



Universidade Federal de Juiz de Fora
Curso de Pos-Graduação em Engenharia Elétrica
Habitação em Sistemas de Energia

Arthur Couto Piazzzi

PLATAFORMA DE PESQUISA REPRODUTÍVEL PARA QUALIDADE DE ENERGIA
ELÉTRICA

Dissertação de Mestrado

Juiz de Fora
2018

Arthur Couto Piazzzi

Plataforma de Pesquisa Reprodutível para Qualidade de Energia Elétrica

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Juiz de Fora, na área de concentração em Sistemas de Energia, como requisito parcial para obtenção título de Mestre em Engenharia Elétrica.

Orientador: Prof. Augusto Santiago Cerqueira

Juiz de Fora
2018

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

Piazzzi, Arthur Couto.

Plataforma de Pesquisa Reprodutível para Qualidade de Energia Elétrica / Arthur Couto Piazzzi. -- 2018.

82 f. : il.

Orientador: Augusto Santiago Cerqueira

Dissertação (mestrado acadêmico) - Universidade Federal de Juiz de Fora, Faculdade de Engenharia. Programa de Pós Graduação em Engenharia Elétrica, 2018.

1. Pesquisa Reprodutível. 2. Qualidade de Energia Elétrica. 3. Plataforma Web. I. Cerqueira, Augusto Santiago , orient. II. Título.

Arthur Couto Piazzzi

Plataforma de Pesquisa Reprodutível para Qualidade de Energia Elétrica

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Juiz de Fora, na área de concentração em Sistemas de Energia, como requisito parcial para obtenção título de Mestre em Engenharia Elétrica.

Aprovada em 22 de Agosto de 2018.

BANCA EXAMINADORA:

Prof. Dr. Augusto Santiago Cerqueira

Universidade Federal de Juiz de Fora, UFJF

Orientador

Prof. Dr. Leandro Rodrigues Manso da Silva

Universidade Federal de Juiz de Fora, UFJF

Prof. Dr Danton Diego Ferreira

Universidade Federal de Lavras, UFLA

À Deus e a minha família

AGRADECIMENTOS

Agradeço primeiramente a Deus, não somente pela vida e saúde mas também pelas de inúmeras bênçãos. Aos meus pais, Luiz e Lorena, pelo amor e carinho incondicional. Aos meus irmãos Thiago e Luís pelo companheirismo e cumplicidade.

Ao meu primo Rafael (*in memoriam*), pelas memórias mais divertidas de toda a minha vida. Sua ausência será sentida eternamente.

A minha namorada Aline, por todos momentos maravilhosos que vivemos juntos e ainda ainda vamos viver.

Aos amigos e companheiros do LAPTEL, pelas ajudas e instantes de descontração.

Aos amigos da graduação e da vida, em especial ao Guido, Lucas, Geruza e Marina. Por me dar tantos motivos, que seria impossível citar somente um.

Ao Professor Augusto Santiago Cerqueira pela orientação, ensinamentos, amizade, atenção e paciência durante todo o desenvolvimento deste trabalho.

Ao CNPq, CAPES, FAPEMIG, à Universidade Federal de Juiz de Fora e à Faculdade de Engenharia pelo suporte financeiro e por prover as ferramentas necessárias para o desenvolvimento deste trabalho.

Quando se trata de luto, as regras normais de troca não se aplicam, porque o luto transcende o valor. Um homem daria nações inteiras para tirar a tristeza do seu coração. E, no entanto, você não pode comprar nada com tristeza, porque tristeza é inútil.

The Counselor

RESUMO

A computação científica está emergindo como o pilar central da ciência. No entanto, as práticas não rigorosas de pesquisa e a própria natureza do trabalho estão levando a uma crise de credibilidade. Mesmo em cenários em que a reprodução é factível, o tempo despendido para reimplementação de pesquisas representa uma expressiva parcela do tempo útil de estudiosos e pesquisadores. A pesquisa computacional reproduzível vem ganhando atração nas últimas décadas, não somente porque beneficia a validação e comprovação dos resultados, tornando uma publicação mais confiável, mas também por diminuir a barreira de entrada, apresentando-a de forma mais compreensível para o público alvo e, conseqüentemente, aumentando impacto. Ao longo das décadas foram desenvolvidas diversas aplicações gerais ou específicas para estimular a cultura de reprodutibilidade. No entanto, as presentes soluções não abrangem todos os campos do conhecimento. Neste trabalho, é apresentada uma plataforma online onde é possível armazenar, compartilhar e executar códigos e suas dependências, para processamento de sinais, aplicados à qualidade da energia elétrica. Desta forma, esta plataforma apresenta uma alternativa para que pesquisadores da área possam realizar e expor seus trabalhos de forma mais dinâmica, transparente e amigável aos leitores. Adicionalmente, a plataforma foi utilizada para avaliação e comparação de técnicas recentemente propostas na literatura para detecção de novidade no sinal elétrico de potência.

Palavras-chave: Pesquisa Reprodutível, Qualidade de Energia Elétrica, Plataforma Web.

ABSTRACT

Scientific computing is emerging as the central pillar of science. However, non-rigorous research practices and the very nature of work are leading to a crisis of credibility. Even in scenarios where reproduction is feasible, the time spent reimplementing research represents a significant portion of the useful time of scholars and researchers. Reproducible computational research has been gaining momentum in recent decades, not only because it benefits the validation and verification of results, making the publication more reliable, but also by reducing the barrier of entry, presenting it in a way that is more understandable to the target audience and, consequently, increasing its impact. Over the decades, several general or specific applications have been developed to stimulate the culture of reproducibility. However, the present solutions do not cover all fields of knowledge. In this work, an online platform is presented where it is possible to store, share and execute codes and their dependencies, for signal processing applied to the electric power system. In this way, this platform presents an alternative for researchers of the area to realize and share their work in a more dynamic, transparent and friendly way for the readers. Additionally, the platform was used to evaluate and compare recently proposed techniques for novelty detection in the electric power system signal.

Keywords: Reproducible Research, Power Quality, Web Platform.

