) e centrais geradoras hidrelétricas (CGHs). Este tipo de usina, apresenta um menor investimento inicial, baixo impacto ambiental e um enorme potencial de geração, principalmente no Brasil. As PCHs e CGHs vêm ganhando destaque nas discussões acadêmicas e industriais, se tornando empreendimentos rentáveis para os investidores, junto a outras formas de energia limpa como usinas eólicas e fotovoltaicas. Com a regulamentação do mercado de energia, possibilitando a compra e venda do excedente produzido em mercados spot, além de contratos firmados entre geradores e consumidores, as PCHs e CGHs vêm ganhando destaque, junto às demais formas de produção de energia limpa.

1.2 Objetivos

Nossa proposta tem como objetivo central a proposta de modelos para o diagnóstico e prognóstico de equipamentos em PCHs. Com a identificação de falhas de forma automatizada e eficiente, alinhada com a previsão do tempo útil até a falha, busca-se a diminuição dos custos operacionais e de manutenção de usinas em operação. Para isso, adotamos técnicas estatísticas e computacionais aplicadas ao grande volume de dados monitorados gerados constantemente pelo sistema de automação industrial das usinas.

Os objetivo principal é desdobrados em diversas hipóteses de pesquisa, que são separadas e respondidas em cada capítulo:

2. Como o setor hidrelétrico tem se beneficiado dos últimos avanços das técnicas de manutenção baseada em condições (MBC)?
   - 2.1. Quais são os modos de falha mais comuns em hidrelétricas, e quais são as variáveis monitoradas associadas a eles?
   - 2.2. Quais ferramentas de extração de atributos foram usadas para aprimorar os sistemas de MBC?
   - 2.3. Que métodos estatísticos e computacionais têm sido aplicados no diagnóstico e prognóstico das hidrelétricas?
3. Qual o perfil de manutenção e operação de PCHs e quais os tipos de falhas que mais frequentemente contribuem para a indisponibilidade forçada
   - 3.1. Como se dá o processo de desenvolvimento e execução do plano de manutenção em PCHs;
   - 3.2. Dentre o tempo de parada, quanto está relacionado a cada estado operacional (parada por falta de água, parada programada, parada forçada, parada por condições externas);
   - 3.3. quais são os principais componentes que contribuem para as paradas forçadas.
4. É possível obter melhores resultados na detecção e diagnóstico de falhas de uma unidade hidrogeradora em uma PCH utilizando o modelo de Floresta de Isolamento?
   - 4.1. Qual a diferença das métricas de desempenho do modelo de Floresta de Isolamento e Floresta de Isolamento estendida no diagnóstico inteligente de falhas em uma unidade geradora de PCH?

4.2. É possível utilizar as métricas de distância temporal e detecção de contagem, na avaliação de modelos no contexto de detecção de anomalias em série temporal?

4.3. Qual a redução nas métricas de distância temporal comparado com os últimos modelos reportados na literatura (PCA e KICA-PCA)?

5. É possível obter resultados satisfatórios no prognóstico de equipamentos hidrelétricos a partir da aplicação do algoritmo TSFRESH de extração seleção de atributos e modelos de análise de sobrevivência?

   5.1. Qual o desempenho de um framework orientado a dados incluindo estratégias de engenharia de atributos e modelos de sobrevivência de aprendizado de máquina para diagnóstico inteligente de falhas da unidade geradora de PCH?

   5.2. Quais atributos são mais importantes de acordo com o método importância de permutação associado ao modelo de sobrevivência floresta randômica de sobrevivência?

   5.3. Qual dos modelos híbridos tem o melhor desempenho avaliando-se o índice de concordância?

1.3 Caracterização da tese de doutorado

Esta pesquisa inova em três aspectos principais. O primeiro é referente à metodologia/tecnologia empregada: a aplicação de ferramentas de métodos de aprendizado de máquina na área da manutenção preditiva tem chamado atenção devido sua capacidade e eficiência em tratar enorme massa de dados, como os registros de sensores em máquinas. No entanto esta aplicação ainda é um desafio para o setor hidrelétrico, devido a alta complexidade do sistema de geração.

No aspecto prático, este trabalho propõe uma ferramenta computacional para auxiliar a rotina de manutenção de diversas usinas monitoradas pela empresa parceira. Espera-se que a ferramenta seja capaz de auxiliar a priorização de recursos no planejamento da manutenção e da geração, aumentando a disponibilidade dos ativos e consequentemente a capacidade de geração e utilização de recursos hídricos.

O terceiro aspecto é no ponto de vista acadêmico. A formação de profissionais e construção e divulgação de conhecimento na área de confiabilidade é um fator crítico para nossa indústria e o desenvolvimento do país. O conhecimento e métodos gerados por este trabalho pode ser reciclado em outros projetos, principalmente no setor de energia limpa como geração eólica ou foto-voltaica.

1.4 Estrutura da tese

Esta tese de doutorado segue o formato de coleção de artigos. Foram dois temas principais abordados na tese. O foco principal refere-se ao desenvolvimento de modelos de machine learning no contexto de PCHs (capítulos 2, 3, 4 e 5). No apêndice se encontram o desenvolvimento de metodologia de previsão de séries temporais Dynamic Time Scan Forecasting (DTSF), que foi desenvolvida pelo grupo de pesquisa e aprimorado conforme no escopo deste trabalho.

No contexto da aplicação, o capítulo 2 traz uma revisão bibliográfica sistemática sobre o tema de manuteção preditiva em usinas hidrelétricas nos últimos 10 anos. O capítulo 3 apresenta a identificação e formulação do problema, a partir do levantamento do perfil da operação e manutenção de PCHs e CGHs. O capítulo 4 propõe a aplicação do método de detecção de anomalias baseado em árvores de decisão para a identificação de falhas. O capítulo 5 traz uma proposta de aplicação de métodos de análise de sobrevivência para estimação da vida útil de unidades geradoras em PCHs, baseado em métodos de extração de atributos de séries temporais e modelos de aprendizado de máquina.

Dentro do contexto metodológico, o apêndice A traz uma análise do desempenho do método DTSF utilizando a base de dados da competição de previsão de séries temporais univariadas M4, frente a outros métodos estatísticos abordados na competição. No apêndice B propomos formas mais eficientes do que a busca exaustiva (*BruteForce*) de se realizar a busca do perfil de correlação entre a última janela observada e todas as demais do passado, através da busca *JustInTime* e da convolução.

1.5 Resultados e publicações

As publicações da pesquisa são apresentadas abaixo, incluindo o ISSN, fator de impacto e última avaliação qualis na área de Engenharias III (quadriênio 2017-2020):

1. **Capítulo 2.** Artigo publicado na revista *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability* (ISSN: 1748-006X; Fator de impacto: 1,602; Qualis: A4) em julho de 2021.

   DE SANTIS, Rodrigo Barbosa; GONTIJO, Tiago Silveira; COSTA, Marcelo Azevedo. Condition-based maintenance in hydroelectric plants: A systematic literature review. Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, v. 236, n. 5, p. 631-646, 2022. <https://doi.org/10.1177/1748006X211035623>

2. **Capítulo 3.** Artigo apresentado na conferência *International Conference on Renewable Energy* (ICREN - Roma, Itália) em novembro de 2020.

3. **Capítulo 4.** Artigo publicado na revista *Sustainability* (ISSN 2071-1050; Fator de impacto: 2,576; Qualis: A2) em agosto de 2020.

   DE SANTIS, R. B.; COSTA, M. A. Extended Isolation Forests for Fault Detection in Small Hydroelectric Plants. *Sustainability*, v. 12(16), p. 6421, 2020. <https://doi.org/10.3390/su12166421>

4. **Capítulo 5.** Artigo publicado na revista *Sensors* (ISSN: 1424-8220; Fator de impacto: 3,847; Qualis: A2) em dezembro de 2022.

   DE SANTIS, R. B.; GONTIJO, T. S.; COSTA, M. A. A Data-Driven Framework for Small Hydroelectric Plant Prognosis Using Tsfresh and Machine Learning Survival Models. *Sensors*, v. 23, n. 1, p. 12, 2022. <https://doi.org/10.3390/s23010012>

5. **Apêndice A.** Artigo publicado na revista *IEEE Latin America Transactions* (ISSN: 1548-0992; Fator de impacto: 0,729; Qualis: B2) em setembro de 2022.

   DE SANTIS, Rodrigo Barbosa; GONTIJO, Tiago Silveira; COSTA, Marcelo Azevedo. Dynamic Time Scan Forecasting: A Benchmark With M4 Competition Data. IEEE Latin America Transactions, v. 21, n. 2, 2023.

6. **Apêndice B.** Artigo publicado na revista *Journal of Renewable and Sustainable Energy* (ISSN: 1941-7012; Fator de impacto: 2,88; Qualis: A4) em outubro de 2020.

   GONTIJO, Tiago Silveira; COSTA, Marcelo Azevedo; DE SANTIS, Rodrigo Barbosa. Similarity search in electricity prices: An ultra-fast method for finding analogs. Journal of Renewable and Sustainable Energy, v. 12, n. 5, p. 056103, 2020. <https://doi.org/10.1063/5.0021557>

2 CONDITION-BASED MAINTENANCE IN HYDROELECTRIC PLANTS: A SYSTEMATIC LITERATURE REVIEW

Abstract: Industrial maintenance has become an essential strategic factor for profit and productivity in industrial systems. In the modern industrial context, condition-based maintenance guides the interventions and repairs according to the machine's health status, calculated from monitoring variables and using statistical and computational techniques. Although several literature reviews address condition-based maintenance, no study discusses the application of these techniques in the hydroelectric sector, a fundamental source of renewable energy. The present study innovates by addressing condition-based maintenance in the hydroelectric sector, a subject still not covered in the literature. To do so, we conducted a systematic literature review of articles published in the area of predictive maintenance in the last ten years. This was followed by quantitative and thematic analyses of the most relevant categories. We identified a research trend in the application of machine learning techniques, both in the diagnosis and the prognosis of the generating unit's assets, this being the primary monitoring variable. Finally, there is a vast field to be explored regarding the application of statistical models to estimate the useful life, and hybrid models based on physical models and specialists' knowledge, of turbine-generators.

Keywords: Condition based maintenance. Hydroelectric. Fault diagnostics. Fault isolation. Fault monitoring. Fault prognostics. System health management.

2.1 Introduction

From time to time, new technologies emerge and revolutionize entire industries, as we know them. Just as the steam engine and weaving loom transformed production in the 18th century, bringing significant productivity gains to the mass industry sectors, today we witness the 4th wave of this revolution with the digitization of processes. New buzzwords such as the internet of things (IoT), cyber-physical systems, cloud solutions, and augmented reality have been gaining popularity in academic and business environments. In this context, the maintenance paradigm changes so that industrial maintenance has become an essential strategic factor, contributing to profit and guaranteeing the productivity of industrial systems. (CACHADA et al., 2018)

Maintenance 4.0 includes a set of advanced data analysis techniques for processing the enormous amount of data produced by shop floor processes. It seeks to detect the occurrence of disturbances in the behaviour of assets. As a result, maintenance managers can develop more effective action plans, maximizing the availability of assets at a lower operating cost. This approach is called condition-based maintenance (CBM), as maintenance is driven by the current state of the machines, measured from principal variables for monitoring the health of the systems and their components. (PENG et al., 2010)

There is a range of reviews in the literature in this area that deal with CBM techniques and their applications in the industry. One of the pioneer reviews to address the topic divided the CBM techniques into three main groups: data acquisition, data processing, and maintenance decision making. (JARDINE et al., 2006) More recently, another review presented a full view of prognosis(PENG et al., 2010), which is the data processing phase related to estimating remaining useful life. Also, a more recent review has restricted analysis to statistical approaches for prognosis. (SI et al., 2011) An update review that includes all stages of a CBM system, from data acquisition to estimating remaining useful life, has been presented recently. (LEI et al., 2018) Yet another review relates the CBM process to maintenance and company management, supporting decision-makers' actions. (BOUSDEKIS et al., 2018) Finally, a review focused on deep-learning methods applied to monitoring machine health is presented. (ZHAO et al., 2019) However, to date, no review has been found specifically addressing the application of these methods in the hydroelectric sector, which has specific characteristics and complexities. Thus, the present article provides a systematic review in an area not yet comprehensively reviewed.

The remainder of the present article is organized as follows. Section 2 describes the study methodology and the systematic literature review process, and presents a qualitative summary of the articles sampled. Section 3 presents the failure modes found most frequently in hydroelectric systems (HS). Section 4 summarizes the monitored variables in CBM applications, associating them with the recurrent failure modes. Sections 5 and 6 discuss the models proposed so far for dealing with the diagnosis and prognosis of HS and their components. Finally, Section 7 presents the conclusions and recommendations for future work.

2.2 Materials and methods

2.2.1 Review methodology

A systematic literature review (SLR) is a replicable and unbiased procedure applied to identify and select representative literature to answer a research question and its sub-questions. (BABATUNDE et al., 2017) The present study aims to answer the following question: “How has the hydroelectric sector been benefiting from the latest advances in CBM techniques?”. The research question comprises three sub-questions:

- Sub-Question 1: What are the most common failure modes in HS, and what are the monitored variables associated with them?

- Sub-Question 2: Which attribute extraction tools have been used to enhance CBM systems?

- Sub-Question 3: What statistical and computational methods have been applied to the diagnosis and prognosis of HS?

The methodology adopted for conducting the SLR consists of a three-step procedure. (GLOCK; HOCHREIN, 2011; HOCHREIN; GLOCK, 2012) The first step is to define the list of relevant keywords that will be used to search peer-reviewed journals in online literature databases. Table 2.1 summarizes the keywords adopted in the present paper. The first set of keywords relates to the context of predictive maintenance; the second set refers to hydroelectric plants and components. The keyword list has been iteratively expanded to include synonyms and frequently used terms. We searched the scholarly databases *Scopus* and *Web of Science* for peer-reviewed articles featuring these keywords, either in their titles, abstracts, or lists of keywords. Only articles published in the English language during the period between 2010 and 2019 were considered. After removing duplicates, the total number of articles is 118.

Keywords
Predictive maintenance keywords: predictive maintenance OR condition-based maintenance OR fault detection OR diagnosis OR remaining useful life OR health monitoring AND *Hydroelectric keywords*: hydroelectric OR hydropower OR hydro generator OR hydro turbine

Tabela 2.1 – List of relevant keywords adopted in searching journal databases.

The second step is to check the articles' relevance by screening their abstracts. If the abstract indicates that the paper might be relevant for this review, a detailed analysis of the entire article is carried out. Articles that do not deal effectively with the topic are removed from the sample at this stage. The third step is to conduct a backward and forward snowball search, examining relevant articles cited in our sample.

Phase	Description	Total
Identification	Records identified through database searching	176
Screening	Records after duplicates removed	118
Eligibility	Full-text articles assessed for eligibility	88
Included	Studies included in quantitative synthesis	80
	Studies included in qualitative synthesis	71

Tabela 2.2 – Review protocol and sample sizes by stages.

2.2.2 Descriptive analysis

Table 2.2 presents the review protocol adopted, with the number of articles at each stage of the SLR process. In the end, the study sample consists of 80 articles. Figure 2.1 shows how the number of publications has been developing over time. There is a significant

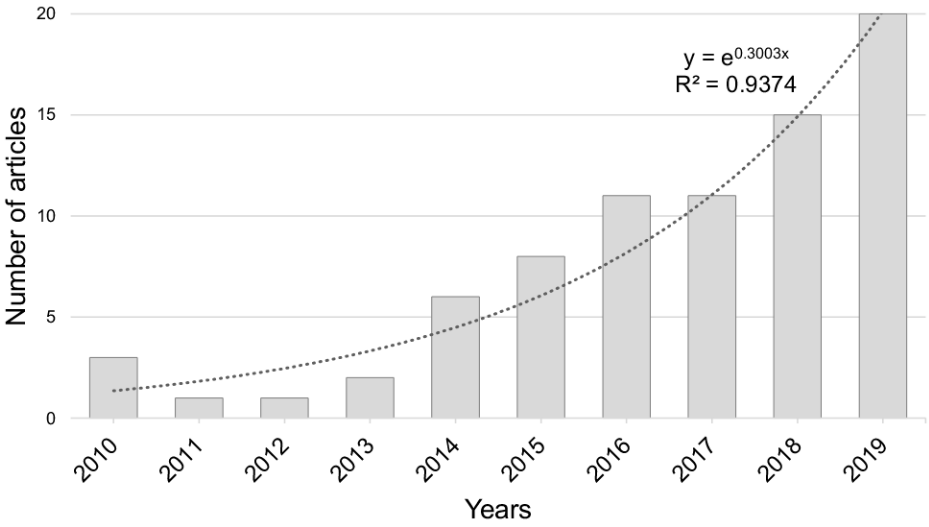


Figura 2.1 – Number of articles published annually.

increase in the number of publications in the sector during the last decade, increasing from fewer than 5 to 20 published articles per year in 2019.

Figure 2.2 presents the total number of publications per journal, grouped by categories. (SCImago, 2020) The categories were grouped into three major clusters, according to their area: the first cluster related to computer science and engineering journals; the second to energy; the third to materials science, mathematics, and physics. In general, the publications are scattered among several journals, with no single journal publishing more than three articles on the subject.

Due to the large number of articles in our literature sample, the present review adopted a strategy of associating the articles with categories belonging to a conceptual framework. This framework was adapted from other literature reviews focused on CBM(BOUSDEKIS et al., 2018; SHIN; JUN, 2015; PENG et al., 2010; JARDINE et al., 2006), in which the categories are consolidated according to the stage of the maintenance process, from data acquisition to useful life estimation. Exemplary publications illustrate the categories, instead of merely listing and discussing the sampled articles. However, the descriptive analysis comprises all articles sampled.

Figure 2.3 presents a temporal word cloud generated from the titles and abstracts of all articles sampled using the VOSviewer 1.6.15 software (ECK; WALTMAN, 2010). From the binary count of multinomials with frequency greater than three, we considered only the 60 % most relevant terms. The circle sizes represent the occurrence of the terms, the arcs denote the strength of the associations between them, and the colour shows the yearly average of the occurrences of the terms. This representation presents a general idea of the categories addressed in the next sessions, providing a global view of the study sample.

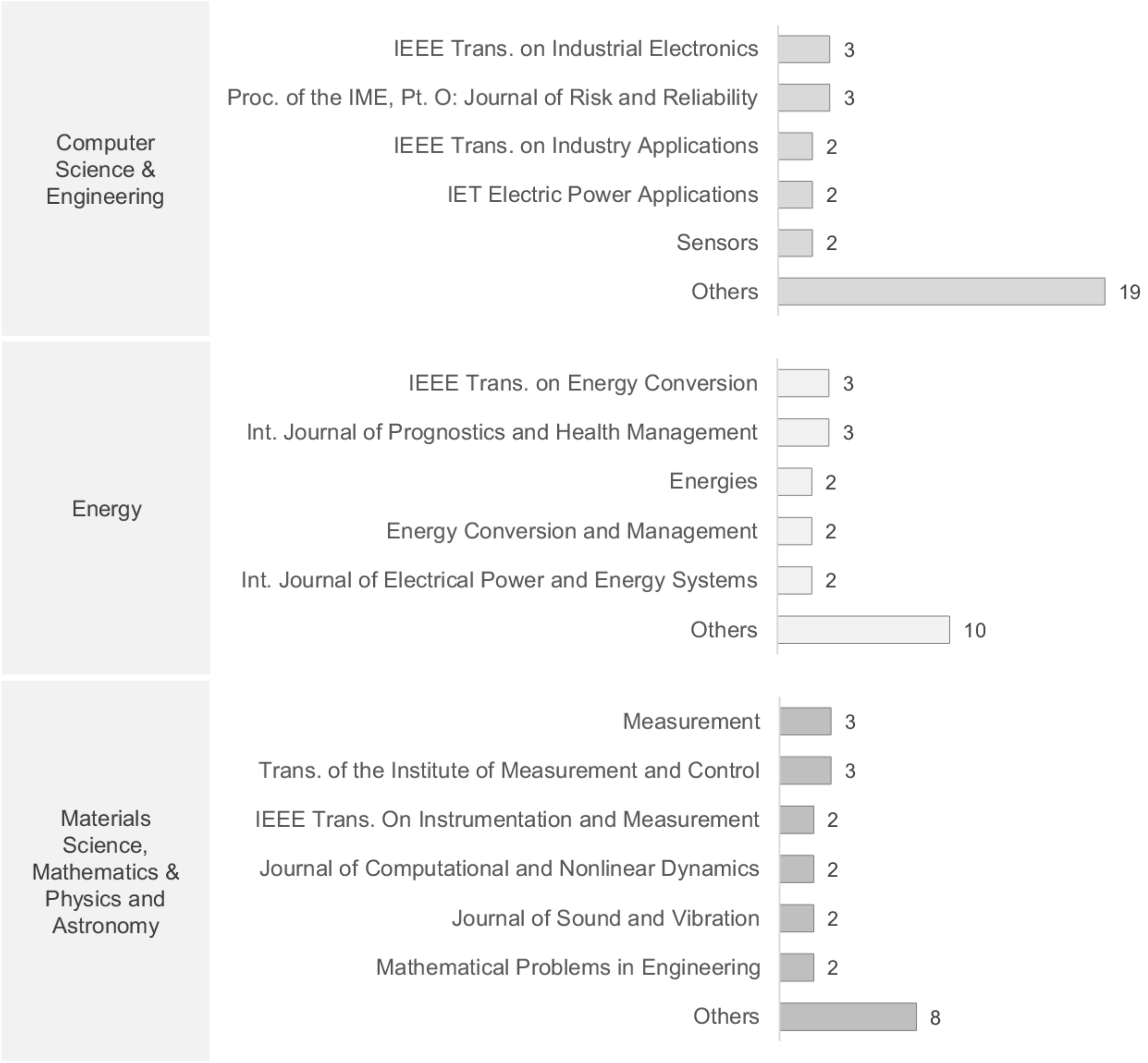


Figura 2.2 – Journals with more publications on the topic, grouped into categories defined by (SCImago, 2020). Journals with only one publication have been omitted.

It is noticed that vibration signal monitoring, applied to feature extraction techniques and computational intelligence models, has been appearing with increasing frequency in the latest publications in the area. On the other hand, mathematical formulation and the finite element method call attention to an additional cluster of publications. In the thematic analysis, we seek to elucidate all these categories in an organized and systematic way.

2.3 Common failure modes in hydroelectric plants

Failure modes differ from plant to plant, related to environmental and design factors, plant requirements, type of turbine, and operation. However, in general, some types of failure are more common in all hydroelectric plants. Below, we present the types discussed most frequently in the sampled literature.

2.3.1 Cavitation

Cavitation is a complex and harmful phenomenon for hydraulic machinery such as turbines, pumps, and valves. Sudden changes in the local pressure of the liquid form bubbles that collapse, radiating acoustic energy waves and causing the erosion of nearby surfaces. (GREGG et al., 2017) Sand erosion increases the likelihood of cavitation, since eroded surfaces increase wall turbulence and, consequently, reduce the local pressure. (EGUSQUIZA et al., 2018)

Cavitation is more likely in Francis turbines and reversible pump turbines than in Kaplan turbines. There are several types of cavitation such as leading-edge, traveling bubbles, draft tube swirl, inter-blade vortex, Karman vortex and tip vortex (only Kaplan turbines). (VALENTÍN et al., 2018)

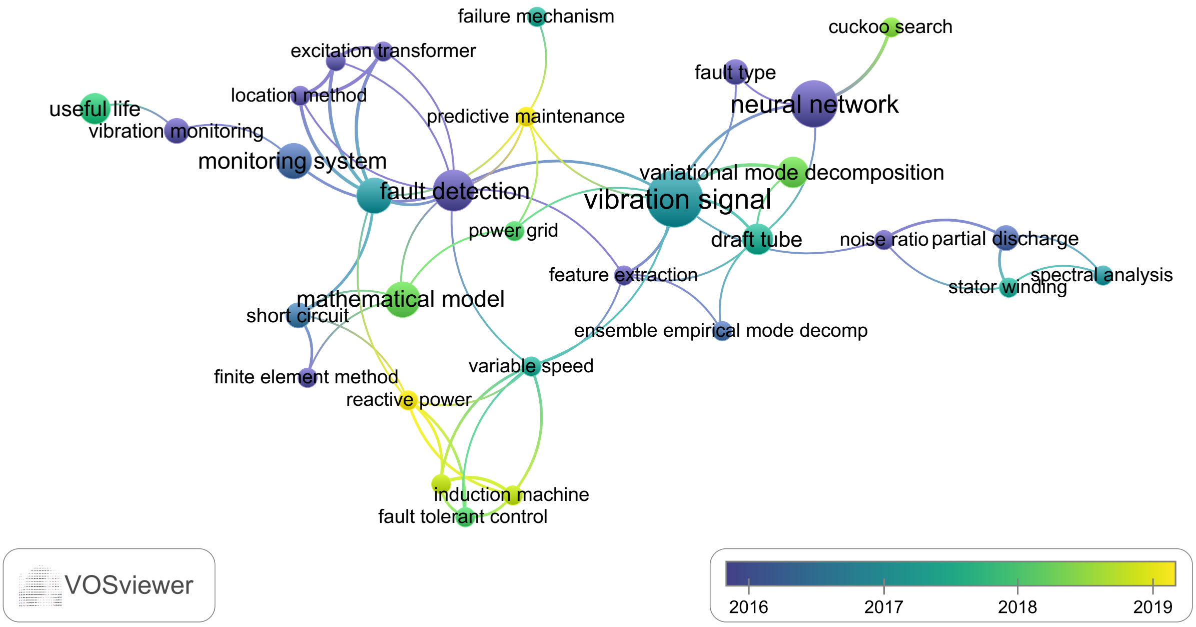


Figura 2.3 – Temporal word cloud created from the titles and abstracts of the sampled articles.

2.3.2 Loss of excitation

Loss of excitation is widespread in synchronous machines and, alone, accounts for 70% of all generator failures. The phenomena are caused by short circuits of the field winding, unexpected field breakers or relay failures. It can increase rotor speed, causing excessive vibration and bearing overheating. Additionally, as the generator operates as an induction machine, the loss of excitation of one piece of equipment can impact the whole system, decreasing active power and increasing reactive power output, which may even result in the collapse of the entire interconnected system. (AZIZ et al., 2017)

Loss of excitation is usually enhanced by short-circuit faults of the rotor winding of the synchronous generator, which can also lead to the rotor grounding and shaft magnetization. While short-circuit failures are frequent and occur in most hydro-generators, in the long run, this type of failure causes an increase of the excitation current and, consequently, of the rotor temperature. These effects cause an unbalanced thermal distribution of the rotor magnetic poles that increase the incidence of short-circuit failures and compromise the reliable operation of the generator. (LI et al., 2019)

2.3.3 Partial discharge

Partial discharge is the name given to electrical micro-discharges generated in the insulating structure when subjected to high-intensity electric fields. The diagnosis of partial discharge allows accurate assessment of the degree of insulation degradation of the generating system. (KANEGAMI et al., 2016) These discharges can partially or entirely break down the insulation between conductors. The phenomena produce physical indicators such as light flashes, acoustic noise, temperature gradients, chemical reactions, and electromagnetic pulses. (OLIVEIRA et al., 2016)

The defects originate from aging deterioration, moisture pollution, or inadequate design. In generators, defects can be due to gaps in the ground-wall insulation or to degradation of the corona shielding. The identification and source separation of these kinds of events are complex tasks, and require intense adoption of pulse shape analysis and statistical/artificial intelligence techniques. (BORGHETTO et al., 2004)

2.3.4 Shaft, bearing, and other components failure modes

Shaft misalignment is a significant problem in hydro-power systems, causing almost 60% of the failures in rotating machines. This defect may lead to a series of vibration patterns that are adverse to steady and safe operation, contributing to accelerated wear of the components, shaft deformation, and deflection of the shaft coupling. (XU et al., 2018) Misalignment is not an exclusive fault of the shaft; it can also be present in guide vanes, runner blades or rotors. (WU et al., 2016)

The bearings and the lubrication system are responsible for absorbing part of the energy from the shaft rotation. If there is misalignment, the absorbed energy will be converted into thermal energy and overheat the bearings, which ends up reducing the life of the components and leading the cooling system to fail. Unbalanced magnetic fields can create a magnetic flux surrounding the shaft which, coupled with the associated bearing currents, increases the wear on the guide bearings.

Each turbine-generator auxiliary system presents specific failure modes and specific monitoring variables such as the cooling and lubrication system (SELAK et al., 2014), turbine governor (GUO et al., 2010), power converter (JOSEPH et al., 2019), servovalve (YU; BREIKIN, 2009) and pressure tubes (MAZZOCCHI et al., 2016). All these failures detrimentally impact the hydroelectric operation. Some studies model failure modes by sub-systems, and present them in an organized and interconnected way through hierarchical models. (JONG; LEU, 2013; MELANI et al., 2016; CHENG et al., 2019b) While we highlight the phenomena recurring the most frequently in the literature, we recommend consulting these studies to comprehend the failure modes by sub-systems and their interactions.

2.4 Data acquisition

Data acquisition is the capturing and storing of monitoring data from several sensors installed in the monitored asset. Below, we list the sensors and variables monitored in the hydroelectric sector, associating them with the types of failure modes.

Table 2.3 presents a detailed summary of CBM systems. The publications were grouped by monitored objects and variables, listing the failure modes that the systems can identify. The systems were assigned to one of the following contexts, depending on the nature of the monitored variables: air gap eccentricity, electrical signature analysis, multi-source, structural health, or vibration monitoring.

The main object of interest for monitoring is the turbine-generator system through vibration. However, the air gap eccentricity and electrical signal variables are intrinsically associated with the health of the rotor and stator winding, which is a specific component of the generator. Most applications are of the diagnosis type, although there are some recent studies of prognosis in vibration monitoring. The following subsections detail the main ways of monitoring and acquiring data in CBM systems in the hydroelectric sector.

2.4.1 Vibration signal

Vibration monitoring is the technique used the most in hydroelectric CBM. It can detect mechanical, hydraulic, and electromagnetic related problems, which impact hydro turbine-generator sets. It is estimated that more than 80% of failures and accidents in generating units are detected through vibration monitoring, making it an essential variable of interest for identifying damage to equipment. (CHENG et al., 2018a)

Authors	Context	Object	Failure mode	Variables monitored	T
(VALAVI et al., 2018)	Air gap	Rotor winding	Inter-turn short circuit	Air gap flux density, phase voltage	D
Griscenko2015EccentricitySpectrum, Babic2017FaultHydrogenerator, Dirani2018ImpactHydro-generator	Air gap	Stator winding	Magnetic unbalance	Magnetic flux and vibration spectrum	D
(RAMÍREZ-NIÑO et al., 2015)	Electrical	Generator	Impendence asymmetry between phases, mechanical defects	Neutral current	D
(AZIZ et al., 2017; JOSEPH et al., 2019)	Electrical	Generator	Loss of excitation, power converter failure	Terminal voltage and stator current	D
(BLANQUEZ et al., 2015; BLANQUEZ et al., 2016; PARDO et al., 2016)	Electrical	Rotor winding	Ground fault	Field-winding voltage and grounding voltage	D
(DALLAS et al., 2011; CARVALHO et al., 2015; OLIVEIRA et al., 2016; SALOMON et al., 2019a)	Electrical	Stator winding	Partial discharge	Voltage from different phases and points of measurement	D
(GUO et al., 2010)	Electrical	Turbine governor	Defective components	Current, frequency, gate displacement	D
(XU, 2013)	Multi-source	Generator	Winding, electromagnetic, structure, oil cooling failures	Current, voltage, power (active, reactive), insulation resistance, temperature, temperature oil, vibration, sound	D
(BLANCKE et al., 2018)	Multi-source	Generator stator	Partial discharge, erosion, insulation degradation, etc.	Expert knowledge and diagnostic data	P
(WU et al., 2016)	Multi-source	Turbine	Cavitation, mass unbalance of the rotor, oil whirl, vortex in draft tube, rotor misalignment, guide vane uneven, and runner blade uneven	Governor, excitation, vibration, ground current, pressure, voltage	D
(CHENG et al., 2019b; XU et al., 2019)	Multi-source	Turbine	Several	Several	D
(SELAK et al., 2014)	Multi-source	Thrust bearing	Overheating, lubrification consumption, cooling system failure, degradation	Output power, rotation frequency, temperature, oil level, oil temperature, velocity	D

Tabela 2.3 – Summary of CBM models applied to the hydroelectric context, including monitored object, variables and type of application - (D) Diagnosis, (P) Prognosis.

(KLUN et al., 2019; MATEJA et al., 2020)	Structural	Dam and bearing structure	Hydraulic faults, fatigue	Vibration signal	D
(MAZZOCCHI et al., 2016)	Structural	Pressure tunnels and shafts	Wall stiffness drop	Pressure wave reflections	D
(MILIC et al., 2013)	Temperature	Rotor poles	Overheating	Temperature by infrared radiation	D
(KANEGAMI et al., 2016)	Temperature	Stator winding	Partial discharge	Resistance-temperature sensor readings	D
(LU et al., 2018; WANG et al., 2019)	Vibration	Draft tube	Vortex strip	Upper/lower guide bearing vibration, turbine guide vibration	D
(PENG et al., 2007; CHENG et al., 2014; ZHU et al., 2014; XIA et al., 2015; XIA; NI, 2016; XIA et al., 2017; CHENG et al., 2018a; FU et al., 2019)	Vibration	Generator	Rotor unbalance, rotor misalignment, rubbing, movement collision, and vortex draft tube, karman vortice	Vibration spectrum	D
(LUO et al., 2010; XU et al., 2018)	Vibration	Generator	Shaft misalignment, mass unbalance	Displacement (orbit), water head, turbine flow, guide vane opening, rotation speed, generator rotor	D
(PINO et al., 2018)	Vibration	Guide bearing	Degradation	Vibration displacement (orbit)	D
(AN et al., 2014)	Vibration	Generator	Degradation	Upper bracket horizontal vibration, active power, working head	P
(GREGG et al., 2017; VALENTÍN et al., 2018)	Vibration	Turbine	Cavitation	Vibration and acoustic emissions	D/P
(XUE et al., 2014; QIAO; CHEN, 2015)	Vibration	Turbine	Mechanical faults	Lower bearing vibration and draft tube pressure	D/P
(ZHANG et al., 2012)	Vibration	Turbine	Several (mechanical, electrical, hydraulic)	Vibration from upper/lower guide bearing, water pilot bearing, upper bracket	D
(AN et al., 2017a; AN et al., 2017b; ZHOU et al., 2019)	Vibration	Turbine	Vortex	Shaft vibration, lower guide vibration	D

Tabela 2.3 – Summary of CBM models applied to the hydroelectric context, including monitored object, variables and type of application - (D) Diagnosis, (P) Prognosis.

Vibration monitoring has a broad range of applications in the generator system, since it can detect anomalies associated with mechanical, hydraulic, and electrical failures. Examples of failure modes usually detected using this technique are cavitation, rotor unbalance, rotor misalignment, rubbing, movement collision, vortex draft tube, vortex strip, and Karman vortices. Nevertheless, the broad range of vibration monitoring applications can be a notable drawback, as it does not clearly inform what type of failure is occurring. For this reason, it is common for other forms of monitoring to be used in conjunction with vibration monitoring.

To measure vibration, accelerometers and acoustic emission sensors are placed in different parts of the machine such as the guide vanes, turbine bearings, draft tubes, or shafts. Each location presents advantages and drawbacks in detecting different types of cavitation. (VALENTÍN et al., 2018) Another form of vibration analysis is shaft orbit monitoring, in which two sensors are placed 90° apart. This arrangement allows description of the movement of the shaft centre, extracting geometric, time-domain, frequency-domain, moment, and angle characteristics. This type of vibration monitoring is usually adopted to identify shaft misalignment, mass imbalance, and degradation, as found in several studies. (LUO et al., 2010; XU et al., 2018; PINO et al., 2018; JABLON et al., 2020)

Frequency decomposition, in harmonics, allows simultaneous evaluation of the health of several components. For example, in a Kaplan turbine composed of thirteen blades, operating at a frequency of 2.73 Hz, each blade is associated with a frequency proportional to $13 \times 2.73 \approx 35.5$ Hz. The vibration data acquisition process must allow an acquisition rate that is high enough to capture the natural frequencies of its components.

When a unit runs under part-load conditions, the turbine cannot achieve optimum flow of the runner inlet and outlet, thus creating a vertex rope in the shaft system. During these unstable operating conditions, the vibration signals are too complex to predict, and damage is more likely to occur to the runner and draft tube system. (AN et al., 2017a) From laboratory testing, it was estimated that each start and stop procedure causes fatigue damage equal to 15-20 hours of stationary operation. (KLUN et al., 2019) To better understand the vibration behaviour during different operating states, parameter conditions variables (or operating conditions) are often adopted, using linear models. Examples of these parameters are active and reactive power, distributor opening, and bearing temperature. (LUCIFREDI et al., 2000) An example of a parameter condition is vibration analysis, together with the rotation speed for diagnosing different failure modes. (XIAO et al., 2014)

2.4.2 Air gap eccentricity

Air gap eccentricity is another object of interest in hydro-power generation. It allows the identification of several causes of failures like unbalanced inner forces, stator core shifts, rotor ovality, defects of stator lamination. This variable measures the space between the spinning rotor and the stationary rotor in a generator unit, through the application of contacting probes or proximity sensors.

The air gap monitoring system assesses rotor eccentricity and can identify shorted turns on the rotor pole winding. Static eccentricity is associated with the wrong positioning of the rotor or stator during operation or assembly. In contrast, dynamic eccentricity is associated with thermal expansion, bearings wear, shaft line bend, and rotor displacement by higher magnetic forces. Before air gap analysis, the standard way to determine the existence of shorted turns was the pole drop test. This test required stopping and partially disassembling the generator unit, and measuring the voltage drop across each pole (BABIĆ et al., 2017). With the recent developments of measuring systems, air gap online monitoring is now possible through the introduction of flux sensors on the stator core teeth.

The main types of measuring systems use: (1) contacting probes in no-load mode which, although precise, is not suitable for continuous monitoring since it requires stopping the generator and running the tests manually; (2) non-contacting capacitive proximity sensors, widely adopted and commercially available; and, (3) non-invasive measuring systems, which present enormous potential but are still in development. Recent experiments with slow-speed generators indicate that the proximity sensors provide measurements almost as precise as the probe sensors. (GRISCENKO; ELMANIS-HELMANIS, 2015)

Air gap monitoring is not proposed as a stand-alone application, but as a complementary source of information in integrated, multi-parameter CBM systems. An algorithm could more efficiently identify excess vibration from magnetic imbalance, rather than only mechanical or hydraulic imbalance, thereby increasing the precision of the system and its false-positive rate. The similarities in the spectra of the variables evidence the connection between the air gap and vibration variables. However, further investigation in future studies and the definition of reliable evaluation criteria of the air gap spectrum is required. (GRISCENKO; ELMANIS-HELMANIS, 2015)

2.4.3 Electrical signature analysis

Electrical signature analysis evaluates the current and voltage profiles of a generator in the frequency domain. It is a non-invasive technique that has been applied increasingly to CBM in hydro-electrics, to detect inter-turn short-circuit, air gap eccentricity and rotating diode failures. As it depends only on electrical measurements, the method has high technical and economic feasibility. (SALOMON et al., 2019a)

Partial discharge was one of the first failure modes associated with electrical monitoring. The voltage spectrum of different phases undergoes cross talk interference, that can be overcome by clustering the partial discharge pulses according to shape similarity. (BORGHETTO et al., 2004) Using signal decomposition techniques, partial discharge pulses can be automatically decomposed and the denoising can be evaluated from the time shift difference and noise threshold levels. (CARVALHO et al., 2015) These approaches can better filter out wide-band noise and significantly reduce background interference with the partial discharge measurement of hydro-generators.

The inter-turn short-circuit diagnosis also benefits from the development of systems based on the electrical signature. The spectral analysis of stator voltage and current can be applied to detect early stage, rotor inter-turn faults. (VALAVI et al., 2018) This is possible since some of the signal harmonics amplitudes increase only when this kind of fault develops.

Several factors such as over speeding, vibration, excessive field currents, reduced cooling, and temperature rise expose the field winding to abnormal mechanical and thermal stresses. These stresses lead to breakdown of the insulation of the field winding and the rotor iron at points where stress is maximum, thereby generating a ground fault. While a single ground fault does not represent any immediate danger, high currents and mechanical imbalances can severely harm or even melt the rotor if a second fault arises. (BLANQUEZ et al., 2016)

2.4.4 Temperature

Temperature sensors are most commonly found in power generation systems, presenting significant advantages such as stability, repeatability, and accuracy. Temperature variations are excellent indicators of impending failure conditions. In generator systems, temperature monitoring is usually associated with bearing monitoring: the bearing being the machine component that supports shaft rotation. In the event of failure of the lubrication system or defect in the shaft (i.e., misalignment, vibration overload, or speed), the bearings absorb the thermal overload and prevent damage to vital components of the system.

In the design of a generating unit, the maximum operating temperatures are defined from technical test simulations. Generation is stopped as soon as the limit is reached. However, temperature monitoring has low latency, which makes it reactive. Frequently, once the temperature trip alarm is triggered, the fault (or set of faults) has already occurred. A recent solution adopted contactless infrared detector measurements for online monitoring of the surfaces of water-cooled rotors poles. (MILIC et al., 2013) From the time-frequency analysis of the resistance-temperature sensor, the stator winding discharge detection can be improved, as the phase angle can aid in distinguishing signals from noise. (KANEGAMI et al., 2016)

Temperature is also adopted as a condition parameter for estimating other variables. A three-dimensional mathematical model relates the temperature, thermal deformation, and thermal stress of magnetic poles fields, in the rotor winding inter-turn short circuits.

The shorted turns decrease the temperature of magnetic poles, indicating that diagnosis can be obtained by monitoring the temperature change of the rotor. (LI et al., 2019)

2.4.5 Structural health monitoring

Structural monitoring assesses the health of the structures that constitute a hydroelectric generation system, such as the powerhouse or the dam. This is vital for preventing structural damage that could collapse the entire system. The effects of dam failure, for instance, have substantial social and environmental costs, which makes structural monitoring so critical and necessary.

Vibration monitoring is commonly associated with structural assessment. In hydroelectric structures, vibration is also applied to diagnose critical structural components. A two-step model can identify the modal order and the characteristic of dams under operation, with the dynamic response of the hydraulic structure excited by fluctuations in flow load. (LIAN et al., 2009) Through modelling the interaction between the unit shaft system and the powerhouse structure during transient, sudden load increasing process, it is concluded that the generator floor structure is more susceptible to the transient process and to excessive vertical vibration. (ZHANG et al., 2019)

The laser Doppler vibrometer (LDV) is a non-contact sensor. It was developed to measure the amplitude and frequency of surface vibration by analysing the reflected laser beam frequency applied to the surface of interest. The use of LDV, under transient conditions within the concrete dam monitoring context, can contribute to the elimination of pseudo-vibrations and noise from measures inherent in the non-stationary process. (KLUN et al., 2019) A low-level reading of instrument noise is obtained by placing the sensor inside the powerhouse, as regular accelerometers are sensitive to magnetic field excitation. Some solutions such as the use of reflective tapes, adoption of standing points that are more rigid than the observation point, and instrument visor shading are proposed to minimize ambient noise. (MATEJA et al., 2020)

2.4.6 Multi-source

While most of the work in the CBM area is related to monitoring a specific type of variable, there is a tendency to develop models that simultaneously monitor variables of different natures. This monitoring process, taking input from multi-modal sensors, is known as sensor fusion. It seeks to develop collaborative distributed systems. (XUE et al., 2008)

Some studies have been successful in applying multivariate monitoring systems in the context of hydroelectric maintenance. An example is the control system based on the combine input of twelve different types of sensors, such as accelerometers, inductive displacement sensors, inductive switches, pressure sensors. In total, 108 attributes were extracted and used to create a classification model of approximately 97% accuracy. (SELAK et al., 2014) Other applications have applied nineteen variables such as tank level, rotor and bearing temperature and vibration, excitation current and voltage, runner speed, among others. The model diagnoses seventeen failure modes, hierarchically grouped in the bearing, rotor, and stator sub-systems, and sequentially grouped in two root nodes: dynamo system and hydro turbine. (CHENG et al., 2019b)

In the context of structural health monitoring, several factors can influence the behaviour of the system. Hydro-power dam dislodging, for instance, is affected by different elements such as dam maturing, store water level, air, water, and stable temperature, which cause complicated, nonlinear behaviour that is hard to foresee. Additionally, natural external factors such as earthquakes and ice pressure interfere with the structural monitoring models and reduce their accuracy. A multivariate approach considered a set of these external variables: air temperature, water temperature, concrete temperature, displacements between dam blocks, inclination of dam blocks, uplift water pressure and underground water pressure. (HAMZIC et al., 2020) The model presented accuracy in the short term; however, the biggest limiter for the long term was the climatic forecast, especially concerning precipitation and air temperature, which directly influence the water level and the concrete temperature.

2.5 Feature extraction

Among the feature extraction techniques found in the literature, fast Fourier transformation (FFT) and wavelet transform (WT) are the most commonly used for feature extraction. They are useful for transforming signals from the time domain to the time-frequency domain. The magnitude and phase signal decomposition of each frequency component can contribute to a set of fault patterns for machine diagnosis: a detection system can promptly identify faults by monitoring the increase of the values of certain higher harmonics in the signal spectrum. Examples of applications that have adopted FFT for feature extraction can be found in the literature. (BABIĆ et al., 2017; WU et al., 2018; KLUN et al., 2019; XIA et al., 2015; XIA; NI, 2016)

Both FFT and WT present a significant limitation, which is the need to determine the specific parameters beforehand. In the case of non-stationary vibration, these parameters are mostly unknown, and more flexible methods have been proposed. Intrinsic time-scale decomposition (ITD), empirical mode decomposition (EMD), and the ensemble of empirical mode decomposition (EEMD) are all self-adaptive signal decomposition methods proposed for analysing nonlinear signals. The application of ITD with a classification algorithm has shown better results than the application of the classification algorithm. (AN et al., 2014)

This result boosted the development of several versions of EEMD, such as the

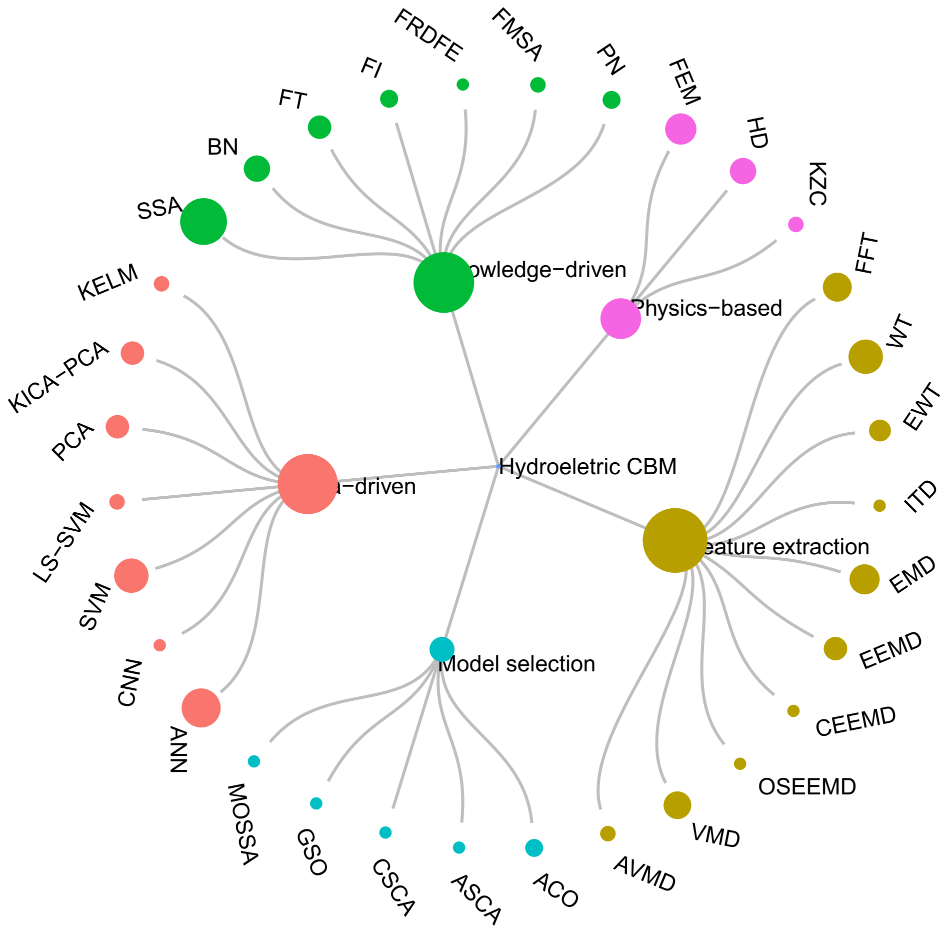


Figura 2.4 – Dendrogram with the most used techniques in CBM models for HS. The size of the nodes represents the binary occurrence of the terms in sampled articles.

noise-assisted method complementary ensemble empirical mode decomposition (CEEMD) and the over-sampling ensemble empirical mode decomposition (OSEEMD), to obtain more accurate decomposition sets while keeping computational costs at a minimum. (XUE et al., 2014) The adaptive local iterative filtering (ALIF) method uses an iterative filtering strategy with an adaptive, data-driven filter length selection to decompose the signal, inhibiting the mixing mode inherent in EMD. (AN et al., 2017a) More recently, empirical wavelet transform (EWT) was adopted to decompose the signal in multiple components. EWT presents higher accuracy mode estimation at significantly reduced computation time, compared to EEMD and EMD. (KEDADOUCHE et al., 2016)

Finally, variational mode decomposition(AN et al., 2017b) (VMD) and adaptive variational mode decomposition(FU et al., 2019) (AVMD) are pre-processing methods used to decompose the signal into a set of intrinsic mode components with limited bandwidth. The AVMD automatically determines the model number, based on the characteristic of intrinsic functions, using a set of indexes: entropy, extreme value, kurtosis criterion, and energy loss coefficient.

2.6 Diagnosis

2.6.1 Data-driven

In fault diagnosis applications using supervised learning algorithms, the data is labelled by specialists as either healthy or faulty. The labels can also be obtained using technical tests in which specialists design specific failure situations that seek to differentiate the algorithms. The algorithms can adopt a multi-class approach, seeking to determine not only if there is a failure, but also what type of failure it is such as misalignment, vortex with eccentricity, or shaft imbalance.

The learning algorithm most frequently found in our literature sample is the artificial neural network (ANN). This is a nonlinear model, widely used in the area of machine learning, that is capable of mapping fault symptoms to a set of source failures. A more elaborate architecture, that considers temporal dependency between observations, the application of 1-dimension convolutional neural networks (CNN), has been proposed. (LIAO et al., 2019)

However, there are some limitations to the application of ANN: the low speed of convergence and the high sensitivity to initial parameters. To circumvent these, some authors propose applying heuristic optimization algorithms such as the ant colony optimization(XIAO et al., 2015) (ACO) and the cuckoo search(CHENG et al., 2018a; CHENG et al., 2018b; CHENG et al., 2019a) (CS). The aim is to decrease the training instability and increase the generalizability and convergence speed of the model. Other machine learning models found in the literature are the support vector machine(XIAO et al., 2014; XIA; NI, 2016) (SVM) and the principal component analysis (GREGG et al., 2017) (PCA).

Failure diagnosis in hydroelectric plants can also be seen as a nonlinear, multivariate process. Conditions are monitored and faults are detected online if the process deviates from normal operating conditions. The kernel independent component analysis and principal component analysis (KICA-PCA) method is used for this, to extract and reduce the dimensionality of independent components. These are combined with the confidence limits of the Hotelling's T^2 and SPE statistics to evaluate the normal condition. (ZHU et al., 2014)

2.6.2 Knowledge-driven

Knowledge-based models are built from the input of experts and technicians, and seek to consolidate tacit knowledge in intelligent decision-making systems.

Spectral signal analysis (SSA) is one of the techniques most frequently applied by specialists to detect anomalies. This technique consists of analysing the harmonics that make up the signal. From their observations, the experts formulate basic operating conditions to be met. The latest developments in the area seek precisely to enable intelligent algorithms to learn to define them, with or without human intervention. The spectral analysis is applicable to vibration signals (AN et al., 2017a; AN et al., 2017b), neutral current (RAMÍREZ-NIÑO et al., 2015), air gap (GRISCENKO; ELMANIS-HELMANIS, 2015) and partial discharge (OLIVEIRA et al., 2016).

Fuzzy inference (FI) systems are capable of assigning a set of reference rules to represent the relationship between the fault phenomenon and the fault reason, in a concise and interpretable way. They can be applied either alone (XU, 2013) or together with machine learning models like, for instance, SVM (ZHANG et al., 2012; XIAO et al., 2014) or ANN (AZIZ et al., 2017). Fuzzy theory is widely applied in the industrial sector, adding artificial intelligence agents to the regulation and control of resource activities with the adoption of the Fuzzy Recursive Decision Feedback Extension(MINO-AGUILAR et al., 2014) (FRDFE) models.

Another knowledge-based approach to multi-fault diagnosis is the construction of system fault trees (FT) and their components. Failure probabilities are interrelated using logical AND and logical OR conditions in a tree hierarchy. The FT starts with the failure mechanism and is grouped into components, sub-systems, and, finally, the whole system. Subsequently, the calculated probabilities feed a Bayesian network (BN) in which the model receives input from maintenance experts. (JONG; LEU, 2013) In this framework, current advances seek to construct the BN model from the perspective of machine learning and the experience of specialists, into a model capable of expanding or reducing according to the size of the hydroelectric station and the requirements of maintenance personnel. (CHENG et al., 2019b)

2.6.3 Physics-based

Physics-based approaches are generally mathematical models built from the premise that there are underlying, deterministic phenomena that influence the generation system. The modelling is focused on a specific component (or group of components). The adoption of simplified models, such as the influence of bearing stiffness (BRITO et al., 2017) and hydraulic dynamics (ZHANG et al., 2019) on the monitored vibration, can generate satisfactory results when the operating condition is appropriately determined.

The stability modelling of a generator system is obtained from the vibration of the unit and conversion efficiency. It seeks to establish bases for the safe and stable operation of hydroelectric stations during the transient processes. A unified mathematical model for the sensitive analysis of turbines is approached from three aspects: hydraulic, mechanical, and electrical. The confidence interval of the variable is estimated from computational simulations. The new observations are monitored using the mathematical model and, if the confidence limit is exceeded, it is considered an anomaly. (XU et al., 2017; XU et al., 2018; XU et al., 2018)

The Kutta-Zhoukowski conditions (KZC) can be applied to the input and output velocity vectors and unbalanced forces to estimate the normality curves of the vibration and efficiency variables. (XU et al., 2019) In this type of model, a challenge arises from the sensitivity influence of the initial conditions on its performance. The Hamiltonian dynamic (HD) can also be used to describe the dynamic evolution of the energy produced, dissipated, and supplied in an operating, multi-generator system. (LI et al., 2018) Finally, a three-dimensional mathematical formulation of the temperature and thermal stress fields of the magnetic poles of the rotor can be used for stability estimation. The model is based on the theory of heat transfer and its resolution is obtained using the finite element method (FEM). Unlike previous models that acted generically, this one is specific to the type of rotor winding inter-turn short circuit failure. (LI et al., 2019)

Among the stability models of hydroelectric units, the application of computational intelligence methods for regression of the vibration and pressure variables, such as ANN and the least square support vector machine(QIAO; CHEN, 2015) (LS-SVM), is becoming more commonplace. The main advantage of these models is their ability to generate nonlinear mapping of the stabilization parameters, providing more accurate models for predicting the output parameters.

2.7 Prognosis

Prognosis seeks to estimate the useful life of an asset and establish a confidence interval for that estimate. In the hydroelectric context, prognosis consists of forecasting a given variable of interest, such as vibration, pressure, or the calculated health index, within a time-frame feasible for interventions in the system. An example of a prognosis system is based on the application of Shepard's interpolation of three variables, bearing vibration, apparent power, and working head, to construct the health index of the generator unit. Applying ITD, the signal is decomposed into a finite number of rotating components. An ANN is trained for each of the temporal components intrinsic to the signal, while the first order gray model predicts the trend of the series. Finally, the individual forecasts of each temporal component are summed together into a single forecast for the original series. (AN et al., 2014)

Later models present a similar framework, with varied individual methods. Signal decomposition can be obtained through the VMD, optimizing the meta-parameters using the least-square error index. The LS-SVM regression model can substitute for the ANN, and the model is fine-tuned using either the chaotic sine cosine algorithm (CSCA) or the adaptive sine cosine algorithm (ASCA). (W. WANG K., 2018; FU et al., 2019) The signal pre-processing, feature selection, and prediction steps can also be condensed into a single, multi-objective optimization framework. The EWT to decompose the signal into several modes, along with an entropy-based sample reconstruction strategy, refactor the modes. Variables are selected using the Gram-Schmidt orthogonal (GSO) process, and each series is extrapolated using the kernel extreme learning machine (KELM) method. A multi-objective salp swarm algorithm (MOSSA) adjusts the hyper-parameters of both the GSO and KELM models from the bias-variance indices. (ZHOU et al., 2019)

Other forms of prognosis can be developed from hybrid models involving knowledge- and data-driven methods. An example is the application of failure mechanism and symptoms analysis (FMSA) and Petri nets (PN) to predict the occurrence of degenerative states. This approach predicts the applicable time interval for maintenance tasks, based on the occurrence and propagation of known failure modes. (BLANCKE et al., 2018)

2.8 Discussion and conclusions

The present paper has provided a systematic overview of the state-of-art of CBM models for the hydroelectric sector. The discussion is summarized according to five categories: common failure modes, data acquisition, feature extraction, diagnosis and prognosis. Machine learning algorithms associated with time-frequency decomposition have been playing an important part in publications in this area in the last decade. The advantage of these models is that they do not require extensive human work or specialist knowledge, since the end-to-end structure is capable of mapping raw data with the associated failure classes. In addition, some research trends and potential future directions are given, as follows:

- *Multi-source data acquisition*: Vibration monitoring clearly predominates in the models proposed in recent years. Nevertheless, combining other variables such as temperature, electrical signature, pressure, and acoustic emission in multi-source systems is a trend in the research, given the capacity of these other variables not only to identify other failure modes that vibration does not capture, but also to help in classifying the type of failure. Studies associating the feature importance of monitored variables with the types of failure, like cavitation(GREGG et al., 2017) and partial discharge(ZEMOURI et al., 2020), can guide the design of new hydroelectric CBM systems.

- *Hybrid models*: The explainability of data-driven models, or machine explainability, offers the potential to provide insights into model behaviour using various methods such as visualization, feature importance scores, counterfactual explanation or influential data.(BHATT et al., 2020) This type of approach requires continuous interaction with specialists who have expertise in the knowledge domain, from the discrimination of attributes to the continuous feedback of the system, to articulate new anomalies as they arise. From the adoption of simple mathematical models and expert judgment, the model shows great improvement in its accuracy.

- *Deep learning techniques*: Machine learning models currently predominate in the hydroelectric CBM models. In the next decade, it is expected that the application of deep learning techniques will become more common in the area. (ZHAO et al., 2019) These techniques may include auto-encoders, restricted Boltzman machines, convolutional neural networks and recurrent neural networks. In recent years, due to their high accuracy in large-scale machinery datasets(SI et al., 2011), these techniques have been widely applied in the context of asset health management.

- *Health management and prognosis:* Reports on the prognosis of hydroelectric generating units are still scarce in the literature. Most studies present a very restricted framework for estimating the useful life of the generating unit. For example, there is a range of statistical methods such as the Wiener and Gamma process, also the stochastic filtering-based, hidden Markov models, that are used in prognosis and could be applied to this specific problem. Another important challenge in the area is to propose approaches that consider the interaction among faults between different generating units and auxiliary systems interconnected in the same generation system.

In conclusion, development of CBM technical applications in the energy sector is a trend that has been evident in recent years. It is gradually transforming the entire sector in the Industry 4.0 context. With the maturing of the different monitoring types, i.e., electrical signature and structural monitoring, it is natural for diagnostic systems to take the next step toward prognosis. The next step in the development of maintenance systems does not depend on the adoption of a single technology but on the interactions between intelligent systems and human specialists, complementing each other’s strengths in striving toward a common goal.

3 MAINTENANCE AND OPERATION PROFILE OF SMALL HYDROELECTRICS: IDENTIFYING THE CAUSES OF UNAVAILABILITY IN POWER GENERATION

Abstract: Developing and maintaining renewable energy sources are vital for sustainable growth in the coming decades. Despite the great increase in the number of publications on predictive maintenance, no study has been found in the literature that presents the application of specific models for small hydroelectric plants. A case study is presented of an energy operator specialized in the operation and maintenance of small hydroelectric plants in Brazil, in which a questionnaire was applied to executives and technicians. An exploratory analysis of a one-year operation was carried out, in order to describe the qualitative and quantitative maintenance profiles of 42 plants in Brazil. The potential to improve the availability of assets was demonstrated by reducing downtime for planned and forced maintenance, which accounts for an average of 47 days (13.0%) per year of plant operation. The components that failed the most frequently were the bearings (15.1%) and hydraulic units (11.2%). The results of the present study show the feasibility of applying predictive maintenance models to plants, which could benefit from lower operating costs and greater equipment availability.

Keywords: Industry 4.0. Predictive maintenance. Small hydroelectric plants. Qualitative profile. Exploratory analysis.

3.1 Introduction

The development of renewable energy sources is essential, in order to guarantee energy supply in the coming decades. According to the World Energy Council (WEC, 2019), energy demand is expected to double by 2060, while renewable energy sources represent three quarters of installed capacity in the same year. Among the clean energy sources already installed, the most used types are those related to the application of water resources. While the growth rate of large hydroelectric plants has maintained a moderate pace over the past fifty years, there has been gigantic growth in small hydroelectric plants (SHPs), which have become economically viable on the world stage. This viability is attributed to the decrease in the initial investment required to build a SHP and the decrease in operating costs, in addition to the development of regulation of energy markets by government and private agents. It is estimated that the potential for the total generation capacity of new SHPs is 78 GW, which represents only 36% of the total potential generation currently existing worldwide (LIU et al., 2016).

Management of the maintenance of a plant is a complex task, that requires a certain level of expertise in order to ensure a satisfactory level of reliability of the asset during its useful life (JARDINE et al., 2006). There are three types of maintenance. The first and most rudimentary is corrective maintenance, in which a component is expected to break and is replaced. The second is preventive maintenance, in which estimates of the expected service life of a component are made, and replacement is done based on a calculated value of operating time. Third, in predictive maintenance, the condition of the system is calculated using data obtained from various sensors, periodically or continuously (BOUSDEKIS et al., 2018; PENG et al., 2010).

Predictive maintenance, also known as condition-based maintenance, consists of two main phases. The first phase is the diagnostic phase, which includes everything from the detection of a fault or abnormal operating condition to the isolation by sub-components and identification of the nature and extent of the fault (PENG et al., 2010). The next phase is the prognostic phase, in which the application of statistical and machine learning models make it possible to estimate the remaining useful life of equipment. This provides a confidence interval of the prediction (SIKORSKA et al., 2011), thereby anticipating the maintenance and increasing the reliability and availability of the energy generation system. Examples of commonly applied methods for estimating remaining useful life are divided among statisticians (SI et al., 2011), which include: regression methods, Wiener process, gamma process, based on Markovian processes; machine learning methods, such as neural networks, vector support machine, principal component analysis (LEI et al., 2018); and, more recently, deep learning techniques, such as auto-encoder, recurrent and convolution neural networks (ZHAO et al., 2019).

Several other energy sectors already make extensive use of these techniques, such as the nuclear (AYO-IMORU; CILLIERS, 2018; LI et al., 2018; WU et al., 2018), wind (MÁRQUEZ et al., 2012; TIAN et al., 2011) and solar sectors (DING et al., 2018; KAID et al., 2018). In the hydroelectric sector, the literature on predictive maintenance is specific and limited to large hydroelectric plants (SELAK et al., 2014). Only one study was found addressing computer monitoring and failure prediction for SHPs. It dealt with two case studies in which specialist systems were developed for monitoring some sensor-detected variables, and issued an alert if some predefined limits were violated (HENDERSON et al., 1998). Therefore, the literature lacks examples reflecting the reality of the SHPs, which present a different operation and maintenance profile than the large hydroelectric plants. With the application of prognostic models, a plant can benefit from increasing its asset availability and its total generation capacity, favouring both the investor and the society.

The general objective of the present article is to identify the maintenance and operation profile of SHPs, and to evaluate the types of failures that contribute most frequently to forced unavailability. Among the specific objectives are: understanding the development process and execution of the maintenance plan; analysis of the time stopped, according to the type of operational state (stopped by water scarcity, planned stop, forced stop, stop by external conditions); and, survey of the main components that contribute to forced downtime.

This article is structured as follows: Section 1 contextualizes and defines the research problem. Section 2 presents the materials and methods applied in the present article. Section 3 reports the results obtained from the methods. Section 4 discusses the results obtained. Section 5 concludes this article with remarks and prospects.

3.2 Materials and methods

3.2.1 Assessment questionnaire

To assist in the development of the present study, a questionnaire was developed related to the maintenance process in SHPs, using the method proposed by (WILLIAMS, 2003). The technique consists of nine steps, from the formulation of the research question and study population to the validation of the results. Figure 3.1 presents the methodology used to design the questionnaire. Following the research area definition, the literature review process on the theme of predictive maintenance in energy engineering systems was developed, to understand the main aspects of a prognostic system and the requirement for its implementation.

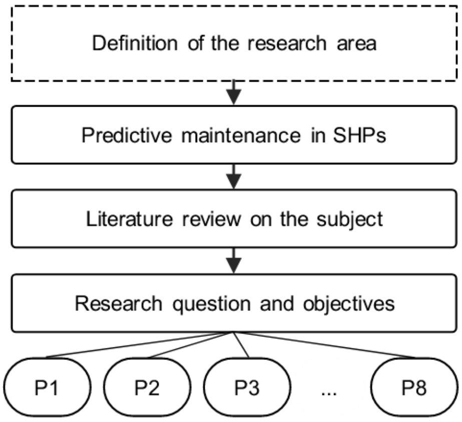

Definition of the research area
Predictive maintenance in SHPs
Literature review on the subject
Research question and objectives
P1
P2
P3
...
P8

Figura 3.1 – Procedure adopted to develop the research questionnaire.

The following research question was raised: “What is the maintenance and operation profile of small hydroelectric plants in Brazil, and where is the greatest potential for gain with the application of a prognostic maintenance model?”. The question was divided into eight sub-questions by the authors, generating the questionnaire that guided the interviews. The questionnaire was applied in the interviews that were conducted with focus groups of executives and technicians. In total, four plant coordinators, two maintenance coordinators and two maintenance technicians were interviewed. The interviews lasted around one hour each, and were conducted by the first author, accompanied by an executive from the company. Four operators were also interviewed, to clarify the operation planning process.

3.2.2 Operational status record base

The operational records dataset of several plants was adopted as a basis, with the operating status defined by the National Electric System Operator (NESO). The records are maintained by the companies' remote operation areas, and are used to generate external operation reports for each client and regulatory agency. These documents contain the records of operating states for 91 generating units (GU), for 43 plants located in the states of Goias, Mato Grosso, Mato Grosso do Sul, Minas Gerais, Rio de Janeiro, and Santa Catarina. The number of GUs per plant varies from 1 to 5 units. Records from January 1, 2018 to December 31, 2018 were used.

3.2.3 Exploratory data analysis

For exploratory data analysis, we adopted the libraries of the graphical representation software Seaborn (v.0.9.0) and Matplotlib (v.3.1.0), and the numerical calculation libraries NumPy (v.1.16.4) and Pandas (v.0.24). The scripts were developed using the Python language (v.3.7.3) on the Jupyter Notebook development platform (v.6.0.0). The analyses were performed by the authors, and iteratively evaluated and validated by company executives and specialists.

3.3 Results

3.3.1 Case study of an energy operator

The current study was conducted at a company responsible for the operation and maintenance of dozens of SHPs and photo-voltaic units in the south, southeast, midwest and northeast regions of Brazil. The company has a remote-operations control centre which monitors all customer plants 24 hours a day, 7 days a week, developing actions to deal with any operational problems that may occur. In addition to the operations and maintenance services, the energy operator also has a team of specialists in automation and retrofitting that provides consultancy services for evaluating energy generating systems in the most diverse types of projects in the area.

3.3.2 Qualitative profile

The qualitative profile is based on the responses to the questionnaires by the specialists and executives involved in the operations and maintenance processes of the plants. This profile is presented, below, in the form of questions and answers, to facilitate the reader's understanding.

1. What is the plant's unavailability and how is it calculated? How is machine breakdown time counted? And for water scarcity?

The process for verifying unavailability is developed by NESO, and is detailed in the Operation Procedures Manual (OPM). The main operating states are: (1) ON - hours in service, connected as a generator; (2) RHD - reserve hours disconnected, the period stopped due to lack of water resources; (3) HSS - hours of scheduled shutdown, planned stops; (4) HFS - hours of forced shutdown, unplanned stops, for example, resulting from machine breakdown; (5) HOEC - hours turned off due to conditions external to the generating unit (GU), normally associated with the energy concessionaire. The equivalent rate of accumulated programmed outage is calculated from HSS divided by the total number of hours in the calculation period considered. The equivalent rate of accumulated forced outage is calculated from HFS divided by the sum of HFS, ON, HOEC and RHD. Full details of the methodology for determining unavailability can be found in the OPM.

2. How is the maintenance plan for a generating unit developed?

General maintenance is performed once a year per GU, usually during the dry, or low generation, period. In this period, the GU is opened and the components are inspected in order to clean them and to identify any problems. This period lasts for two weeks. The aim is to mitigate all potential risks to the equipment and to prepare it for the rainy season, when the equipment will need to operate with the greatest possible reliability. All components of the generator system are mapped on an equipment tree by the maintenance team. Each piece of equipment has its own maintenance policy, defined in the manufacturer's manual or by the engineering team. Minor maintenance is performed weekly by the maintenance team.

3. Who carries out maintenance on the SHPs? Is there a team that moves to each SHP when there are failures or is maintenance done locally by someone who operates the SHP? Is this team segmented with people with greater expertise in one segment than the other, i.e., a specialist in hydraulics vs an expert in electrical?

Maintenance is divided into three levels, according to complexity. First level maintenance includes inspections or simple interventions that can be performed by the plant's own maintenance team. Second level maintenance is more specific (i.e., oil changes, cleaning, equipment measurements) and requires a group of mechanical and/or electrician specialists. These teams can be linked to a single plant or to a group of plants, depending on the level of demand and the proximity between plants, to reduce fixed costs. Third-level maintenance is that which cannot be performed internally by the company, such as opening machines, machining components, for which external companies are hired.

4. What are the main causes of generation failure?

Each plant has a different fault pattern. Some of the main causes identified were: (1) the lack of silting up of the water, which obstructs the industrial system and filtration grids; (2) voltage drop and fluctuation, which generates disconnection between the system and the transmission line; (3) communication failure between the remote operation control centre and the plant, due to the unavailability of either the VPN or the internet; (4) automation system generating an incorrect trip, usually caused by an incorrect sensor reading due to external factors or waste accumulation.

5. Is there redundancy in the generating units to avoid shortages? Are there auxiliary systems whose failure can interfere with generation?

Most plants have at least two turbines, although there are some older plants that have only one turbine. Auxiliary systems common to the generation systems are the: compressed air, direct current and alternating current, industrial water and drainage systems. Generally, these systems are shared among the generating units, and their failure impacts the total generation of the plant.

6. What variables are sensed in real time? How often is the data recorded

There are some basic variables that are monitored in almost all generating units, such as the energy generated and the temperature of the bearings. The water levels upstream, downstream and in the dam are found only in larger plants, while some smaller plants usually do not have a dam (run-of-river plants). The flow of oil and water that feed the GU is also commonly monitored. Some information, however, is more difficult to find, which includes vibration and noise sensing, and inlet and outlet oil temperature. The data readings are saved in the database every 5 minutes. There are some measurements that are made periodically by specialists, such as thermographic tests.

7. Is the event data recorded? How reliable is this data?

The operating states of each generating plant are recorded by the post-operation area, and are disclosed to NESO and to the customer who owns the plant. The reliability of this information is high, as the information must be generated according to the normative instructions. There are also records of all maintenance performed in a maintenance system. These records identify all service orders that were executed, by equipment, with details of the maintenance team and materials that were used.

8. How is the operation plan developed?

Each GU has an operating parameter that is defined based on engineering tests. For plants with a dam, the responsible planner uses the maintenance plan, the water levels (when available to read) and the rain forecast of the region as input information. That information comes from company experts and external weather services. The production plan seeks to optimize generation, keep plants at more efficient levels of operation and avoid the loss of water resources as much as possible. There are groups of plants positioned along the same water system, and these require coordinated planning because the water output from one system is the input of the subsequent system.

3.3.3 Quantitative profile

The first decision to be made relates to the frequency and duration of each state, as shown in Figure 3.2. On average, a GU is connected 237 days per year, which represents 65% of the total time. It is turned off for 78 days due to water outages. Dividing the generation time by the available time with water resources, we were able to estimate the average availability of the generating units – 237 / (365-78-3) = 83.45%.

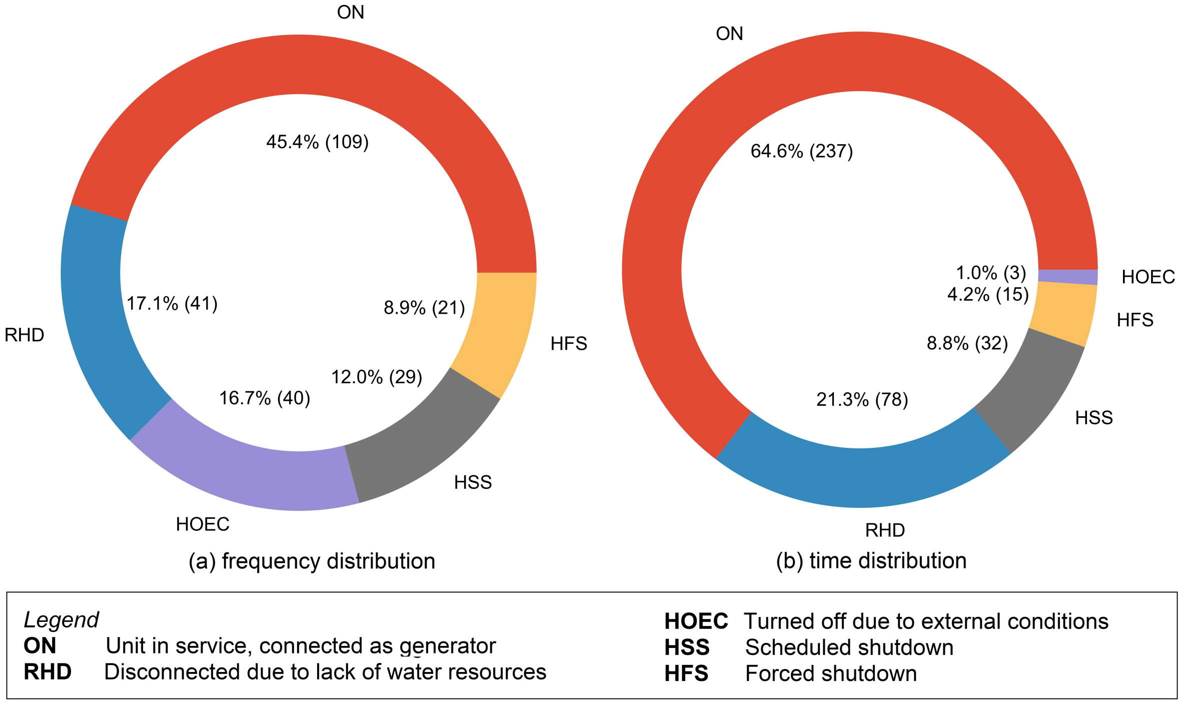


Figura 3.2 – Distribution of average operating states. On the left, we find the average number of occurrences of each operating state per GU, while on the right, we find the average number of days in each state per GU.

Analysing the downtime for each operational state, we found that the most frequent average number of failures is due to external conditions (HOEC). With about 40 occurrences per year for each GU, and considering that HOEC, HSS and HFS are the states related to failure, it is evident that HOEC is the problem faced most frequently by the company during the operation. However, analysing the graph on the right, we find that the time in this state represents only 1% of the time stopped, about 3 days per year on average, for each GU.

On the other hand, although HFS and HSS occur less frequently, they require more time to be resolved. Together, they cause 13% of downtime, while HOEC causes only 1%. Thus, the potential gains from decreasing HFS / HSS times for total availability are much greater than the gains from improvements to the transmission system.

The comparison of HFS and HSS allows assessment of the maturity level of the maintenance planning. In general, it is more effective to plan maintenance than for it to happen of necessity, so that the company can mobilize human and material resources in advance and optimize costs. The unexpected failures (HFS) occur less frequently than the planned shutdowns. However, the total time spent in the scheduled shutdowns (HSS) is greater. This is an indication that the failures that require more intervention time are closely monitored by the company's technicians and analysts, while the unexpected failures are less familiar and receive more rapid intervention.

Figure 3.3 presents a strip graph representing the occurrences of each status of all generators, segmented by operating state. The water shortage situation is distributed in the interval between 0 to 50 days, with most occurrences lasting fewer than 10 days. The period distribution of machines connected without intermittence is similar, with most occurrences fewer than 20 days.

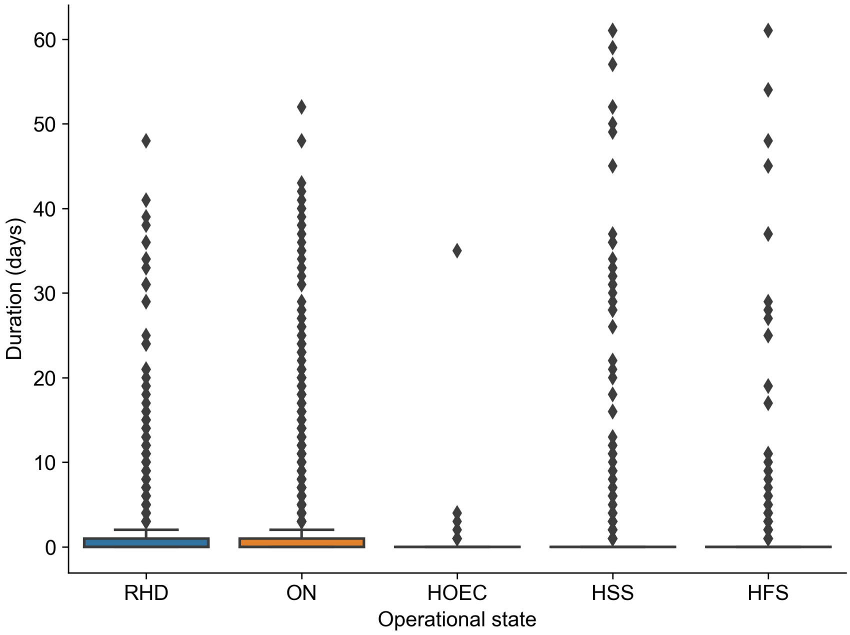


Figura 3.3 – Distribution of records by operating status, according to duration in days.

It is possible to observe that shutdowns due to external conditions (HOEC) are resolved quickly, with a single, isolated exception that took 35 days to resolve. However, most occurrences of this state did not last more than two days.

Finally, the durations of the programmed and forced shutdown states also show close distributions, with intervals ranging from 0 and 60 days. Most of the data is concentrated in the period of 7-10 days, which indicates the average time of intervention in the system for both scheduled and corrective maintenance. However, these interventions can be spread over longer periods, as shown in the representation of the categorical variable.

Figure 3.4 shows the distributions of the durations of the operating states by the months of the year. Average values are represented by circles and confidence intervals (CI = 0.95) are represented by bars. In this visualization, it is noteworthy that the longest planned downtime (HSS) is carried out during the time of drought, between May and July.

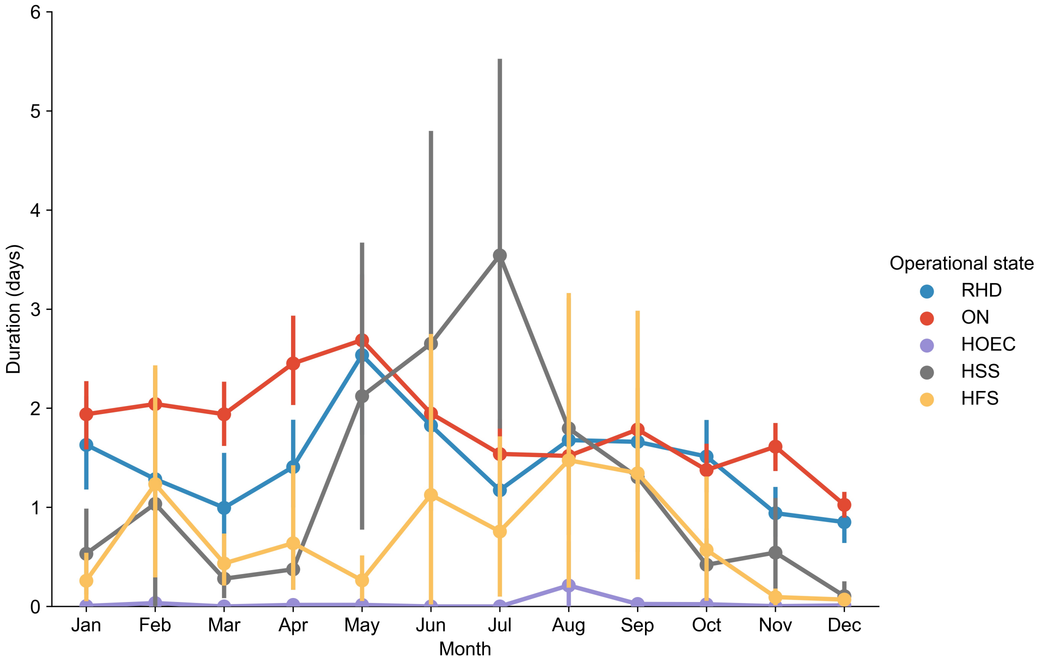


Figura 3.4 – Distribution of duration in each operating state by months. The points represent the mean values, while the bar represents the value for the confidence interval of 0.95.