LISTA DE ILUSTRAÇÕES

Figura 1	Arquitetura Smart Grid	23
Figura 2	Ciclo Hipotético-Empírico.	25
Figura 3	Diagrama do tradicional modelo de publicação.	27
Figura 4	Espectro da reprodutibilidade.	28
Figura 5	Fluxograma de detecção de novidade baseada no domínio do tempo.	38
Figura 6	Fluxograma de detecção de novidade baseada no domínio da frequência.	39
Figura 7	Exemplo de curvas Característica de Operação do Receptor, do inglês <i>Receiver Operating Characteristic</i> (ROC).	41
Figura 8	Visão geral do plataforma.	44
Figura 9	Página de Login.	44
Figura 10	Visualização da página inicial.	45
Figura 11	Visualização da página de algoritmos.	45
Figura 12	Visualização de uma rotina na página de algoritmos.	46
Figura 13	<i>Template</i> de um script incorporado à plataforma.	47
Figura 14	Visualização da página de exibição do <i>Dataset</i> .	49
Figura 15	Visualização da página contendo um evento elétrico.	49
Figura 16	Página de exibição de experimentos.	51
Figura 17	Página de exibição de experimentos.	51
Figura 18	Passo a passo da adição de um novo experimento.	52
Figura 19	Página de exibição de resultados.	53
Figura 20	Passo a passo da adição de uma nova comparação.	54
Figura 21	Página de exibição de comparação.	55
Figura 22	Busca realizadas no Google sobre linguagens de programação.	56
Figura 23	Perguntas realizadas no Stack OverFlow sobre linguagens de programação.	57

Figura 24	Estatísticas referente a linguagens utilizadas no site Github em 2017.	57
Figura 25	Busca realizadas no Google sobre <i>Frameworks Back-End</i>	58
Figura 26	Perguntas realizadas no Stack OverFlow sobre <i>Frameworks Back-End</i>	59
Figura 27	Arquitetura MVC.	59
Figura 28	Busca realizadas no Google sobre <i>Frameworks Front-End</i>	60
Figura 29	Perguntas realizadas no Stack OverFlow sobre <i>Frameworks Front-End</i> . ..	61
Figura 30	Explicação da ferramenta AJAX.	61
Figura 31	Bootstrap vs Foundation: Pesquisas no Google.	62
Figura 32	Bootstrap vs Foundation: Perguntas no Stack Overflow.	62
Figura 33	Visão esquemática do site evidenciando as tecnologias utilizadas.	64
Figura 34	Análise classificadores baixo ruído.	67
Figura 35	Comparação entre DEF e ER SNR 30.	67
Figura 36	Comparação entre DEF e ER SNR 20.	68
Figura 37	Comparação entre métodos temporais com filtro SNR 30 dB.	69
Figura 38	Comparação entre métodos temporais com filtro SNR 20.	70
Figura 39	Comparação entre métodos espectrais SNR 30.	71
Figura 40	Comparação entre métodos espectrais SNR 20 dB.	71

LISTA DE TABELAS

Tabela 1	Níveis de reprodutibilidade.	29
Tabela 2	Comparação entre as funcionalidades dos recursos em prol da RR.	63
Tabela 3	Composição do banco de dados	66
Tabela 4	Recapitulação das siglas	66
Tabela 5	Resultados de probabilidade de detecção e falso alarme.	72
Tabela 6	Resultados de área embaixo do gráfico e precisão média.	73

LISTA DE ABREVIATURAS E SIGLAS

QEE Qualidade de Energia Elétrica

MVC Modelo Visão Controle, do inglês Model View Controller

JS JavaScript

CSS Folhas de Estilo em Cascata ,do inglês Cascading Style Sheets

HTML Linguagem de Marcação de Hipertexto, do inglês HyperText Markup Language

SGs Redes Inteligentes, do inglês *Smart Grids*

SEP Sistema Elétrico de Potência

RMS Raiz do Valor Quadrático Médio, do inglês *Root Mean Square*

FT Transformada de Fourier, do inglês *Fourier Transform*

WT *Wavelet Transform*

STFT Transformada de Fourier de curta duração, do inglês *Short-Time Fourier Transform*

SNR Relação Sinal Ruído, do inglês *Signal to Noise Ratio*

FPGA Arranjo de Portas Programáveis em Campo, do inglês *Field Programmable Gate Array*

MM Morfologia Matemática

SVDD Descrição de Dados por Vetores de Suporte, do inglês *Support Vector Data Description*

PDE Estimador de Densidade Parzen, do inglês *Parzen Density Estimator*

HH Hilbert Huang

NN Redes Neurais, do inglês *Neural Network*

ICA Análise de componente independente, do inglês *Independent Component Analysis*

SVM Máquina de Vetor Suporte, do inglês *Support Vector Machine*

RR Pesquisa Reprodutível, do inglês *Reproducible Research*

DEF Diferença de Energia Entre Frames

ER Energia do Resíduo

FFT Transformada de Fourier Rápida, do inglês *Fast Fourier Transform*

VMCF Variação Máxima em Componentes de Frequência

VMMC Variação Máxima na Magnitude das Componentes

PRESEP Plataforma de Pesquisa Reprodutível para o Processamento de Sinais Aplicado ao SEP

FIR Resposta Finita ao Impulso, do inglês *Finite Impulse Response*

ROC Característica de Operação do Receptor, do inglês *Receiver Operating Characteristic*

PR Precisão e revocação, do inglês *Precision Recall*

AUC Área Embaixo do Gráfico, do inglês *Area Under the Curve*

SP Soma-Produto

CSV Valores separados por vírgula, do inglês *Comma Separated Values*

JSON Notação de Objetos *JavaScript*, do inglês *JavaScript Object Notation*

GPU Unidade de Processamento Gráfico, do inglês *Graphics Processing Unit*

JMLR Jornal de Pesquisas em Aprendizado de Máquinas, do inglês *Journal of Machine Learning Research*

IEEE Instituto de Engenheiros Eletricistas e Eletrônicos, do inglês *Institute of Electrical and Electronics Engineers*

SUMÁRIO

1 Introdução	18
1.1 Identificação do problema	19
1.2 Objetivos	20
1.2.1 Objetivos Gerais	20
1.2.2 Objetivos Específicos	20
1.3 Estrutura da dissertação	21
2 Processamento de Sinais Aplicado ao SEP e Pesquisa Reprodutível	22
2.1 Processamento de Sinais Aplicados ao SEP	22
2.2 Pesquisa Computacional e Pesquisa Reprodutível	23
2.3 Pesquisa Reprodutível hoje	30
2.3.1 Ferramentas	30
Knitr	30
Notebooks	30
VisTrails	31
2.3.2 Plataformas	31
RunMyCode.org	31
Acervos online de pesquisas	32
Code Ocean	33
BEAT Platform	33
2.4 Detecção de Novidade no Sinal do SEP	34
2.4.1 Técnicas Promissoras que Serão Avaliadas	37
Método Baseado no Domínio do Tempo (<i>SILVA, 2016</i>)	38

Método Baseado no Domínio da Frequência (HEYDT et al., 1999;ELSPEC, 2007) . . .	39
Estratégias de comparação e definição de limiar ótimo em detecção	40
2.5 Conclusões parciais	42
3 Plataforma de Pesquisa Reprodutível para Processamento de Sinais Aplicado ao SEP	43
3.1 Componentes da Plataforma	43
3.1.1 Repositório de algoritmos	45
3.1.2 Banco de dados	47
3.1.3 Repositório de experimentos	49
3.1.4 Comparações e Análises	53
3.2 Tecnologias utilizadas	54
3.2.1 Python	55
3.2.2 Django	56
3.2.3 jQuery	59
3.2.4 Bootstrap	61
3.3 Conclusões parciais	63
4 Utilização da PRESEP para Avaliação de Técnicas para Detecção de Novidades no Sinal SEP	65
4.1 Detecção de novidades	65
4.1.1 Exposição de resultados através da PRESEP	65
Métodos no domínio do tempo	66
Métodos no domínio da frequência	70
4.2 Conclusões parciais	72
5 CONCLUSÕES	74
5.1 Trabalhos futuros	75

Referências

77

Apêndice A – Produção Científica

82

1 INTRODUÇÃO

O uso de energia elétrica é de fundamental importância para todos os segmentos da sociedade. A crescente demanda por energia elétrica impulsiona o constante desenvolvimento de novas tecnologias de geração, transmissão e distribuição. Juntamente com os benefícios alcançados por esses avanços, surgem os desafios de operação, controle e condicionamento de um sistema cada vez mais complexo. Faz-se necessário mencionar algumas mudanças ocorridas nas últimas décadas, não somente pelo benefício que promoveram, mas também pelas consequências que incitaram.

A ampla e crescente utilização de eletrônica de potência no sistema elétrico de potência (SEP) tem incorporado diversas vantagens e funcionalidades. Em contrapartida, estes equipamentos, devido às suas características não-lineares, muitas vezes deterioram a Qualidade de Energia Elétrica (QEE), inserindo diferentes tipos de distúrbios na rede de alimentação. Tais deformações geram efeitos depreciativos em componentes do SEP, e portanto, devem ser minimizadas. A monitoração e o processamento de sinais elétricos tornaram-se importantes campos de estudos nas últimas três décadas, por serem parte essencial na busca de soluções para os problemas correntes.

Atualmente, a rede de energia está passando por uma outra revolução com o advento de **SG!** (**SG!**). Essas redes são compostas pela integração de diversas novas tecnologias, como sensores e medidores inteligentes, sistemas de gestão de dados, distribuição automatizada e sistemas de armazenamento de energia. Tais sistemas habilitam aplicações que anteriormente eram impraticáveis, como otimização operacional, integração eficiente de fontes renováveis de energia à rede elétrica, comunicação bidirecional de informação, redes auto recuperáveis, gestão domiciliar de energia, entre outras. Essas alterações são necessárias para suprir a crescente demanda de energia. Além disso, elas tendem a reduzir problemas atuais relacionados à geração de energia, como a degradação ambiental ocasionada pela produção de energia e a diminuição das reservas de combustíveis fósseis. Todas essas alterações, produzem um novo sistema que ainda não é totalmente compreendido pelos cientistas e engenheiros.

1.1 IDENTIFICAÇÃO DO PROBLEMA

Como exposto anteriormente, o processamento de sinais do Sistema Elétrico de Potência (SEP) caracteriza uma área de crescente importância para a sociedade, sobretudo no contexto das redes inteligentes. Por outro lado, diversas áreas de pesquisa, entre elas o processamento digital de sinais, são passíveis de limitações no que se diz respeito à sua correta disseminação e reutilização dos avanços desenvolvidos. Diversos empecilhos que serão expostos nos próximos capítulos, diminuem a transparência, entendimento e a difusão de novas técnicas, criando entraves ao processo de evolução do conhecimento.

O progresso científico é baseado na utilização e posterior aperfeiçoamento dos conhecimentos já existentes. Uma das frases mais célebres associada à esse conceito foi escrita por Sir Isaac Newton em 1675: “Se eu enxerguei mais longe, foi me apoiando no ombro de gigantes.”

Dessa forma, a disseminação efetiva do conhecimento é tão importante quanto o desenvolvimento do mesmo. Para garantir a qualidade do conteúdo transmitido nos veículos de comunicação científicos como jornais, revista e congressos, a prática de revisão por pares (do inglês *peer-review*), que consiste na avaliação de publicações por especialistas da área, é adotada desde o século XVII. Apesar desse processo ter se consolidado no início dos anos 70, o processo de revisão foi primeiramente adotado em 1665 pela *Royal Society* em Londres (MOXHAM & FYFE, 2017).

Publicações baseadas em ciência computacional expõem descrição de técnicas, métodos e seus resultados, mas muitas vezes falham na divulgação dos demais elementos, como código, parametrização e base de dados utilizados. Essa apresentação parcial das inovações desenvolvidas dificultam o processo de atestamento de resultados e reutilização em pesquisas de terceiros, desacelerando o processo de desenvolvimento na área.

Para fomentar avanços em áreas que se apoiam em algoritmos computacionais, criando pesquisas que podem ser facilmente replicadas, atestadas e utilizadas, emergiu nas últimas décadas o conceito de Pesquisa Reprodutível, do inglês *Reproducible Research* (RR). Diversos cientistas defendem que tal prática deve ser adotada como padrão mínimo de rigor em pesquisas. Estudos em diferentes áreas apontam que as atuais condutas aliadas às dificuldades inerentes ao ramo computacional estão gerando uma crise de reprodutibilidade.

A fim de promover pesquisas mais transparentes, diversos recursos foram desenvolvidos para possibilitar e facilitar a prática. O capítulo posterior apresentará uma revi-

são sobre os avanços já obtidos. Embora uma solução geral ao problema seja desejada, essa pode não ser factível por diferentes requerimentos de cada campo, além de onerado custo de desenvolvimento e manutenção de um suposto repositório abrangente a todas as áreas. Soluções específicas às linhas de pesquisa oferecem vantagens, por atender diferentes necessidades e facilidades de indexação e busca de conteúdos. Avanços podem ser observados no campo da Biomédica, Genética, Inteligencia Artificial e Segurança Biométrica, entre outros, mas até o presente momento pesquisadores que atuam com processamento de sinais aplicado ao sistema de potência não possuem um recurso específico que auxilie na prática de RR.