The forced stop time (HFS) is longer between the months of June and October. The duration of the states connected as a generator (ON) and stopped due to lack of water resources (RHD) are relatively stable throughout the year, with the longest time in the months of April and May. Finally, downtime due to external conditions (HOEC) is close to zero during all months of the year, with a small increase in August.

Assessing the confidence intervals, it is noted that the states RHD and ON have smaller standard deviations when compared to the states HSS and HFS. From these results, it appears that the operation of the plants seeks to maintain the generation and availability of water resources because, even in the dry months, the states ON and RHD are balanced and have shorter confidence intervals.

Another feature that stands out in the graph in Figure 3.4, is that the frequency of failures is greater after the period of planned maintenance, where the system should be operating with greater reliability. However, during the rainy season, between November and March, the downtime for both planned and corrective maintenance is low.

Figure 3.5 shows the 10 generator system components that contribute the most to the forced stop time (HFS). The red bar represents how much the component failures contribute to the total forced stop time, while the blue bar shows the frequency with which the components fail in relation to all the failures that occurred during the same period.

The first component that contributed the most to forced downtime was the bearing, with 15.1% of the total time, although it is not the item that failed the most frequently. It

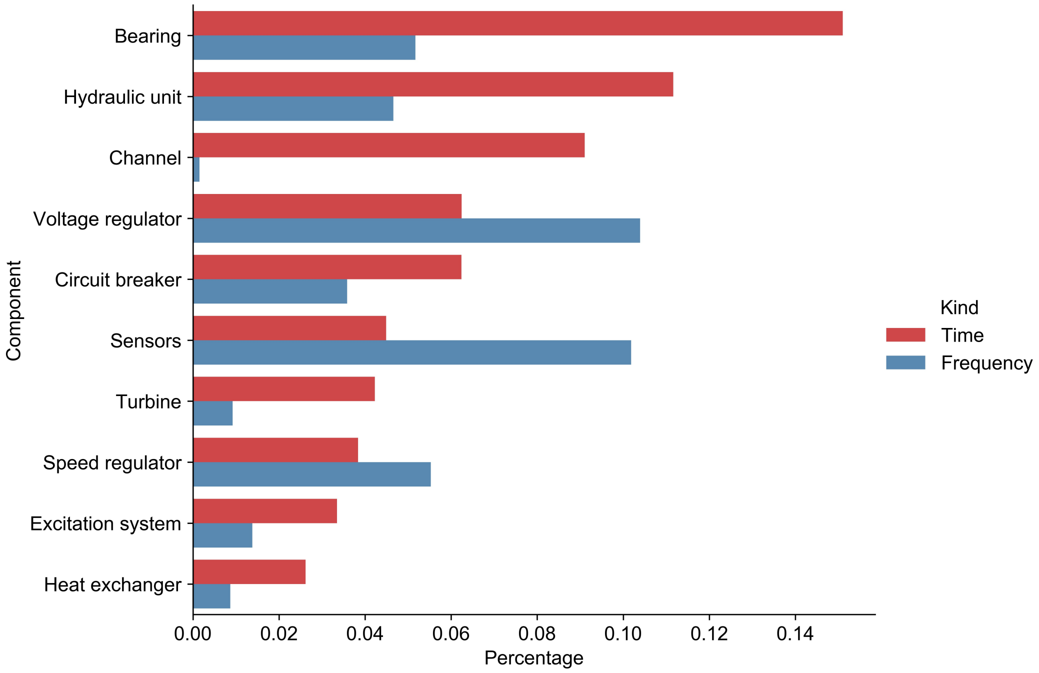


Figura 3.5 – Unavailability rate (red) and failure rate (blue) by component, filtered among the 10 most significant.

represents 5.2% of the total failure occurrences. The second component that contributed the most to forced downtime was the hydraulic unit, with 11.2% of the time and 4.7% of the total failure occurrences. The third component that contributed the most to forced downtime was channel obstruction, with 9.1% of the entire HFS time but only 0.2% of the total failure occurrences. This indicates an isolated occurrence, in which a single problem resulted in the operation failing for too long.

Other components that stand out are protection systems of the system itself, such as circuit breakers and sensors, which together represent 10.7% of the total time and 13.8% of the occurrences. In general, these stops are due to the action of the system if any system variable (e.g., power generated, temperature, etc.) exceeds a predetermined limit in the maintenance design.

The turbine is one of the main components of a GU and, although it has a relatively low occurrence rate (0.9%), it contributed 4.2% of the total downtime. This indicates that the turbine failure requires more repair time, given the complexity and accessibility of the component. The speed regulator is associated with the turbine, as it regulates the flow of the water that feeds the turbine. This represents 3.8% of the total forced downtime and 5.5% of the failure occurrences.

3.4 Discussion

When compared to the large hydroelectric plants, it appears that the SHPs have some restrictions such as decentralized maintenance teams and a lower level of sensing.

Furthermore, many are very old and have been operating since the early 20th century. However, most SHPs have undergone recent automation and now show satisfactory levels of sensing in both real-time and remote operations.

However, it appears that the level of maintenance management is relatively high, with the definition of the equipment tree and the preventive maintenance plan. NESO regulation and event registration activity, associated with the history of the order of services performed, provide fundamental information for the implementation of a prognostic system.

There is coherence between the results found and the variables sensed by the automation systems: the bearing is the most critical component and, in most GUs, there are temperature sensors in these components. The hydraulic unit is the second most critical component, as it is responsible for feeding the system with oil. For this reason, most GUs have flow switches for measuring flow, and temperature sensors for measuring the temperature of oil inlet and outlet. Finally, the turbines are monitored using vibration sensors and thermographic measurements.

The adoption of a prognostic maintenance system based on condition monitoring could come to contribute greatly to the increase in system availability, primarily associated with the main components of the generator system that fail the most frequently. Those components are the bearings, hydraulic unit, turbine and speed regulators. Separately, they represent 34.3% of all the forced downtime, which yields an average of 3 days per year of generation loss in SHPs.

3.5 Conclusion

The present study addressed the theme of maintenance in SHPs. It raised the profile of the maintenance planning and execution processes, based on the practical experience of executives and specialists in the area and by analysing historical records of operational states in order to identify the main causes of unavailability in more than 40 SHPs operating in Brazil. From the results obtained, we identified several configurations according to the plant project: some are more sensed, others have less monitoring data; dams may or may not exist; the operation can be local or remote; the teams may or may not be decentralized, with part of the activities being carried out by specialized, external companies. The main causes of failure can vary from plant to plant, but in general the components that generate the most downtime are bearings, hydraulic units and others related to the main generator system.

Although many interruptions in supply are due to failures in the transmission line, it was found that the total time of these stoppages is insignificant because the repair time is exceptionally low. However, there is potential for improvement in reducing corrective and preventive maintenance times, based on adopting predictive maintenance. Further study still needs to be done to verify which components are the most critical to monitor, in order to prioritize the design of the prognosis system.

The results found in the present study show the feasibility of applying predictive maintenance models to plants, which could benefit from lower operating costs and greater availability of assets. For future work, it is suggested to build the equipment failure tree in a practical case study of a plant. In such a study, the authors suggest selecting a pilot plant and applying the failure tree analysis method (MÁRQUEZ et al., 2012; LEE et al., 1985) to the generation system, combining expert support and the plant's failure history. This diagnostic step is essential for the development of a prognostic model based on the history of reading sensors, and operational and maintenance data.

4 EXTENDED ISOLATION FORESTS FOR FAULT DETECTION IN SMALL HYDROELECTRIC PLANTS

Abstract: Maintenance in small hydroelectric plants is fundamental for guaranteeing the expansion of clean energy sources and supplying the energy estimated to be necessary for the coming years. Most fault diagnosis models for hydroelectric generating units, proposed so far, are based on the distance between the normal operating profile and newly observed values. The extended isolation forest model is a model, based on binary trees, that has been gaining prominence in anomaly detection applications. However, no study so far has reported the application of the algorithm in the context of hydroelectric power generation. We compared this model with the PCA and KICA-PCA models, using one-year operating data in a small hydroelectric plant with time-series anomaly detection metrics. The algorithm showed satisfactory results with less variance than the others; therefore, a suitable candidate for online fault detection applications in the sector.

Keywords: Hydroelectric power plant. Condition-based maintenance. Machine learning. Early fault detection. Decision tree algorithm.

4.1 Introduction

With energy demand expected to double by 2060, the development of clean energy sources is essential for guaranteeing an energy supply in the coming decades. Renewable energy already represents three quarters of yearly new installed capacity (WEC, 2019), and those related to water resources are the most applied. In this group, the construction of small hydroelectric plants (SHPs) has grown worldwide due to the lower initial investment, low operating costs and increasing regulation of energy markets. The potential total energy generation capacity of these SHPs is twice the current total capacity of the presently installed energy plants (UNIDO, 2016).

Several case studies are reported in recent literature addressing the energy potential and importance of developing SHPs in emerging countries like Brazil (FERREIRA et al., 2016), Turkey (DURSUN; GOKCOL, 2011), Nigeria (OHUNAKIN et al., 2011) and other sub-Saharan African (KAUNDA et al., 2012) countries. Overall life cycle assessment is applied for quantitative economic evaluation of this type of undertaking in India (BHAT; PRAKASH, 2014) and Thailand (SUWANIT; GHEEWALA, 2011). Economic models of viability sensitivity analysis of SHPs stations are presented and applied to the energy context in Spain (ALONSO-TRISTÁN et al., 2011) and Greece (KALDELLIS et al., 2005). A common factor among all these models of economic viability is the cost of operation and maintenance, which is a determining variable for the development of new stations.

Maintenance of a hydroelectric generating plant is a complex task, though. It requires a certain level of expertise to ensure a satisfactory level of reliability of the asset during its useful life. There are three types of maintenance. The first, and most rudimentary, is corrective maintenance, in which a component is expected to break, and is then replaced. Preventive maintenance estimates the expected service life of a component, and replacement is done when the operating lifetime is reached. Last, in predictive maintenance, the condition of the system is calculated from data periodically or continuously obtained from various sensors (BOUSDEKIS et al., 2018; PENG et al., 2010). A predictive, or condition-based maintenance (CBM) system, consists of two main phases. The first phase is diagnosis, which comprises fault detection or abnormal operating conditions, fault isolation by sub-components, and identification of the nature and extent of the failure (PENG et al., 2010). The next phase is prognosis, applying statistical and machine learning models to estimate the useful life of the equipment and the confidence interval of the prediction (SIKORSKA et al., 2011), to anticipate maintenance, and to increase the reliability and availability of the generation system.

Examples of commonly applied methods for estimating useful life are divided between statistics (SI et al., 2011), which includes the regression methods, Wiener process, gamma process, based on Markovian processes; machine learning methods such as neural networks, vector support machine, electrical signature analysis (SALOMON et al., 2019b), principal component analysis (LEI et al., 2018); and more recently, deep learning techniques such as auto-encoder, recurrent and convolution neural networks (ZHAO et al., 2019). Several other energy sectors already make extensive use of these techniques: nuclear (AYO-IMORU; CILLIERS, 2018; LI et al., 2018; WU et al., 2018), wind (MÁRQUEZ et al., 2012; TIAN et al., 2011) and solar (DING et al., 2018; KAID et al., 2018).

Multivariate statistical methods such as Principal Component Analysis (PCA) (LIU et al., 2009), Independent Component Analysis (ICA) (ŽVOKELJ et al., 2016) and Least Square - Support Vector Machine (LS-SVM) (FU et al., 2019; QIAO; CHEN, 2015; VU et al., 2013; PENG et al., 2007), have been widely applied for fault detection and diagnosis in hydro-generating systems. For instance, PCA decomposition is applied to aid experts in identifying and selecting the main features which contribute to cavitation in hydro-turbines (GREGG et al., 2017). Recent studies have proposed a new monitoring method, based on ICA-PCA, that can extract both non-Gaussian and Gaussian information of process data for fault detection and diagnosis (GE; SONG, 2007). Later, the ICA-PCA was extended with the adoption of a non-linear kernel transformation prior to the application of the decomposition method, which became known as the Kernel ICA-PCA (KICA-PCA) (ZHU et al., 2014). They reported its application in the hydroelectric generation context with higher success rates and lower fault detection delays than either the PCA or ICA-PCA applications.

The isolation forest (iForest) (LIU et al., 2012; LIU et al., 2008) is an anomaly detection model based on decision trees which, recently, is appearing in several case studies of anomaly detection in the business(SUN et al., 2016a), industrial (SUSTO et al., 2017) and virtual security (RIERA et al., 2020; VARTOUNI et al., 2018) areas. Briefly, the iForest method provides a non-parametric density estimate of the data. The non-parametric density can be estimated using data under normal operating conditions. After fitting the iForest model, density estimates or anomaly scores are calculated using online data. Faults are detected when the anomaly scores are higher than a pre-defined upper bound, indicating that the system under monitoring is no longer operating under normal conditions. The meta-heuristic model has some interesting advantages when compared to the classical linear decomposition models: it can handle an enormous amount of data and heterogeneous variables, without needing a data labeling process. It can, thus, develop non-linear models of learning based on random, decision tree ensembles.

The most recent version of the algorithm, the extended isolation Forest (EIF), adopts hyperplanes with random slopes to separate the data, solving problems related to how the algorithm calculates the anomaly score (HARIRI et al., 2019). The EIF can build scores with less variance and obtain better accuracy in the area under the receiver operating characteristics metric, compared to the original algorithm, without sacrificing computational efficiency (HARIRI et al., 2019; SUN et al., 2016b). However, no study has been found reporting the application of the iForest or EIF in hydroelectric turbines.

In this context, the present paper proposes the application of iForest and EIF to support fault detection and diagnosis of a hydro-generating unit (HGU) in an SHP. We compared the algorithms with PCA and KICA-PCA, using specific metrics for anomaly detection in time series (KOVÁCS et al., 2020). The main findings and contributions of the current paper are:

- Application of iForest and EIF for intelligent fault diagnosis in an SHP generating unit.
- Proposal of the application of time distance and count detection metrics, most appropriate for the evaluation of models in the context of anomalies detection in time series.
- EIF presented reductions of 40.62% and 7.28% in the temporal distance, compared to the PCA and KICA-PCA.

The remainder of the present article is organized as follows. Section 2 defines the study methodology, describing the methods, algorithms and data set applied. Section 3 presents the results and discussions of the simulations of the models, in addition to the outputs of the forest committee with illustrative examples of imminent failures. Finally, Section 4 presents the conclusions and recommendations for future work.

4.2 Materials and methods

4.2.1 Dataset

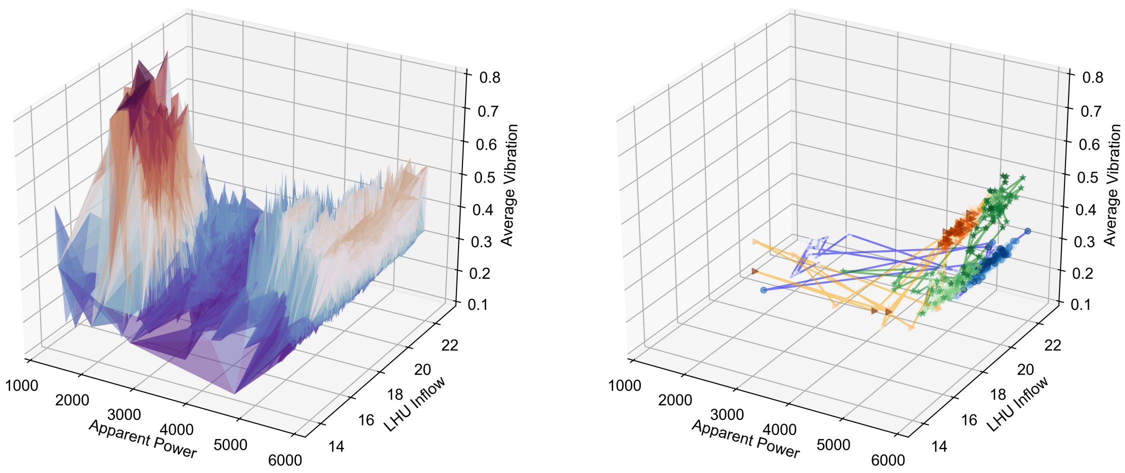
The current study was developed in Ado Popinhak, an SHP situated in the southern region of Brazil. With an installed capacity of 22.6 MW, the plant supplies energy to 50,000 residences. Condition monitoring data from the main single HGU is registered every 5 minutes, and the scope of the study period is from 8/13/2018 to 8/9/2019. We filtered out from the dataset the periods of maintenance, planned stop, operator intervention, or another status not associated to normal operation. Event data related to the asset is gathered from status reports and the maintenance management system to identify past failures. Fifty-nine faults were registered in the disclosed period, totaling 123 hours of downtime. Six monitored variables are used: generator apparent power; bearing hydraulic lubrication unit (HLU) inflow; and, bearing vibration from four different positions: axial, vertical radial, horizontal radial and coupled.

Figure 4.1 presents the interaction between the variables in the dataset, in two different visualizations. The vibration variables are replaced by the average of the variables at the different measuring points. Figure 4.1-a indicates a low-apparent power region, where the average vibration is higher than in the rest of the observations. These present a transient period in which the generator unit operates with imbalanced water inflow inside the runner. In such a state, the wear damage to the system and the fault risks are more serious.

Figure 4.1-b presents three excerpts of the time series before failure. The analysis of the representation indicates that the failures generally occur in regions where the vibration and apparent power are at their maximum, and there may be significant fluctuations in the power and flow of HLU before they occur. The figure presents only a sub-sample, of 3 out of the 59 faults in the entire database, to avoid overload of information in the representation, which would make it difficult for the reader to analyze.

A fixed period, from 12h prior to the failure up to the failure, splits the full data set into a training set and a test set. The training set corresponds to the healthy state, or normal operation, as long as the test set is linked to abnormal operation. In this way, the algorithm focuses its training on the positive class related to normal operating conditions, thus becoming a density estimator of the class of interest (HEMPSTALK et al., 2008). This type of approach is common in problems of unbalanced classes, in the context of anomaly detection, in which negative cases (our outliers) are absent or not adequately sampled (KHAN; MADDEN, 2009).

After separation, the training and test set sizes were 47857 and 4897, respectively. The ratio between training and test sets is about 10:1, which is an appropriate ratio when compared to reference studies on failure detection in hydroelectric plants that have



(a) (b)

Figura 4.1 – Graphical representation of the data set in 3 dimensions. In (a), the entire data set is presented, regardless of the temporal relationship between data points. In (b), three excerpts from the series with imminent failures are presented, each in a different color. The darker the marker, the closer the fault. Points connected by lines represent sequential states.

adopted proportions of 8:1 (LIAO et al., 2019) and 1140:100 ∼ 10:1 (ZHU et al., 2014). Anomaly detection algorithms, reported sequentially, are trained using only training data.

4.2.2 PCA

Principal Component Analysis (PCA) is a linear decomposition technique, effective for data dimensionality reduction, that projects the correlated variables onto smaller sets of new variables that are orthogonal and retain most of the original variance. PCA is the most widely used data-driven technique for process monitoring, due to its capacity to deal with high-dimensional, noisy and correlated data variance (NAVI et al., 2018).

Let $\mathbf{X} \in R^{n\times m}$ be an observation matrix, where n is the number of samples, and m is the number of monitor variables. $\mathbf{X}$ can be decomposed by the function

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{4.1}$$

where $\mathbf{E}$ is the residual matrix, $\mathbf{T} \in R^{nxa}$ is the score matrix, and $\mathbf{P} \in R^{mxa}$ is loading matrix. The measure of PCA variance can be obtained by Hotelling's T^2 statistic representing the sum of the normalized squared scores

$$T^2 = \mathbf{t}^T\mathbf{D}^{-1}\mathbf{t} \tag{4.2}$$

, where $\mathbf{D}$ is the diagonal matrix of the eigenvalues with the retained principal components and $\mathbf{t} = \mathbf{P}^T x$, is the score of PCA, calculated from the multiplication of each element x and the loading matrix $\mathbf{P}$.

The T^2 index is used for monitoring processing, detecting a systematic variation of the process every time an observation exceeds the confidence limit T^2_α, given by

$$T^2_\alpha = \frac{(n^2-1)\alpha}{n(n-\alpha)} F_\alpha(\alpha, n-\alpha) \tag{4.3}$$

where n is the number of samples, α is the number of sensed variables, F_α is the upper 100% critical point of F-distribution with α and $n-\alpha$ degree of freedom. As to a classification, a set of class labels C is set as 1 if $T^2_i > T^2_\alpha$ or else 0, if condition not met, for $T^2_1, T^2_2, ..., T^2_n$.

4.2.3 KICA-PCA

The KICA-PCA method provides a kernel transformation of data into higher dimensional data, prior to the application of decomposition. Thus, the method is capable of handling non-linear multivariate processes, such as SHP condition monitoring (ZHU et al., 2014).

In this application we adopted the explicitly mapping to a low-dimensional Euclidean inner product space using a randomized feature map $z : R^{nxm} \rightarrow R^{nxd}$ proposed by (RAHIMI; RECHT, 2009), so that the inner product between a pair of transformed points approximates their kernel evaluation:

$$k(x,y) = \langle \Phi(x), \Phi(y) \rangle \approx z(x)'z(y). \tag{4.4}$$

Contrary to kernel's lifting Φ, z is low dimensional and k is the radial basis function $k(x,y) = exp(-||x-y||^2/\sigma)$ and σ is the standard deviation.

The z mapping competes favorably in speed and accuracy, as evidenced by (RING; ESKOFIER, 2016; SENECHAL et al., 2015; RAHIMI; RECHT, 2009), being capable of handling the large training matrix of this study without exceeding computational resources of a standard personal computer.

The transformed matrix $\mathbf{X}'$ is calculated by the kernel approximation $\mathbf{z}^T\mathbf{z}$, such as each element

$$k(x_i,x_j) = x'_{ij} = z(x_i)^T z(x_j), \tag{4.5}$$

where x_i and x_j are the ith and jth columns of $\mathbf{X}$ respectively.

Before the application of ICA, the matrix $\mathbf{X}'$ should be whitened to eliminate the cross-relations among random variables. One popular method for whitening is to use the eigenvalue decomposition, considering $x'(k)$ with its co-variance $R'_x = E\{x'(k)x'(k)^T\}$, as described in (GE; SONG, 2007). The association of the kernel transform and the ICA is known in the literature as KICA.

ICA is a statistical, computational technique originally proposed to solve the blind source separation problem by revealing patterns hidden in signals, variable sets, or measurements (GE; SONG, 2007).

$$\bar{\mathbf{X}}' = \mathbf{AS} + \mathbf{E} \tag{4.6}$$

where $\bar{\mathbf{X}}'$ is the the whitened transformed matrix, $\mathbf{A}$ is the mixing matrix, $\mathbf{S}$ is the independent component matrix and $\mathbf{E}$ is the residual matrix. The basic problem is estimating the original component $\mathbf{S}$ and the matrix $\mathbf{A}$ from $\bar{\mathbf{X}}'$. ICA calculates a separating matrix $\mathbf{W}$ such that the components of the reconstructed matrix $\mathbf{S}$ become as independent of each other as possible, given as

$$\hat{\mathbf{S}} = \mathbf{W}\bar{\mathbf{X}}'. \tag{4.7}$$

From the multiplication of $x'' = \mathbf{S}^T \hat{x}'$ is obtained a new matrix $\mathbf{X}''$ which represents the independent components (ICs) from the sensed data. These matrices are used as input for the PCA monitoring algorithm in equation (1) and used to calculate the T^2 score and classification set from the comparison with the T^2_α threshold.

4.2.4 iForest and EIF

While most anomaly detection approaches are based on normal instance profiling, iForest is an anomaly detection algorithm that explicitly isolates anomalies. The method exploits two particularities of anomalies: they represent fewer instances in the observed set, and, compared to healthy instances, they have discrepant attribute-values (LIU et al., 2008).

The method does not apply any distance or density measures, thereby eliminating the major computational cost of distance calculation. Also, the algorithm scales up linearly while keeping memory usage low and constant, which aligns with parallel computing, making the model suitable for handling large, high-dimensional data sets.

The anomaly detection procedure using iForest is a two-stage procedure: the training stage constructs the isolation trees (iTree), using sub-samples from the training set; the subsequent evaluation stage calculates the anomaly score for each instance of test set (LIU et al., 2008; LIU et al., 2012).

The iForest builds an ensemble of binary trees individually trained using a sub-sample $\mathbf{X}^s$ randomly drawn from $\mathbf{X}$, $\mathbf{X}^s \subset \mathbf{X}$. There are two control parameters in the algorithm: (1) the sub-sampling rate ψ, sets the number of samples used for each tree training, and (2) the number of trees nt, related to the complexity of the model.

Algorithm 1: *iForest* ($\mathbf{X}$, nt, ψ)

```
Input: $\mathbf{X}$ - input data, $nt$ - number of trees, $\psi$ - sub-sampling size
Output: a set of $t$ iTrees
Initialize Forest
for $i \leftarrow 1$ to $nt$ do
    $\mathbf{X}^s \leftarrow sample(\mathbf{X}, \psi)$;
    $Forest \leftarrow Forest \cup iTree(\mathbf{X}^s)$;
end
return Forest
```

The normal points tend to be isolated at the deeper end of the tree, whereas anomalies are closer to the tree root, due to their singularity nature. The shorter the average path length, the higher the chances to be anomalies. Hence, the anomaly score s is then defined by:

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}} \tag{4.8}$$

where n is the number of samples in the dataset, $E(h(x))$ is the average of path length $h(x)$ from a group of isolation tree, and $c(n)$ is the average of $h(x)$ given n, used for normalizing the path length. If an instance returns an anomaly score s very close to 1, it is very likely one represents an anomaly; if it is much smaller than 0.5, it is safe to say the instance is normal; if the instance returns $s \approx 0.5$, the sample does not present any distinct anomaly (LIU et al., 2012).

Although the standard iForest algorithm is computationally efficient, there is a limitation as to how the anomaly score aggregates tree branches' length. Branch cuts are always horizontal or vertical, which introduces a bias in the anomaly score map.

The EIF algorithm can overcome this limitation by adopting random slopes along with the branching process. The selection of the branch cut then requires a random slope and a random intercept chosen from the range of values available in the training data. Each random slope is drawn from a random number for each coordinate of a vector $\vec{m}$ of size equal to the number of variables in a normal distribution $N(0,1)$. The intercept is obtained from the uniform distribution of a range of values present at the branching point. The splitting criterion for a point x is given by $(\vec{x} - \vec{p}).\vec{m} \leq 0$.

The property of concentration of data in clusters is maintained with the algorithm, as the intercept points $\vec{p}$ tend to accumulate where the data is, while the score maps are free of previously observed artifacts. EIF implementation modifies the lines of the original formulation 2 that describes the choice of the random value and intercept and adds an inequality condition test. The algorithm 3 is modified accordingly to receive the regular observation and intercept point of each tree, and to calculate the path depth if the

```
Algorithm 2: iTree ($\mathbf{X}$, $e$, $hl$)
  Input: $\mathbf{X}$ - input data, $e$ - current tree height, $hl$ - height limit
  Output: an iTree
  if $e \geq hl$ or $|\boldsymbol{X}| \leq 1$ then
  |  return $exNode\{Size \leftarrow |\mathbf{X}|\}$
  else
  |  randomly select a normal vector $\vec{m} \in IR^{|\mathbf{X}|}$ by drawing each coordinate of $\vec{m}$
  |    from a standard Gaussian distribution.
  |  randomly selects an intercept point $\vec{p} \in IR^{|\mathbf{X}|}$ in the range of $\mathbf{X}$
  |  set coordinates of $\vec{m}$ to zero according to extension level
  |  $\mathbf{X}_{hl} \leftarrow filter(\mathbf{X}, (\mathbf{X} - \vec{p}).\vec{m} \leq 0)$
  |  $\mathbf{X}_r \leftarrow filter(\mathbf{X}, (\mathbf{X} - \vec{p}).\vec{m} > 0)$
  |  return $inNode\{Left \leftarrow iTree(\mathbf{X}_{hl}, e+1, el), Right \leftarrow iTree(\mathbf{X}_r, e+1, el),$
  |                      $Normal \leftarrow \vec{m}, Intercept \leftarrow \vec{p}\}$
  end
```

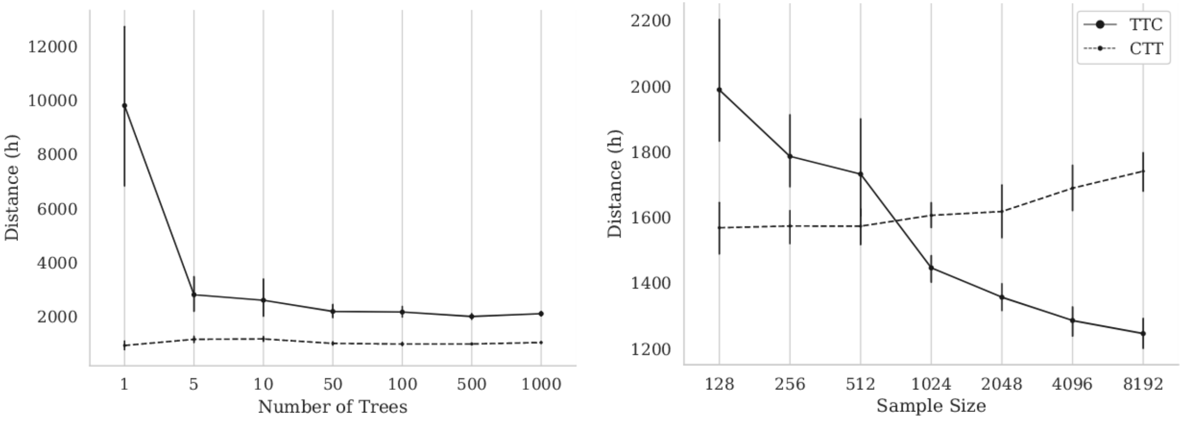
condition test is valid (HARIRI et al., 2019).

```
Algorithm 3: PathLength ($\vec{x}$, $T$, $e$)
  Input: $\vec{x}$ - an instance, $T$ - an iTree, $e$ - current path; to be initialized to zero
         when first called
  Output: path length of $\vec{x}$
  if $T$ is an external node then
  |  return $e + c(T.size)$ {$c(.)$ is defined in Equation (8)}
  $\vec{m} \leftarrow T.Normal$
  $\vec{p} \leftarrow T.Intercept$
  if $(\vec{x} - \vec{p}).\vec{m} \leq 0$ then
  |  return $PathLength(\vec{x}, T.left, e+1)$
  else
  |  return $PathLength(\vec{x}, T.right, e+1)$
  end
```

Contamination is the parameters that estimate the number of outliers in a given set. The value is set near the confidence interval of 0.95, adapted for the Hotelling's distance-based models. Proposed values for the number of trees nt and the size of the ψ sub-sample are 100 and 256 respectively (LIU et al., 2012; LIU et al., 2008). Although, these parameters may vary according to the size and complexity of the dataset.

We carried out 50 simulations varying one parameter at a time while keeping the other fixed at its standard value. Figure 5.4 summarizes the results of these simulations, in which the points represent the average values calculated by the metric, while the bars represent the confidence interval of 0.95. The nt search parameter space is defined as 1, 5, 10, 50, 100, 500 and 1000, while ψ the sample space followed the power of 2, from 2^7 to 2^{13}. By varying the nt, we find that with the increase in the number of trees, the variance and average of the TTC and CTT decrease, with the model showing excellent stability with

500 trees. Performing the same analysis for ψ, we found that for values above 2048, CTT starts to increase gradually. Table 4.1 presents the parameters adopted for EIF based on these observations.



(a) (b)

Figura 4.2 – Graphical representation for 25 simulations of the EIF model varying the control parameters of the algorithm. The points represent the average values and the bars the confidence interval of 0.95 calculated for the CTT and TTC metrics. In (a), the value of nt is changed and ψ is fixed at 256, while in (b) the value of ψ is changed and nt is fixed at 100.

Tabela 4.1 – Parameters adopted for iForest and EIF models.

Parameter	**Value**
Contamination	0.06
Number of trees – nt	500
Sub-sampling size – ψ	2048

4.2.5 Temporal distance metric

Whereas traditional anomaly detection adopts classification evaluation techniques, such as confusion matrix or one of its derived metrics (ZEMOURI et al., 2020), these metrics may be deficient in a time series context. Specific metrics, first developed for evaluating time series segmentation, are adopted in time series evaluation. In the present work, we adopt the average detection count and the absolute detection distance in order to evaluate the different methods (KOVÁCS et al., 2020).

Let $\mathbf{Y}$ be a time series of ordered set of real values indexed by natural numbers $\mathbf{Y} = \{y_0,y_1,y_2,...,y_n\}, y_t \in R$ and $\mathbf{C}$ a set of class labels or classification $\mathbf{C} = \{c_0, c_1, c_2, ..., c_n\}, c_t \in \{0,1\}$, similar to $\mathbf{Y}$ but consisted of binary values $\{0,1\}$. Values labelled 0 represent normal values, and values labelled as 1 stand for anomalous values.

The average detection count l is simply given by the difference between the number of anomalies in the target classification and the candidate classification (KOVÁCS et al., 2020):

$$l_{\downarrow}(\mathbf{C}_i) = |\mathrm{count}(\mathbf{C}_0) - \mathrm{count}(\mathbf{C}_i)| \tag{4.9}$$

where $\mathrm{count}(\mathbf{C}_i) = \sum_{j=1}^{n} c_j, c_j \in C_i$.

The absolute temporal distance (TD) method consists on calculating the sum of all distances between anomalies from two classifications by (KOVÁCS et al., 2020):

$$\mathrm{TD}_{\downarrow}(C_i) = \mathrm{TTC} + \mathrm{CTT} \tag{4.10}$$

where TTC appears for Target To Candidate and is calculated by $\mathrm{f}_{\mathrm{closest}}(\mathbf{C}_0, \mathbf{C}_i)$, and CTT means Candidate To Target given by $\mathrm{f}_{\mathrm{closest}}(\mathbf{C}_i, \mathbf{C}_0)$. $\mathbf{C}_0$ and $\mathbf{C}_i$ denote target classification and candidate classification, respectively.

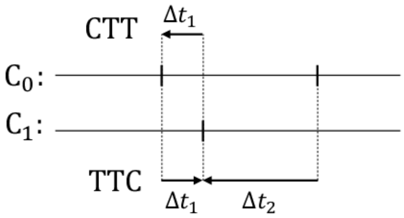


Figura 4.3 – Temporal distance metric calculation. Adapted from (KOVÁCS et al., 2020).

Note that lower values scored in each of individual metrics are better than higher ones. Figure 4.3 graphically represents the calculation of the metric. In the given example, $\mathrm{TTC} = \Delta t_1 + \Delta t_2$ and $\mathrm{CTT} = \Delta t_1$, thus the absolute temporal distance $\mathrm{TD} = 2\Delta t_1 + \Delta t_2$. The best possible value for the metric is zero, comprising a perfect detection system.

4.2.6 Software and hardware

The simulation was developed using the Python language, version 3.7.6, adopting common scientific libraries SciPy 1.4.1, Pandas 1.0.1 and NumPy 1.18.1. Scikit-learn 0.22.1 (VAROQUAUX et al., 2015) presents efficient and reliable implementations of machine learning algorithms, such as PCA and Isolation forest. KICA-PCA was implemented by sequentially declaring the kernel approximation, ICA and PCA into a pipeline of transformers. We applied the authors' original EIF algorithm implementation, available at the isotree 0.1.16 package (HARIRI et al., 2019).

Hardware specifications adopted to perform the simulation are: CPU Intel Core i7-8550U, 1.80 GHz, 16 GB RAM installed, and Windows 10 v.1909 operating system.

The amount of time necessary to perform all 150 simulations is around 2 hours. All data and scripts are available in the researcher's public repository [1].

4.3 Results and discussion

The temporal distance and anomaly detection count are measured in test sets for each method, using equations 4.9 and 4.10 shown in section 4.2.5. The methods were trained 150 times using the training set, with unique random seeds. Table 4.2 exhibits the average and standard deviation of the calculated metrics. KICA-PCA obtained the smallest difference between real fault detection, in general. It was followed by PCA, EIF, and iForest. The methods with the lowest scores are shown in bold. As a linear model, PCA converges equally to the same solution when applied to a single training set. Thus, the standard deviation, unaffected by randomness, is not calculated for this particular method.

Tabela 4.2 – Results of temporal distance and detection count obtained in simulations of anomaly detection models – standard deviation in parentheses.

	Temporal distance (hours)			
Model	TTC	CTT	TD	Detection count – l
PCA	3270	**738**	4008	118
KICA-PCA	1633 (363)	933 (94)	2567 (375)	**111.7 (15.7)**
iForest	1541 (182)	934 (33)	2476 (185)	156.5 (6.1)
EIF	**1474 (208)**	906 (32)	**2380 (210)**	150.2 (5.0)

PCA is the method with the lowest CTT distance. This means that, since the distance between the detections and anomalies is the lowest, it is less likely to raise false alarms than the others. On the other hand, the TTC distance is higher with PCA than with all other methods. This means that the linear approach is less effective in detecting all anomalies. The total TD, obtained from the sum of the two individual components, is higher due to the TTC score.

Combining the PCA with the kernel trick and the ICA increased the accuracy, compared to the PCA alone. KICA-PCA presented intermediate distances when compared to all other methods. When compared to PCA, the non-linear method improved the TTC, TD, and l. However, the main drawback is that its variance is higher than the others, meaning the method is more susceptible to randomness.

iForest presented the second-lowest TTC distance, which is the distance between the anomalies and the closest detection. Thus, this model was able to detect anomalies closest to their occurrence and, having the lowest TD and standard deviation, to present a suitable method for adoption in an online detection system. The main drawback with this method is the detection count: the number of detections is higher than the other methods.

[1] Github repository: https://github.com/rodrigosantis1/shp_anomaly

EIF algorithm boosted the results obtained by iForest, with a TD reduction of 3.88% and an l reduction of 4.02% compared to its original implementation.

Figure 4.4 shows the anomaly score calculated using the EIF model for the entire test period. The moving average of the anomaly score illustrates that the observations are time-related, noting that the level of health tends to fluctuate over time. Factors like wear level, time of operation, maintenance, and shutdown directly impact the health index (HI).

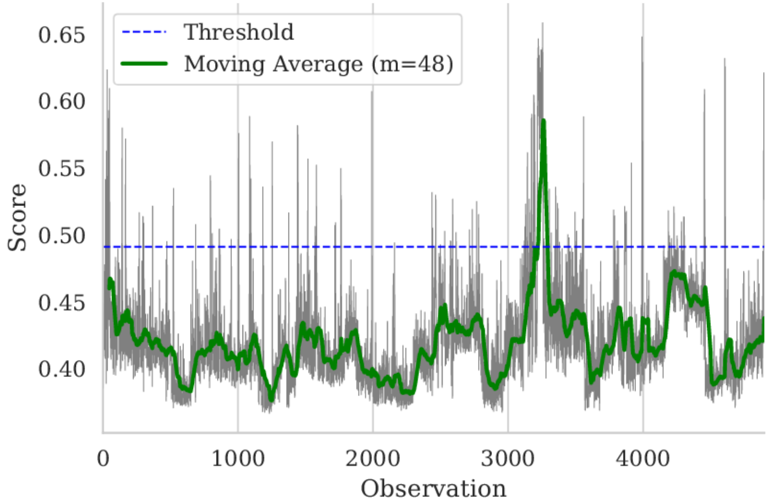


Figura 4.4 – iForest anomaly score for the entire test period. The upper limit (threshold) represents the 0.95th quartile of the training data anomaly score and is shown in blue, while the moving average of 48 periods shows the trend of the series.

Analyzing some sample cases give us a better visualization of the model outcomes in practice. Figure 4.5 displays three selected time series excerpts, which comprise five faults out of the 59 registered in the dataset. iForest learned from the training set, and we analyzed the behavior of the model in a prominent failure situation. The s score represents the average count length of all iTrees. The threshold is set based on a confidence interval of 0.95. Every moment that the score exceeds the threshold is considered an anomaly. The actual anomalies are represented by red crosses, while the model detections are represented by blue dots. The observed excerpt periods are 12/17/2018, 1/18/2019, and 10/17/2018, respectively. Unconnected line periods represent missing data, which may occur when communication with the supervisory system is interrupted by network connection problems.

In the first period, the iForest model detected two anomalies before the first registered system failure. The first detection occurred 10 hours before the failure, and the second detection occurred 2 hours before the failure. In this case, an intervention in the HGU system could have avoided the failure. A late detection was raised just after the first failure, which is considered a near detection. One hour before the second failure, the EIF model detected an anomaly. In total, the model detected four anomalies while two faults registered. The anomaly score profile calculated in this first excerpt is unsteady,

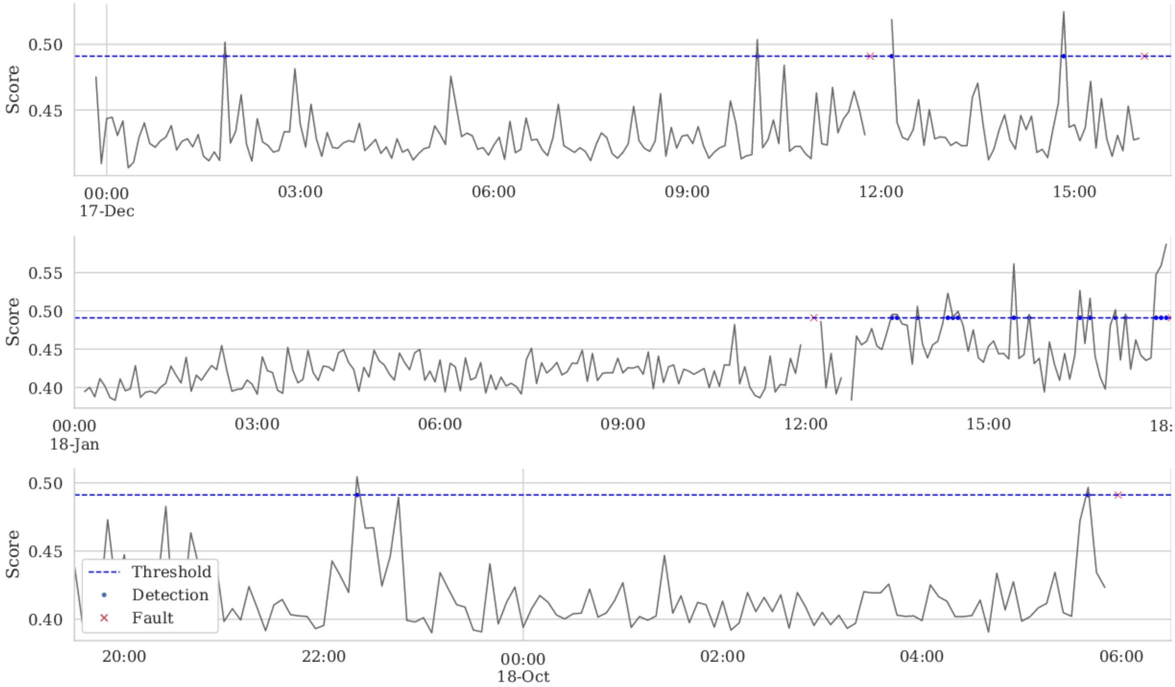


Figura 4.5 – Three series excerpts, illustrating process monitoring time using EIF. Whenever the value calculated by the model exceeds the established limit, the measurement is considered an anomaly.

with sudden fluctuations in the HI throughout the plotted period, making the detection task even more difficult.

The anomaly score increased over time in the second observed period, resulting in the first fault around noon. The HGU was rebooted and continued to operate, while iForest detected several anomalies. Six hours later, another fault occurred. In this example, the system could not detect any anomaly before the first failure, although it identified an abnormal state of operation between failures. It is possible to visually identify a positive growth trend in the anomaly score. The number of detections, in this example, associated with iForest results in Table 4.2, can explain the higher value of l and TTC.

In the third example, the anomaly score calculated between 22:00 and 23:00 detected an unusual operation, identifying two anomalies and presenting a situation in which the system nearly failed. However, the online monitoring system registered no fault until the next day. The model detects an anomaly minutes before the real fault. The damage is severe, and the system is shutdown. Throughout the rest of the observed period, the anomaly score was relatively low, close to 0.40.

Prognosis usually adopts the anomaly score to forecast the HGU behavior. This approach is commonly used to predict vibration trends and, since the score is a non-linear combination of the monitored variables, it is likely that it also reflects the frequencies and amplitudes of the original signals. Some examples of vibration prediction models for prognosis can be found in (FU et al., 2019) and (ZHOU et al., 2019).

The model has the limitation of not allowing the analysis of the importance of attributes, as do other models based on decision trees. The choice of the separation attributes in each node is random, and not generated from an explicit rule. However, machine understanding models, such as permutation-based and depth-based isolation forest, feature importance that can be used to circumvent this model limitation (CARLETTI et al., 2019).

From our simulations, we found that EIF obtained an average TD reduction of 1628 (40.62%) compared to PCA, and of 187 (7.28%) compared to KICA-PCA. These results indicate that the anomaly detection algorithms are efficient and suitable for dealing with the problem of intelligent fault detection in hydroelectric plants, as indicated in the qualitative analysis of imminent failure. In some cases, the anomaly score depicts the trend in the risk of failure. In other cases, the anomaly score identifies regions of at-risk operation, even though no fault is registered.

Continuous improvement of the model is found in associating the detected fault patterns with known failure modes, using fault analysis techniques such as fault trees (JONG; LEU, 2013; MELANI et al., 2016; CHENG et al., 2019b). The anomaly score that is calculated can be used in future work to develop forecasting systems. The adoption of a single dimension HI simplifies the process control and the design of the predictive system. Instead of predicting each variable in isolation, one can focus on analyzing a single time series, which carries the individual characteristics of each of the individual measurement variables.

4.4 Conclusions

In the present paper, we propose the application of iForest for fault diagnosis in a small hydro-electric plant in CBM. The observed period is approximately one year, and the main input variables are vibration, oil inflow and apparent power. The model benchmarks, in the recently reported hydro-power fault diagnosis literature, are PCA and KICA-PCA , using the specific metrics of time series anomaly detection, temporal distance, and average detection count. The tree ensembles presented promising results, with lower error levels and variance than KICA-PCA. Another significant advantage of adopting iForest and EIF is their capability for parallel computing, which speeds up model training while keeping memory usage low, and fixed to a known limit.

Identifying failures before they occur is vital to allowing better management of asset maintenance, reducing operating costs and, in the case of SHPs, enabling the expansion of renewable energy sources in the energy matrix (ZHANG et al., 2017a). With the application of machine learning models such as iForest and EIF, the aim is to improve the health of the equipment and reduce power generation downtime.

Future studies should include investigating feature and model selection through exhaustive searching, Bayesian or evolutionary optimization, as parameters manually adjusted. Fine-tuning the models can contribute even more to increasing model accuracy. A step towards the prognostic model can be taken from the prediction of the anomaly score by decomposing the signal into components in the time and frequency spectrum, and combining methods of extracting attributes with uni- or multi-variate forecasting (QIAO; CHEN, 2015; ZHOU et al., 2019).

Another essential beneficial area of the present study is identifying feature importance in a SHP diagnosis system. This knowledge can guide the development of CBM systems by prioritizing the installation of critical sensors in SHP automation projects. EIF, since it is a generalization of iForest, can be combined with forward selection component analysis (PUGGINI; MCLOONE, 2018) for automatic variable selection.

Finally, the present study contributes to the improvement of SHP maintenance, a vital renewable power resource with huge potential for energy supply worldwide. By identifying faults before failure, management can take actions to avoid further damage caused to joint systems and further aggravation of the components, lowering the operating costs of power plants.

Nomenclature

The following nomenclature is adopted in this chapter:

$\mathbf{A}$	mixing component matrix
$\mathbf{C}$	classification vector
CTT	candidate to target distance
$\mathbf{D}$	diagonal matrix of eigenvalues
$\mathbf{E}$	residual matrix
$f_{closest}$	closest distance
F_α	F-distribution
h	average of path length
k	radial basis function
l	detection count
m	number of variables
$\vec{m}$	random number with m dimension
n	number of observations
nt	number of trees
$\vec{p}$	intercept
$\mathbf{P}$	loading matrix
$\mathbf{R}$	co-variance matrix
s	anomaly score
$\mathbf{S}$	independent component matrix
t	score
$\mathbf{T}$	score matrix
T^2	Hotelling's score vector
T^2_α	threshold
TD	temporal distance
TTC	target to candidate distance
$\vec{x}$	vector of observations
$\mathbf{X}$	matrix of observations
$\mathbf{X}'$	matrix of observations transformed
$\bar{\mathbf{X}}'$	matrix of observations transformed and whitened
$\mathbf{Y}$	observations time series
z	low dimensional
α	degree of freedom
Δt	time difference
ψ	sub-sampling size
Φ	kernel's lifting

5 A DATA-DRIVEN FRAMEWORK FOR SMALL HYDROELECTRIC PLANTS PROGNOSIS USING TSFRESH AND MACHINE LEARNING SURVIVAL MODELS

Abstract: Maintenance in small hydroelectric plants (SHPs) is essential for securing the expansion of clean energy sources and supplying the energy estimated to be required for the coming years. Identifying failures in SHPs before they happen is crucial for allowing better management of asset maintenance, lowering operating costs, and enabling the expansion of renewable energy sources. Most fault prognosis models proposed thus far for hydroelectric generating units are based on signal decomposition and regression models. In the specific case of SHPs, there is a high occurrence of data being censored, since the operation is not consistently steady and can be repeatedly interrupted due to transmission problems or scarcity of water resources. To overcome this, we propose a two-step, data-driven framework for SHP prognosis based on time series feature engineering and survival modeling. We compared two different strategies for feature engineering: one using higher-order statistics and the other using the Tsfresh algorithm. We adjusted three machine learning survival models—CoxNet, survival random forests, and gradient boosting survival analysis—for estimating the concordance index of these approaches. The best model presented a significant concordance index of 77.44%. We further investigated and discussed the importance of the monitored sensors and the feature extraction aggregations. The kurtosis and variance were the most relevant aggregations in the higher-order statistics domain, while the fast Fourier transform and continuous wavelet transform were the most frequent transformations when using Tsfresh. The most important sensors were related to the temperature at several points, such as the bearing generator, oil hydraulic unit, and turbine radial bushing.

Keywords: Hydroelectric power plant. Condition-based maintenance. Prognosis. Survival analysis. Time series feature engineering. Survival random forest.

5.1 Introduction

The expansion of renewable energy sources is vital for ensuring the energy supply of a fast-paced market growing in the coming decades, with expectations for it to double by 2060 (WEC, 2019). Clean energy already accounts for three quarters of newly installed capacity annually (WEC, 2019), and those related to water resources are the most-used ones. The building of small hydroelectric plants (SHPs), which accounts for a significant share of this group, has increased worldwide due to the lower initial investment, lower operating costs, and expanding regulation of energy markets. The potential total energy generation capacity of these SHPs is twice the total capacity of the currently installed energy plants (UNIDO, 2016).

The maintenance of a hydropower plant is a complex task. It demands a specific level of skill to ensure an adequate level of dependability of the asset through its useful life. There are three kinds of maintenance. The first and most basic is corrective maintenance, in which a component is replaced after a failure occurs. The second is preventive maintenance, which estimates the service life of a component and realizes a replacement once the operating lifetime is reached. Finally, there is predictive maintenance, in which the system condition is assessed from data periodically or continually acquired from various sensors (BOUSDEKIS et al., 2018; PENG et al., 2010). A predictive or condition-based maintenance system consists of two stages. The first stage is the diagnosis, which incorporates fault detection or anomalous operating conditions, fault isolation by subcomponents, and identification of the character and degree of the failure (PENG et al., 2010). The second stage is the prognosis, which involves using statistical and machine learning models in order to calculate the use life of the assets and the confidence interval of the estimation (SIKORSKA et al., 2011), foresee maintenance, and increase the dependability and availability of the generation units.

Many data-driven models have been proposed for fault detection and diagnosis in hydroelectric plants. These models include principal component analysis (PCA) (LIU et al., 2009), independent component analysis (ICA) (ŽVOKELJ et al., 2016), and a least square support vector machine (FU et al., 2019; QIAO; CHEN, 2015; VU et al., 2013; PENG et al., 2007). PCA decomposition is used to assist specialists in determining and selecting the principal features which contribute to cavitation in hydro-turbines (GREGG et al., 2017). Current studies have presented a new monitoring method based on ICA-PCA that can extract both non-Gaussian and Gaussian information from operating data for fault detection and diagnosis (GE; SONG, 2007). This ICA-PCA method has been expanded with the adoption of a nonlinear kernel transformation prior to the application of the decomposition method, which has become known as kernel ICA-PCA (ZHU et al., 2014). Zhu et al. applied this method in the hydropower generation context with increased success rates and lower fault detection delays than either the PCA or ICA-PCA applications. While most models rely on signal processing, De Souza Gomes et al. proposed functional analysis and computational intelligence models for fault classification in power transmission lines (GOMES et al., 2013). Santis and Costa proposed the application of isolation iForest for small hydroelectric monitoring, where iForest isolates anomalous sensor readings by creating a health index based on the average distance of the points to the tree root (SANTIS; COSTA, 2020). Hara et al. extended iForest's performance by implementing a preliminary step of feature selection using the Hilbert–Schmidt independence criterion (HARA et al., 2021). It is worth emphasizing that in addition to data-driven models, there is the application of analytic model-based methods, which have been presenting significant design results in the context of fault diagnosis in power systems, such as in (WU et al., 2020).

For prognoses, the techniques generally applied for estimating the use life are classified into statistical techniques, comprising regression techniques (SI et al., 2011), Wiener-, Gamma-, and Markovian-based processes such as machine learning techniques, comprising neural networks, vector support machines, and electrical signature analysis (SALOMON et al., 2019b), and principal component analysis (LEI et al., 2018), as well as deep learning techniques more recently, comprising auto-encoder, recurrent, and convolutional neural networks (ZHAO et al., 2019).

Reports on the prognoses of hydroelectric generating units are scarcer than publications related to their diagnosis (SANTIS et al., 2021). A great challenge in the area is proposing procedures that contemplate faults between different generating units and auxiliary interconnected systems (SANTIS et al., 2021). An et al. presented a prognosis model based on the application of Shepard's interpolation of three variables: bearing vibration, apparent power, and working head (AN et al., 2014). The signal is decomposed by applying intrinsic time-scale decomposition to a limited number of rotating components, and the artificial neural network is trained for each of the temporal components of the signal. Thereafter, the models present a similar framework, with varied individual methods for signal decomposition and regression models. Fu et al. applied variational mode decomposition for signal decomposition and a least square support vector machine regression model fine-tuned using an adaptive sine cosine algorithm (W. WANG K., 2018; FU et al., 2019). Zhou et al. combined a feature strategy using empirical wavelet decomposition for decomposing and Gram–Schmidt orthogonal process feature selection combined with kernel extreme learning machine regression (ZHOU et al., 2019).

Since feature extraction is a key factor in the success of data-driven diagnosis and prognosis systems, the Time Series Feature Extraction Based on Scalable Hypothesis Tests (TSFRESH, or TSF for short) algorithm has gained prominent attention in the literature, leading to better results than physical and statistical features alone (DINDORF et al., 2020). The algorithm is capable of generating hundreds of new features while reducing collinearity through its hypothesis test-integrated selection procedure. Tan et al. adopted TSF along with a probability-based forest for bearing diagnosis (TAM et al., 2020). A two-stage feature learning approach combining TSF and a multi-layer perceptron classifier was adopted for anomaly detection in machinery processes by Tnani et al. [44] and for earthquake detection by Khan et al. (KHAN et al., 2020).

Finally, the random survival forest (RSF) is a survival analysis model that has recently been adapted for data-driven maintenance prognosis systems. Voronov et al. proposed the application of RSF for heavy vehicle battery prognosis (VORONOV et al., 2020), an important part of the electrical system and mostly affected by lead-acid during the engine starting. Gurung adopted the RSF along with histogram data for interpretive modeling and prediction of the remaining survival time of components of heavy-duty trucks, aiming to improve operation and maintenance processes (GURUNG, 2020). Snider and McBean proposed an RSF-based model for the water main pipe replacement model, expecting savings of USD 26 million, or 14% (SNIDER; MCBEAN, 2021) of the total cost of the ductile iron pipe, over the next 50 years.

In this context, the present paper innovates by proposing a framework for the prognosis of hydroelectric plants, based on the TSF feature extraction and selection algorithm and survival analysis models. The authors did not find any evidence or study that has adopted a similar approach in the literature thus far. We compare the different strategies of feature engineering associated with three survival model analyses, evaluating the models using the concordance index metric. The main findings and contributions of the current paper are the following:

- The proposal of a data-oriented framework including feature engineering strategies and machine learning survival models for intelligent fault diagnosis of the SHP generating unit;
- Evaluation of the importance of attributes using the permutation importance method associated with the RSF survival model;

- Affirmation that the RSF survival analysis model associated with the TSF feature engineering hybrid model obtained the highest concordance index score (77.44%).

The remainder of the present article is organized as follows. Section 2 defines the study methodology, describing the methods, algorithms, and dataset applied. Section 3 presents the results and discussions of the simulations of the models, in addition to the outputs of the feature engineering strategies and survival analysis models, with illustrative examples of those models' inference. Finally, Section 4 presents the conclusions and recommendations for future work.

5.2 Problem Formulation

The prognosis problem was formulated as an inference problem based on historical data, specialist knowledge, external factors, and future usage profiles. Prognosis is a condition-based maintenance (CBM) practice widely applied to reduce costs incurred during inefficient schedule-based maintenance. In mechanical systems, the repetitive stresses from rotating machinery vibration temperature cycles leads to structural failures. Since mechanical parts commonly move slowly to a critical level, monitoring the growth of these failures permits evaluating the degradation and estimating the remaining component life over a period of time (MATHUR et al., 2001).

The current study was developed in Ado Popinhak, an SHP situated in the southern region of Brazil. With an installed capacity of 22.6 MW, the plant supplies energy to 50,000 residences. Monitoring data from the main single hydro generator unit were registered every 5 min, and the study period was from 13 August 2018 to 9 August 2019. Table 5.1 describes the number of runs by the generators contained in the dataset, the number of runs that ended due to failure, the average cycle time per run, and the longest cycle time.

The objective was to predict the remaining useful life (RUL) of a power system based on multiple component-level sensor and event data. The RUL information allows decision makers to better plan maintenance and interventions, improving availability and reducing costs.

Tabela 5.1 – Descriptive information of the runs by generators contained in the dataset.

Generator	No. of Total Runs	No. of Faulty Runs	Avg. Cycle Time	Max. Cycle Time
1	133	50	691	4067
2	64	20	1270	6162
3	157	89	972	6835
4	130	40	764	3026

Raw sensor reading data and event data, such as interventions, shutdowns, and planned and corrective maintenance, were curated and merged. The data registered were classified and split into runs, which are periods from the moment the generating unit is turned on until it is shut down, whether due to failure or not. Runs that ended because of failure were labeled with a true or false failure label. The time distribution plot until the end of the runs that ended in failure and those that were interrupted for another reason is shown in Figure 5.1.

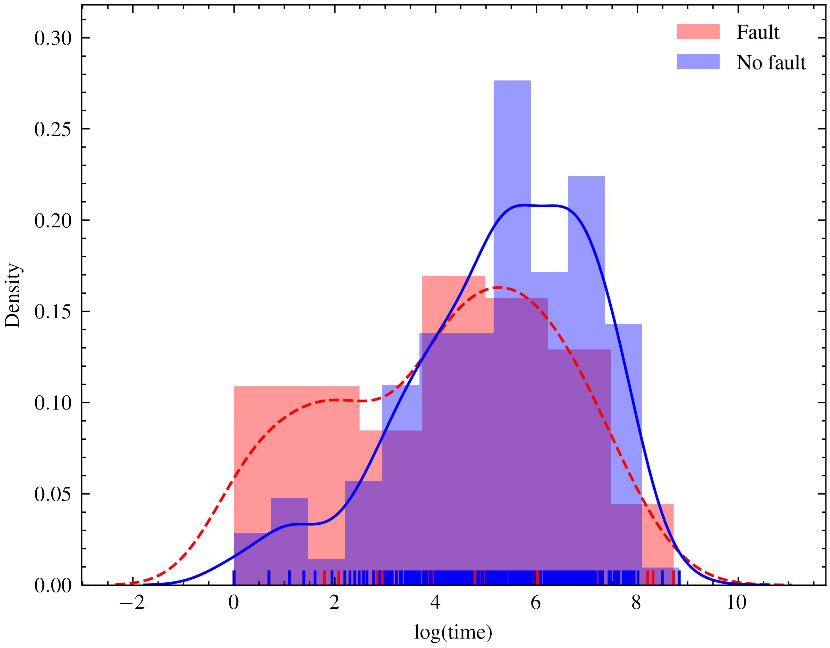


Figura 5.1 – Distribution chart of the last operating cycle time registered for runs with and without faults. The time cycle scale was converted to the logarithmic scale in order to better show the distribution of the variable. The average maximum cycle time of failed runs is less than that of normal operations. The distribution of both presents a bimodal characteristic, with two different concentration points more clearly verified in the faulty series.

The nature of the problem is interpreted as a problem of survival analysis, given that the system does not run to failure and can be shut down due to a lack of water resources for generation, failures in the transmission system, or the execution of scheduled maintenance. A summary of the characteristics of the problem and dataset is as follows:

- Data were collected for four generator units of the same manufacturer, model, and age;
- Fifty-four variables were monitored, and readings were registered in the transactional database every 5 min;
- Data were heterogeneous, including either control settings or monitored variables, both numerical and categorical;
- Missing data represented around 5% of total registrations, mostly caused by loss of a network connection between the remote plants and the operations center;
- The runs in which the subsequent state was a forced stop were labeled, and the last reading was registered as the logged time of failure;
- There were many runs where no failure was registered during the time (i.e., data were right-censored).

The runs were considered independent, given that the systems could be turned off for a long time and undergo modifications, such as routine maintenance, during this period, and because machine start-up is the biggest cause of system deterioration. For this reason, the Kaplan–Meier model was adjusted and presented in order to describe the survival function of each run of the four generators in Figure 5.2.

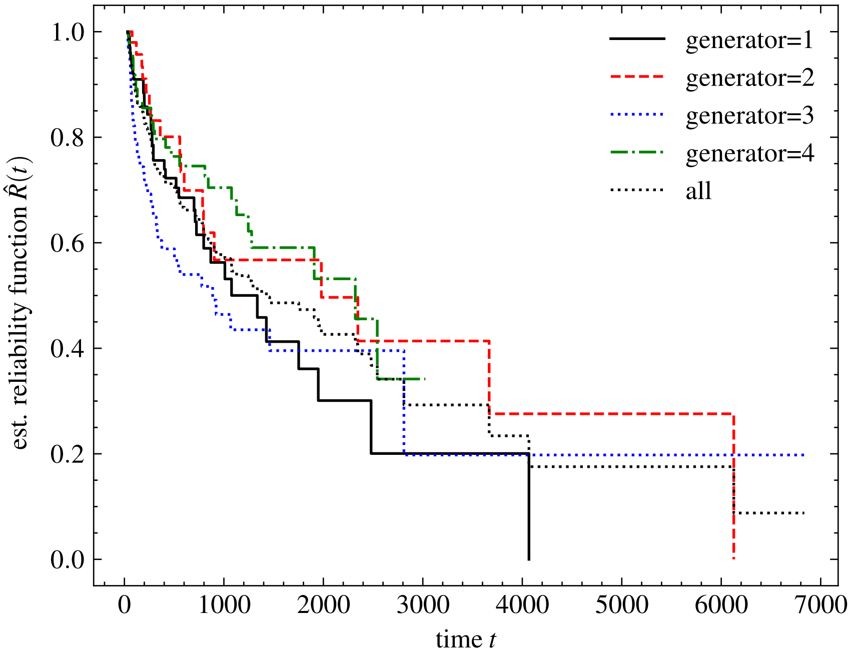


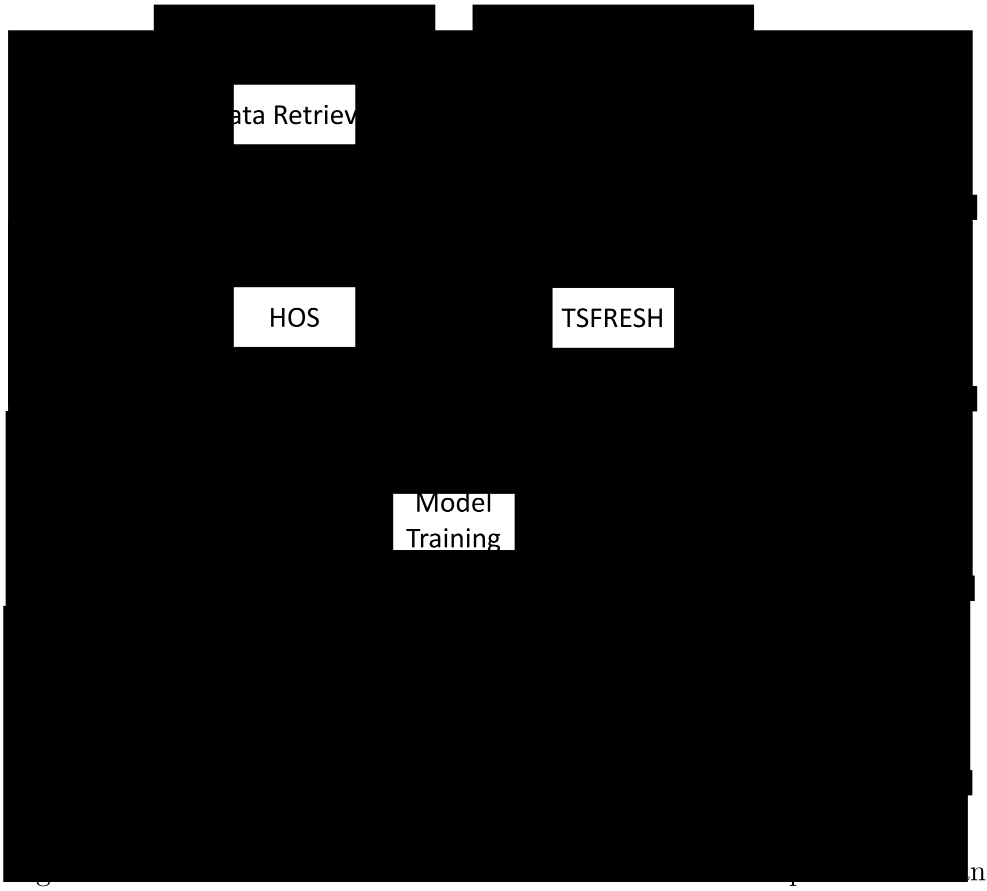
Figura 5.2 – Kaplan–Meier model adjusted for each of the generators. The decay rate of the survival function of the adjusted model for each of the generators was similar, indicating the correlated behavior of the health of the generating units.

The data transformation workflow is described in Figure 5.3. Sensor and event data were collected from the transactional database of telemetric systems and stored in text files. In the data-wrangling phase, the record tables were parsed and joined with the event tables, and the records were resampled into 5 minute periods. While still in this stage, the imputation of the missing data and the classification of the runs were made if they ended due to failure or programmed shutdowns (censoring).

In the feature extraction and selection step, the features were extracted from each of the time series of the sensors during the first 30 5 minute time units (150 min of operation) using the feature engineering strategies described in Section 2.2. The fixed period of the first 30 cycle times was selected from each run to extract features and adjust the survival models. Runs with a cycle time of fewer than 30 seconds were excluded from the training base. This approach was adopted to avoid data leakage in model training, where size-related features can contribute to models readily predicting the estimated total time to failure. These features were recorded in a text file and zipped due to the size of the generated tables.

In the next step, the runs were randomly divided into training and test sets, using proportions of 90% of the runs for training and 10% for testing. In each of the simulations, the partitioning was performed using a different random seed. The models were fitted to the training set, and metrics were calculated on the training set. The computational time was calculated for each fit of the models and saved for later analysis.

Finally, the model metrics were compared using a set of statistical tests in order to identify if there was a difference in the average scores for different groups of models or feature strategies.



nalysis.

5.3 Materials and Methods

5.3.1 Time Series Feature Engineering

5.3.1.1 Higher-Order Statistics (HOS)

Higher-order statistics (HOS) have been applied in different fields which require separation and characterization of non-Gaussian signals against a Gaussian background. Moments and cumulants are widely used to quantify certain probability distributions, such as location (first moment) and scale (second moment). Several authors have used HOS in signal processing. For example, De La Rosa and Muñoz reported the application of higher-order cumulants via signal processing using HOS for early detection of subterranean termites (ROSA; MUÑOZ, 2008), while Nemer et al. presented an algorithm for robust voice activity based on third- and fourth-order cumulants of speech (NEMER et al., 2001).

Let $X = [x(t)], t = 0, 1, 2, 3, \cdots$ be a real stationary discrete-time signal and its moments up to order p exist. Then, its pth-order moment can be given by (WELLING, 2005; NEMER et al., 2001)

$$m_p(\tau_1, \tau_2, \dots, \tau_{p-1}) \equiv E\{x(t)x(t+\tau_1)\cdots x(t+\tau_{p-1})\} \tag{5.1}$$

depending solely on the time differences $\tau_1, \tau_2, \dots, \tau_{p-1}$ for all i. $E(.)$ represents the statistical expectation for a deterministic signal. If the signal has zero mean as well, then its cumulant functions are given by (NEMER et al., 2001)

$$\text{second-order cumulant: } C_2(\tau_1) = m_2(\tau_1) \tag{5.2}$$

$$\text{third-order cumulant: } C_2(\tau_1,\tau_2) = m_3(\tau_1,\tau_2) \tag{5.3}$$

$$\begin{gathered}\text{fourth-order cumulant: } C_4(\tau_1,\tau_2,\tau_3) = m_4(\tau_1,\tau_2,\tau_3) - \\ m_2(\tau_1) * m_2(\tau_3-\tau_2) - m_2(\tau_2).m_2(\tau_3-\tau_1) - m_2(\tau_3).m_2(\tau_2-\tau_1)\end{gathered} \tag{5.4}$$

By setting all the lags to zero in the cumulant expressions and normalizing the input data to have a unity variance, we obtained the variance, normalized skewness, and normalized kurtosis:

$$\text{variance: } \gamma_2 \equiv C_2(0) = E\{x^2(n)\} \tag{5.5}$$

$$\text{normalized skewness: } \gamma_3 \equiv \frac{C_3(0,0)}{[C_2(0)]^{1.5}} = \frac{E\{x^3(n)\}}{[E\{x^2(n)\}]^{1.5}} \tag{5.6}$$

$$\text{normalized kurtosis: } \gamma_4 \equiv \frac{C_4(0,0,0)}{[C_2(0)]^{2}} = \frac{E\{x^4(n)\}}{[E\{x^2(n)\}]^{2}} \tag{5.7}$$

The skewness indicates to which side of the distribution the data are concentrated for unimodal distributions, so a positive skew indicates that the tail is to the right, and a negative skew indicates that it is to the left. The kurtosis is usually associated with the measure of the “peakedness” of the probability distribution of a real-valued random variable. Higher kurtosis means that more of the variance is due to infrequent extreme deviations, as opposed to frequent, modestly sized deviations. The first four moments were calculated for each of the runs in order to extract the basic descriptive variables of the sensor signals:

5.3.1.2 Tsfresh (TSF)

TSF is an algorithm presented for time series feature engineering, which accelerates this procedure by combining 63 time series characterization methods. Features are chosen based on automatically configured hypotheses (CHRIST et al., 2018).

Given a set of time series $D = \{X_i\}_{i=1}^{N}$, each time series X_i is mapped into a feature space with a problem-specific dimensionality M and feature vector $\overrightarrow{x}_i = (x_{i,1}, x_{i,w}, \cdots, x_{i,M})$. The feature vector $\overrightarrow{x}_i$ is built by applying time series characterization methods $f_j : X_i \rightarrow x_{i,j}$ to the respective time series X_i, which results in the feature vector (CHRIST et al., 2018)

$$\overrightarrow{x}_i = (f_1(X_i), f_2(X_i), \cdots, f_M(X_i) \tag{5.8}$$

The feature vector might be extended by additional univariate attributes $\{a_{i,1}, a_{i,2}, \cdots, a_{i,U}\}_{i=1}^{N}$ and feature vectors from other kinds of time series. For a machine learning system with K different time series and U univariate variables per sample i, the resulting design matrix would have i rows and $(K \cdot M + U)$ columns (CHRIST et al., 2018).

From the set of 63 characterization methods f_j available in the algorithm, we illustrate two of the most important ones based on our feature analysis, which are the fast Fourier transform (FFT) and the continuous wavelet transform (CWT). Both methods are time–frequency decomposition methods often applied in signal analysis.

The discrete Fourier transform (DFT) is a signal decomposition technique adequate for discrete and periodic signals. Let a signal a_n for $n = 0, \ldots, N-1$ and $a_n = a_{n+jN}$ for all n and j. The discrete Fourier transform of a, also known as the spectrum of a, is described by (HECKBERT, 1995)

$$A_k = \sum_{n=0}^{N-1} W_N^{kn} a_n \tag{5.9}$$

where $W_N = e^{-i\frac{2\pi}{N}}$ and W_N^k are called the Nth roots of unity. The sequence A_k is the DFT of the sequence a_n, where each is a sequence of N complex numbers. The FFT is a fast algorithm for computing the DFT into $log_2 N$ states, each of which consists of fewer computations (HECKBERT, 1995).

The CWT of a signal a with the wavelet ψ is defined as (MUNOZ et al., 2002)

$$W_\psi a(s,t) = \frac{1}{\sqrt{s}} \int_{-\infty}^{+\infty} a(x) \psi \frac{t-x}{s} dx \tag{5.10}$$

where the scale s is inversely proportional to the central frequency of the rescaled wavelet $\psi_s(x) = \psi x/s$, which is a bandpass, and t represents the time location of the signal analysis. The larger the scale s, the wider the analyzing function $\psi(x)$, and therefore the smaller the corresponding examined frequency. The main advantage over the Fourier transform methods is that the frequency description is localized in time, and that window size varies. It gives more flexibility and effectiveness than fixed-size analysis since low frequencies can be analyzed over wide time windows, while high frequencies can be analyzed over narrow time windows (MUNOZ et al., 2002).

Finally, the feature selection of TSF is used to filter out irrelevant features based on automated statistical hypothesis tests (CHRIST et al., 2018). Feature selection is crucial to reducing the number of variables, which increases generalization and prevents overfitting, in addition to bringing speed gains and less complexity to the estimator (ATTALLAH et al., 2017).

5.3.2 Survival Analysis

5.3.2.1 Evaluation Metrics

The most employed evaluation metric of survival models is the concordance index (C-index or C-statistic) (JR et al., 1996). It reflects a model's capacity of ranking the survival times based on the individual risk scores, and it can be expressed by the formula (UNO et al., 2011)

$$\text{C-index} = \frac{\sum_{i,j} 1_{T_j<T_i} \cdot 1_{\eta_j>\eta_i} \cdot \delta_j}{\sum_{i,j} 1_{T_j<T_i} \cdot \delta_j} \tag{5.11}$$

where η_i is the risk score of a unit, $1_{T_j<T_i} = 1$ if $T_j < T_i$ and is otherwise 0, and $1_{\eta_j>\eta_i} = 1$ if $\eta_j > \eta_i$. A C-index score equivalent to 1 corresponds to a perfect model estimator, while a C-index score of 0.5 represents a random estimator (UNO et al., 2011).

The C-index score can compare pairs in which the predictions and outputs are concordant, which means that the one with a higher risk score has a shorter actual survival time. If two instances experience an event at different times, or if one experiences an event and is outlasted by the other, we say that they are comparable. In contrast, a pair is said to not be comparable when they experience events at the same time (JR et al., 1996; UNO et al., 2011).

5.3.2.2 CoxNet (CN)

The Cox proportional hazard (CPH) assumes that the hazard is proportional to the instantaneous probability of an event at a particular time. In this case, the effect of the covariates is multiplying the hazard function by a function of the exploratory covariates. This means that two units of observation have a ratio of the constant of their hazards, and it depends on their covariate values (FISHER et al., 1999).

Let $Xi = (X_{i1}, \ldots, X_{ip})$ be the realized values of the covariates for a subject i. The hazard function for the CPH model is described by (COX, 1972)

$$\lambda(t|X_i) = \lambda_0(t) \exp(\beta_1 X_{i1} + \cdots + \beta_p X_{ip}) = \lambda_0(t) \exp(X_i \cdot \beta) \tag{5.12}$$

where $\lambda(t|X_i)$ is the hazard function at time t for subject i with a covariate vector $X)i$, $\lambda_0(t)$ is the baseline hazard, and β_i represents the effect parameters.

The CPH model is especially interpretive since the regression coefficients represent the hazard ratio, providing useful insights into the problem. However, in applications with a large set of features, the standard CPH fails due to the fact that the model convergence relies on inverting the matrix that becomes non-singular due to correlation among features (SIMON et al., 2011).

The CoxNet (CN) overcomes these problems by implementing an Elastic Net regression with a weighted combination of the l_1 and l_2 penalty by solving (SIMON et al., 2011)

$$\arg\max_{\beta} \ \log PL(\beta) - \alpha \left(r \sum_{j=1}^{p} |\beta_j| + \frac{1-r}{2} \sum_{j=1}^{p} \beta_j^2 \right) \tag{5.13}$$

where PL is the partial likelihood function of the Cox model, $\beta_1, \ldots, \beta_p$ are the coefficients for p features, $\alpha \geq 0$ is a hyperparameter that controls the amount of shrinkage, and $r \in [0; 1[$ is the relative weight of the $l1$ and $l2$ penalty. The l_1 penalty helps the model select only a subset of features, while l_2 leads to better stability through regularization. In this paper, we adopted the default value proposed for $r = 0.5$ and an automatic procedure for selecting $\alpha \geq 0.01$.

5.3.2.3 Random Survival Forest (RSF)

The random survival forest (RSF) model is an adaption of the random survival regressor for the analysis of right-censored survival data. The main components of the RSF algorithm are the growing of the survival trees and the forming of the ensemble cumulative hazard function. Survival trees are binary trees grown by the recursive splitting of tree nodes using a predetermined survival criterion. The splitting into nodes maximizes the survival distinction between the nodes, and eventually, each node of the tree becomes homogeneous and populated by cases with similar survival. Once training is complete, the cumulative hazard function estimation $\lambda(t|X_i)$ of each survival tree is described by the function (ISHWARAN et al., 2008)

$$\lambda(t|X_i) = \hat{\lambda}_h(t) = \sum_{t_{l,h} \leq t} \frac{d_{l,h}}{Y_{l,h}} \tag{5.14}$$

where $d_{l,h}$ is the number of failures and $Y_{l,h}$ is the operation run at risk at time $t_{l,h}$. The ensemble cumulative function estimation $\lambda_e^*(t|X_i)$ is the simple average of the M base estimators and is given by (ISHWARAN et al., 2008)

$$\lambda_e^*(t|X_i) = \frac{1}{M} \sum_{j=1}^{M} \lambda_b^*(t|X_i) \tag{5.15}$$

where $\lambda_b^*(t|X_i)$ represents the cumulative hazard function estimation of jth survival trees in the ensemble and M is the total number of survival trees. The base estimator hyperparameters were chosen from the convergence analysis performed on our data (shown in Figure 5.4). We used the parameter of the number of base estimators $M = 100$, as it was a value close to the smallest error observed. For the minimum value of the samples in each node, the value adopted was 15 samples, selected because it presented the lowest error for ensembles with 100 trees.

5.3.2.4 Gradient Boosting Survival Analysis (GBS)

The gradient boosting survival analysis (GBS) model was constructed using the gradient boosting framework for optimizing a specified loss function. The model was built on the principle of additively combining the predictions of multiple base learners into a powerful overall model (FRIEDMAN, 2002). GBS is an ensemble model similar to RSF, since it relies on multiple base learners to produce an overall prediction. The main difference between the two approaches is that while RSF independently fits the base learners and averages their predictions, the GBS model is assembled sequentially in a greedy, stage-wise manner. The GBS overall additive model f can be described by (FRIEDMAN, 2002)

$$f(X_i) = \sum_{m=1}^{M} \beta_m g(X_i; \theta_m) \tag{5.16}$$

where $M > 0$ represents the number of base learners, β_M is the weighting term, the function g refers to a base learner outcome, and θ is the parameterized vector.

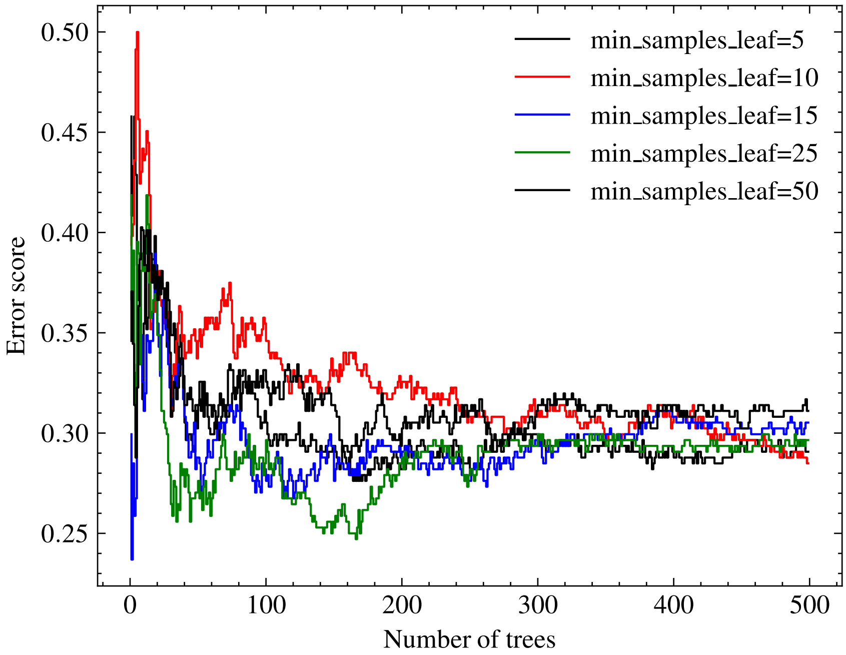


Figura 5.4 – Analysis of the convergence of the RSF model, varying the parameters of the number of base estimators M and the minimum of samples in each node. We adopted a standard sample count of 15, since it achieved the lowest error for ensembles with $M = 100$ estimators.