1.2 OBJETIVOS

1.2.1 OBJETIVOS GERAIS

Partindo do pressuposto de que a introdução de transparência, rapidez e confiabilidade do compartilhamento de conhecimento impacta positivamente a área e através da observação da não existência de uma ferramenta específica que facilite essa prática na área de processamento de sinais do sistema aplicado ao SEP, foi elaborado o presente trabalho. Portanto, este trabalho tem como objetivo contribuir para a aplicação dos conceitos de pesquisa reprodutível na área de processamento de sinais aplicado ao SEP.

1.2.2 OBJETIVOS ESPECÍFICOS

1. Desenvolvimento de uma plataforma de pesquisa reprodutível para processamento de sinais aplicado ao SEP, contendo as seguintes funcionalidades:
 - Repositório de algoritmos de análise de sinais elétricos;
 - Repositório de dados ligados ao sistema de potência;
 - Execução em servidor de agrupamento de rotinas, dados e parâmetros de configuração, gerando resultados atesteáveis;
 - Criação de rotinas de análises comparativas entre métodos que compartilham o mesmo propósito, possibilitando uma maneira alternativa de revisão de estado da arte e estabelecimento de *benchmark*.
2. Estudo de métodos de detecção de inovação no sinal do SEP através da plataforma;
3. Análise comparativa entre diferentes métodos de detecção de inovação através da plataforma.

1.3 ESTRUTURA DA DISSERTAÇÃO

Além deste, o presente trabalho possui outros quatro capítulos que são descritos a seguir.

O Capítulo 2 é destinado a definição conceitos e revisão bibliográfica de técnicas de detecção de inovação. Também será definido e contextualizado o tema pesquisa reprodutível e serão expostas ferramentas e plataformas que auxiliam a disseminação de pesquisas de forma mais transparente e rápida ao público alvo.

No capítulo 3, serão apresentadas as funcionalidades oferecidas pela plataforma desenvolvida juntamente com as motivações de suas implementações. Em seguida serão discutidas as tecnologias utilizadas para conceber a plataforma.

No capítulo 4, serão apresentadas duas técnicas de detecção de inovação, uma apoiada no domínio do tempo e outra no domínio da frequência, de forma a ilustrar a utilização da plataforma desenvolvida. Os resultados são também comparados através do uso da plataforma, mostrando a possibilidade de automatização na avaliação entre métodos por meio da ferramenta desenvolvida.

Por fim, o capítulo 5 apresenta as conclusões gerais do trabalho e propostas para a continuidade do trabalho.

2 PROCESSAMENTO DE SINAIS APLICADO AO SEP E PESQUISA REPRODUTÍVEL

A crescente utilização de técnicas processamento digital de sinais e inteligência computacional no SEP vem demandando cada vez mais a incorporação de procedimentos relacionados a pesquisas reprodutíveis na área. Este capítulo contextualiza este cenário e apresenta em maiores detalhes os conceitos e requisitos atuais da pesquisa reprodutível.

2.1 PROCESSAMENTO DE SINAIS APLICADOS AO SEP

A crescente complexidade da rede elétrica requer um monitoramento mais intenso e abrangente. Nesse contexto, a utilização de técnicas de processamento de sinais é necessária para caracterizar, identificar, diagnosticar, proteger e também para melhor entender a natureza de certos fenômenos e eventos. Processamento de Sinais também pode ser utilizado para prever e antecipar comportamentos do sistema, sendo portanto, uma ferramenta vital para separar, decompor e revelar diferentes aspectos e dimensões da operação de sistemas elétricos, nos quais fenômenos são intrínseca e intrincadamente agregados, não podendo ser resolvidos de maneira trivial (SILVA; DUQUE & RIBEIRO, 2015).

À medida que a rede elétrica se torna multifacetada, sua modelagem e simulação tornam-se menos capazes de captar a influência da multiplicidade de componentes independentes dentro da rede. Como as técnicas de processamento de sinais podem lidar com o sistema atual (sem abstração ou redução de modelagem, embora possa ser usado em conexão com simulações), é possível detectar e revelar a verdadeira natureza do comportamento da rede elétrica (SILVA; DUQUE & RIBEIRO, 2015).

O advento e implementação de novas tecnologias habilitaram o surgimento de **SGI**, com infra-estrutura composta por medidores, equipamentos e recursos inteligentes com uma combinação de tecnologias como controle, instrumentação e comunicação, conforme ilustrado na Figura 1. Em tal cenário, técnicas de processamento de sinais são essenciais para entender, planejar, proteger e operar as complexas redes elétricas inteligentes (JIANG et al., 2009). O estudo dos sinais elétricos do SEP leva a uma maior compreensão

das consequências da integração desses novos equipamentos e tecnologias com a rede elétrica. O monitoramento contínuo das características desses sistemas é necessário para capturar diversos eventos e variações. Para atender as demandas atuais e futuras, métodos e técnicas devem ser desenvolvidos para explorar toda a gama de sinais derivados da interação continuamente mais complexa entre fornecedores, consumidores e operadores da rede (LI et al., 2010).