The loss function set for GBS is the partial likelihood loss of the CPH model. Therefore, the model maximizes the log partial likelihood function with the additive model f(X) such that (RIDGEWAY, 1999)

$$\arg\min_{f} \sum_{i=1}^{n} \delta_i \left[f(X_i) - \log \left(\sum_{j} \exp(f(X_j)) \right) \right]. \tag{5.17}$$

The base estimator of GBS, as in the RSF model, is the survival tree. In this way, we adopted the same control parameters for the estimator number M and a minimum number of samples in each node for both models.

5.3.3 Software and Hardware

All the routines, including data preparation, simulation, and result analysis, were developed using the Python language version 3.9.7 (ROSSUM; DRAKE, 2009), adopting the following common scientific libraries: scipy 1.4.1 (VIRTANEN et al., 2020) for statistical analysis and hypothesis testing, pandas 1.2.4 (MCKINNEY, 2010) for data wrangling, numpy 1.20.2 (HARRIS et al., 2020) for array manipulation, scikit-learn 1.0.2 (PEDREGOSA et al., 2011) for general data science functions, scikit-survival 0.17.2 (PÖLSTERL, 2020) for survival model implementation, tsfresh 0.19.0 (CHRIST et al., 2018) for the TSF feature extraction model implementation, matplotlib 3.4.2 (HUNTER, 2007) and seaborn 0.11.2 (WASKOM, 2021) for plots and visualization, and eli5 0.13.0 for permutation importance testing.

The specifications of the hardware used to perform the simulation were as follows: CPU Intel Core i9 2.30 GHz, 16 GB of RAM installed, and the macOS v.12.5 operating system. The approximate amount of time necessary to perform the data preparation, feature selection, and all 100 simulations was around three hours (one hour for feature extraction and two hours for simulation) without any parallelization. All scripts are available from the researcher's public repository (Github repository: <https://github.com/rodrigosantis1/shp_prognosis> accessed on 1 December 2022) for reproducibility and replicability. Data have not been made publicly available by the SHP but can be shared upon request.

5.4 Results and Discussion

5.4.1 Simulation Results

Figure 5.5 shows the C-index scores calculated for each of the 100 randomized simulations with different training and testing sets. This visualization format provides better understanding of the metric distribution of each of the CN, RSF, and GBS survival analysis models when combined with HOS and TSF feature engineering.

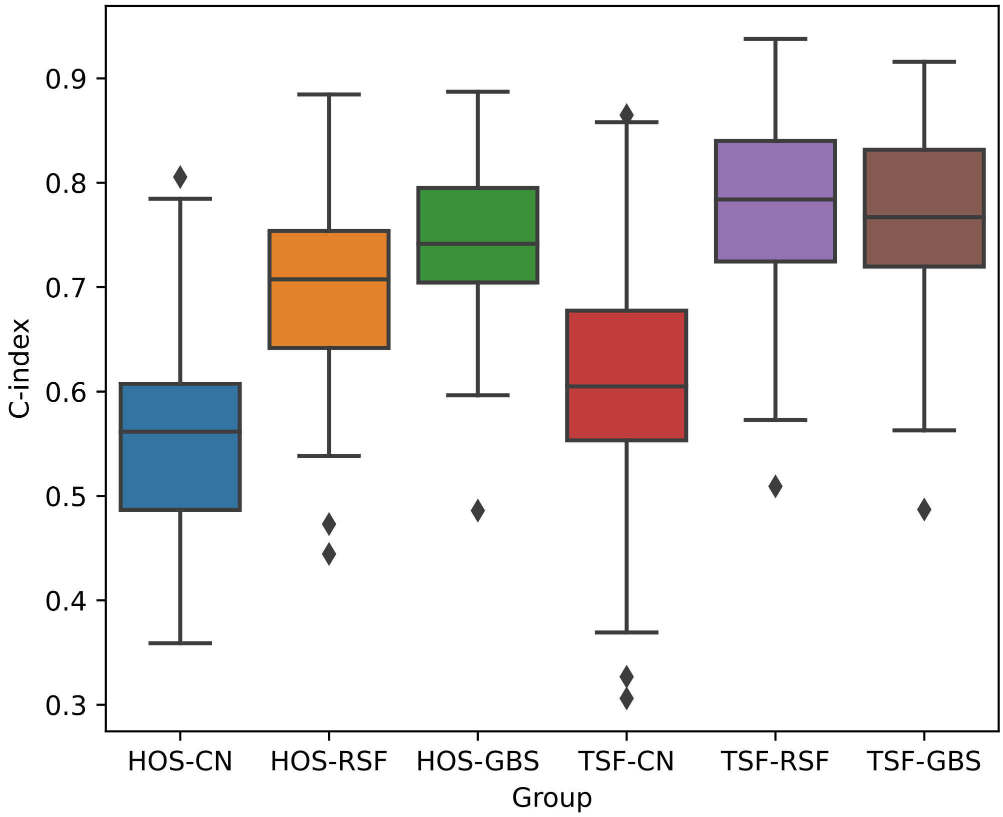


Figura 5.5 – Box plot representation of the C-index scores by group of feature strategies and survival models for the 100 simulations performed. The groups with the highest accuracy were TSF-RSF and TSF-GBS, while the lowest was HOS-CN.

From the box plot analysis, we observed that the HOS-CN group obtained the lowest accuracy, while the TSF-RSF and TSF-GBS groups obtained the highest accuracies. The variance of CN was higher than those for the other survival models, especially when adopted with TSF feature engineering.

There were a few outliers in all models which were mostly in the lower bound, indicating possible convergence problems. A suggestion for both improving the variances and reducing outliers is to adopt a model selection schema for tuning and adjusting the models. The TSF-RSF and TSF-GBA groups presented close distribution in terms of both median and variance. In general, most of the RSF and GBS groups presented close variance.

Table 5.2 presents the average and standard deviation of the C-index score and fitting time, highlighting in bold the model with the highest score and the one with the lowest fitting time. The nonlinear models RSF and GSA, which require more computational time for training, achieved better accuracy scores than the linear model with regularization (CN). This trade-off between accuracy and computational time is expected in machine learning applications. When comparing RSF and GBS, RSF required up to 10 times more fitting time than GBS. However, it is worth mentioning that RSF, a bagging ensemble, can be more easily parallelized than GBS, a boosting ensemble. The fitting time difference between the TSF-CN and the nonlinear models was significant, requiring more than 1000 times less time than TSF-RSF for training.

Tabela 5.2 – C-index score and average computational time. TSF-RSF obtained the best average score, while HOS-CN achieved the lowest score at a reasonably lower fitting time.

Model	**Description**	**C-Index**		**Fitting Time (s)**	
HOS-CN	Higher-Order Statistics + CoxNet	0.5562	0.0898	0.0053	0.0061
HOS-GBS	Higher-Order Statistics + Gradient Boosting Survival	0.7440	0.0736	1.7514	1.5064
HOS-RSF	Higher-Order Statistics + Random Survival Forest	0.7026	0.0843	14.6232	12.1850
TSF-CN	Tsfresh + CoxNet	0.6060	0.1060	0.0193	0.0023
TSF-GBS	Tsfresh + Gradient Boosting Survival	0.7644	0.0854	9.7241	2.2749
TSF-RSF	Tsfresh + Random Survival Forest	0.7744	0.0903	27.5949	23.2420

Table 5.3 presents the total time necessary to execute both feature engineering strategies. As this is a step preceding model adjustment, it is worth considering its time when evaluating the models.

The computational time required to extract and select attributes using the TSF method was about 20 times greater than the time required using HOS. This is a significant difference that must be taken into account, especially for real-time applications of the prognosis model. However, it is interesting to point out that the TSF library offers the possibility of implementing cluster parallel computing. Furthermore, the time required for inference was lower, given that only the features previously selected by the feature hypothesis tests and applied in the model training needed to be calculated.

Tabela 5.3 – Preprocessing time for feature engineering strategies. TSF technique requires about 20× more computational time than HOS.

Feature Engineering Strategy	**Preprocessing Time (s)**
Higher-Order Statistics (HOS)	4.68 s
Tsfresh (TSF)	67 s (extraction) + 26.6 s (selection)

One-way ANOVA (MCDONALD, 2014) was applied to test the null hypothesis that the groups had the same mean C-index score. Table 5.4 displays the F_S statistics, which represent the ratio of the variance among score means divided by the average variance within groups, and the p value calculated for the statistics. By adopting a confidence level of 0.95, we rejected the hypothesis that the score was equal between all groups since the p-value was lower than $alpha = 0.05$. Normality was checked using a Q-Q plot. The homogeneity of the variance when checking the ratio of the largest to the smallest sample standard deviations was less than 2 (1.44).

Tabela 5.4 – One-way ANOVA F_S statistics and P value for the null hypothesis of scores being equal for all groups. The hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_6$ was rejected at the significance level of $1 - \alpha > 0.95$.

F_S Statistics	**P Value**	**Reject H_0**
103.144	3.0036e-78	True

Sequentially, a pairwise Tukey test (TUKEY et al., 1977) was applied, and the results are presented in Table 5.5. The pair of groups in which the mean difference of the scores was not significant at the 0.95 level is highlighted in bold. The results show that the mean score metrics of the HOS-GBS, TSF-GBS, and TSF-RSF groups were statistically different. These results indicate that, from our experimentation, it is not possible to verify significant differences among the scores in these models. A reasonable model for the dataset we simulated was the HOS-GBS group, since it presented the least computational time for both preprocessing and fitting and was among the top three models.

Tabela 5.5 – The pairwise Tukey test q_S statistics and lower and upper bounds for mean differences between each pair of groups. The null hypothesis $H_0 : \mu_1 = \mu_2$ was rejected at a significance level of $1 - \alpha > 0.95$.

Group 1	**Group 2**	$\boldsymbol{\mu_1 - \mu_2}$	$\boldsymbol{q_S}$ **Statistics**	**Lower**	**Upper**	**Reject $\boldsymbol{H_0}$**
HOS-CN	HOS-GBS	0.1878	0.001	0.1519	0.2238	True
HOS-CN	HOS-RSF	0.1464	0.001	0.1105	0.1824	True
HOS-CN	TSF-CN	0.0498	0.0012	0.0139	0.0857	True
HOS-CN	TSF-GBS	0.2082	0.001	0.1723	0.2441	True
HOS-CN	TSF-RSF	0.2182	0.001	0.1823	0.2541	True
HOS-GBS	HOS-RSF	−0.0414	0.0132	-0.0773	−0.0055	True
HOS-GBS	TSF-CN	−0.138	0.001	-0.1739	−0.1021	True
HOS-GBS	TSF-GBS	0.0203	0.5743	−0.0156	0.0562	False
HOS-GBS	TSF-RSF	0.0304	0.152	−0.0056	0.0663	False
HOS-RSF	TSF-CN	−0.0966	0.001	−0.1325	−0.0607	True
HOS-RSF	TSF-GBS	0.0617	0.001	0.0258	0.0977	True
HOS-RSF	TSF-RSF	0.0718	0.001	0.0359	0.1077	True
TSF-CN	TSF-GBS	0.1584	0.001	0.1225	0.1943	True
TSF-CN	TSF-RSF	0.1684	0.001	0.1325	0.2043	True
TSF-GBS	TSF-RSF	0.01	0.9	−0.0259	0.0459	False

5.4.2 Feature Importance Analysis

Feature importance was evaluated using the permutation importance method, which measures how the score decreases when a feature is not available (BREIMAN, 2001). The score adopted for evaluation was the C-index, the base estimator was the RSF model, and the number of permutation iterations was equal to 15.

Table 5.6 presents the 20 most important features of the HOS-RSF combination, detailed by the sensor and the statistic used for aggregation into the feature used to train the RSF model.

The features associated with the speed registered in the speed regulator and the temperature of the oil were the most important features, contributing to an average increase of 2.8% in the C-index score. The features related to the coupled side bearing temperature of the generator were the most frequent ones (4/20), the oil temperature of the hydraulic unit was the second-most-frequent feature (3/20), and the uncoupled side bearing temperature was the third-most-frequent feature (2/20). When analyzing the type of aggregation, kurtosis and variance were the most frequent types (6/20), and skewness and average were the least frequent types (4/20), although there was a balance among all four types.

Table 5.7 presents the top 20 most important features of the TSF-RSF combination, detailed by the sensor and the type of feature extraction technique applied.

The most important features were the CWT coefficient of the radial bushing temperature in the turbine and the absolute energy of the voltage in the bar, which contributed to an increase of 1.1% in the C-index score. Since more features were extracted using the TSF strategy than the HOS, it was expected that the weight of each individual feature would be lower. The features related to the radial bushing temperature in the turbine were most frequent (4/20), followed by those related to the coupled bearing temperature in the generator (2/20) and the bar voltage (2/20). The features originating from CWT were most frequent (9/20), followed by the FFT (3/20). The dominance of the CWT and FFT indicates the importance and efficiency of the time–frequency decomposition methods in this type of application.

Tabela 5.6 – The 20 most important features of HOS-RSF using permutation importance. Column weight represents the average increase in the C-index when the feature is available.

Sensor	**Aggregation**	**Weight**	
Speed Regulator: Speed	Kurtosis	0.0285	0.0346
Hydraulic Unit: Oil Temperature	Skewness	0.0281	0.0156
Generator: Coupled Side Bearing Temperature	Skewness	0.0180	0.0189
Turbine: Downstream Shaft Sealing Temperature	Average	0.0149	0.0241
Voltage Regulator: Excitation Voltage	Variance	0.0120	0.0120
Turbine: Radial Bushing Temperature	Average	0.0116	0.0223
Spiral Case: Pressure	Kurtosis	0.0116	0.0064
Generator: Current T	Variance	0.0114	0.0083
Generator: Current S	Variance	0.0110	0.0048
Speed Regulator: Distributor	Variance	0.0107	0.0055
Generator: Coupled Side Bearing Temperature	Average	0.0103	0.0092
Generator: Uncoupled Side Bearing Temperature	Skewness	0.0093	0.0117
Coupled Side Bearing Vibration	Kurtosis	0.0091	0.0127
Generator: Uncoupled Side Bearing Temperature	Kurtosis	0.0085	0.0157
Generator: Voltage RN	Average	0.0081	0.0141
Hydraulic Unit: Oil Temperature	Kurtosis	0.0079	0.0096
Generator: Coupled Side Bearing Temperature	Kurtosis	0.0078	0.0127
Hydraulic Unit: Flow Switch	Variance	0.0076	0.0101
Hydraulic Unit: Oil Temperature	Variance	0.0074	0.0085
Generator: Coupled Side Bearing Temperature	Skewness	0.0074	0.0070

Tabela 5.7 – The 20 most important features of the TSF-RSF group by permutation importance. Column weight represents the average increase in the C-index when the feature is available.

Sensor	Extraction	Weight	
Turbine: Radial Bushing Temperature	CWT Coefficient	0.0122	0.0211
Bar: Voltage	Abs. Energy	0.0112	0.0256
Generator: Coupled Bearing Temperature	CWT Coefficient	0.0093	0.0147
Speed Regulator: Speed	Energy Ratio	0.0083	0.0144
Generator: Voltage RN	CWT Coefficient	0.0074	0.0125
Generator: T-Phase Winding Temperature	Autocorrelation	0.0070	0.0063
Turbine: Radial Bushing Temperature	CWT Coefficient	0.0066	0.0204
Generator: Voltage TS	CWT Coefficient	0.0064	0.0098
Generator: S-Phase Winding Temperature	CWT Coefficient	0.0064	0.0044
Generator: Voltage ST	CWT Coefficient	0.0064	0.0138
Generator: Reactive Power	Quantiles Change	0.0052	0.0080
Bearing: Vertical Radial Vibration	Index Max Quantile	0.0052	0.0044
Bar: Frequency	FFT Coefficient	0.0052	0.0166
Turbine: Radial Bushing Temperature	Lempel Ziv Complexity	0.0052	0.0064
Hydraulic Unit: Flow Switch	FFT Coefficient	0.0052	0.0088
Bar: Voltage	CWT Coefficient	0.0052	0.0182
Generator: Frequency	Quantiles Change	0.0050	0.0054
Generator: Coupled Bearing Temperature	CWT Coefficient	0.0050	0.0103
Turbine: Radial Bushing Temperature	Longest Strike above Mean	0.0050	0.0172
Turbine: Downstream Shaft Sealing Temperature	FFT Coefficient	0.0050	0.0078

It is important to note that the TSF algorithm includes the statistical aggregations of kurtosis, skewness, mean, and variance from the HOS feature engineering strategy. Additionally, none of these aggregations were present in the 20 most important attributes after the inclusion of more complex features, such as the FFT and CWT.

5.4.3 Model Application Analysis

In this section, we present a deeper look at the model which presented the highest mean score in the simulation: TSF-RSF. The C-index of the model was 0.8139. It is worth noting that the maximum value for the C-index is 1, which indicates the order of observed events followed the same order as all predicted events, and a C-index value of 0.5 indicates the prediction was no better than a random guess (SNIDER; MCBEAN, 2021). For comparative purposes, the application of the RSF method on the remaining service life of water mains obtained a C-index of 0.88 (SNIDER; MCBEAN, 2021), while for modeling the disruption durations of a subway service, the metric was 0.672 (WANG et al., 2022).

Figure 5.6 presents the reliability, and Figure 5.7 presents the cumulative hazard function plots predicted by the model for 20 instances randomly selected from the test set. When analyzing the representations, we can identify three operation cycles with a reliability pitfall in the earliest minutes of operation. These indicate some cases in which there was an intrinsic problem in the generator-turbine system prior to or during start-up, and those systems must be stopped as soon as possible for maintenance. There was a second group of four instances in which the reliability dropped by half in the first 1000 5 min time units. This behavior might be related to some operating conditions that were observed in the operation of the machine. Finally, there was a third group containing the other instances with a steadier rhythm of reliability decay, in which more than half of the systems were expected to fail after 2000 5 min time units.

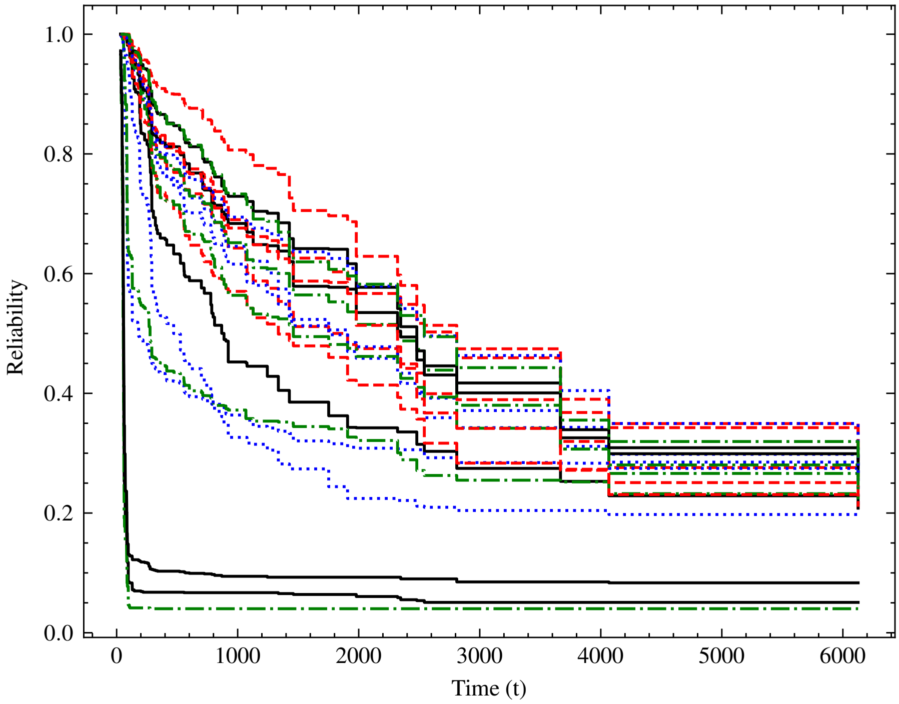


Figura 5.6 – Reliability function estimate of test samples ($n = 20$) using TSF-RSF. With the passage of time units (t), the probability of the system not failing declines. According to the measured variables, the model estimates whether the reliability decays more abruptly or not. After 400 5 minute time units, stability in the operation of the generating units is expected.

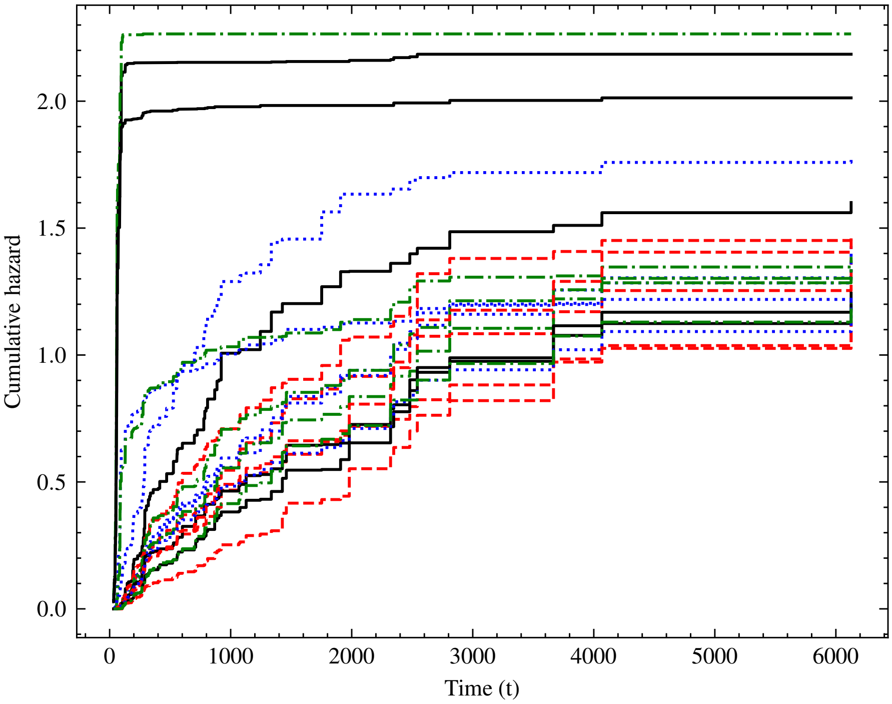


Figura 5.7 – Cumulative hazard function estimate of test samples ($n = 20$) using TSF-RSF. The cumulative hazard risk increased as the operation time increased, with the highest rate occurring in the first 3000 5 minute time units of operation. In unstable start-up operating cycles, this increase happened drastically in the first moments of operation.

In practical applications, the survival model can be used to evaluate both the current and previous runs of a generator unit, returning both the risks and the expected remaining useful life. Maintenance teams may want to keep all their systems with a reliability function closer to the third group described before, especially right before the rainy periods. During these periods, the generation is higher, and the stopped periods are rarer, making it more difficult and less desirable to execute maintenance procedures on the machines, which may lead to a loss in power generation.

The model can also be extended for a prescriptive perspective combined with the parameters of the start-up process, aiming to optimize the start-up process in order to achieve the highest reliability level possible. With this, a longer lifetime of the assets and greater time between failures can be expected.

5.5 Conclusions

In the present paper, we presented a structured modeling pipeline for survival analysis and remaining useful life estimation in a small hydroelectric plant in CBM. The available period of operations was approximately 1 year, and the 54 variables were monitored in 4 generating units of the same model and manufacturer. The HOS-GBS, TSF-RSF, and TSF-GBS models presented the highest C-index scores in our simulation. All three are suitable for production deployment.

Identifying failures before they happen is crucial for allowing better management of asset maintenance, lowering operating costs, and in the case of SHPs, promoting the expansion of renewable energy sources in the energy matrix (ZHANG et al., 2017a). Applying time series feature engineering and machine learning survival models, such as a framework, aims to enhance the health of the equipment and decrease power generation downtime.

Looking at variable importance, variance and kurtosis represented the most frequent transformation functions in HOS feature engineering, while the FFT and CWT were the most frequent transformations in TSF feature engineering. The sensors that contributed the most to the model accuracy were the generator bearing temperature, hydraulic unit oil temperature, and turbine radial bushing temperature. The data-driven framework presents generalities, and thus it can be reused to model generator units with different types of sensors.

Future studies should examine feature and model selection through exhaustive searching and Bayesian or evolutionary optimization, as the parameters were manually adjusted. Fine-tuning the models can contribute even more to improving the model accuracy. From the point of the modeling assumptions, runs are set to be independent, but features can be crafted to include times from other runs and from the last imperfect or perfect repairs. Additionally, the predictive model opens a path for prescriptive optimization of the machine operation parameters, aiming to minimize wear, operational wear, and risk over time. Reinforcement learning approaches are a prominent course of action, since they have been adopted for dynamically developing maintenance policies for multi-component systems such as the power system of our object of study. (YOUSEFI et al., 2020)

Finally, the present study contributes to the advancement of SHP maintenance, a crucial renewable power resource with enormous potential for supplying energy worldwide. By determining the faults before failure, management can carry out actions to avoid additional damage caused to combined systems and additional aggravation of the components, thus reducing the operating costs of power plants.

6 CONCLUSÃO

Nesta proposta de qualificação, apresentamos seis trabalhos desenvolvidos e publicados nos temas centrais da tese: análise e de séries temporais e a manutenção preditiva em PCHs e CGHs. Os trabalhos são interligados e apresentados de forma sequencial ao seu desenvolvimento. Enquanto obtivemos sucesso em adotar técnicas estatísticas e computacionais aos dados monitorados para detecção e previsão de falhas, não foi possível propor um modelo de detecção de falhas em séries temporais baseado no DTSF. Destacamos aqui algumas conclusões relevantes de cada capítulo.

Da revisão bibliográfica, no capítulo 2, verificamos o aumento exponencial das publicações sobre aplicações de técnicas de manutenção preditiva no setor hidrelétrico. A aplicação de modelos de aprendizado de máquina vem sido amplamente defendida por pesquisadores nos estudos de caso relatados. Em especial as técnicas de aprendizado profundo, dada sua eficiência em desenvolver modelos de alta acurácia e capacidade de generalização, indicam uma oportunidade enorme de avanço neste tipo de sistemas preditivos. Com o desenvolvimento das plataformas de nuvem e seu respectivo poder computacional, estes algoritmos são capazes de lidar com banco de dados gigantescos, como os encontrados no contexto de monitoramento da saúde de máquinas. Estes avanços recentes, em harmonia com novas técnicas de entendimento de máquina (machine understanding), indicam um futuro em que a interação entre especialistas e sistemas inteligente será cada vez mais próxima e dinâmica.

No capítulo 3, as análises dos perfis de manutenção e operação e do perfil de operação das usinas nos elucidou quanto à realidade das usinas em funcionamento e as principais causas de indisponibilidade dos ativos de geração. Verificamos uma alta frequência de falhas na transmissão, porém a maior contribuição para o tempo total parado das unidades geradoras foi dada pelos componentes do sistema gerador-turbina. Em especial o sistema de sustentação dos rotores, representado pelos mancais e unidade hidráulica de lubrificação, representam mais de 25% do tempo parado por indisponibilidade forçada. A partir destas análises, e em conjunto com os estudos de revisão bibliográfica, direcionamos nossos esforços para um sistema de diagnóstico de falhas nestes componentes críticos.

No capítulo 4, propomos a aplicação de um método de detecção de anomalia conhecido como floresta de isolamento estendida, para o diagnóstico inteligente de falhas em PCHs e CGHs. O modelo não-supervisionado é capaz de isolar observações distantes da massa central de dados, utilizando como métrica a profundidade média nas árvores individuais do comitê. A saída do modelo reflete um índice de saúde único para cada unidade geradora, e com a definição de um ponto limítrofe de controle, os gestores de manutenção são capazes de acompanhar o risco de falha do sistema.

No capítulo 5, propomos a combinação de técnicas de engenharia de atributos de séries temporais (TsFresh) com modelos de sobrevivência para estimação da curva de sobrevivência da unidade geradora em relação ao tempo de operação. O modelo foi ajustado utilizando como métrica a o índice de concordância (C-index), e a predição é realizada nos primeiros 150 minutos de operação (que inclui a partida de máquina) e indica a probabilidade de falha nas próximas X unidades de tempo (1 unidade = 5 minutos).

Como próximos passos, sugerimos o aprofundamento na discussão no tema de diagnóstico inteligente de falhas (COSTA et al., 2019), realizando a classificação de falhas por tipos (falhas elétricas, hidráulicas e mecânicas) em um contexto de aprendizado supervisionado desbalanceado. Os dados são rotulados a partir das falhas acusadas pelo sistema supervisório. Serão avaliados modelos baseados em árvores de decisão, como florestas randômicas, gradient boosting; modelos estatísticos, como regressão logística com e sem penalização; e modelos baseados em redes neurais artificiais. A partir da análise da importância dos atributos dos mais apropriado para o problema, esperamos poder avaliar quais variáveis de monitoramento mais impactam cada tipo de falha.

Na linha de prognóstico, acreditamos que o modelo possa obter melhores resultados se forem exploradas novos atributos que modelem a dependência entre as corridas das máquinas e o histórico de reparos realizados. Isso porque o modelo proposto assume independência entre as corridas, enquanto na prática pode existir dependência. Ainda, destacamos a possibilidade do desenvolvimentos de modelos prescritivos por exemplo otimizando a operação da máquina de forma a minimizar o desgaste sofrido pelas unidades geradores principalmente durante a partida.

Quanto ao modelo DTSF para detecção de falhas, foram realizados alguns testes criando um índice de saúde a partir da média do perfil de correlação dos N análogos mais próximos. Embora a ideia tenha sido promissora, na prática o índice se tornou sensível a escolha do tamanho da janela de busca, enquanto grandes janelas levaram a modelos com pouca sensibilidade no decorrer do tempo. Em busca de melhorar o método proposto, foi utilizado a técnica de controle estatístico CUMSUM. Porém, novamente uma grande quantidade de parâmetros era necessário ser ajustado para se obter resultados satisfatórios.

Outra linha de desenvolvimento proposta é na área de prognóstico, que busca prever o comportamento em momentos futuros de uma dada variável de interesse no sistema. Para isso iremos aplicar métodos de previsão de séries temporais univariadas (COSTA et al., 2019), como por exemplo a suavização exponencial, modelos de auto-regressão, e baseados em análogos. No contexto multi-variado, os modelos xgboost e de redes neurais recorrentes serão aplicados. Os métodos serão avaliados utilizando métricas clássicas de de erro de regressão, como a média absoluta percentual entre valores previstos e observados, erro quadrado médio e coeficiente de determinação. Iremos testar modelos de decomposição em tempo-frequência, que vêm sido recentemente aplicados junto aos métodos univariados e demonstrado melhores resultados do que a aplicação isolada dos métodos de previsão temporal.

APÊNDICE: DEMAIS PUBLICAÇÕES

A DYNAMIC TIME SCAN FORECASTING: A BENCHMARK WITH M4 COMPETITION DATA

Abstract: Univariate forecasting methods are fundamental for many different application areas. M-competitions provide important benchmarks for scientists, researchers, statisticians, and engineers in the field, for evaluating and guiding the development of new forecasting techniques. In this paper, the Dynamic Time Scan Forecasting (DTSF), a new univariate forecasting method based on scan statistics, is presented. DTSF scans an entire time series, identifies past patterns which are similar to the last available observations and forecasts based on the median of the subsequent observations of the most similar windows in past. In order to evaluate the performance of this method, a comparison with other statistical forecasting methods, applied in the M4 competition, is provided. In the hourly time domain, an average sMAPE of 12.9% was achieved using the method with the default parameters, while the baseline competition – the simple average of the forecasts of Holt, Damped, and Theta methods – was 22.1%. The method proved to be competitive in longer time series, with high repeatability.

Keywords: Univariate methods. M4 competition. Benchmarking. Dynamic time scan forecasting.

A.1 Introduction

The development of predictive models is widely debated in the literature (HILL et al., 1994; PAI; LIN, 2005; DUDEK, 2016; SHANMUGAM, 2006), since it assists the control of associated uncertainty intrinsic to random variables. Given the above, there are several categories of predictive models based on this physical knowledge (such as spectral analysis (TCHRAKIAN et al., 2011)) of intensive machine learning and statistical approaches (VOYANT et al., 2017). Forecasting models associated with a single random variable as a function of time support univariate forecasting, which is a very important area given its application in various sectors such as (HASSANI; SILVA, 2018; CAI et al., 2017; BERNARDINI; CUBADDA, 2015), business (JAFFUR et al., 2017; ZHANG et al., 2017b; TULARAM; SAEED, 2016), energy (GIRISH et al., 2016; RANA et al., 2016; RAVIV et al., 2015), among others. In this context, it is fundamentally valuable to develop meticulous criteria for selecting the models (BILLAH et al., 2005).

The M-competition (MAKRIDAKIS et al., 2018; MAKRIDAKIS; HIBON, 2000; MAKRIDAKIS et al., 1993; MAKRIDAKIS; HIBON, 1979) is the most important forecasting competition in academia, in which researchers from all around the world test their methods on real-life, anonymous time series from distinct areas of industry. The 4th edition took place in 2018 (MAKRIDAKIS et al., 2018), and 17 methods based on combinations of statistical- and machine-learning or hybrids were tested on 100,000-time series. Outputs from these events are registered in review articles, pointing out the directions of development and refinement of the most promising forecasting techniques (FLORES et al., 2019). The 5th edition took place in 2020, and focused on a retail sales application with 42,850 unit sales hierarchical series, with the objective to produce the most accurate point forecast as well as the most accurate estimation of the uncertainty of these forecasts (MAKRIDAKIS et al., 2021). The 6th competition will take place this year and it will focus on predicting the overall market returns of individual stocks.

Whereas most well-known forecasting methods are based on identifying intrinsic components of the time series, such as level, trend, or seasonality, a particular group of methods based on similarity searches have been arousing interest in the areas of meteorology and renewable energy (YANG; ALESSANDRINI, 2019a; HOELTGEBAUM et al., 2021). These methods consist of identifying past weather patterns ("analogs") that closely resemble the current state. These methods are capable of handling lengthy historical time series in order to produce accurate and interpretive forecasts.

Among these methods, Dynamic Time Scan Forecasting (DTSF) consists of a new and simple analog-based forecasting technique (COSTA et al., 2021). It generates forecasts based on similar patterns, those with the highest R2 scores, calculated from the last available window.

The accuracy of analog-based methods is scarcely reported in areas other than energy prediction and is mostly limited to wind and solar energy forecasting applications (GONTIJO et al., 2020; GONTIJO et al., 2021), which begs the question: "are analog-search-based models competitive compared to classical statistical prediction methods?

. Additionally, no research was found that compared analog search methods and statistical methods.

To fill this gap, the current paper describes the DTSF forecasting method and discloses its performance on the M4 competition time series. We compare DTSF with eight classical statistical methods (Naive, Seasonal Naive, Simple Exponential Smoothing, Holt, Damped, Theta, AutoRegressive Integrated Moving Average (ARIMA), and ExponenTial Smoothing state space model (ETS)) and a combination of the outcomes of 3 individual methods (Holt, Damped, and Theta), which compose the baseline of the M4 competition. The M4 benchmark dataset was selected for this research because: (1) it consists of a reliable and curated benchmark base, adopted by other researchers and practitioners for developing and testing forecasting methods; (2) it has a significant number of series: 100,000 time series, with different frequencies (hourly, daily, monthly, weekly, quarterly, yearly); (3) it has been mostly predominated by statistical methods of forecasting; (4) and it is composed of univariate and independent series.

The major contributions of the present paper can be summarized as follows:

- the study applies a new method to M4 competition for benchmark purposes;
- the method is compared with nine classical statistical methods and a combination of the outcomes of three individual methods, which compose the baseline of the competition;
- in addition to applying the method, along with its default parameters, an exhaustive search with hold-out validation is adopted for model selection.

The major conclusions are:

- in the hourly time domain, an average error of 12.9% was obtained using the method with the default parameters, while the competition baseline was 22.1%;

- through the automatic selection of parameters, we boosted the accuracy of the method by 12.31% compared to the method application without parameters selection;
- the method proved to be competitive, both in terms of accuracy and computational cost, over long time series and with high repeatability.

The present paper is organized into 5 sections. Following this Introduction, Section 2 provides a review of the proposed forecasting method. Section 3 provides a background of the datasets and methods applied in this study. Section 4 presents the results and discussions obtained from the application of the methods. Finally, Section 5 concludes the present paper and includes some recommendations for future studies.

A.2 Materials and methods

A.2.1 M4 competition dataset

The data used in the current study comes from the M4 competition dataset (MAKRIDAKIS et al., 2018). It is composed of 100,000 time series, taken from different domains such as Economics, Finance, Demographics, and Industry, among others. The time series show different periods: yearly, quarterly, monthly, weekly, daily, or hourly.

Table A.1 summarizes the information about the competition's dataset. Domain refers to the time period from which the data have been extracted, ranging from hourly to yearly. The number of Series shows how many time series are available, in total. The dataset is mostly composed of a collection of time series from yearly, quarterly or monthly domains - 95,000 time series. The minimum length is the shorter time series in the given domain: the more aggregated the domain, like yearly, the more difficult it is to retrieve data. For example, hourly time series are longer, having at least 700 available observation points. Horizon refers to how many steps are being predicted in the future and are being used for metric computation. Seasonality represents the expected recurrence of an event in a given time domain.

Tabela A.1 – Summary of M4 competition dataset, including time-frequency, minimum length of time series, and forecast horizon of each time series.

Domain	Number of series	Min. length	Horizon	Seasonality
Yearly	23,000	13	6	1
Quarterly	24,000	16	8	4
Monthly	48,000	42	18	12
Weekly	359	80	13	52
Daily	4,227	93	14	7
Hourly	414	700	48	24

The dataset provides a public and reliable source for comparing statistical, machine learning, or hybrid methods on univariate time series forecasting (BONTEMPI, 2020). It is internationally recognized by researchers and data scientists as the most important competition in this area (FILDES; MAKRIDAKIS, 1995).

A.2.2 Dynamic time scan forecasting

DTSF is a forecasting method based on scan statistics (GLAZ; BALAKRISHNAN, 2012) and was originally developed to address the problem of wind forecasting for Brazilian power generation plants. It consists of scanning a time series and identifying past patterns (called "analogs") similar to the last observations available of the time series (called "query") (COSTA et al., 2021).

Let y_t be a time series of length N, $t = 1, ..., N$. Firstly, let vector $\mathbf{y}^{[w]}$ be defined as the last w observations of the series:

$$\mathbf{y}^{[w]} = [y_{N-w+1}, ..., y_N]. \tag{A.1}$$

The goal of DTSF is to identify analogs in the time series which are greatly correlated with vector $\mathbf{y}^{[w]}$. Hence, the set of candidate vectors can be defined by:

$$\mathbf{x}_t^{[w]} = [y_{t-w+1}, ..., y_{t-w}] \tag{A.2}$$

where $t = 1, ..., N - 2 \cdot w$. The upper limit of the time sequence $(N - 2 \cdot w)$ guarantees that vector $\mathbf{x}_t^{[w]}$ does not overlap with vector $\mathbf{y}^{[w]}$. Fig. A.1 presents the DTSF procedure. Given the last w observed values, which comprises vector $\mathbf{y}^{[w]}$, a rolling window with the same size $(\mathbf{x}_t^{w})$ is used for scanning previous values of the series.

Lastly, DTSF provides a k steps ahead forecast of the time series, $y_{N+1}, ..., y_{N+k}$. To produce this outcome, the DTSF scans the series to find the closest analogs $\mathbf{x}_t^{[w]}$. The subsequent values of the time series are used as the forecast values:

$$y_{N+i} = f_{\mathbf{x}_t^{[w]}}(y_{t-w+i}) \tag{A.3}$$

where $f_{\mathbf{x}_t^{[w]}}$ is a function which correlates the elements of vector $\mathbf{x}_t^{[w]}$ and the elements of vector $\mathbf{y}^{[w]}$.

According to that, a first constraint can be set on $k : 1 \leq k \leq w$. This constraint guarantees that if the most correlated time series window comprises the most recent values, prior to vector $\mathbf{y}^{[w]}$, then the forecast values are a function of vector $\mathbf{y}^{[w]}$,

$$y_{N+i} = f_{\mathbf{x}_{N-2w}^{[w]}}(y_{N-w+i}). \tag{A.4}$$

As stated in Equations (3) and (4), forecast values depend on the window length w and the function $f_{\mathbf{x}_t^{[w]}}(.)$. A intuitive proposal for function $f_{\mathbf{x}_t^{[w]}}(.)$ is a linear scaling of the elements of vector $\mathbf{x}_t^{[w]}$, i.e., a linear model. This occurs due to the fact that previous values are likely similar to the last observations, except for a scale and/or offset shift. So, the method searches for values that may be similar to the last values, after applying a similarity function (COSTA et al., 2021).