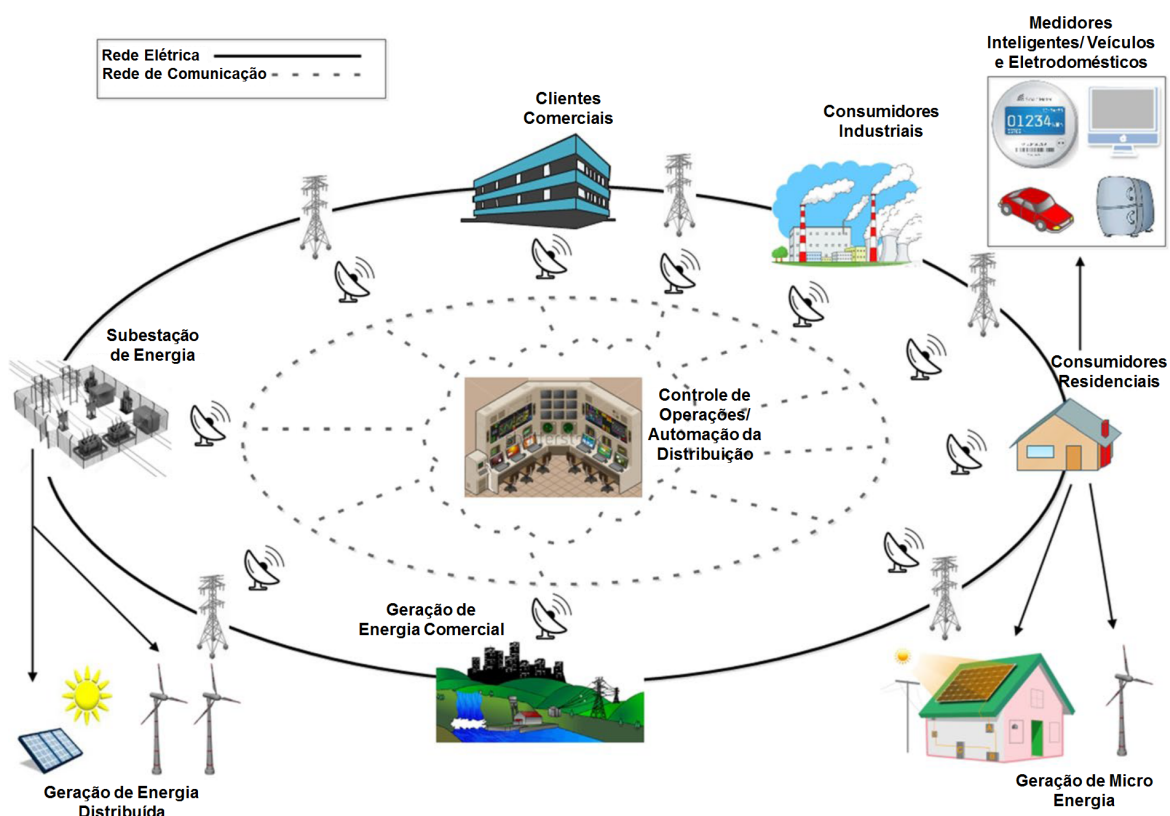


Figura 1: Arquitetura Smart Grid. Fonte: Adaptado de Uddin et al. (2018)

Uma das tecnologias base desse novo contexto é o crescente número de pontos de medição instalados na rede elétrica, o monitoramento contínuo de todos esses medidores impõem desafios quanto ao armazenamento e comunicação dessa massiva quantidade de dados gerada. Em Gu, Bollen e Le (2011) é proposto uma estrutura e diretrizes do sistema de monitoramento do SEP no novo contexto da **SGI**, dentre os quesitos desse novo sistema de monitoramento inteligente é a determinação do instante de ocorrência de anomalias no sistema.

2.2 PESQUISA COMPUTACIONAL E PESQUISA REPRODUTÍVEL

Com o advento da Revolução Técnico-Científica-Informacional, o uso de computadores foi integrado em praticamente todas as atividades desenvolvidas pelo ser humano.

Em destaque no campo científico, a tecnologia dedicada ao assunto deu início a uma nova metodologia de atuação, a pesquisa computacional. Tal ramo desempenha um essencial papel na ciência moderna, abrangendo totalmente ou parcialmente todos os campos do conhecimento atual.

O aparecimento dessa nova ferramenta possibilitou o surgimento ou aprimoramento de diversas áreas. Sua posterior popularização, possibilitou que avanços científicos pudessem ser realizados de forma mais ágil e por um número maior de pessoas, não se restringindo apenas a grandes faculdades e centros de pesquisas. Tal evento acarretou uma dinamicidade sem precedentes ao conhecimento científico, em contra partida, o modo ao qual as pesquisas são desenvolvidas e até mesmo a própria natureza da pesquisa científica computacional ocasionou uma crise de reprodutibilidade ou até mesmo de credibilidade (IOANNIDIS, 2005).

Em tempos de pre-informatização, a metodologia científica podia ser dividida em dois campos: metodologia descritiva e metodologia empírica. A metodologia descritiva, também chamada de dedutiva, abrange áreas como a matemática, a qual busca formular regras gerais para explicar relações específicas a partir de hipóteses já existentes. Uma publicação relacionada a pesquisa dedutiva consiste na própria dedução de novos teoremas, mostrando todos os passos necessários para atingir tal resultado.

Já a pesquisa empírica, que nesse caso pode-se citar a física experimental, segue o caminho inverso da investigação dedutiva, que após considerar um número suficiente de casos particulares, conclui uma verdade geral (COPI; COHEN & FLAGE, 2016). Por sua vez, publicações neste meio incluem detalhada descrição dos experimentos realizados para basear as hipóteses gerais, garantindo uma certa transparência à metodologia. Naturalmente, manter transparência não garante qualidade, exatidão nem validade dos resultados (PENG, 2011), mesmo ramos muito disciplinados e consolidados da ciência sofrem notavelmente com o problema de erros em conclusões finais publicadas (IOANNIDIS, 2005). Tal problema é agravado no âmbito da informática, onde as atuais práticas são relaxadas em comparação a outros campos de pesquisa (DONOHO et al., 2009).

Ciências computacionais, como ciência da computação, estatística, processamento de sinais e alguns ramos da engenharia começaram a deixar sua marca na segunda metade do século passado. Elas formam um híbrido entre o dedutível e empírico descritos acima: desenvolvimentos teóricos, como novas teorias, teoremas e simulações, são combinados com experimentos práticos e desenvolvimento de novos algoritmos. Enquanto essa integração pode resultar em uma combinação do melhor dos dois lados (provas rigo-

rosas para os teoremas e experimentos que são corretamente configurados e descritos), em algumas situações pode resultar em uma combinação pobre dos dois. Muitas vezes depara-se com rigor reduzido nas provas com relação ao que um matemático consideraria aceitável e descreve-se configurações experimentais menos precisas em relação ao ensaio de cientista experimental (VANDEWALLE; KOVACEVIC & VETTERLI, 2009). Levando isso em consideração, a pesquisa computacional se difere dos métodos anteriores, e apesar de ser considerada por muitos como o terceiro campo da metodologia científica, este ainda não merece esta elevação, pois as atuais práticas científicas não geram rotineiramente um conhecimento verificável (DONOHO et al., 2009).

A Figura 2 ilustra o ciclo Hipotético-Empírico ideal. Ele descreve, em linhas gerais, o processo de evolução do conhecimento em azul e as possíveis práticas que deterioram sua evolução em vermelho. Tal ciclo apresenta o panorama geral, sendo que os erros em vermelho aprestam efeitos negativos em maior ou menor intensidade dependendo do campo de pesquisa. Por exemplo o **HARKing**, que em linhas gerais consiste em apresentar hipóteses depois dos resultados serem conhecidos como se as mesmas fossem propostas a priori, não atribui grandes malefícios em algumas áreas, mas podem ser prejudiciais em outras, como na psicologia (KERR, 1998), enquanto **p-hacking** (SIMMONS; NELSON & SIMONSOHN, 2011), referente ao grau de liberdade do pesquisador na elaboração da pesquisa, preprocessamento dos dados e escolha de métricas de avaliação entre outras inevitáveis práticas, pode atingir de forma negativa todas as áreas.

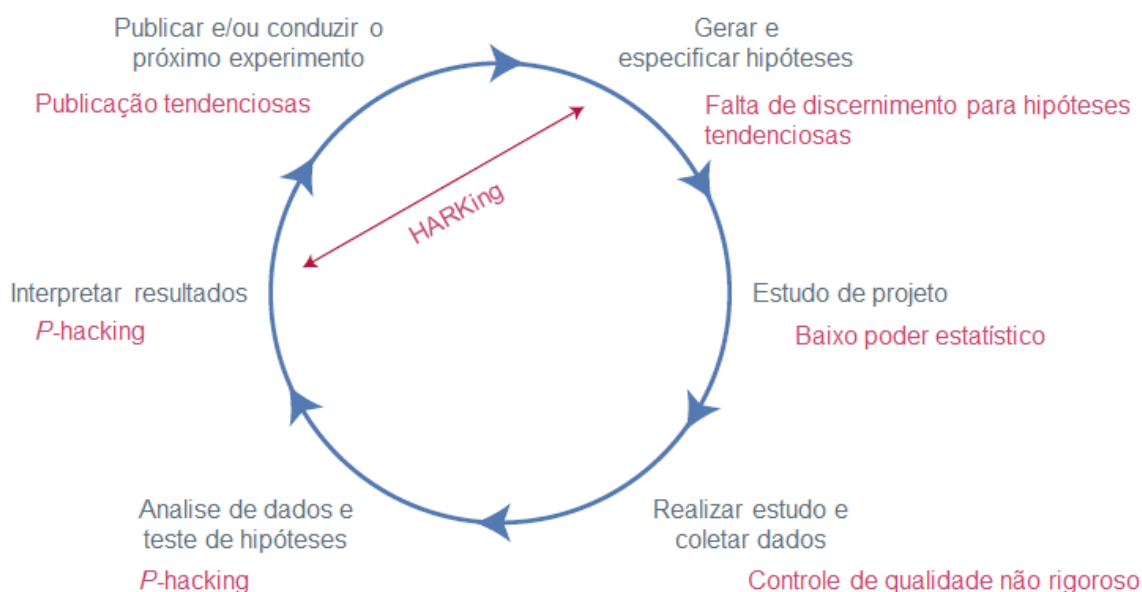


Figura 2: Ciclo Hipotético-Empírico. Fonte: Adaptado de Chambers (2014)

Como prevenção a essas vulnerabilidades do desenvolvimento científico, a reprodução

completa de estudos através de métodos, dados, equipamentos e protocolos independentes é o padrão no qual novas descobertas são avaliadas (PENG & ECKEL, 2009). No entanto, em muitos campos, reprodução completa não é praticável por falta de tempo, recursos ou até mesmo interesse. Em tais situações, há a necessidade do estabelecimento de um padrão mínimo que possa ser praticado a fim de garantir validação e utilização de pesquisas por terceiros. Um candidato para este padrão é a ideia da RR, que requer que se faça disponíveis elementos essenciais para a realização da pesquisa, como códigos fontes e conjuntos de dados usados na publicação.

Vários importantes jornais e revistas vêm clamando por mais transparência na ciência. No encontro convocado pelo Instituto Nacional de Saúde dos Estados Unidos, com mais de 30 importantes jornais, como os periódicos *Nature* e *Science* e representantes de órgãos financiadores (MCNUTT, 2014), foram discutidas medidas em prol da reprodutibilidade. O assunto é recorrente e o resultado de uma enquete realizada em 2016 pela revista *Nature* (BAKER, 2016) abrangendo mais de 1500 pesquisadores de diversos campos do conhecimento, mostra que a crise é real. O questionário apontou que 52% dos entrevistados acham que existe uma séria crise e 38% uma crise fraca, contra somente 3% se posicionando contra a existência e os 7% restantes não possuindo opinião formada sobre o assunto. A mesma pesquisa mostrou ainda que mais de 70% dos pesquisadores falharam em reproduzir trabalhos publicados por terceiros e mais da metade já falhou em reproduzir os próprios resultados.

É importante apontar que a divulgação de uma pesquisa de forma reprodutível não garante relevância, exatidão ou qualidade, do mesmo modo uma pesquisa irreproduzível não significa que os resultados são fraudulentos nem que seus métodos sejam inválidos (KENALL et al., 2015), porém a evolução da ciência e um contínuo processo de auto-melhoramento e auto-correção ficam comprometidos com a impossibilidade de reprodução de pesquisas.

Tanto na indústria como na academia, existem diversas pesquisas que desenvolvem produtos com a intenção de serem totalmente ou parcialmente restritos ao público, por serem passíveis de monetarização, futura patente ou usarem bancos de dados com informações sensíveis. O conceito de RR vai de encontro a essas práticas, portanto os argumentos apresentados nesse capítulo não contemplam esses tipos de pesquisas, garantindo, todavia, a parcela de pesquisa que é disseminada pela comunidade científica.

Uma publicação científica tem como propósito primordial expor os avanços alcançados em determinada área. Um pesquisador ao desenvolver algo significativo em seu

campo, tem a intenção de mostrar seus achados da maneira mais eficiente possível, almejando a adoção de suas técnicas pelo maior número de instituições e pessoas possíveis, criando assim uma publicação de impacto, que beneficia a sociedade. De certa forma, a disseminação do conhecimento é tão importante quanto a própria pesquisa. RR tem como objetivo mudar o paradigma atual de como a pesquisa atual é exposta ao público.

O tradicional modelo de publicação é exemplificado na Figura 3. Nesse caso, o leitor só tem acesso aos resultados obtidos, a compreensão do método em si, fica dependente de outros fatores, como a capacidade do autor de descrever e sintetizar sua pesquisa de forma a não exceder do tamanho limitado de um artigo, e também ao fato que a linguagem natural é passível de ambiguidade, ao contrário de linguagens de programação.