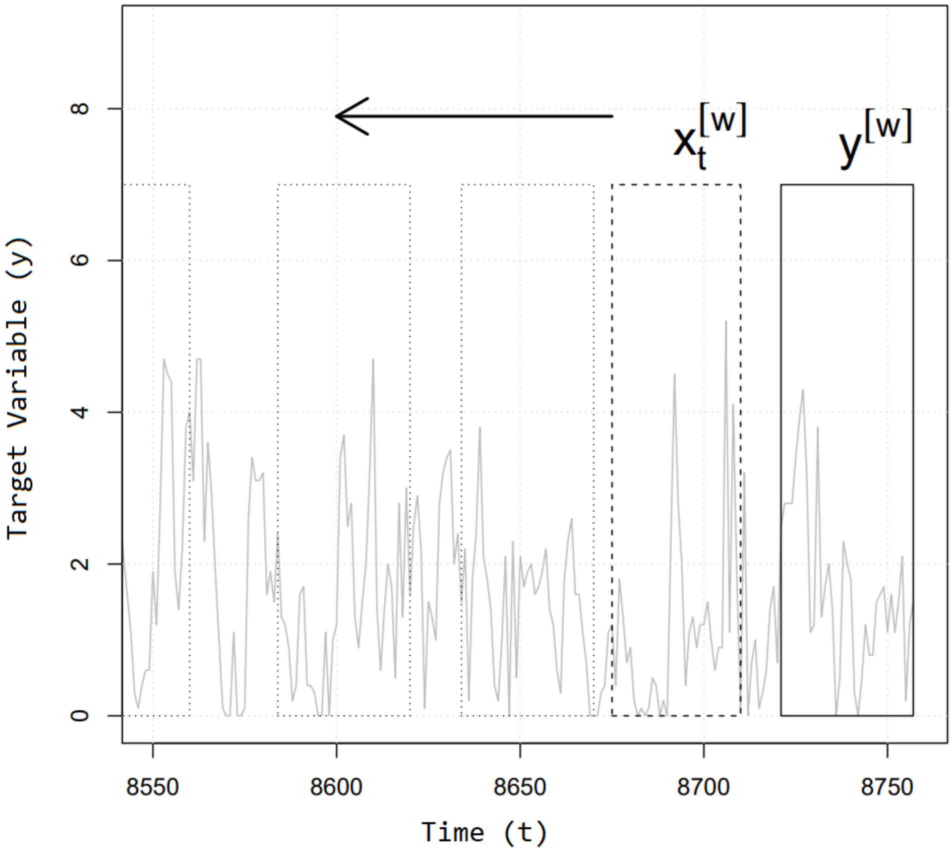


Figura A.1 – Illustration of the DTSF time series scan procedure.

By taking a linear function as the similarity function, the parameters of the model can be estimated to minimize the sum of squares between the elements of vector $\mathbf{y}^{[w]}$ and the linear equation: $\beta_0^{[t]} + \beta_1^{[t]} \times \mathbf{x}_t^{[w]}$. Moreover, the similarity statistic can be assumed as the linear regression coefficient of determination R^2 (COSTA et al., 2021; MONTGOMERY et al., 2021):

$$R^2 = 1 - \frac{\sum_j \left(\mathbf{y}_j^{[w]} - \hat{\mathbf{y}}_j^{[w]}\right)^2}{\sum_j \left(\mathbf{y}_j^{[w]} - \bar{y}_j^{[w]}\right)^2} \tag{A.5}$$

where $\mathbf{y}_j^{[w]}$ is the j-th value of vector $\mathbf{y}^{[w]}$ and $\hat{\mathbf{y}}_j^{[w]}$ is the j-th predicted value using the estimated linear function. Finally, the method calculates a similarity profile based on the R^2 score resulting from the comparison of the query with previous windows. The analogs with higher R^2 scores are considered closer analogs. Predictions of future steps are calculated from a predefined number of analogs using aggregation functions, such as median (COSTA et al., 2021).

The DTSF model requires three parameters to be selected by the user: the length of the query window, the similarity function specification, and the number of analogs to be considered. The original implementation of DTSF is available on the R package, DTScanF. In the present study, the original implementation is the extent to which the aggregation function applied to analogs can be either the median or the mean, according to the user or the model selection procedure.

Fig. A.2 illustrates the forecasting procedure, using time scanning in a given hourly time

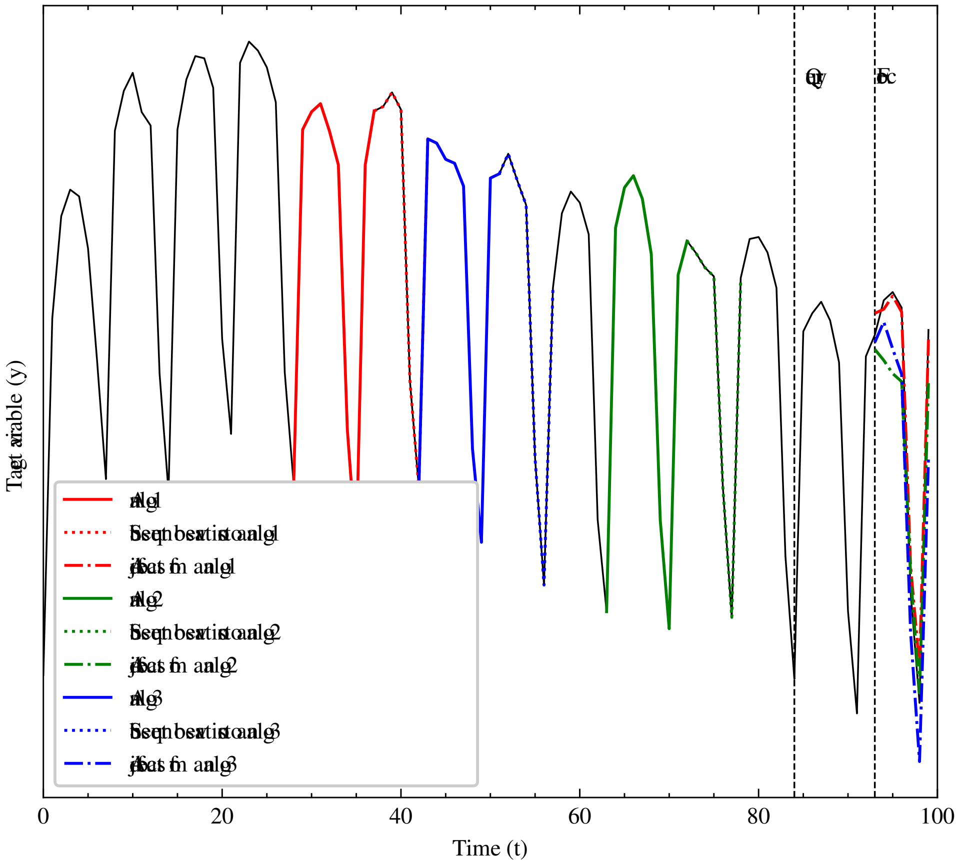


Figura A.2 – Example of DTSF application to forecasting a time series. The three colored lines represent the top three analogs correlated to the queried period. The dashed lines are the subsequent observations of the analogs. The forecast is given by the median of the adjusted forecast from the subsequent observations of the top analogs.

series, adopting a window with a length equal to 48 hours, a linear similarity function (degree equal to 1), and the three analogs. Windows 1, 2, and 3 are the ones most similar to the last window of available data. The forecast is given by the median (but other statistics can be used such as the mean) of the subsequential observations of the analogs.

As a data-driven method, DTSF usually performs better on time series with large numbers of observations and it can also be extended to search the patterns of secondary series related to the prediction. The main disadvantage of the method is the computational cost of scanning the entire time series and calculating the similarity profile. However, more efficient methods, such as the Maureen's Algorithm of Similarity Search (MASS) which applies convolution, have been applied for speeding up this task (GONTIJO et al., 2020). To keep it feasible, the linear similarity functions commonly adopted are from the first to the third-degree polynomials.

A.2.3 Statistical forecasting methods

A univariate forecasting method is a procedure for estimating a point. The forecast is based on past and present values of a given time series (CHATFIELD, 2000). This method is generally applied when there is a large number of series to forecast, or when multivariate methods require forecasts for each explanatory variable. Given the advantage of simplicity and high usage, univariate forecasting methods are employed in most of the forecast applications in areas such as business, energy, and finance. The following methods are selected from the latest M4 competition benchmark (MAKRIDAKIS et al., 2018), and a simple explanation is given for each one, as follows:

1. *Naive*: the simplest, yet still powerful forecasting method; assumes that the next steps to be predicted are equal to the last available observation (MAKRIDAKIS; HIBON, 1979).
2. *Seasonal Naive (sNaive)*: the same concept as Naive, with the adaptation that the time series is deseasonalized; method adjusted and forecast later, re-adjusted with the seasonal component (MAKRIDAKIS; HIBON, 1979).
3. *Naive2*: each time series uses the forecast of either Naive or sNaive, based on their score on the validation set.
4. *Simple Exponential Smoothing (SES)*: classic statistical method which applies an exponentially weighted average (HYNDMAN et al., 2008).
5. *Holt*: exponential smoothing with level and linear trend components (HYNDMAN et al., 2008).
6. *Damped*: exponential smoothing with dampened parameters for flattening trends, after a given period (GARDNER; MCKENZIE, 2011).
7. *Theta*: method based on a coefficient of curvature of the time-series, applied to the second difference of the data (ASSIMAKOPOULOS; NIKOLOPOULOS, 2000).
8. *Combined (Comb)*: the simple average of the forecasts of the previous three models: Holt, Damped and Theta.
9. *ARIMA*: general forecast method estimated from the autoregressive, moving average and integration components from the time series analysis (BOX; PIERCE, 1970).
10. *ETS*: automatic forecasting based on an extended range of exponential smoothing methods (HYNDMAN et al., 2002).
11. *DTSF*: the proposed method, adopting the defined default parameters, which are: (i) polynomial function degree equal to 1, (ii) analogs equal to 10, (iii) window size equal to length of forecast horizon, and (iv) median as aggregation function (COSTA et al., 2021).

Table A.2 presents the range adopted for the parameters of the proposed method. The polynomial degree is the degree of the function used for approximation, analogs are the number of analogs to be used to estimate the forecast, window size defines the length of the scan window, and aggregation function is the one that transforms the projection of the analogs into the final forecast.

Tabela A.2 – Parameters range adopted for DTSF.

Parameters	**Range**
Polynomial degree	1
Analogs	10
Window size	48
Aggregation function	Median

A.2.4 Model selection procedure

The split of the data into training sets and test sets split is predefined and given by the competition organizers. The data come from different files for each of the time series domains. The test set has a fixed horizon for all the time series, and it is used only for computing the final scores. The evaluation metrics adopted are the same ones that are applied in the M4 Competition, and are those most used in literature (AL-ALAWI; ISLAM, 1996; AZADEH et al., 2008): the Symmetric Mean Absolute Percentage Error (sMAPE), Mean Absolute Scaled Error (MASE) and Overall Weighted Average (OWA). The formula for calculating the metrics is given:

$$sMAPE = \frac{1}{h}\sum_{t=1}^{h}\frac{2|Y_t - \hat{Y}_t|}{|Y_t| + |\hat{Y}_t|} \tag{A.6}$$

$$MASE = \frac{1}{h}\frac{(n-m)\sum_{t=1}^{h}|Y_t - \hat{Y}_t|}{\sum_{t=m+1}^{n}|Y_t - Y_{t-m}|} \tag{A.7}$$

$$OWA = \frac{sMAPE_k/sMAPE_{base} + MASE_k/MASE_{base}}{2} \tag{A.8}$$

where Y_t is the post sample value of the time series at point t, $\hat{Y}_t$ is the estimated forecast, h is the forecasting horizon, m is the frequency of the data, k is a given regressor, and *base* is the sNaive estimator.

A hold-out cross-validation scheme is adopted to evaluate and select the best parameters for the methods, in which the last k observations are kept as the validation set, k being equal to the forecast horizon. All possible parameter combinations are enumerated within the defined ranges, and the methods are tuned using an exhaustive grid search procedure with sMAPE as the scorer.

A.2.5 Software and hardware

Routines were implemented using the R 3.6.0 programming language with the official benchmarks and evaluation script of M4 Competition, available at the GitHub repository (https://github.com/M4Competition/M4-methods). The *Forecast 8.7* package is used for the SES, Holt, Damped, ARIMA, and ETS methods. DTSF comes from the official implementation of the method in R and C++, available from the public repository (https://rdrr.io/github/leandromineti/DTScanF/) All data and scripts are available from the authors upon request.

Computer specifications used to execute the algorithms and calculate the forecasts are as follows: CPU 8-core Intel Core i9 2.3 GHz, 16 GB of RAM, and macOS 12.5 operating system.

Once the predictions are calculated, the error arrays are next calculated and saved as RDS files, allowing analysis of the results. Fitting time is computed from the time delta of the system, before and after each execution of the methods.

A.3 Results and discussion

Table A.3 presents the average sMAPE achieved by each of the statistical methods and by the proposed method, computed for each of the time domains. The Theta method achieved the best scores for the yearly and monthly frequencies (14.603 and 13.003), which composed more than 70% of the total of the series, thus contributing to this particular method outperforming the other methods in the overall average (12.312). In the individual domains, Comb achieved the lowest error for both the daily (10.197) and the quarterly (10.197) domains, while the ARIMA method scored the lowest error on the weekly frequency (8.593).

The average error of all methods is the lowest for daily frequency (close to 3.00), and there seems to exist a trend toward increasing as the time domain becomes broader: the weekly average error is around 9, the monthly is around 13, and so on. The exception is for the hourly frequency, in which most of the statistical methods scored errors from 13.912 to 43.003.

DTSF exhibited errors considerably fewer errors methods considered for benchmarking in this particular kind of time series (12.927). This makes the DTSF method interesting for studying applications in which competitive estimators are sought.

Table A.4 presents the evaluation of the methods using OWA. This metric is understood as showing how one method is more accurate when compared to Naive2. If OWA is lower than 1 the method is more adequate than Naive2. Otherwise, Naive2 provides better forecasting performance. The DTSF scores for the hourly series imply a meaningful increase in accuracy over the Naive method (0.552). Moreover, when applying fine-tuning, the gain increases to nearly 50%. For all other domains, the only ones in which the method performed worse than Naive2 were the yearly and the daily, both of which have in common the longer term forecast period and the lowest seasonality traits in common.

The outcome of the experiment can be explained by the intrinsic design of the DTSF method, which was originally conceived to deal with very long time series with recurrent patterns, such as its original application to 30-min frequency wind speed forecasting. Comparing results to Table A.1, which presents the seasonality, length, and forecast horizon of each time domain, it is shown that the DTSF accuracy is greater when the number of available data points is also greater.

Fig. A.3 displays the average sMAPE for each one of the 414 hourly time series available in the competition database, listed in ascending order according to the calculated error of the DTSF method. The methods Naive, sNaive and SES methods were holdouts of the graphical representation. The y-axis is presented using the base-10 logarithmic scale in order to facilitate visual analysis.

In the first 170 time series with the lowest sMAPE – one-third of the total available – the

Tabela A.3 – The performance of DTSF compared to M4 benchmark statistical methods – sMAPE metric.

	sMAPE						
Method	Yearly (23k)	Quarterly (24k)	Monthly (48k)	Weekly (359)	Daily (4,227)	Hourly (414)	Average (100k)
Naive	16.342	11.610	15.255	9.161	3.405	43.003	14.207
sNaive	16.342	12.521	15.994	9.161	3.405	13.912	14.660
Naive2	16.342	11.012	14.429	9.161	3.405	18.383	13.565
SES	16.398	10.600	13.620	9.012	3.405	18.094	13.089
Holt	16.535	10.955	14.833	9.706	3.070	29.474	13.839
Damped	15.162	10.243	13.475	8.867	3.063	19.277	12.655
Theta	**14.603**	10.312	**13.003**	9.094	3.053	18.138	**12.312**
Comb	14.874	**10.197**	13.436	8.947	**2.985**	22.114	12.567
ARIMA	15.150	10.408	13.486	**8.593**	3.185	14.081	12.679
ETS	15.356	10.291	13.525	8.727	3.046	17.307	12.725
DTSF	16.816	11.006	13.823	8.983	3.313	**12.927**	13.370

Tabela A.4 – The performance of DTSF compared to M4 benchmark statistical methods – OWA metric.

	OWA						
Method	Yearly (23k)	Quarterly (24k)	Monthly (48k)	Weekly (359)	Daily (4,227)	Hourly (414)	Average (100k)
Naive	1.000	1.066	1.095	1.000	1.000	3.593	1.072
sNaive	1.000	1.153	1.147	1.000	1.000	0.628	1.106
Naive2	1.000	1.000	1.000	1.000	1.000	1.000	1.000
SES	1.003	0.970	0.951	0.975	1.000	0.990	0.970
Holt	0.956	0.935	0.989	0.964	0.997	2.760	0.976
Damped	0.888	0.893	0.924	0.916	0.996	1.140	0.912
Theta	0.872	0.917	0.907	0.971	0.999	1.006	**0.906**
Comb	**0.868**	0.891	0.920	0.926	**0.979**	1.559	**0.906**
ARIMA	0.891	0.898	**0.904**	0.927	1.041	0.950	**0.906**
ETS	0.903	**0.890**	0.914	0.931	0.996	1.824	0.913
DTSF	1.002	0.961	0.950	**0.914**	1.092	**0.552**	0.969

method proposed in the present article achieved errors close to 10^{-2}, while most of the others obtained errors between $10^{0.5}$ and 10^{2}. This shows the enormous predictive power in this specific type of series, and the great gain in accuracy that explains the best performance of this method, on average. Analyzing the sets between the 170th and 300th time series with the smallest error, there is less distinction between all the methods which, in general, presented errors very close to each other. Other methods have shown a lower errors than DTSF along all time series, specially the methods ARIMA and ETS. In the set between 300th and 414th, DTSF again marginally outperformed the other benchmark methods in most of the series.

Table A.5 presents the average sMAPE detailed by the forecast horizon, grouped by 6-hour periods. DTSF obtained lower errors, for all horizons than the other compared methods. Furthermore, the average error is 12.9%, and the highest errors were obtained during the periods between the hours from 19 to 30 and the hours from 43 to 48.

To provide better visualization of error evolution over time, Fig. A.4 presents the mean errors per step of each method (excluding the three from the previous figure), for all hourly time series. An increase in error over time, according to the phenomenon of error propagation, is

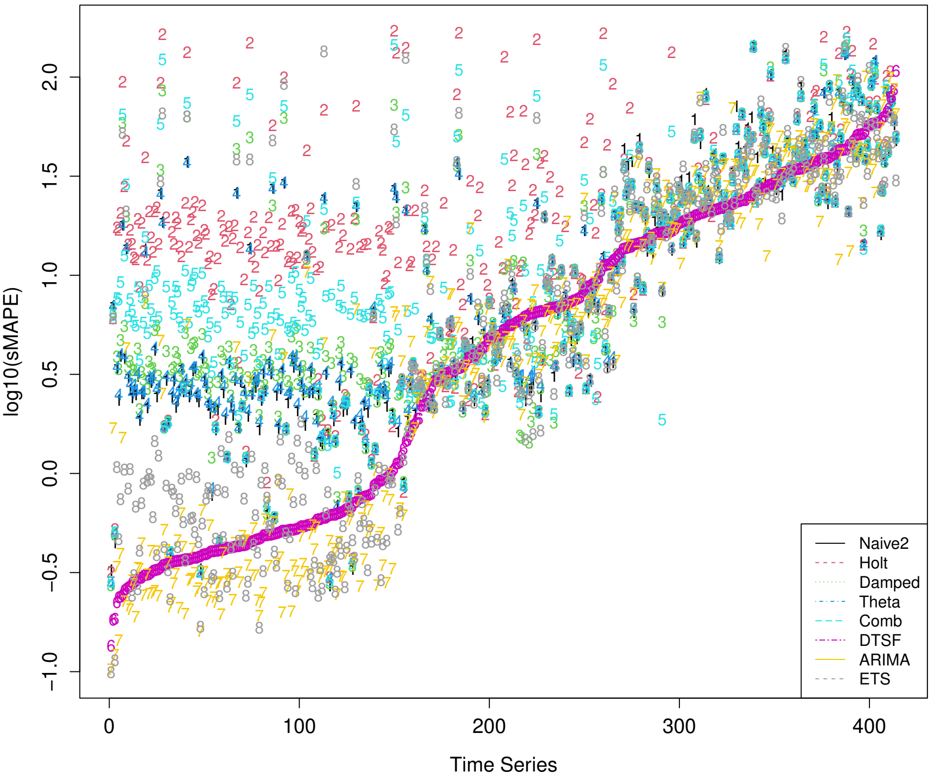


Figura A.3 – Forecasting methods average sMAPE for each of the 414 hourly time series, ordered by the accuracy of the DTSF method. The proposed method obtained fewer errors for most of the time series in this particular domain of application.

Tabela A.5 – Average sMAPE obtained in the 414 hourly time series by the predicted steps, grouped in 6-hour periods.

	Steps								
Methods	1-6	7-12	13-18	19-24	25-30	33-36	37-42	43-48	1-48
Naive2	16.3	20.1	18.8	15.7	18.2	20.7	19.3	18.0	18.4
Naive2	16.3	20.1	18.8	15.7	18.2	20.7	19.3	18.0	18.1
Holt	15.7	23.0	27.1	27.5	29.9	34.9	37.9	39.8	29.5
Damped	15.5	20.3	20.5	17.5	18.1	21.2	21.2	19.9	19.3
Theta	16.1	19.9	18.5	15.3	17.8	20.5	19.2	17.8	18.1
Comb	15.6	20.6	21.8	19.7	20.8	24.9	26.7	26.8	22.1
ARIMA	14.2	11.4	11.2	15.8	15.4	13.9	13.4	17.0	14.1
ETS	13.6	16.5	16.4	16.6	16.5	19.0	18.9	17.4	17.3
DTSF	**12.6**	**10.7**	**10.2**	**15.0**	**14.8**	**11.6**	**11.6**	**11.6**	**12.9**

expected. This is better observed in the Holt method, in which error varied from 10% at the first step to 40% at the last step. Moreover, in such a visual representation, the Theta model is perceived to have been more accurate, on average, than the DTSF model for the 1st and 24th hours.

Most statistical methods presented a pattern of very similar curves, with the exception of the DTSF method. In DTSF, the errors presented a different pattern, alternating peaks, and valleys with the patterns of the other statistical methods. In general, DTSF appeared to remain

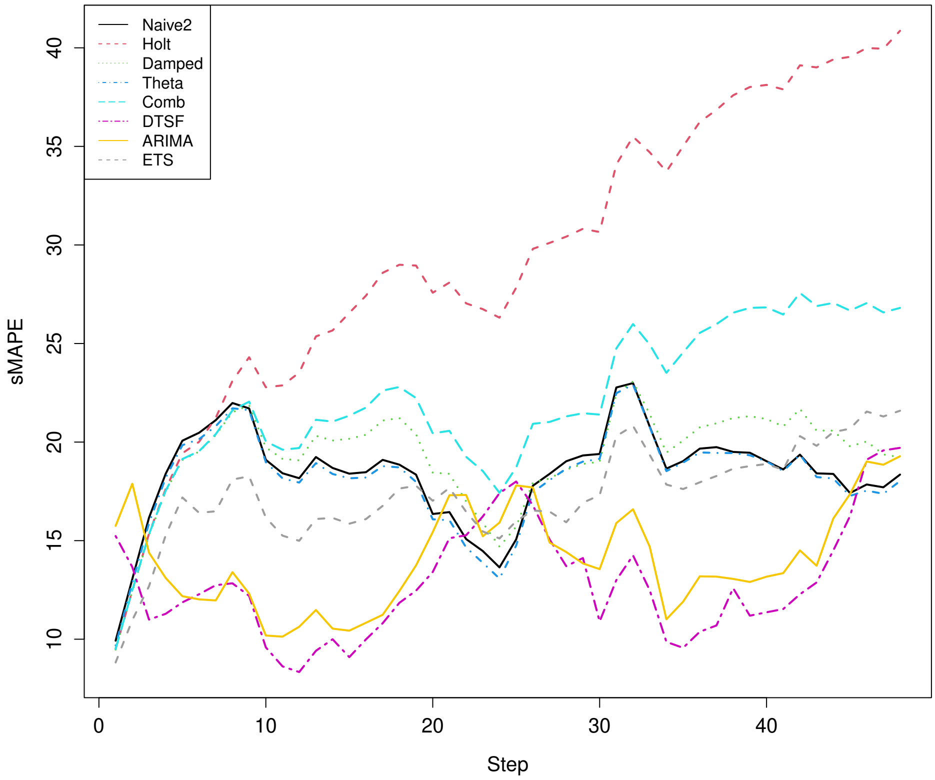


Figura A.4 – Average sMAPE (obtained in the 414 hourly time series by all the methods for each step of the prediction, up to 48 hours – forecast horizon).

more stable throughout the period, experiencing less of the error propagation effect and not exceeding the limit of 20%. These are more examples that explain the better performance of the DTSF method, compared to the benchmark, in the hourly domain.

Table A.6 shows the time necessary to fit the methods for all of the 100,000 time series. The methods Naive2 and Comb have been omitted as these two are a combination/selection of individual methods. Total fitting time is given in seconds, while the average time per series is given in microseconds. The Ratio Naive column compares the average time of a particular method compared to the execution time of the Naive method.

Tabela A.6 – Total and average times necessary for fitting the methods.

Methods	**Total fitting time (s)**	**Average time per series (ms)**	**Ratio to naive**
Naive	0.458	1.106	1.00
sNaive	0.656	1.584	1.43
SES	2.219	5.360	4.85
Holt	5.947	14.365	12.99
Damped	12.789	30.892	27.94
Theta	2.964	7.159	6.47
ARIMA	18437.598	44535.261	40278.22
ETS	1838.638	4441.155	4016.63
DTSF	6.241	15.074	13.63

DTSF was the method that consumed the most computational time, almost 9 times more than Naive. It is worth mentioning that the default parameters for DTSF adopt 10 analogs to estimate the forecast. Also, part of the method is executed in the C compiled language, and part of it is executed in R.

A.4 Conclusions

The current paper presents the results of applying the dynamic time scan forecasting method with the M4 competition data and compares it with statistical methods used as baselines in the same competition. The results point to a significant gain in accuracy in hourly time domain problems, compared to the reference, which justifies adopting this method for problems of this particular nature.

Since the method was developed for problems with long time series and high repeatability, DTSF has been proved competitive. In the present experiment, the DTSF method reduced the sMAPE by 12.13%.

Furthermore, the dissemination of this method may be interesting for other researchers who wish to extend it to existing methods, either by combining it with other techniques or by adapting its operation to other applications.

Future research should extend the method to multivariate forecasting problems and hierarchical time series and should assess its performance in other applications with this characteristic (the M5 competition, for instance). Also, some extensions of the method itself are foreseen, in order to improve its accuracy on time series for which its performance was less satisfactory than the performance of other statistical methods, for example, adopting k-fold instead of hold-out cross-validation for model selection (BERGMEIR et al., 2018).

B SIMILARITY SEARCH IN ELECTRICITY PRICES: AN ULTRA-FAST METHOD FOR FINDING ANALOGS

Abstract: Accurately predicting electricity prices allows us to minimize risks and establish more reliable decision support mechanisms. In particular, the theory of analogs has gained increasing prominence in this area. The analog approach is constructed from the similarity measurement, using fast search methods in time series. The present paper introduces a rapid method for finding analogs. Specifically, we intend to: (i) simplify the leading algorithms for similarity searching, and (ii) present a case study with data from electricity prices in the Nordic market. To do so, Pearson's distance correlation coefficient was rewritten in simplified notation. This new metric was implemented in the main similarity search algorithms, namely: Brute Force, JustInTime, and Mass. Next, the results were compared to the Euclidean distance approach. Pearson's correlation, as an instrument for detecting similarity patterns in time series, has shown promising results. The present study provides innovation in that Pearson's distance correlation notation could reduce the computational time of similarity profiles by an average of 17.5%. It is noteworthy that computational time was reduced in both short and long time series. For future research, we suggest testing the impact of other distance measurements, e.g., Cosine correlation distance and Manhattan distances.

Keywords: Analog. Ensemble forecasting. Similarity search. Electricity prices.

B.1 Introduction

The construction of predictive models is gaining prominence in the literature (GEISSER, 2017), since economic agents deal with uncertainty and aim to achieve the best results using available resources (CHOI, 1993). Therefore, developing models with acceptable accuracy presents a meaningful challenge to researchers. George Box stated, "All models are wrong, but some are useful"(BOX, 1976). In other words, prediction is a technique that deals with risk, and there will always be a fundamental error associated with it. The best model is the one that most adequately represents the phenomenon of interest.

In relation to the object of our study, electricity prices, there are several forecasting applications: (i) classical time series models like the autoregressive moving average, autoregressive integrated moving average, generalized autoregressive conditional heteroscedastic, among others (LIU; SHI, 2013); (ii) pre-processing techniques like spectrum analysis, wavelets and Fourier analysis (MIRANIAN et al., 2013); and, (iii) machine learning approaches such as neural networks, fuzzy systems and support vector machine (BUI et al., 2016). Additionally, an alternative class known as hybrid models aims to combine machine learning representations with different methods. Instances of these methods are focused time-delay neural networks (CHEN et al., 2019), neural networks with fuzzy inputs (LIU et al., 2015), finite-impulse response neural networks (PIR et al., 2017), local feedback dynamic fuzzy neural networks (NAGARAJA et al., 2016), type recurrent fuzzy networks (JAIN et al., 2014), neuro-fuzzy inference systems (MORENO; COELHO, 2018), among others.

The energy market is known for being an industry with high-frequency data (MADADI et al., 2018), for several reasons. First, sensor usage is widespread in energy (JARADAT et al., 2015). Second, high-frequency data can better represent specific weather conditions, enabling the improvement of energy modeling (AIGNER et al., 2007). Examples are diverse, such as: (i) solar radiation, which can be collected in minutes (ASSUNCAO et al., 2003); and, (ii) air humidity, atmospheric pressure, temperature and wind speed, which can also be measured in minutes (LONGMAN et al., 2018).

In particular, the pricing of electricity also has significant volumes of information, in most cases, arranged on an hourly scale (VORONIN; PARTANEN, 2014). Although the literature on this question is extensive, there is academic interest in the construction of nonparametric models applied to electricity prices, as they have presented promising predictive results. In general, these models are designed to deal with long-time series and are chiefly based on analog ensemble (AnEn) searches (YANG; ALESSANDRINI, 2019b; YANG et al., 2018a) and scan-clustering methodologies (Azevedo Costa et al., 2019).

Due to both the complexity and the high volume of information, finding patterns in time series is a data science challenge. Given that, similarity analysis has been studied since the 1960s (Lorenz, 1969. In addition to the complexity of creating highly accurate models, significant volumes of information lead to developing algorithms with low computational time. As a result, the literature reflects efforts in mathematical and computational solutions to this problem (MUEEN et al., 2017; YANG et al., 2018b).

In general, similarity and analog studies are based on searches of similarity patterns between the latest available observations and the old observations through a scanning process on data (GENSLER et al., 2016). This methodology is widely used in climatology studies, where an AnEn is developed by first matching up the actual prediction from a numerical weather prediction (NWP) model with similar past projections (ECKEL; MONACHE, 2016).

As an example, some research in this area deserves special mention. Yang et al. (YANG et al., 2018a) presented a dual NWP model approach, boy jointing the AnEn and the bias-corrected analog ensemble (BCAnEn) procedure and demonstrated that by combining different NWP models, it is possible to improve the storm wind speed prediction. Another critical study was carried out by Yang (YANG, 2019), which pointed out that using the kd-tree in AnEn, it could be possible to save computational time when necessary to test different model adjustments. Still, in this context, research on the forecast of solar irradiation is frequent, and (YANG et al., 2018b) presented a substantive review of this area's main procedures.

Although relevant, previous work on the similarity search is mainly aimed at climatological research. This article innovates, as it addresses this methodology in the energy commercialization sector. Also, it is highlighted that previous analog forecasting studies are based on Euclidean distance as a metric of similarity (MUEEN et al., 2017; YANG, 2019). McDermott Wikle (MCDERMOTT; WIKLE, 2016) show that this procedure may present trouble. Since searches of analogs rely on embedding vectors being spatially similar over time, it is not certain that Euclidean distance ever leads to first-rate analogs, particularly for the spatiotemporal state processes. Pearson distance has mathematical similarities to the Euclidean approach (IMMINK; WEBER, 2015), and could be a simplified way of rewriting its notation.

A research gap still needs to be addressed: finding alternative measures for the similarity pattern to reduce the computation time of analog searches. The research question is formulated: how it is possible to rewrite the classical analog ensemble models, based on the Euclidian distance profile, into simplified Pearson distance notation to obtain computational gains in the main analog algorithms? Therefore, our goal is to simplify the notation of the analog procedure to achieve the same distance profile with less computational time. The present paper contributes to the debate about electricity since it introduces a new predictive instrument based on the analog procedure, using the Nord Pool prices of electricity as a case study.

This paper is structured as follows: Section 1 outlines the objectives of this paper. Section 2 presents the materials and methods employed in preparing this paper. Section 3 presents the results obtained. Finally, section 4 discusses the implications of this research as well as possibilities for future research.

B.2 Materials and Methods

The algorithms used in this paper are: (i) Brute Force, (ii) JustInTime, and (iii) Mass. Usually, the Euclidean formula is presented in the literature on analogs to calculate the distance between length-m query (X_i) and each length-m subsequence (Y_i) in a given time series (RADACK; BADLER, 1989; YANG; ALESSANDRINI, 2019b; ZHU et al., 2019). Generally, this approach calculates Euclidean distance d=(Y,X), based on the normalized values of Y_i^* and X_i^*, as $d = \sqrt{(Y_i^* - X_i^*)^2}$. If we perform z-score normalization on each object, the Euclidean Distance behaves similarly to the Pearson correlation coefficient (HöPPNER; KLAWONN, 2009). Finally, the use of the Pearson correlation can produce simplified mathematical expressions, as shown in Equation (1).

B.2.1 Similarity profile computation based on the Pearson correlation distance

The Pearson coefficient, ρ, measures the degree of correlation and the direction of this correlation, positive or negative, between two random variables. The Pearson correlation coefficient is defined as follows (PEARSON, 1895):

$$\rho_{xy} = \frac{\sum_{i=1}^{m}(X_i - \mu_X).(Y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{m}(X_i - \mu_X)^2 . \sum_{i=1}^{m}(Y_i - \mu_Y)^2}} \tag{B.1}$$

Equation (1) represents a single-pass algorithm for calculating the Pearson correlation. However, depending on the amount of data, it can demand considerable computational time. Using a little algebra, we can rearrange Equation (1) as follows, obtaining the Pearson product-moment correlation coefficient (KELLEY, 1925):

$$\rho_{xy} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{X_i - \mu_X}{\sigma_X}\right)\left(\frac{Y_i - \mu_Y}{\sigma_Y}\right) \tag{B.2}$$

where $\sigma_X = \sqrt{\sum_{i=1}^{m}(X_i - \mu_X)^2/m}$ and $\sigma_Y = \sqrt{\sum_{i=1}^{m}(Y_i - \mu_Y)^2/m}$.

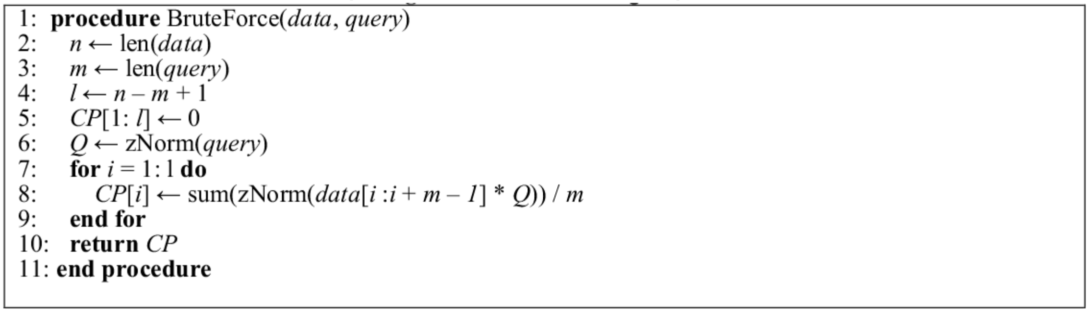
Note that Equation (2) presents a simplified way for calculating the correlation between two sets, which reduces the computational time. It is noteworthy that this approach will be tested in the calculation algorithm called "Brute Force". Finally, Equation (2) can be written in the abbreviated notation, illustrated below:

$$\rho_{x^*y^*} = \frac{1}{m}\sum_{i=1}^{m}(X_i^*.Y_i^*) \tag{B.3}$$

where $X_i^* = (X_i - \mu_X)/\sigma_X$ and $Y_i^* = (Y_i - \mu_Y)/\sigma_Y$.

The Brute Force algorithm was based on the Pearson formulation product-moment correlation coefficient. Figure 1 details the steps of this procedure, which consists of calculating the normalized values of the last observations (query) and the rest of the data.

Figura B.1 – Brute Force algorithm



```
procedure BruteForce(data, query)
    $n \leftarrow$ len(data)
    $m \leftarrow$ len(query)
    $l \leftarrow n - m + 1$
    $CP[1: l] \leftarrow 0$
    $Q \leftarrow$ zNorm(query)
    for $i = 1: l$ do
        $CP[i] \leftarrow$ sum(zNorm(data[$i : i + m - 1$] * $Q$)) / $m$
    end for
    return CP
end procedure
```

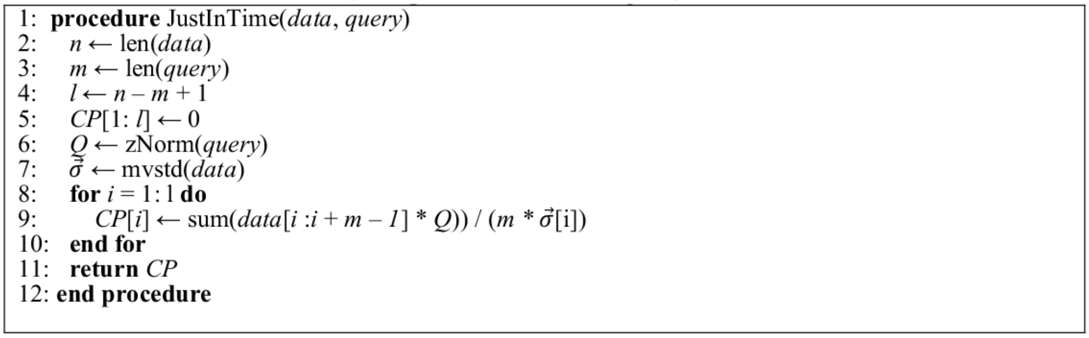
Source: adapted by authors from: Yang & Alessandrini (2019).

An essential principle of a given similarity measure should be the invariance, under some specific conditions, e.g., data manipulation without changing the scale (STREHL et al., 2000). Thus, we highlight that the Pearson correlation coefficient is invariant to scaling (ORANG; SHIRI, 2012). This means that, when multiplying all elements by a non-zero constant, the correlation remains the same. The same is valid when adding any constant to all the elements. This is a fundamental property, since the main goal of correlation is not to verify if two vectors are similar in absolute terms, but if they vary in the same direction:

$$\rho_{XY} = \rho_{X^*Y^*} = \rho_{(}X^*Y) =_{(} XY^*) \tag{B.4}$$

According to (YANG; ALESSANDRINI, 2019b), a valid research strategy is to measure the degree of association between one normalized variable and one without normalization. This procedure assists in simplifying notations and will be utilized in the "JustInTime"algorithm. However, the main difference from the Brute Force procedure is that JustInTime only normalizes the latest information (query), leaving the rest of the series without any transformation (Figure 2).

Figura B.2 – JustInTime algorithm



```
procedure JustInTime(data, query)
    $n \leftarrow$ len(data)
    $m \leftarrow$ len(query)
    $l \leftarrow n - m + 1$
    $CP[1:l] \leftarrow 0$
    $Q \leftarrow$ zNorm(query)
    $\vec{\sigma} \leftarrow$ mvstd(data)
    for $i = 1:1$ do
        $CP[i] \leftarrow$ sum(data[$i : i + m - 1$] * $Q$)) / ($m$ * $\vec{\sigma}$[i])
    end for
    return CP
end procedure
```

Source: adapted by authors from: Yang & Alessandrini (2019).

Considering the correlation between X_i^* and Y_i, where $X_i^* \approx N(0{,}1)$, Equation (2) can be rewritten as:

$$\rho_{X^*Y} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{X_i^*.Y_i}{\sigma_Y}\right) - \frac{1}{m}\sum_{i=1}^{m}\left(\frac{X_i^*.\mu_Y}{\sigma_Y}\right) \tag{B.5}$$

since $\frac{1}{m}\sum_{i=1}^{m}\left(\frac{X_i^*.\mu_Y}{\sigma_Y}\right) = \frac{m.\bar{Y}}{m.\sigma_Y}\sum_{i=1}^{m}\left(\frac{X_i^*}{m}\right) = \frac{m^2.\mu_Y.\mu_{X_i^*}}{m.\sigma_Y} = 0$, then:

$$\rho_{X^*Y} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{X_i^*.Y_i}{\sigma_Y}\right) \tag{B.6}$$

Additionally, adopting the correlation coefficient as a distance metric has other advantages. For example, it is possible to develop a regression model relating query X_i to the last observations Y_i. Assuming that the joint distribution of X_i and Y_i is the bivariate normal distribution, that μ_Y and σ_Y^2 are the mean and variance of Y, that μ_X and σ_X^2 are the mean and variance of X, and that ρ is the correlation coefficient between Y and X (MONTGOMERY et al., 2012). The conditional distribution of Y for a given value of $X = x$ is:

$$f_{Y|x}(Y) = \frac{1}{\sqrt{2\pi\sigma_{Y|x}}}\exp\left[\frac{-1}{2}\left(\frac{y - (\beta_0 + \beta_1 x)}{\sigma_{Y|x}}\right)^2\right] \tag{B.7}$$

where

$$\beta_0 = \mu_Y - \mu_X\rho\frac{\sigma_Y}{\sigma_X} \tag{B.8}$$

and

$$\beta_1 = \frac{\sigma_Y}{\sigma_X}\rho \tag{B.9}$$

and the variance of the conditional distribution of Y given $X = x$ is

$$\sigma^2_{Y|x} = \sigma^2_Y(1 - \rho^2) \tag{B.10}$$

For additional details on computational procedures, see the Attachment section.