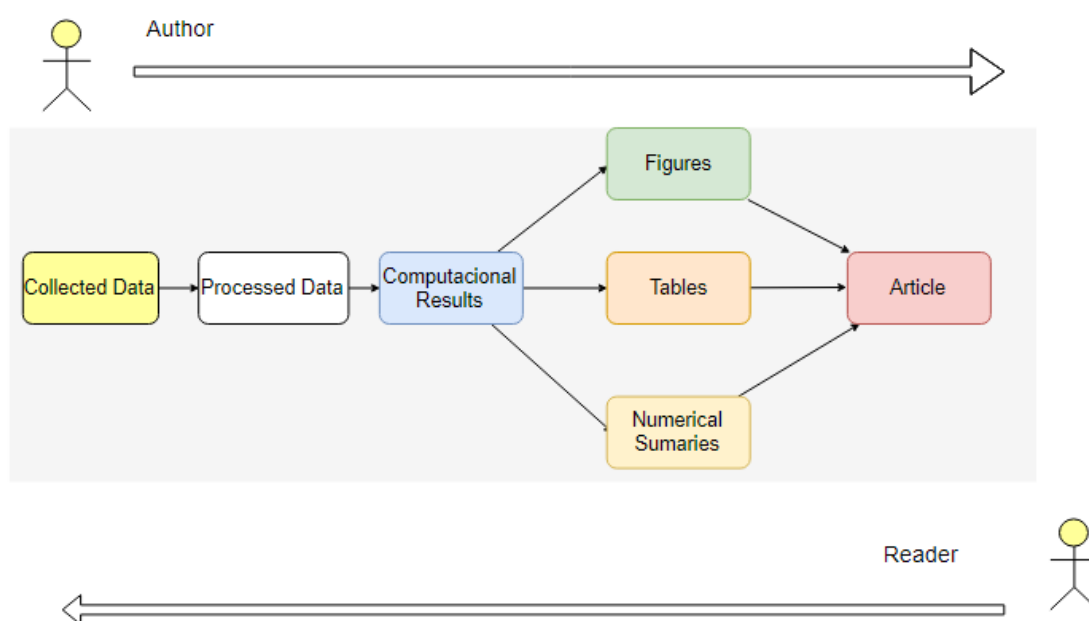


Figura 3: Diagrama do tradicional modelo de publicação. Fonte: Adaptado de Peng e Eckel (2009)

Porém, como aludido por Buckheit e Donoho (1995), um artigo sobre pesquisa computacional não é a pesquisa em si, é apenas uma propaganda para a pesquisa. A real pesquisa é o completo ambiente virtual e códigos que geraram as figuras e os resultados expostos. Logo, tal modelo de comunicação entre autor e leitor muitas vezes se torna ineficiente, isso devido ao fato de que as análises realizadas são cada vez mais complexas, a descrição detalhada de todos os passos necessários para compreensão completa muitas vezes excedem o limite de páginas permitido pelos jornais e revistas, além de tornar a escrita e a leitura massante, mais descritiva do que informativa, não atendendo de forma completa nem o autor nem o leitor.

A alternativa lógica é compartilhar de forma paralela ao artigo todos os elementos da pesquisa, mantendo a carga informativa da publicação e possibilitando entendimento aprofundado, facilidade de reutilização e possíveis melhorias com base na sua pesquisa de forma mais rápida e fácil.

Apesar do incentivo da maioria dos jornais e revistas por pesquisas reproduzíveis, a maioria ainda não exige tal prática como requisito obrigatório. Tal exigência nem seria justa, uma vez que não existe uma ferramenta consolidada para realizar essa função. O campo atual da RR é bem heterogêneo, algumas áreas com grandes repositórios como Mendeley¹, criado pela editora *Elsevier*, inicialmente focado em dados da biomedicina, e outras com praticamente nenhuma ferramenta específica.

A classificação de uma publicação não é binária em relação sua reprodutibilidade, existem várias possibilidades entre disponibilizar somente a publicação e a reprodução completa, conforme ilustrado na Figura 4, e a apresentação de uma pesquisa parcialmente reproduzível já é considerada melhor que uma irreproduzível.

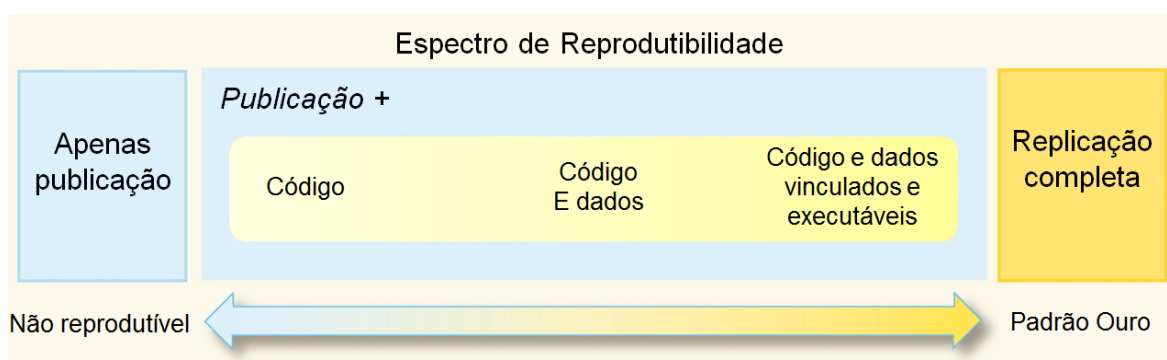


Figura 4: Espectro da reprodutibilidade. Fonte: Adaptado de Peng (2011)

Vandewalle, Kovacevic e Vetterli (2009) definiram uma escala para quantificar o nível de reprodutibilidade de uma pesquisa, que pode ser observada pela Tabela 1. Essa classificação ratifica a sinergia existente entre a utilização de ferramentas de código livre e aberto e a RR, cuja classificação de reprodutibilidade máxima exige o emprego de tais tecnologias. De forma geral, a divulgação de uma pesquisa científica por meios habituais receberia classificação 3 na hipótese mais otimista. Conforme pode ser notado, através da escala, essa categorização exige um esforço considerável para sua reprodução, o que acarreta em um desperdício de tempo útil dos pesquisadores.

Outra classificação quanto aos aspectos de reprodutibilidade foi definida por Freire, Bonnet e Shasha (2012). Considerando a importância da reprodução e arquivamento de

¹<https://data.mendeley.com/>

Tabela 1: Níveis de reprodutibilidade.