B.2.2 Similarity profile computation based on Euclidean distance

Equation (11) presents the mathematical formulation of the Euclidean distance between the elements of two vectors. Note that the formula below illustrates the case where the two vectors have previously been normalized. This is the formulation used in the Brute Force method.

$$d(X,Y) = \sqrt{\sum_{i=1}^{m}\left(\frac{X_i - \mu_X}{\sigma_X} - \frac{Y_i - \mu_Y}{\sigma_Y}\right)^2} = \sqrt{\sum_{i=1}^{m}(X_i^* - Y_i^*)^2} \tag{B.11}$$

where $X_i^* = (X_i - \mu_X/\sigma_X)$ and $Y_i^* = (Y_i - \mu_Y/\sigma_Y)$.

The JustInTime method can be considered a rewrite of Equation (11), above. However, it takes one normalized variable and one without normalization. Equation (12) uses some algebra steps to demonstrate how to determine the adjusted equation for Euclidean distance.

$$= \sqrt{\textstyle\sum_{i=1}^{m=1} \frac{(X_i - \mu_X)^2}{\sigma_X^2} + \sum_{i=1}^{m=1} \frac{(Y_i - \mu_Y)^2}{\sigma_Y^2} - 2\sum_{i=1}^{m=1} \frac{\sum_{i=1}^{m} X_iY_i - m.\mu_X.\mu_Y}{\sigma_X.\sigma_Y}}$$

$$= \sqrt{2\left(m - \frac{\sum_{i=1}^{m} X_iY_i - m.\mu_X.\mu_Y}{\sigma_X.\sigma_Y}\right)} \text{ (B.12)}$$

Assuming the normalization of variable X (query), we can simplify Equation (12):

$$d(X,Y) = \sqrt{2\left(m - \frac{\sum_{i=1}^{m} X_iY_i}{\sigma_Y}\right)} \tag{B.13}$$

Equation (14) uses some algebra to demonstrate the existence of a relationship between Pearson's correlation coefficient and the Euclidean distance formula. Thus, since the correlation coefficient values vary between minus one and one, the smaller the distance between the vectors, the greater the force (correlation) between them:

$$\rho_{X^*Y} = 1 - \frac{d(X,Y)^2}{2m} \tag{B.14}$$

The Euclidean distance incorporates the Pearson correlation function. Thus, the present paper will search for similarity patterns considering the Pearson coefficient, since this approach will return similar results but with lower computational costs. To illustrate this advantage, these results will be compared with those obtained using the Euclidean distance method.

According to (YANG; ALESSANDRINI, 2019b), there are alternative ways to calculate the correlation between a pair of vectors using of convolution procedure. Suppose A is the set of six data points, $A = A_1{,}A_2{,}A_3{,}A_4{,}A_5{,}A_6$, and F represents 4 forecasts to be matched, $F = F_1{,}F_2{,}F_3{,}F_4$. The full convolution between these two vectors is given by:

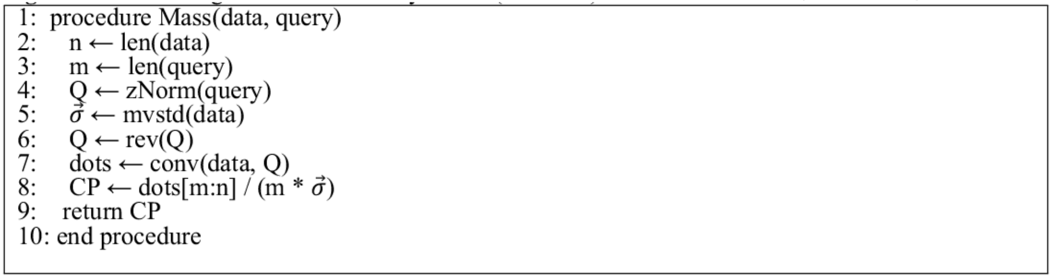
$$(A) \otimes (F) = \begin{bmatrix} A_1F_1 \\ A_1F_3 + A_2F_4 \\ A_1F_2 + A_2F_3 + A_3F4 \\ A_1F_1 + A_2F_2 + A_{3_F}3 + A_4F_4 \\ A_2F_1 + A_3F_2 + A_{4_F}3 + A_5F_4 \\ A_3F_1 + A_4F_2 + A_5F_3 + A_6F_4 \\ A_4F_1 + A_5F_2 + A_6F_3 \\ A_5F_1 + A_6F_2 \\ A_6F_1 \\ 0 \\ 0 \end{bmatrix} \tag{B.15}$$

Thus, the convolution formulation for the correlation coefficient is presented by Equation (16).

$$\rho_{X^*Y} = \frac{(X^*) \otimes (Y)}{m.\sigma_Y} \tag{B.16}$$

Finally, we present the Mass algorithm proposed by (MUEEN et al., 2017) (Figure 3). This procedure uses the concept of "Convolution", i.e., a mathematical method between two sets that produces a third one, expressing how the shape of one is modified by the other. Convolution refers to both the resulting function and the computing of it (BURRUS; PARKS, 1985).

Figura B.3 – Mueen's algorithm for similarity search (modified)



```
procedure Mass(data, query)
    n ← len(data)
    m ← len(query)
    Q ← zNorm(query)
    $\vec{\sigma}$ ← mvstd(data)
    Q ← rev(Q)
    dots ← conv(data, Q)
    CP ← dots[m:n] / (m * $\vec{\sigma}$)
  return CP
end procedure
```

Source: adapted by authors from: Mueen et al. (2017).

The next section presents the dataset used, refers to the Nordic electricity market in the short term (Nord Pool), and outlines the simulation procedures employed.

B.2.3 Dataset and simulation procedures

The data used in the present study were obtained from the Nord Pool, the leading power market in Europe (JANKE et al., 2020). The dataset includes the hourly average electricity price for seven different countries, segregated into market areas (POOL, 2020).

The period of data analysis ranges from January 1st, 2014, 00:00, to September 2nd, 2019 00:00, totaling 49,709 registers for each time series. There were six missing data points, from hour 02:00 to 03:00, at the end of March of each year. Missing data were computed using the average price of the preceding and subsequent hours. The time series utilized, including the number of time series per country and their acronyms, are presented in Table 1.

Tabela B.1 – Nord pool energy submarkets analyzed

Time series	*Acronyms*	*# Time Series*
System reference	SYS	1
Sweden	SE1, SE2, SE3, SE4	4
Finland	FI	1
Denmark	DK1, DK2	2
Norway	Oslo, Krsand, Bergen, Molde, Trhein, Tromso	6
Estonia	EE	1
Latvia	LV	1
Lithuania	LT	1

From each of the time series, 30 samples of size n equal to 720, 2,400, 7,200, 12,000 and 24,000 are randomly drawn. These values are associated with time series lengths of 30, 100, 300, 500 and 1000 days. The values of m adopted for this simulation were, 6, 9, 24 and 48 hours.

Figure 4 shows the flowchart with the detailed simulation process used to compare the similarity search algorithms based on Pearson's correlation and those based on normalized distance. The validation of the analysis is obtained by comparing the computational times calculated for the different methods in carrying out the same task, building the similarity profile. As the similarity profile is deterministic, the accuracy is the same as long as the model reaches its objective.

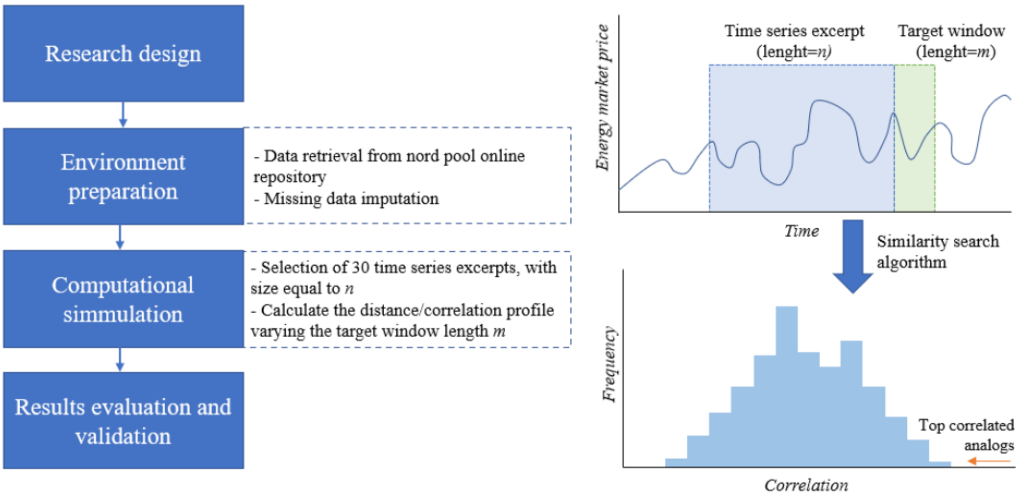
Routines were implemented using the R® 3.6.0 programming language, adapting algorithms from (MUEEN et al., 2017) and (YANG; ALESSANDRINI, 2019b). The R-package RollingWindow was used to calculate the standard deviation of the data, considering fixed-width subsets of observations, called windows. This package is available from the GitHub repository at: https://github.com/andrewuhl/RollingWindow.

The computer used to execute the algorithms and to calculate the correlation and distance profiles had: CPU Intel Core i5-4570 3.20GHz, 16 GB of RAM and operating system Windows 10 x64. Computational time was calculated from the system's time delta before and after each execution of the methods.

B.3 Results and discussion

Each of the search algorithms (Brute Force, JustInTime, and Mass) were properly calculated using both Pearson correlation metrics and Euclidian distance formulation. To make

Figura B.4 – Flowfchart with the detailed process of simulation and analysis of similarity search algorithms


Research design
Environment preparation
- Data retrieval from nord pool online repository
- Missing data imputation
Computational simmulation
- Selection of 30 time series excerpts, with size equal to *n*
- Calculate the distance/correlation profile varying the target window length *m*
Results evaluation and validation
Energy market price
Time series excerpt (lenght=*n*)
Target window (lenght=*m*)
Time
Similarity search algorithm
Frequency
Top correlated analogs
Correlation

the simulation more robust, results were obtained considering different samples, namely 2400 and 24000 observations. The search criteria used variable windows for data regarding the last observations (query) of sizes equal to 6, 9, 24 and 48 hours.

Table 2 presents the results of the 100 simulations performed, considering the sample formed from the last 2400 observations. The best methods, in terms of computational time, are highlighted in bold. The standard deviation of the simulations is shown in parentheses. Note that the Pearson's distance-based similarity search methods, especially JustInTime and Mass, respectively, had shorter computational times.

Tabela B.2 – Sample data length of 2,400 (100 days) [hours]. The query length m varies from 3 to 48 [hours]. Each scenario is repeated 100 times, the mean computational times (in ms) are shown in the table.

	Avg. time (ms)					
	Correlation similarity profile			Euclidean distance profile		
m	BruteForce	JustInTime	MASS	BruteForce	JustInTime	MASS
m = 6	37.488 (7.582)	1.658 (4.232)	0.423 (2.138)	38.348 (7.587)	1.920 (4.487)	0.450 (2.230)
m = 9	38.401 (7.526)	1.704 (4.228)	0.445 (2.198)	38.218 (7.266)	2.022 (4.442)	0.533 (2.470)
m = 24	39.293 (7.647)	1.780 (4.138)	6.642 (6.740)	39.437 (7.635)	2.198 (4.603)	6.726 (6.704)
m = 48	39.869 (7.548)	2.136 (4.684)	6.765 (6.781)	40.220 (7.695)	2.387 (4.802)	7.141 (6.861)

There is a notable difference between the performance of the BruteForce algorithm and the others. Computational time is about 20 times greater than the JustInTime model, and 10 to 100 times greater than the MASS model. The computational time of MASS especially draws attention, as it reaches optimal values when m is equal to 6 and 9. Then, it presents computational times up to 4 times less than the JustInTime model.

However, when m is greater than 24, the computational time tends to be 3 times greater than the time calculated using JustInTime. This greater time variation found for the MASS model is intrinsic to the nature of the MASS, which computational time is mainly affected by the convolution procedure. In many cases, this is an advantage over the JustInTime algorithm.

In other cases, however, the former is more efficient. It is up to the user of the algorithm to assess which model best suits their specific situation and data types.

Here, a similar analysis is presented (Table 3). However, the sample universe was substantially increased (n = 24,000 hours). Again, the Pearson correlation-based models stood out concerning computational time, with the JustInTime algorithm as the most promising method for computing long time series (the most abundant sample universe).

Tabela B.3 – Sample data length of 24,000 (1,000 days) [hours]. The query length m varies from 6 to 48 [hours]. Each scenario is repeated 100 times, the mean computational times (in ms) are shown in the table.

	Avg. time (ms)					
	Correlation similarity profile			Euclidean distance profile		
m	BruteForce	JustInTime	MASS	BruteForce	JustInTime	MASS
m =6	377.402 (12.924)	15.678 (4.842)	136.023 (8.606)	379.560 (14.351)	19.249 (6.436)	135.817 (8.815)
m =9	381.169 (15.298)	16.321 (5.009)	84.343 (8.274)	384.333 (19.447)	20.051 (6.644)	84.518 (7.871)
m =24	394.380 (14.623)	18.377 (5.699)	1087.231 (29.833)	396.693 (14.785)	22.121 (7.312)	1086.358 (28.683)
m =48	406.765 (14.676)	20.495 (6.714)	10.859 (6.644)	408.716 (14.300)	24.404 (7.587)	11.140 (6.748)

By increasing the sample size of the available period by ten times, we obtained proportional increases in computational times for almost all models. The standard deviation increase, however, was limited to twice its original value. Thus, the computational advantage of models based on the correlation similarity profile becomes more evident.

The JustInTime algorithm, using the similarity profile, showed a 17.5% reduction in computational time. With the BruteForce and MASS algorithms, the gains were more discrete due to the greater variability of computational times. Again, the computational time of the MASS algorithm showed high sensitivity to the parameter m, assuming values 0.5 to 60 times the average time value of the JustInTime algorithm.

Finally, conclusions of the present paper are presented, emphasizing the time saving of the proposed formulation as well as suggesting potential studies to be developed in the future.

B.4 Conclusions

The electricity energy market is known for having high-frequency data. The examples are numerous, as the large-scale use of sensors across a wide range of processes provides a robust set of data. Thus, as the amount of information stored continuously increases over time, the search for statistical solutions that model this data is remarkable. Regarding predictive models, the range of approaches is broad. In particular, the literature has highlighted the relevance of predictive methods based on similarity or analogous searches. These methods scan a time series and, from the most recent observations, define moments where there is a high degree of affinity.

The main work on the methodology of analogs ensamble (AnEn) has made use of the Euclidean distance function. Our methodology revealed a high degree of similarity between the Euclidean formulation and Pearson's method. Thus, the present study is innovative in that, by rewriting Pearson's correlation equation, it was able to obtain the same results as the traditional approach but using less computational time. Therefore, the results of the present study are expected to provide a fast and robust tool for finding patterns in long time series, contributing to different actors in the energy planning sector.

The present study contributes to the energy planning processes of different agents, given that understanding price patterns has singular importance for minimizing risks and supporting reliable production planning. Good forecasts for future energy pricing can support operational arrangements, e.g., when the energy price is high, it may be more valuable for an industry to delay part of its production temporarily, trade the surplus electricity, and carry out preventive maintenance on machines and accessories.

There are no disadvantages in applying Pearson's correlation in the search for analogs, as the correlation profile is a mathematical simplification of the normalized distance: the temporal analog with the shortest normalized distance is also the one with the most significant correlation with the search. The proposition is valid for the other windows: the analog with the second shortest distance has the second-largest correlation, and so on. The same is not observed; however, for the search algorithms: the JustInTime algorithm presented the lowest computational times in most excerpts of the series; however, the Mass algorithm obtained the best efficiency in others.

Future research should test the effect of different probability distributions on the data standardization process. A study of other measurement functions, such as distance from Manhattan, is recommended. Finally, yet no less importantly, we suggest the analysis of the impact of using different coefficient approaches such as entropy, Kendall, and Spearman.

REFERÊNCIAS

AIGNER, W.; MIKSCH, S.; MüLLER, W.; SCHUMANN, H.; TOMINSKI, C. Visual methods for analyzing time-oriented data. *IEEE transactions on visualization and computer graphics*, v. 14, n. 1, p. 47–60, 2007.

AL-ALAWI, S. M.; ISLAM, S. M. Principles of electricity demand forecasting. i. methodologies. *Power Engineering Journal*, IET, v. 10, n. 3, p. 139–143, 1996.

ALONSO-TRISTÁN, C.; GONZÁLEZ-PEÑA, D.; DÍEZ-MEDIAVILLA, M.; RODRÍGUEZ-AMIGO, M.; GARCÍA-CALDERÓN, T. Small hydropower plants in Spain: A case study. *Renewable and Sustainable Energy Reviews*, v. 15, n. 6, p. 2729–2735, 2011. ISSN 13640321.

AN, X.; PAN, L.; YANG, L. Condition parameter degradation assessment and prediction for hydropower units using Shepard surface and ITD. *Transactions of the Institute of Measurement and Control*, v. 36, n. 8, p. 1074–1082, 2014. ISSN 14770369.

AN, X.; PAN, L.; ZHANG, F. Analysis of hydropower unit vibration signals based on variational mode decomposition. *JVC/Journal of Vibration and Control*, v. 23, n. 12, p. 1938–1953, 2017. ISSN 17412986.

AN, X.; YANG, W.; AN, X. Vibration signal analysis of a hydropower unit based on adaptive local iterative filtering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, v. 231, n. 7, p. 1339–1353, 2017. ISSN 20412983.

ASSIMAKOPOULOS, V.; NIKOLOPOULOS, K. The theta model: a decomposition approach to forecasting. *International journal of forecasting*, Elsevier, v. 16, n. 4, p. 521–530, 2000.

ASSUNCAO, H.; ESCOBEDO, J.; OLIVEIRA, A. Modelling frequency distributions of 5 minute-averaged solar radiation indexes using beta probability functions. *Theoretical and Applied Climatology*, v. 75, n. 3-4, p. 213–224, 2003.

ATTALLAH, O.; KARTHIKESALINGAM, A.; HOLT, P. J.; THOMPSON, M. M.; SAYERS, R.; BOWN, M. J.; CHOKE, E. C.; MA, X. Using multiple classifiers for predicting the risk of endovascular aortic aneurysm repair re-intervention through hybrid feature selection. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, SAGE Publications Sage UK: London, England, v. 231, n. 11, p. 1048–1063, 2017.

AYO-IMORU, R. M.; CILLIERS, A. C. Continuous machine learning for abnormality identification to aid condition-based maintenance in nuclear power plant. *Annals of Nuclear Energy*, v. 118, p. 61–70, 2018. ISSN 18732100.

AZADEH, A.; GHADERI, S.; SOHRABKHANI, S. Annual electricity consumption forecasting by neural network in high energy consuming industrial sectors. *Energy Conversion and management*, Elsevier, v. 49, n. 8, p. 2272–2278, 2008.

Azevedo Costa, M.; Brioschi Mineti, L.; Oliveira Prates, M.; Ruiz Cardenas, R. Dynamic Time Scan Forecasting. *arXiv e-prints*, p. arXiv:1906.05399, Jun 2019.

AZIZ, M. S. A.; ELSAMAHY, M.; HASSAN, M. A. M.; BENDARY, F. M. A novel study for hydro-generators loss of excitation faults detection using ANFIS. *International Journal of Modelling and Simulation*, Taylor & Francis, v. 37, n. 1, p. 36–45, 2017. ISSN 19257082. Disponível em: <http://dx.doi.org/10.1080/02286203.2016.1232956>.

BABATUNDE, K. A.; BEGUM, R. A.; SAID, F. F. Application of computable general equilibrium (cge) to climate change mitigation policy: a systematic review. *Renewable and Sustainable Energy Reviews*, Elsevier, v. 78, p. 61–71, 2017.

BABIĆ, B. M.; MILIĆ, S. D.; RAKIĆ, A. Fault detection algorithm used in a magnetic monitoring system of the hydrogenerator. *IET Electric Power Applications*, v. 11, n. 1, p. 63–71, 2017. ISSN 17518679.

BERGMEIR, C.; HYNDMAN, R. J.; KOO, B. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, Elsevier, v. 120, p. 70–83, 2018.

BERNARDINI, E.; CUBADDA, G. Macroeconomic forecasting and structural analysis through regularized reduced-rank regression. *International Journal of Forecasting*, Elsevier, v. 31, n. 3, p. 682–691, 2015.

BHAT, V. I. K.; PRAKASH, R. Life Cycle Analysis of Run-of River Small Hydro Power Plants in India. *The Open Renewable Energy Journal*, Bentham Science Publishers Ltd., v. 1, n. 1, p. 11–16, 3 2014. ISSN 18763871.

BHATT, U.; XIANG, A.; SHARMA, S.; WELLER, A.; TALY, A.; JIA, Y.; GHOSH, J.; PURI, R.; MOURA, J. M.; ECKERSLEY, P. Explainable machine learning in deployment. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* [S.l.: s.n.], 2020. p. 648–657.

BILLAH, B.; HYNDMAN, R. J.; KOEHLER, A. B. Empirical information criteria for time series forecasting model selection. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 75, n. 10, p. 831–840, 2005.

BLANCKE, O.; TAHAN, A.; KOMLJENOVIC, D.; AMYOT, N.; LÉVESQUE, M.; HUDON, C. A holistic multi-failure mode prognosis approach for complex equipment. *Reliability Engineering and System Safety*, v. 180, p. 136–151, 2018. ISSN 09518320.

BLANQUEZ, F. R.; ARANDA, M.; REBOLLO, E.; BLAZQUEZ, F.; PLATERO, C. A. New fault-resistance estimation algorithm for rotor-winding ground-fault online location in synchronous machines with static excitation. *IEEE Transactions on Industrial Electronics*, v. 62, n. 3, p. 1901–1911, 2015. ISSN 02780046.

BLANQUEZ, F. R.; PLATERO, C. A.; REBOLLO, E.; BLAZQUEZ, F. Novel rotor ground-fault detection algorithm for synchronous machines with static excitation based on third-harmonic voltage-phasor comparison. *IEEE Transactions on Industrial Electronics*, v. 63, n. 4, p. 2548–2558, 2016. ISSN 02780046.

BONTEMPI, G. Comments on m4 competition. *International Journal of Forecasting*, Elsevier, v. 36, n. 1, p. 201–202, 2020.

BORGHETTO, J.; CAVALLINI, A.; CONTIN, A.; MONTANARI, G. C.; NIGRIS, M. D.; PASINI, G.; PASSAGLIA, R. Partial discharge inference by an advanced system. Analysis of online measurements performed on hydrogenerator. *IEEE Transactions on Energy Conversion*, v. 19, n. 2, p. 333–339, 2004. ISSN 08858969.

BOUSDEKIS, A.; MAGOUTAS, B.; APOSTOLOU, D.; MENTZAS, G. Review, analysis and synthesis of prognostic-based decision support methods for condition based maintenance. *Journal of Intelligent Manufacturing*, Springer, v. 29, n. 6, p. 1303–1316, 2018.

BOX, G. Science and statistics. *Journal of the American Statistical Association*, v. 71, n. 356, p. 791–799, 1976.

BOX, G. E.; PIERCE, D. A. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American statistical Association*, Taylor & Francis, v. 65, n. 332, p. 1509–1526, 1970.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

BRITO, G. C.; MACHADO, R. D.; NETO, A. C. Using Simplified Models to Assist Fault Detection and Diagnosis in Large Hydrogenerators. *International Journal of Rotating Machinery*, v. 2017, 2017. ISSN 15423034.

BUI, D.; TUAN, T.; KLEMPE, H.; PRADHAN, B.; REVHAUG, I. Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. *Landslides*, v. 13, n. 2, p. 361–378, 2016.

BURRUS, C.; PARKS, T. *and Convolution Algorithms.* New York: John Wiley and Sons, 1985.

CACHADA, A.; BARBOSA, J.; LEITÑO, P.; GCRALDCS, C. A.; DEUSDADO, L.; COSTA, J.; TEIXEIRA, C.; TEIXEIRA, J.; MOREIRA, A. H.; MOREIRA, P. M. et al. Maintenance 4.0: Intelligent and predictive maintenance system architecture. In: IEEE. *2018 IEEE 23rd International Conference on Emerging Technologies and Factory Automation (ETFA).* [S.l.], 2018. v. 1, p. 139–146.

CAI, W.; CHEN, J.; HONG, J.; JIANG, F. Forecasting chinese stock market volatility with economic variables. *Emerging Markets Finance and Trade*, Taylor & Francis, v. 53, n. 3, p. 521–533, 2017.

CARLETTI, M.; MASIERO, C.; BEGHI, A.; SUSTO, G. A. Explainable machine learning in industry 4.0: Evaluating feature importance in anomaly detection to enable root cause analysis. *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, IEEE, v. 2019-Octob, p. 21–26, 2019. ISSN 1062922X.

CARVALHO, A. T.; LIMA, A. C.; CUNHA, C. F.; PETRAGLIA, M. Identification of partial discharges immersed in noise in large hydro-generators based on improved wavelet selection methods. *Measurement: Journal of the International Measurement Confederation*, Elsevier Ltd, v. 75, p. 122–133, 2015. ISSN 02632241. Disponível em: <http://dx.doi.org/10.1016/j.measurement.2015.07.050>.

CHATFIELD, C. *Time-series forecasting.* [S.l.]: Chapman and Hall/CRC, 2000.

CHEN, Y.; HE, Z.; SHANG, Z.; LI, C.; LI, L.; XU, M. A novel combined model based on echo state network for multi-step ahead wind speed forecasting: A case study of nrel. *Energy conversion and management*, v. 179, p. 13–29, 2019.

CHENG, J.; AI, L.; DUAN, Z.; XIONG, Y. Fault classification of hydroelectric generator unit based on improved evidence theory. *Open Fuels and Energy Science Journal*, v. 7, n. 1, p. 78–83, 2014. ISSN 1876973X.

CHENG, J.; DUAN, Z.; XIONG, Y. Cuckoo Search Algorithm with Quantum Mechanism and its Application in the Fault Diagnosis of a Hydroelectric Generating Unit. *Recent Patents on Computer Science*, v. 11, n. 1, p. 70–76, 2018. ISSN 22132759.

CHENG, J.; WANG, L.; XIONG, Y. An improved cuckoo search algorithm and its application in vibration fault diagnosis for a hydroelectric generating unit. *Engineering Optimization*, v. 50, n. 9, p. 1593–1608, 2018. ISSN 10290273. Disponível em: <https://doi.org/10.1080/0305215X.2017.1401067>.

CHENG, J.; WANG, L.; XIONG, Y. Cuckoo search algorithm with memory and the vibrant fault diagnosis for hydroelectric generating unit. *Engineering with Computers*, Springer London, v. 35, n. 2, p. 687–702, 2019. ISSN 14355663. Disponível em: <http://dx.doi.org/10.1007/s00366-018-0627-1>.

CHENG, J.; ZHU, C.; FU, W.; WANG, C.; SUN, J. An Imitation medical diagnosis method of hydro-turbine generating unit based on Bayesian network. *Transactions of the Institute of Measurement and Control*, v. 41, n. 12, p. 3406–3420, 2019. ISSN 01423312.

CHOI, Y. *Paradigms and conventions: Uncertainty, decision making, and entrepreneurship.* [S.l.]: University of Michigan Press, 1993.

CHRIST, M.; BRAUN, N.; NEUFFER, J.; KEMPA-LIEHR, A. W. Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). *Neurocomputing*, Elsevier, v. 307, p. 72–77, 2018.

COSTA, M. A.; MINETI, L. B.; PRATES, M. O.; CARDENAS, R. R. Dynamic time scan forecasting. *arXiv preprint arXiv:1906.05399*, 2019.

COSTA, M. A.; RUIZ-CÁRDENAS, R.; MINETI, L. B.; PRATES, M. O. Dynamic time scan forecasting for multi-step wind speed prediction. *Renewable Energy*, Elsevier, v. 177, p. 584–595, 2021.

COSTA, M. A.; WULLT, B.; NORRLÖF, M.; GUNNARSSON, S. Failure detection in robotic arms using statistical modeling, machine learning and hybrid gradient boosting. *Measurement*, Elsevier, v. 146, p. 425–436, 2019.

COX, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 34, n. 2, p. 187–202, 1972.

DALLAS, S. E.; SAFACAS, A. N.; KAPPATOU, J. C. Interturn stator faults analysis of a 200-MVA hydrogenerator during transient operation using FEM. *IEEE Transactions on Energy Conversion*, v. 26, n. 4, p. 1151–1160, 2011. ISSN 08858969.

DINDORF, C.; TEUFL, W.; TAETZ, B.; BLESER, G.; FRÖHLICH, M. Interpretability of input representations for gait classification in patients after total hip arthroplasty. *Sensors*, MDPI, v. 20, n. 16, p. 4385, 2020.

DING, H.; DING, K.; ZHANG, J.; WANG, Y.; GAO, L.; LI, Y.; CHEN, F.; SHAO, Z.; LAI, W. Local outlier factor-based fault detection and evaluation of photovoltaic system. *Solar Energy*, v. 164, p. 139–148, 2018. ISSN 0038092X.

DUDEK, G. Pattern-based local linear regression models for short-term load forecasting. *Electric Power Systems Research*, Elsevier, v. 130, p. 139–147, 2016.

DURSUN, B.; GOKCOL, C. The role of hydroelectric power and contribution of small hydropower plants for sustainable development in Turkey. *Renewable Energy*, Elsevier Ltd, v. 36, n. 4, p. 1227–1235, 2011. ISSN 09601481. Disponível em: <http://dx.doi.org/10.1016/j.renene.2010.10.001>.

ECK, N. J. V.; WALTMAN, L. Software survey: Vosviewer, a computer program for bibliometric mapping. *scientometrics*, Springer, v. 84, n. 2, p. 523–538, 2010.

ECKEL, F.; MONACHE, L. D. A hybrid nwp–analog ensemble. *Monthly Weather Review*, v. 144, n. 3, p. 897–911, 2016.

EGUSQUIZA, M.; EGUSQUIZA, E.; VALERO, C.; PRESAS, A.; VALENTÍN, D.; BOSSIO, M. Advanced condition monitoring of Pelton turbines. *Measurement: Journal of the International Measurement Confederation*, Elsevier, v. 119, n. October 2017, p. 46–55, 2018. ISSN 02632241. Disponível em: <https://doi.org/10.1016/j.measurement.2018.01.030>.

FERREIRA, J. H. I.; CAMACHO, J. R.; MALAGOLI, J. A.; JÚNIOR, S. C. G. Assessment of the potential of small hydropower development in Brazil. *Renewable and Sustainable Energy Reviews*, Elsevier Ltd, v. 56, p. 380–387, 2016. ISSN 18790690.

FILDES, R.; MAKRIDAKIS, S. The impact of empirical accuracy studies on time series analysis and forecasting. *International Statistical Review/Revue Internationale de Statistique*, JSTOR, p. 289–308, 1995.

FISHER, L. D.; LIN, D. Y. et al. Time-dependent covariates in the cox proportional-hazards regression model. *Annual review of public health*, v. 20, n. 1, p. 145–157, 1999.

FLORES, M. P. P.; SOLÍS, J. F.; VALDEZ, G. C.; BARBOSA, J. J. G.; ORTEGA, J. P.; VILLANUEVA, J. D. T. Hurst exponent with arima and simple exponential smoothing for measuring persistency of m3-competition series. *IEEE Latin America Transactions*, IEEE, v. 17, n. 05, p. 815–822, 2019.

FRIEDMAN, J. H. Stochastic gradient boosting. *Computational statistics & data analysis*, Elsevier, v. 38, n. 4, p. 367–378, 2002.

FU, W.; WANG, K.; ZHANG, C.; TAN, J. A hybrid approach for measuring the vibrational trend of hydroelectric unit with enhanced multi-scale chaotic series analysis and optimized least squares support vector machine. *Transactions of the Institute of Measurement and Control*, v. 41, n. 15, p. 4436–4449, 2019. ISSN 01423312.

GARDNER, E. S.; MCKENZIE, E. Why the damped trend works. *Journal of the Operational Research Society*, Springer, v. 62, n. 6, p. 1177–1180, 2011.

GE, Z.; SONG, Z. Process monitoring based on independent Component Analysis-Principal Component Analysis (ICA-PCA) and similarity factors. *Industrial and Engineering Chemistry Research*, v. 46, n. 7, p. 2054–2063, 2007. ISSN 08885885.

GEISSER, S. *Predictive inference*. [S.l.]: Routledge, 2017.

GENSLER, A.; SICK, B.; PANKRAZ, V. An analog ensemble-based similarity search technique for solar power forecasting. In: *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC*. [S.l.]: IEEE, 2016. p. 002850–002857.

GIRISH, G.; TIWARI, A. K. et al. A comparison of different univariate forecasting models forspot electricity price in india. *Economics Bulletin*, AccessEcon, v. 36, n. 2, p. 1039–1057, 2016.

GLAZ, J.; BALAKRISHNAN, N. *Scan statistics and applications*. [S.l.]: Springer Science & Business Media, 2012.

GLOCK, C. H.; HOCHREIN, S. Purchasing organization and design: A literature review. *Business Research*, v. 4, n. 2, p. 149–191, 2011. Disponível em: <https://link.springer.com/content/pdf/10.1007/BF03342754>.

GOMES, A. de S.; COSTA, M. A.; FARIA, T. G. A. de; CAMINHAS, W. M. Detection and classification of faults in power transmission lines using functional analysis and computational intelligence. *IEEE Transactions on Power Delivery*, IEEE, v. 28, n. 3, p. 1402–1413, 2013.

GONTIJO, T. S.; COSTA, M. A.; SANTIS, R. B. de. Similarity search in electricity prices: An ultra-fast method for finding analogs. *Journal of Renewable and Sustainable Energy*, AIP Publishing LLC, v. 12, n. 5, p. 056103, 2020.

GONTIJO, T. S.; COSTA, M. A.; SANTIS, R. B. de. Electricity price forecasting on electricity spot market: a case study based on the brazilian difference settlement price. In: EDP SCIENCES. *E3S Web of Conferences*. [S.l.], 2021. v. 239.

GREGG, S. W.; STEELE, J. P.; BOSSUYT, D. L. V. Feature selection for monitoring erosive cavitation on a hydroturbine. *International Journal of Prognostics and Health Management*, v. 8, n. 1, 2017. ISSN 21532648.

GRISCENKO, M.; ELMANIS-HELMANIS, R. Eccentricity of slow-speed salient-pole generator: Analysis based on air gap spectrum. *Latvian Journal of Physics and Technical Sciences*, v. 52, n. 1, p. 26–37, 2015. ISSN 08688257.

GUO, J.; LIU, Y.; XU, X.; CHEN, Q. Integrated distributed bond graph modeling and neural network for fault diagnosis system of hydro turbine governors. *Kybernetes*, v. 39, n. 6, p. 925–934, 2010. ISSN 0368492X.

GURUNG, R. B. *Random Forest for Histogram Data: An application in data-driven prognostic models for heavy-duty trucks.* Tese (Doutorado) — Department of Computer and Systems Sciences, Stockholm University, 2020.

HAMZIC, A.; AVDAGIC, Z.; BESIC, I. *Multistage cascade predictor of structural elements movement in the deformation analysis of large objects based on time series influencing factors.* [S.l.: s.n.], 2020. v. 9. ISSN 22209964. ISBN 3876399661.

HARA, Y.; FUKUYAMA, Y.; ARAI, K.; SHIMASAKI, Y.; OSADA, Y.; MURAKAMI, K.; IIZAKA, T.; MATSUI, T. Fault detection of hydroelectric generators by robust random cut forest with feature selection using hilbert-schmidt independence criterion. In: IEEE. *2021 IEEE International Conference on Smart Internet of Things (SmartIoT).* [S.l.], 2021. p. 136–143.

HARIRI, S.; KIND, M. C.; BRUNNER, R. J. Extended isolation forest. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, 2019.

HARRIS, C. R.; MILLMAN, K. J.; WALT, S. J. van der; GOMMERS, R.; VIRTANEN, P.; COURNAPEAU, D.; WIESER, E.; TAYLOR, J.; BERG, S.; SMITH, N. J.; KERN, R.; PICUS, M.; HOYER, S.; KERKWIJK, M. H. van; BRETT, M.; HALDANE, A.; RÍO, J. F. del; WIEBE, M.; PETERSON, P.; GÉRARD-MARCHANT, P.; SHEPPARD, K.; REDDY, T.; WECKESSER, W.; ABBASI, H.; GOHLKE, C.; OLIPHANT, T. E. Array programming with NumPy. *Nature*, Springer Science and Business Media LLC, v. 585, n. 7825, p. 357–362, set. 2020. Disponível em: <https://doi.org/10.1038/s41586-020-2649-2>.

HASSANI, H.; SILVA, E. S. Forecasting uk consumer price inflation using inflation forecasts. *Research in Economics*, Elsevier, v. 72, n. 3, p. 367–378, 2018.

HECKBERT, P. Fourier transforms and the fast fourier transform (fft) algorithm. *Computer Graphics*, v. 2, p. 15–463, 1995.

HEMPSTALK, K.; FRANK, E.; WITTEN, I. H. One-class classification by combining density and class probability estimation. In: SPRINGER. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases.* [S.l.], 2008. p. 505–519.

HENDERSON, D. S.; LOTHIAN, K.; PRIEST, J. PC Based monitoring and fault prediction for small hydroelectric plants. In: *IEE Conference Publication.* [S.l.]: IEE, 1998. p. 28–31. ISSN 05379989.

HILL, T.; MARQUEZ, L.; O'CONNOR, M.; REMUS, W. Artificial neural network models for forecasting and decision making. *International journal of forecasting*, Elsevier, v. 10, n. 1, p. 5–15, 1994.

HOCHREIN, S.; GLOCK, C. H. Systematic literature reviews in purchasing and supply management research: A tertiary study. *International Journal of Integrated Supply Management*, v. 7, n. 4, p. 215–245, 2012. ISSN 17418097.

HOELTGEBAUM, L. E. B.; DIAS, N. L.; COSTA, M. A. An analog period method for gap-filling of latent heat flux measurements. *Hydrological Processes*, Wiley Online Library, v. 35, n. 4, p. e14105, 2021.

HUNTER, J. D. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, IEEE COMPUTER SOC, v. 9, n. 3, p. 90–95, 2007.

HYNDMAN, R.; KOEHLER, A. B.; ORD, J. K.; SNYDER, R. D. *Forecasting with exponential smoothing: the state space approach.* [S.l.]: Springer Science & Business Media, 2008.

HYNDMAN, R. J.; KOEHLER, A. B.; SNYDER, R. D.; GROSE, S. A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of forecasting*, Elsevier, v. 18, n. 3, p. 439–454, 2002.

HöPPNER, F.; KLAWONN, F. Compensation of translational displacement in time series clustering using cross correlation. In: *International Symposium on Intelligent Data Analysis.* Berlin, Heidelberg: Springer, 2009. p. 71–82.

IMMINK, K.; WEBER, J. Hybrid minimum pearson and euclidean distance detection. *IEEE Transactions on Communications*, v. 63, n. 9, p. 3290–3298, 2015.

ISHWARAN, H.; KOGALUR, U. B.; BLACKSTONE, E. H.; LAUER, M. S. Random survival forests. *The annals of applied statistics*, Institute of Mathematical Statistics, v. 2, n. 3, p. 841–860, 2008.

JABLON, L. S.; AVILA, S. L.; BORBA, B.; MOURÃO, G. L.; FREITAS, F. L.; PENZ, C. A. Diagnosis of rotating machine unbalance using machine learning algorithms on vibration orbital features. n. September 2019, 2020.

JAFFUR, Z. R. K.; SOOKIA, N.-U.-H.; GONPOT, P. N.; SEETANAH, B. Out-of-sample forecasting of the canadian unemployment rates using univariate models. *Applied Economics Letters*, Taylor & Francis, v. 24, n. 15, p. 1097–1101, 2017.

JAIN, L.; SEERA, M.; LIM, C.; BALASUBRAMANIAM, P. A review of online learning in supervised neural networks. *Neural computing and applications*, v. 25, n. 3-4, p. 491–509, 2014.

JANKE, L.; MCDONAGH, S.; WEINRICH, S.; MURPHY, J.; NILSSON, D.; HANSSON, P.; NORDBERG, Optimizing power-to-h2 participation in the nord pool electricity market: Effects of different bidding strategies on plant operation. *Renewable Energy*, 2020.

JARADAT, M.; JARRAH, M.; BOUSSELHAM, A.; JARARWEH, Y.; AL-AYYOUB, M. The internet of energy: smart sensor networks and big data management for smart grid. *Procedia Computer Science*, v. 56, p. 592–597, 2015.

JARDINE, A. K. S.; LIN, D.; BANJEVIC, D. A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical systems and signal processing*, Elsevier, v. 20, n. 7, p. 1483–1510, 2006.

JONG, C. G.; LEU, S. S. Bayesian-network-based hydro-power fault diagnosis system development by fault tree transformation. *Journal of Marine Science and Technology (Taiwan)*, v. 21, n. 4, p. 367–379, 2013. ISSN 10232796.

JOSEPH, A.; CHELLIAH, T. R.; SELVARAJ, R.; LEE, K. B. Fault Diagnosis and Fault-Tolerant Control of Megawatt Power Electronic Converter-Fed Large-Rated Asynchronous Hydrogenerator. *IEEE Journal of Emerging and Selected Topics in Power Electronics*, IEEE, v. 7, n. 4, p. 2403–2416, 2019. ISSN 21686785.

JR, F. E. H.; LEE, K. L.; MARK, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, Wiley Online Library, v. 15, n. 4, p. 361–387, 1996.

KAID, I.; HAFAIFA, A.; GUEMANA, M.; HADROUG, N.; KOUZOU, A.; MAZOUZ, L. Photovoltaic system failure diagnosis based on adaptive neuro fuzzy inference approach: South Algeria solar power plant. *Journal of Cleaner Production*, Elsevier Ltd, v. 204, p. 169–182, 2018. ISSN 09596526. Disponível em: <https://doi.org/10.1016/j.jclepro.2018.09.023>.

KALDELLIS, J. K.; VLACHOU, D. S.; KORBAKIS, G. Techno-economic evaluation of small hydro power plants in Greece: A complete sensitivity analysis. *Energy Policy*, v. 33, n. 15, p. 1969–1985, 10 2005. ISSN 03014215.

KANEGAMI, M.; MIYAZAKI, S.; MIYAKE, K. Partial Discharge Detection with High-Frequency Band through Resistance-Temperature Sensor of Hydropower Generator Stator Windings. *Electrical Engineering in Japan (English translation of Denki Gakkai Ronbunshi)*, v. 195, n. 4, p. 9–15, 2016. ISSN 15206416.

KAUNDA, C. S.; KIMAMBO, C. Z.; NIELSEN, T. K. Potential of Small-Scale Hydropower for Electricity Generation in Sub-Saharan Africa. *ISRN Renewable Energy*, v. 2012, p. 1–15, 2012. ISSN 2090-746X.

KEDADOUCHE, M.; THOMAS, M.; TAHAN, A. A comparative study between Empirical Wavelet Transforms and Empirical Mode Decomposition Methods: Application to bearing defect diagnosis. *Mechanical Systems and Signal Processing*, Elsevier, v. 81, p. 88–107, 2016. ISSN 10961216. Disponível em: <http://dx.doi.org/10.1016/j.ymssp.2016.02.049>.

KELLEY, T. Measures of correlation determined from groups of varying homogeneity. *Journal of the American Statistical Association*, v. 20, n. 152, p. 512–521, 1925.

KHAN, I.; CHOI, S.; KWON, Y.-W. Earthquake detection in a static and dynamic environment using supervised machine learning and a novel feature extraction method. *Sensors*, MDPI, v. 20, n. 3, p. 800, 2020.

KHAN, S. S.; MADDEN, M. G. A survey of recent trends in one class classification. In: SPRINGER. *Irish conference on artificial intelligence and cognitive science*. [S.l.], 2009. p. 188–197.

KLUN, M.; ZUPAN, D.; LOPATIČ, J.; KRYŽANOWSKI, A. On the application of laser vibrometry to perform structural health monitoring in non-stationary conditions of a hydropower dam. *Sensors (Switzerland)*, v. 19, n. 17, 2019. ISSN 14248220.

KOVÁCS, G.; SEBESTYEN, G.; HANGAN, A. Evaluation metrics for anomaly detection algorithms in time-series. *Acta Universitatis Sapientiae, Informatica*, v. 11, n. 2, p. 113–130, 2020. ISSN 2066-7760.

LEE, W. S.; GROSH, D. L.; TILLMAN, F. A.; LIE, C. H. Fault Tree Analysis, Methods, and Applications - A Review. *IEEE Transactions on Reliability*, R-34, n. 3, p. 194–203, 1985. ISSN 15581721.

LEI, Y.; LI, N.; GUO, L.; LI, N.; YAN, T.; LIN, J. Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mechanical Systems and Signal Processing*, Elsevier, v. 104, p. 799–834, 2018.

LI, H.; CHEN, D.; TOLO, S.; XU, B.; PATELLI, E. Hamiltonian Formulation and Analysis for Transient Dynamics of Multi-Unit Hydropower System. *Journal of Computational and Nonlinear Dynamics*, v. 13, n. 10, p. 1–10, 2018. ISSN 15551423.

LI, Y. g.; WANG, L.; MA, M. h. Diagnosis of rotor winding inter-turn short-circuit of hydro-generator based on no-load curve reverse calculation. *IEEJ Transactions on Electrical and Electronic Engineering*, v. 14, n. 1, p. 130–137, 2019. ISSN 19314981.

LIAN, J.; LI, H.; ZHANG, J. ERA modal identification method for hydraulic structures based on order determination and noise reduction of singular entropy. *Science in China, Series E: Technological Sciences*, v. 52, n. 2, p. 400–412, 2009. ISSN 10069321.

LIAO, G. P.; GAO, W.; YANG, G. J.; GUO, M. F. Hydroelectric Generating Unit Fault Diagnosis Using 1-D Convolutional Neural Network and Gated Recurrent Unit in Small Hydro. *IEEE Sensors Journal*, v. 19, n. 20, p. 9352–9363, 2019. ISSN 15581748.

LIU, F. T.; TING, K. M.; ZHOU, Z.-H. Isolation Forest ICDM08. *Icdm*, 2008. Disponível em: <https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf%0Ahttps://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf?q=isolation-forest>.

LIU, F. T.; TING, K. M.; ZHOU, Z. H. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data*, v. 6, n. 1, 2012. ISSN 15564681.

LIU, H.; SHI, J. Applying arma–garch approaches to forecasting short-term electricity prices. *Energy Economics*, v. 37, p. 152–166, 2013.

LIU, H.; TIAN, H.; LIANG, X.; LI, Y. Wind speed forecasting approach using secondary decomposition algorithm and elman neural networks. *Applied Energy*, v. 157, p. 183–194, 2015.

LIU, X.; KRUGER, U.; LITTLER, T.; XIE, L.; WANG, S. Moving window kernel PCA for adaptive monitoring of nonlinear processes. *Chemometrics and Intelligent Laboratory Systems*, Elsevier B.V., v. 96, n. 2, p. 132–143, 2009. ISSN 01697439. Disponível em: <http://dx.doi.org/10.1016/j.chemolab.2009.01.002>.

LIU, X.; LUO, Y.; WANG, Z. A review on fatigue damage mechanism in hydro turbines. *Renewable and Sustainable Energy Reviews*, Elsevier, v. 54, p. 1–14, 2016. ISSN 18790690. Disponível em: <http://dx.doi.org/10.1016/j.rser.2015.09.025>.

LONGMAN, R.; GIAMBELLUCA, T.; NULLET, M.; FRAZIER, A.; KODAMA, K.; CRAUSBAY, S.; ARNOLD, J. Compilation of climate data from heterogeneous networks across the hawaiian islands. *Scientific data*, v. 5, p. 180012, 2018.

LU, S.; ZHANG, X.; SHANG, Y.; LI, W.; SKITMORE, M.; JIANG, S.; XUE, Y. Improving Hilbert–Huang transform for energy-correlation fluctuation in hydraulic engineering. *Energy*, Elsevier B.V., v. 164, p. 1341–1350, 2018. ISSN 03605442.

LUCIFREDI, A.; MAZZIERI, C.; ROSSI, M. Application of multiregressive linear models, dynamic kriging models and neural network models to predictive maintenance of hydroelectric power systems. *Mechanical Systems and Signal Processing*, v. 14, n. 3, p. 471–494, 2000. ISSN 08883270.

LUO, Z.; ZHOU, J.; XIANG, X.; HE, Y.; PENG, S. A new method for automatically identifying the shaft orbit moving direction of hydroelectric generating set. *Sensor Review*, v. 30, n. 3, p. 197–203, 2010. ISSN 02602288.

MADADI, S.; NAZARI-HERIS, M.; MOHAMMADI-IVATLOO, B.; TOHIDI, S. Application of big data analysis to operation of smart power systems. In: *Big Data in Engineering Applications*. Singapore: Springer, 2018. p. 347–362.

MAKRIDAKIS, S.; CHATFIELD, C.; HIBON, M.; LAWRENCE, M.; MILLS, T.; ORD, K.; SIMMONS, L. F. The m2-competition: A real-time judgmentally based forecasting study. *International Journal of Forecasting*, Elsevier, v. 9, n. 1, p. 5–22, 1993.

MAKRIDAKIS, S.; HIBON, M. Accuracy of forecasting: An empirical investigation. *Journal of the Royal Statistical Society: Series A (General)*, Wiley Online Library, v. 142, n. 2, p. 97–125, 1979.

MAKRIDAKIS, S.; HIBON, M. The m3-competition: results, conclusions and implications. *International journal of forecasting*, Elsevier, v. 16, n. 4, p. 451–476, 2000.

MAKRIDAKIS, S.; SPILIOTIS, E.; ASSIMAKOPOULOS, V. The m4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, Elsevier, v. 34, n. 4, p. 802–808, 2018.

MAKRIDAKIS, S.; SPILIOTIS, E.; ASSIMAKOPOULOS, V. The m5 competition: Background, organization, and implementation. *International Journal of Forecasting*, Elsevier, 2021.

MÁRQUEZ, F. P. G.; TOBIAS, A. M.; PÉREZ, J. M. P.; PAPAELIAS, M. Condition monitoring of wind turbines: Techniques and methods. *Renewable Energy*, Elsevier Ltd, v. 46, p. 169–178, 2012. ISSN 09601481. Disponível em: <http://dx.doi.org/10.1016/j.renene.2012.03.003>.

MATEJA, K.; DEJAN, Z.; ANDREJ, K. Vibrations of a hydropower plant under operational loads. *Journal of Civil Structural Health Monitoring*, Springer Berlin Heidelberg, v. 10, n. 1, p. 29–42, 2020. ISSN 21905479. Disponível em: <https://doi.org/10.1007/s13349-019-00367-2>.

MATHUR, A.; CAVANAUGH, K. F.; PATTIPATI, K. R.; WILLETT, P. K.; GALIE, T. R. Reasoning and modeling systems in diagnosis and prognosis. In: SPIE. *Component and Systems Diagnostics, Prognosis, and Health Management.* [S.l.], 2001. v. 4389, p. 194–203.

MAZZOCCHI, E.; PACHOUD, A. J.; FARHAT, M.; HACHEM, F. E.; CESARE, G. D.; SCHLEISS, A. J. Signal analysis of an actively generated cavitation bubble in pressurized pipes for detection of wall stiffness drops. *Journal of Fluids and Structures*, Elsevier, v. 65, p. 60–75, 2016. ISSN 10958622. Disponível em: <http://dx.doi.org/10.1016/j.jfluidstructs.2016.05.009>.

MCDERMOTT, P.; WIKLE, C. A model-based approach for analog spatio-temporal dynamic forecasting. *Environmetrics*, v. 27, n. 2, p. 70–82, 2016.

MCDONALD, J. *Handbook of Biological Statistics (3rd ed.).* [S.l.]: Sparky House Publishing, Baltimore, Maryland, 2014.

MCKINNEY Wes. Data Structures for Statistical Computing in Python. In: WALT Stéfan van der; MILLMAN Jarrod (Ed.). *Proceedings of the 9th Python in Science Conference.* [S.l.: s.n.], 2010. p. 56 – 61.

MELANI, A. H.; SILVA, J. M.; SOUZA, G. F. de; SILVA, J. R. Fault diagnosis based on Petri Nets: the case study of a hydropower plant. *IFAC-PapersOnLine*, v. 49, n. 31, p. 1–6, 2016. ISSN 24058963.

MILIC, S. D.; ZIGIC, A. D.; PONJAVIC, M. M. Online temperature monitoring, fault detection, and a novel heat run test of a water-cooled rotor of a hydrogenerator. *IEEE Transactions on Energy Conversion*, v. 28, n. 3, p. 698–706, 2013. ISSN 08858969.

MINO-AGUILAR, G.; MUÑOZ-HERNÁNDEZ, G. A.; GUERRERO-CASTELLANOS, J. F.; MOLINA-FLORES, E.; DÍAZ-SÁNCHEZ, A.; DOMINGUEZ-RAMIREZ, O. A.; GRACIÓS-MARIN, C. A. Alternative soft fault model of the cross-coupling effect correlated at hydroelectric power energy system. *International Journal of Electrical Power and Energy Systems*, Elsevier Ltd, v. 58, p. 274–280, 2014. ISSN 01420615. Disponível em: <http://dx.doi.org/10.1016/j.ijepes.2014.01.030>.

MIRANIAN, A.; ABDOLLAHZADE, M.; HASSANI, H. Day-ahead electricity price analysis and forecasting by singular spectrum analysis. *IET Generation, Transmission Distribution*, v. 7, n. 4, p. 337–346, 2013.

MONTGOMERY, D.; PECK, E.; VINING, G. *Introduction to linear regression analysis*. [S.l.]: John Wiley Sons, 2012. v. 821.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. *Introduction to linear regression analysis*. [S.l.]: John Wiley & Sons, 2021.

MORENO, S.; COELHO, L. S. Wind speed forecasting approach based on singular spectrum analysis and adaptive neuro fuzzy inference system. *Renewable energy*, v. 126, p. 736–754, 2018.

MUEEN, A.; ZHU, Y.; YEH, M.; KAMGAR, K.; VISWANATHAN, K.; GUPTA, C.; KEOGH, E. *The fastest similarity search algorithm for time series subsequences under Euclidean distance*. 2017. Accessed 14 November, 2019). Disponível em: <https://www.cs.unm.edu/~mueen/FastestSimilaritySearch.html>.

MUNOZ, A.; ERTLÉ, R.; UNSER, M. Continuous wavelet transform with arbitrary scales and o (n) complexity. *Signal processing*, Elsevier, v. 82, n. 5, p. 749–757, 2002.

NAGARAJA, Y.; DEVARAJU, T.; KUMAR, M.; MADICHETTY, S. A survey on wind energy, load and price forecasting: (forecasting methods. In: *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT*. [S.l.]: IEEE, 2016. p. 783–788.

NAVI, M.; MESKIN, N.; DAVOODI, M. Sensor fault detection and isolation of an industrial gas turbine using partial adaptive KPCA. *Journal of Process Control*, Elsevier Ltd, v. 64, p. 37–48, 2018. ISSN 09591524. Disponível em: <https://doi.org/10.1016/j.jprocont.2018.02.002>.

NEMER, E.; GOUBRAN, R.; MAHMOUD, S. Robust voice activity detection using higher-order statistics in the lpc residual domain. *IEEE Transactions on Speech and Audio Processing*, IEEE, v. 9, n. 3, p. 217–231, 2001.

OHUNAKIN, O. S.; OJOLO, S. J.; AJAYI, O. O. Small hydropower (SHP) development in Nigeria: An assessment. *Renewable and Sustainable Energy Reviews*, Elsevier Ltd, v. 15, n. 4, p. 2006–2013, 2011. ISSN 13640321. Disponível em: <http://dx.doi.org/10.1016/j.rser.2011.01.003>.

OLIVEIRA, R. M. D.; MODESTO, J. F.; DMITRIEV, V.; BRASIL, F. S.; VILHENA, P. R. D. Spectral method for localization of multiple partial discharges in dielectric insulation of hydro-generator coils: Simulation and experimental results. *Journal of Microwaves, Optoelectronics and Electromagnetic Applications*, v. 15, n. 3, p. 170–190, 2016. ISSN 21791074.

ORANG, M.; SHIRI, N. A probabilistic approach to correlation queries in uncertain time series data. In: *Proceedings of the 21st ACM international conference on Information and knowledge management*. [S.l.]: ACM, 2012. p. 2229–2233.

PAI, P.-F.; LIN, C.-S. A hybrid arima and support vector machines model in stock price forecasting. *Omega*, Elsevier, v. 33, n. 6, p. 497–505, 2005.

PARDO, M.; BLÁNQUEZ, F. R.; PLATERO, C. A.; REBOLLO, E.; BLÁZQUEZ, F. Detection and location of a ground-fault in the excitation circuit of a 106 MVA synchronous generator by a new on-line method. *Electric Power Systems Research*, Elsevier B.V., v. 140, p. 303–311, 2016. ISSN 03787796. Disponível em: <http://dx.doi.org/10.1016/j.epsr.2016.06.011>.

PEARSON, K. Correlation coefficient. In: *Royal Society Proceedings*. [S.l.: s.n.], 1895. v. 58, p. 214.

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.

PENG, W. J.; LUO, X. Q.; GUO, P. C.; LU, P. Vibration fault diagnosis of hydroelectric unit based on LS-SVM and information fusion technology. *Zhongguo Dianji Gongcheng Xuebao/Proceedings of the Chinese Society of Electrical Engineering*, v. 27, n. 23, p. 86–92, 2007. ISSN 02588013.

PENG, Y.; DONG, M.; ZUO, M. J. Current status of machine prognostics in condition-based maintenance: a review. *The International Journal of Advanced Manufacturing Technology*, Springer, v. 50, n. 1-4, p. 297–313, 2010.

PINO, G.; RIBAS, J. R.; GUIMARÃES, L. F. Bearing Diagnostics of Hydro Power Plants Using Wavelet Packet Transform and a Hidden Markov Model with Orbit Curves. *Shock and Vibration*, v. 2018, 2018. ISSN 10709622.

PIR, M.; SHAH, F.; ASGER, M. Comparative study of different wavelet based neural network models for iip growth forecasting using different yield spreads. *International Journal of Electrical Electronics Computer Science Engineering*, v. 4, n. 6, p. 5–13, 2017.

PÖLSTERL, S. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, v. 21, n. 212, p. 1–6, 2020. Disponível em: <http://jmlr.org/papers/v21/20-729.html>.

POOL, N. *Historical market data (Report of spot prices*. 2020. Retrieved from. Disponível em: <https://www.nordpoolgroup.com/historical-market-data/.>

PUGGINI, L.; MCLOONE, S. An enhanced variable selection and isolation forest based methodology for anomaly detection with oes data. *Engineering Applications of Artificial Intelligence*, Elsevier, v. 67, p. 126–135, 2018.

QIAO, L.; CHEN, Q. Forecasting Models for Hydropower Unit Stability Using LS-SVM. *Mathematical Problems in Engineering*, v. 2015, 2015. ISSN 15635147.

RADACK, G.; BADLER, N. Local matching of surfaces using a boundary-centered radial decomposition. *Computer vision, graphics, and image processing*, v. 45, n. 3, p. 380–396, 1989.

RAHIMI, A.; RECHT, B. Random features for large-scale kernel machines. *Advances in Neural Information Processing Systems 20 - Proceedings of the 2007 Conference*, n. 1, p. 1–8, 2009.

RAMÍREZ-NIÑO, J.; PASCACIO, A.; MIJAREZ, R.; RODRÍGUEZ-RODRÍGUEZ, J. On-line fault monitoring system for hydroelectric generators based on spectrum analysis of the neutral current. *IET Generation, Transmission and Distribution*, v. 9, n. 16, p. 2509–2516, 2015. ISSN 17518687.

RANA, M.; KOPRINSKA, I.; AGELIDIS, V. G. Univariate and multivariate methods for very short-term solar photovoltaic power forecasting. *Energy Conversion and Management*, Elsevier, v. 121, p. 380–390, 2016.

RAVIV, E.; BOUWMAN, K. E.; DIJK, D. V. Forecasting day-ahead electricity prices: Utilizing hourly prices. *Energy Economics*, Elsevier, v. 50, p. 227–239, 2015.

RIDGEWAY, G. The state of boosting. *Computing science and statistics*, p. 172–181, 1999.

RIERA, T. S.; HIGUERA, J.-R. B.; HIGUERA, J. B.; HERRAIZ, J.-J. M.; MONTALVO, J.-A. S. Prevention and fighting against web attacks through anomaly detection technology. a systematic review. *Sustainability*, Multidisciplinary Digital Publishing Institute, v. 12, n. 12, p. 4945, 2020.

RING, M.; ESKOFIER, B. M. An approximation of the Gaussian RBF kernel for efficient classification with SVMs. *Pattern Recognition Letters*, v. 84, p. 1339–1351, 2016. ISSN 01678655.

ROSA, J. J. G. de la; MUÑOZ, A. M. Higher-order cumulants and spectral kurtosis for early detection of subterranean termites. *Mechanical Systems and Signal Processing*, Elsevier, v. 22, n. 2, p. 279–294, 2008.

ROSSUM, G. V.; DRAKE, F. L. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN 1441412697.

SALOMON, C. P.; FERREIRA, C.; SANT'ANA, W. C.; LAMBERT-TORRES, G.; SILVA, L. E. B. da; BONALDI, E. L.; OLIVEIRA, L. E. de Lacerda de; TORRES, B. S. A study of fault diagnosis based on electrical signature analysis for synchronous generators predictive maintenance in bulk electric systems. *Energies*, v. 12, n. 8, 2019. ISSN 19961073.

SALOMON, C. P.; FERREIRA, C.; SANT'ANA, W. C.; LAMBERT-TORRES, G.; SILVA, L. E. B. da; BONALDI, E. L.; OLIVEIRA, L. E. de Lacerda de; TORRES, B. S. A study of fault diagnosis based on electrical signature analysis for synchronous generators predictive maintenance in bulk electric systems. *Energies*, v. 12, n. 8, 2019. ISSN 19961073.

SANTIS, R. B. de; COSTA, M. A. Extended isolation forests for fault detection in small hydroelectric plants. *Sustainability*, MDPI, v. 12, n. 16, p. 6421, 2020.

SANTIS, R. B. de; GONTIJO, T. S.; COSTA, M. A. Condition-based maintenance in hydroelectric plants: A systematic literature review. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, SAGE Publications Sage UK: London, England, p. 1748006X211035623, 2021.

SCImago. [S.l.], 2020.

SELAK, L.; BUTALA, P.; SLUGA, A. Condition monitoring and fault diagnostics for hydropower plants. *Computers in Industry*, Elsevier B.V., v. 65, n. 6, p. 924–936, 2014. ISSN 01663615. Disponível em: <http://dx.doi.org/10.1016/j.compind.2014.02.006>.

SENECHAL, T.; MCDUFF, D.; KALIOUBY, R. E. Facial Action Unit Detection Using Active Learning and an Efficient Non-linear Kernel Approximation. *Proceedings of the IEEE International Conference on Computer Vision*, v. 2015-Febru, p. 10–18, 2015. ISSN 15505499.

SHANMUGAM, R. Book review. *Journal of Statistical Computation and Simulation*, Taylor Francis, v. 76, n. 10, p. 935–940, 2006. Disponível em: <https://doi.org/10.1080/00949650412331321034>.

SHIN, J.-h.; JUN, H.-b. On condition based maintenance policy. *Journal of Computational Design and Engineering*, Elsevier, v. 2, n. 2, p. 119–127, 2015. ISSN 2288-4300. Disponível em: <http://dx.doi.org/10.1016/j.jcde.2014.12.006>.

SI, X.-S.; WANG, W.; HU, C.-H.; ZHOU, D.-H. Remaining useful life estimation–a review on the statistical data driven approaches. *European journal of operational research*, Elsevier, v. 213, n. 1, p. 1–14, 2011.

SIKORSKA, J. Z.; HODKIEWICZ, M.; MA, L. Prognostic modelling options for remaining useful life estimation by industry. *Mechanical Systems and Signal Processing*, v. 25, n. 5, p. 1803–1836, 7 2011. ISSN 08883270.

SIMON, N.; FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software*, NIH Public Access, v. 39, n. 5, p. 1, 2011.

SNIDER, B.; MCBEAN, E. A. Combining machine learning and survival statistics to predict remaining service life of watermains. *Journal of Infrastructure Systems*, American Society of Civil Engineers, v. 27, n. 3, p. 04021019, 2021.

STREHL, A.; GHOSH, J.; MOONEY, R. Impact of similarity measures on web-page clustering. In: *Workshop on artificial intelligence for web search (AAAI 2000*. [S.l.: s.n.], 2000. v. 58, p. 64.

SUN, L.; VERSTEEG, S.; BOZTAS, S.; RAO, A. Detecting Anomalous User Behavior Using an Extended Isolation Forest Algorithm: An Enterprise Case Study. 2016. Disponível em: <http://arxiv.org/abs/1609.06676>.

SUN, L.; VERSTEEG, S.; BOZTAS, S.; RAO, A. Detecting anomalous user behavior using an extended isolation forest algorithm: an enterprise case study. *arXiv preprint arXiv:1609.06676*, 2016.

SUSTO, G. A.; BEGHI, A.; MCLOONE, S. Anomaly Detection through on-line Isolation Forest: An application to plasma etching. p. 89–94, 2017.

SUWANIT, W.; GHEEWALA, S. H. Life cycle assessment of mini-hydropower plants in Thailand. *International Journal of Life Cycle Assessment*, Springer Verlag, v. 16, n. 9, p. 849–858, 2011. ISSN 16147502.

TAM, I.; KALECH, M.; ROKACH, L.; MADAR, E.; BORTMAN, J.; KLEIN, R. Probability-based algorithm for bearing diagnosis with untrained spall sizes. *Sensors*, MDPI, v. 20, n. 5, p. 1298, 2020.

TCHRAKIAN, T. T.; BASU, B.; O'MAHONY, M. Real-time traffic flow forecasting using spectral analysis. *IEEE Transactions on Intelligent Transportation Systems*, IEEE, v. 13, n. 2, p. 519–526, 2011.

TIAN, Z.; JIN, T.; WU, B.; DING, F. Condition based maintenance optimization for wind power generation systems under continuous monitoring. *Renewable Energy*, Elsevier Ltd, v. 36, n. 5, p. 1502–1509, 2011. ISSN 09601481. Disponível em: <http://dx.doi.org/10.1016/j.renene.2010.10.028>.

TUKEY, J. W. et al. *Exploratory data analysis*. [S.l.]: Reading, MA, 1977. v. 2.

TULARAM, G. A.; SAEED, T. Oil-price forecasting based on various univariate time-series models. *American Journal of Operations Research*, Scientific Research Publishing, v. 6, n. 03, p. 226, 2016.

UNIDO. *World Small Hydropower Development Report 2016*. [S.l.], 2016. Disponível em: <www.smallhydroworld.org.>

UNO, H.; CAI, T.; PENCINA, M. J.; D'AGOSTINO, R. B.; WEI, L.-J. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, Wiley Online Library, v. 30, n. 10, p. 1105–1117, 2011.

VALAVI, M.; JORSTAD, K. G.; NYSVEEN, A. Electromagnetic analysis and electrical signature-based detection of rotor inter-turn faults in salient-pole synchronous machine. *IEEE Transactions on Magnetics*, v. 54, n. 9, p. 1–9, 2018. ISSN 00189464.

VALENTÍN, D.; PRESAS, A.; EGUSQUIZA, M.; VALERO, C.; EGUSQUIZA, E. Transmission of high frequency vibrations in rotating systems. Application to cavitation detection in hydraulic turbines. *Applied Sciences (Switzerland)*, v. 8, n. 3, p. 1–18, 2018. ISSN 20763417.

VAROQUAUX, G.; BUITINCK, L.; LOUPPE, G.; GRISEL, O.; PEDREGOSA, F.; MUELLER, A. Scikit-learn. *GetMobile: Mobile Computing and Communications*, v. 19, n. 1, p. 29–33, 2015. ISSN 2375-0529.

VARTOUNI, A. M.; KASHI, S. S.; TESHNEHLAB, M. An anomaly detection method to detect web attacks using Stacked Auto-Encoder. *2018 6th Iranian Joint Congress on Fuzzy and Intelligent Systems, CFIS 2018*, IEEE, v. 2018-Janua, p. 131–134, 2018.

VIRTANEN, P.; GOMMERS, R.; OLIPHANT, T. E.; HABERLAND, M.; REDDY, T.; COURNAPEAU, D.; BUROVSKI, E.; PETERSON, P.; WECKESSER, W.; BRIGHT, J.; van der Walt, S. J.; BRETT, M.; WILSON, J.; MILLMAN, K. J.; MAYOROV, N.; NELSON, A. R. J.; JONES, E.; KERN, R.; LARSON, E.; CAREY, C. J.; POLAT, İ.; FENG, Y.; MOORE, E. W.; VanderPlas, J.; LAXALDE, D.; PERKTOLD, J.; CIMRMAN, R.; HENRIKSEN, I.; QUINTERO, E. A.; HARRIS, C. R.; ARCHIBALD, A. M.; RIBEIRO, A. H.; PEDREGOSA, F.; van Mulbregt, P.; SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, v. 17, p. 261–272, 2020.

VORONIN, S.; PARTANEN, J. Forecasting electricity price and demand using a hybrid approach based on wavelet transform, arima and neural networks. *International Journal of Energy Research*, v. 38, n. 5, p. 626–637, 2014.

VORONOV, S.; KRYSANDER, M.; FRISK, E. Predictive maintenance of lead-acid batteries with sparse vehicle operational data. *International Journal of Prognostics and Health Management*, v. 11, n. 1, 2020.

VOYANT, C.; NOTTON, G.; KALOGIROU, S.; NIVET, M.-L.; PAOLI, C.; MOTTE, F.; FOUILLOY, A. Machine learning methods for solar radiation forecasting: A review. *Renewable Energy*, Elsevier, v. 105, p. 569–582, 2017.

VU, V. H.; THOMAS, M.; LAFLEUR, F.; MARCOUILLER, L. Towards an automatic spectral and modal identification from operational modal analysis. *Journal of Sound and Vibration*, v. 332, n. 1, p. 213–227, 2013. ISSN 10958568.

W. WANG K., L. C. L. X. L. Y. Z. H. F. Vibration trend measurement for hydropower generator based on optimal variational mode decomposition and LSSVM improved with chaotic sine cosine algorithm optimization. *Measurement Science And Technology*, p. 1361–6463, 2018. ISSN 1524-8372.

WANG, W.; CHEN, Q.; YAN, D.; GENG, D. A novel comprehensive evaluation method of the draft tube pressure pulsation of Francis turbine based on EEMD and information entropy. *Mechanical Systems and Signal Processing*, Elsevier Ltd, v. 116, p. 772–786, 2019. ISSN 10961216. Disponível em: <https://doi.org/10.1016/j.ymssp.2018.07.033>.

WANG, X.; LI, J.; YU, R. Modeling disruption durations of subway service via random survival forests: The case of shanghai. *Journal of Transportation Safety & Security*, Taylor & Francis, p. 1–23, 2022.

WASKOM, M. L. seaborn: statistical data visualization. *Journal of Open Source Software*, The Open Journal, v. 6, n. 60, p. 3021, 2021. Disponível em: <https://doi.org/10.21105/joss.03021>.

WEC. *World Energy Insights Brief*. London, 2019. 35 p. Disponível em: <https://www.worldenergy.org/assets/downloads/WEInsights-Brief-Global-Energy-Scenarios-Comparison-Review-R02.pdf>.

WELLING, M. Robust higher order statistics. In: PMLR. *International Workshop on Artificial Intelligence and Statistics*. [S.l.], 2005. p. 405–412.

WILLIAMS, A. How to write and analyse a questionnaire. *Journal of Orthodontics*, v. 30, n. 3, p. 245–252, 2003. ISSN 14653125.

WU, G.; TONG, J.; ZHANG, L.; ZHAO, Y.; DUAN, Z. Framework for fault diagnosis with multi-source sensor nodes in nuclear power plants based on a Bayesian network. *Annals of Nuclear Energy*, v. 122, p. 297–308, 2018. ISSN 18732100.

WU, Y.; JIANG, B.; WANG, Y. Incipient winding fault detection and diagnosis for squirrel-cage induction motors equipped on crh trains. *ISA transactions*, Elsevier, v. 99, p. 488–495, 2020.

WU, Y.; LI, Z.; LI, B.; CHU, F.; CUI, X. Vibration monitoring approach in integrated system of hydroelectric generating sets for smart power stations. *International Journal of Power and Energy Systems*, v. 36, n. 4, p. 147–155, 2016. ISSN 10783466.

XIA, X.; NI, W. 2145. A novel failure analysis and diagnosis method for hydraulic-turbine generator unit. *Journal of Vibroengineering*, v. 18, n. 6, p. 3568–3580, 2016. ISSN 13928716.

XIA, X.; NI, W.; SANG, Y. A novel analysis method for fault diagnosis of hydro-turbine governing system. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, v. 231, n. 2, p. 164–171, 2017. ISSN 17480078.

XIA, X.; ZHOU, J.; LI, C.; ZHU, W. A novel method for fault diagnosis of hydro generator based on NOFRFs. *International Journal of Electrical Power and Energy Systems*, Elsevier Ltd, v. 71, p. 60–67, 2015. ISSN 01420615. Disponível em: <http://dx.doi.org/10.1016/j.ijepes.2015.02.022>.

XIAO, H.; ZHOU, J.; XIAO, J.; FU, W.; XIA, X.; ZHANG, W. Identification of vibration-speed curve for hydroelectric generator unit using statistical fuzzy vector chain code and support vector machine. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, v. 228, n. 3, p. 291–300, 2014. ISSN 17480078.

XIAO, Z.; HE, X.; FU, X.; MALIK, O. P. ACO-Initialized Wavelet Neural Network for Vibration Fault Diagnosis of Hydroturbine Generating Unit. *Mathematical Problems in Engineering*, v. 2015, 2015. ISSN 15635147.

XU, B.; CHEN, D.; BEHRENS, P.; YE, W.; GUO, P.; LUO, X. Modeling oscillation modal interaction in a hydroelectric generating system. *Energy Conversion and Management*, Elsevier, v. 174, n. August, p. 208–217, 2018. ISSN 01968904. Disponível em: <https://doi.org/10.1016/j.enconman.2018.08.034>.

XU, B.; CHEN, D.; PATELLI, E.; SHEN, H.; PARK, J. H. Mathematical model and parametric uncertainty analysis of a hydraulic generating system. *Renewable Energy*, Elsevier Ltd, v. 136, p. 1217–1230, 2019. ISSN 18790682. Disponível em: <https://doi.org/10.1016/j.renene.2018.09.095>.

XU, B.; CHEN, D.; ZHANG, H.; LI, C.; ZHOU, J. Shaft mis-alignment induced vibration of a hydraulic turbine generating system considering parametric uncertainties. *Journal of Sound and Vibration*, Elsevier Ltd, v. 435, p. 74–90, 2018. ISSN 10958568. Disponível em: <https://doi.org/10.1016/j.jsv.2018.08.008>.

XU, B.; LI, H.; PANG, W.; CHEN, D.; TIAN, Y.; LEI, X.; GAO, X.; WU, C.; PATELLI, E. Bayesian network approach to fault diagnosis of a hydroelectric generation system. *Energy Science and Engineering*, v. 7, n. 5, p. 1669–1677, 2019. ISSN 20500505.

XU, B.; YAN, D.; CHEN, D.; GAO, X.; WU, C. Sensitivity analysis of a Pelton hydropower station based on a novel approach of turbine torque. *Energy Conversion and Management*, Elsevier Ltd, v. 148, p. 785–800, 2017. ISSN 01968904. Disponível em: <http://dx.doi.org/10.1016/j.enconman.2017.06.019>.

XU, Y. Intelligent fault diagnosis of house transformer simulation system in hydro-electricity factory by fuzzy reasoning. *International Journal of Internet Manufacturing and Services*, v. 3, n. 2, p. 87–98, 2013. ISSN 17516048.

XUE, X.; SUNDARARAJAN, V.; GONZALEZ-ARGUETA, L. Sensor fusion for machine condition monitoring. *Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2008*, v. 6932, n. September, p. 69321D, 2008. ISSN 0277786X.

XUE, X.; ZHOU, J.; ZHANG, Y.; ZHANG, W.; ZHU, W. An improved ensemble empirical mode decomposition method and its application to pressure pulsation analysis of hydroelectric generator unit. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, v. 228, n. 6, p. 543–557, 2014. ISSN 17480078.

YANG, D. Ultra-fast analog ensemble using kd-tree. *Journal of Renewable and Sustainable Energy*, v. 11, n. 5, p. 053703, 2019.

YANG, D.; ALESSANDRINI, S. An ultra-fast way of searching weather analogs for renewable energy forecasting. *Solar Energy*, Elsevier, v. 185, p. 255–261, 2019.

YANG, D.; ALESSANDRINI, S. An ultra-fast way of searching weather analogs for renewable energy forecasting. *Solar Energy*, v. 185, p. 255–261, 2019.

YANG, D.; KLEISSL, J.; GUEYMARD, C.; PEDRO, H.; COIMBRA, C. History and trends in solar irradiance and pv power forecasting: A preliminary assessment and review using text mining. *Solar Energy*, v. 168, p. 60–101, 2018.

YANG, J.; ASTITHA, M.; MONACHE, L. D.; ALESSANDRINI, S. An analog technique to improve storm wind speed prediction using a dual nwp model approach. *Monthly Weather Review*, v. 146, n. 12, p. 4057–4077, 2018.

YOUSEFI, N.; TSIANIKAS, S.; COIT, D. W. Reinforcement learning for dynamic condition-based maintenance of a system with individually repairable components. *Quality Engineering*, Taylor & Francis, v. 32, n. 3, p. 388–408, 2020.

YU, R.; BREIKIN, T. Hybrid neural network-based fault diagnosis and fault-tolerance design with application in electro-hydraulic Servovalve. *Open Automation and Control Systems Journal*, v. 2, n. 1, p. 62–68, 2009. ISSN 18744443.

ZEMOURI, R.; LEVESQUE, M.; AMYOT, N.; HUDON, C.; KOKOKO, O.; TAHAN, S. A. Deep Convolutional Variational Autoencoder as a 2D-Visualization Tool for Partial Discharge Source Classification in Hydrogenerators. *IEEE Access*, v. 8, p. 5438–5454, 2020. ISSN 21693536.

ZHANG, L.; WU, Q.; MA, Z.; WANG, X. Transient vibration analysis of unit-plant structure for hydropower station in sudden load increasing process. *Mechanical Systems and Signal Processing*, Elsevier Ltd, v. 120, n. 79, p. 486–504, 2019. ISSN 10961216. Disponível em: <https://doi.org/10.1016/j.ymssp.2018.10.037>.

ZHANG, X.; ZHOU, J.; GUO, J.; ZOU, Q.; HUANG, Z. Vibrant fault diagnosis for hydroelectric generator units with a new combination of rough sets and support vector machine. *Expert Systems with Applications*, Elsevier Ltd, v. 39, n. 3, p. 2621–2628, 2012. ISSN 09574174. Disponível em: <http://dx.doi.org/10.1016/j.eswa.2011.08.117>.

ZHANG, Y.; ZHAO, X.; ZUO, Y.; REN, L.; WANG, L. The development of the renewable energy power industry under feed-in tariff and renewable portfolio standard: A case study of china's photovoltaic power industry. *Sustainability*, Multidisciplinary Digital Publishing Institute, v. 9, n. 4, p. 532, 2017.

ZHANG, Y.; ZHONG, M.; GENG, N.; JIANG, Y. Forecasting electric vehicles sales with univariate and multivariate time series models: The case of china. *PloS one*, Public Library of Science, v. 12, n. 5, p. e0176729, 2017.

ZHAO, R.; YAN, R.; CHEN, Z.; MAO, K.; WANG, P.; GAO, R. X. *Deep learning and its applications to machine health monitoring.* [S.l.]: Elsevier, 2019. 213–237 p.

ZHOU, K. B.; ZHANG, J. Y.; SHAN, Y.; GE, M. F.; GE, Z. Y.; CAO, G. N. A hybrid multi-objective optimization model for vibration tendency prediction of hydropower generators. *Sensors (Switzerland)*, v. 19, n. 9, 2019. ISSN 14248220.

ZHU, W.; ZHOU, J.; XIA, X.; LI, C.; XIAO, J.; XIAO, H.; ZHANG, X. A novel KICA-PCA fault detection model for condition process of hydroelectric generating unit. *Measurement: Journal of the International Measurement Confederation*, Elsevier Ltd, v. 58, p. 197–206, 2014. ISSN 02632241. Disponível em: <http://dx.doi.org/10.1016/j.measurement.2014.08.026>.

ZHU, Y.; IMAMURA, M.; NIKOVSKI, D.; KEOGH, E. Introducing time series chains: a new primitive for time series data mining. *Knowledge and Information Systems*, v. 60, n. 2, p. 1135–1161, 2019.

ŽVOKELJ, M.; ZUPAN, S.; PREBIL, I. EEMD-based multiscale ICA method for slewing bearing fault detection and diagnosis. *Journal of Sound and Vibration*, v. 370, p. 394–423, 2016. ISSN 10958568.