Nível	Requisitos
5	Os resultados podem ser facilmente reproduzidos por um pesquisador independente com no máximo 15 minutos de esforço do usuário, exigindo apenas ferramentas padronizadas e gratuitamente disponíveis (compilador C, interpretador Python, etc.)
4	Os resultados podem ser facilmente reproduzidos por um pesquisador independente com no máximo 15 minutos de esforço do usuário, exigindo alguma ferramenta proprietária (MATLAB, etc.)
3	Os resultados podem ser reproduzidos por um pesquisador independente, exigindo um esforço considerável.
2	Os resultados poderiam ser reproduzidos por um pesquisador independente, exigindo um esforço extremo.
1	Aparentemente os resultados não podem ser reproduzidos por um pesquisador independente.
0	Os resultados não podem ser reproduzidos por um pesquisador independente

experimentos computacionais, bem como a falta de plataformas que encapsulem todos os componentes necessários para reproduzir um resultado e que os revisores verifiquem os resultados. Os autores categorizaram a pesquisa computacional em três perspectivas com relação ao nível de reprodutibilidade:

1. **Profundidade:** Referente ao quanto da pesquisa é disponibilizada. Variando do padrão mínimo atual de inclusão de figuras e tabelas ao manuscrito. Profundidades mais altas podem ser obtidas incluindo: código fonte, banco de dados, dados brutos e resultados intermediários, características do experimento (configuração do sistema, protocolo de medição) que produziram os dados brutos e software, quer seja como caixa branca (fonte, arquivos de configuração, ambiente de compilação) ou caixa preta (executável).
2. **Portabilidade:** O nível de portabilidade indica se o experimento pode ser reproduzido:
 - (a) Somente no ambiente original (basicamente sem portabilidade, os experimentos só são executados com sucesso pelo autor em sua máquina);
 - (b) Em ambiente semelhante ao original (e.g.: mesmo sistema operacional, mesma versão de softwares);
 - (c) Ambiente diferente (e.g.: sistemas operacionais diferentes).
3. **Cobertura:** Reflete se o experimento pode ser parcialmente ou totalmente reproduzido (e.g.: experimentos que requerem hardwares especiais para derivar dados, somente reprodução parcial dos resultados pode ser alcançada através da disponibilização dos dados gerados pelo hardware em associação com demais elementos da pesquisa).

2.3 PESQUISA REPRODUTÍVEL HOJE

Com a crescente preocupação e necessidade de desenvolver pesquisas com maior grau de reutilização e confiabilidade, reconhecendo as limitações atuais atreladas a disseminação de pesquisas computacionais, o conceito de RR vem ganhando atenção na comunidade científica como solução para os atuais desafios. Através dos anos surgiram ferramentas e técnicas a fim de ajudar a comunidade a realizar seus estudos de maneira mais transparente e com maior acessibilidade a terceiros. A seguir serão apresentados alguns dos avanços realizados neste contexto. Esta subseção será dividida em duas partes, ferramentas e plataformas. A primeira se refere a recursos que podem ser usados para auxiliar a apresentação de resultados de forma mais transparente, mas que por si só, não se configura como reprodutibilidade. Já a segunda se refere a ambientes onde é possível apresentar resultados mais completos, com funcionalidades específicas para RR.

2.3.1 FERRAMENTAS

KNITR

Na perspectiva de ciência da computação, pesquisa reproduzível é frequentemente relacionada à programação documentada (XIE, 2014), um antigo paradigma concebido por Knuth (1984), que consiste em combinar a código e documentação do software no mesmo documento. Knitr é um pacote *open-source*, de código aberto, para a linguagem R, porém pode ser integrado com outras linguagens também como Python. Knitr é a evolução do pacote Sweave (LEISCH, 2002), que combina a documentação na linguagem \LaTeX com a linguagem R. Knitr é extensível e possui maior flexibilidade, também faz uso das linguagens \LaTeX , Linguagem de Marcação de Hipertexto, do inglês HyperText Markup Language (HTML) e *Markdown* para documentação. O resultado final do uso de tal pacote é um documento em formato legível (i.e. PDF) ou arquivos de códigos, onde todos os gráficos tabelas e resultados são gerados dinamicamente.

NOTEBOOKS

Os notebooks são projetados para suportar o fluxo de trabalho da computação científica, da exploração interativa à publicação de um registro detalhado de computação. O código em um Notebook é organizado em células, partes de códigos que podem ser modificadas e executadas individualmente. A interface de Notebook primeiramente se tornou popular entre os matemáticos. Através de programas como Mathematica, Maple e Sa-

geMath (KLUYVER et al., 2016). Atualmente a principal interface de *Notebook* é o Jupyter Notebook, projeto que evoluiu do IPython (PÉREZ & GRANGER, 2007) e agora agrega mais de 50 linguagens.

Um dos mais notáveis exemplos de como Jupyter Notebook pode ser usado para realizar reprodução em publicação é o trabalho apresentado por Abbott et al. (2016), em que pesquisadores postaram um Notebook contendo detalhadamente todos os processos de filtragem e processamento de dados de ondas gravitacionais que registraram a fusão de dois buracos negros.

VISTRAILS

VisTrails é um programa de código aberto de gestão de versão baseado na criação *workflows* que foi projetado para auxiliar o processo de descoberta científica (FREIRE et al., 2008). VisTrails captura todas as entradas e saídas usadas na pesquisa, bem como as versões dos códigos, módulos utilizados, dados intermediários e alteração de parâmetros. Acompanhar a evolução do *workflow* é um elemento-chave na manutenção de resultados reproduzíveis, capturando as mudanças que são feitas a partir de explorações iniciais, testes e extensões para os resultados finais publicados. Como nenhuma versão é excluída ou substituída, todos os resultados podem ser recuperados, reproduzidos e comparados. Ao mesmo tempo, saber como uma computação foi construída pode ajudar os outros a entender o processo, estender os resultados ou ajustá-los de maneira significativa (FREIRE et al., 2014).

2.3.2 PLATAFORMAS

RUNMYCODE.ORG

RunMyCode.org (STODDEN; HURLIN & PÉRIGNON, 2012) é um web site que fornece a capacidade de compartilhamento de pesquisas multi-disciplinar. De modo simplificado, o site consiste em proporcionar aos usuários a funcionalidade de rapidamente criar os chamados *companion website*, que consiste em uma página web da publicação. Em tal página é possível realizar o download de códigos e dados utilizados em uma publicação, além disso, a página apresenta exibição do artigo ou link para tal, links para demais *companion website* de mesmo autores e publicações relacionadas, constituindo uma interessante ferramenta didática de revisão teórica. Apesar de publicações antigas sobre a plataforma (STODDEN; HURLIN & PÉRIGNON, 2012) e (HURLIN et al., 2014) apresentam a funcionalidade de execução na nuvem, uma avaliação das funcionalidades atuais do web site

mostrou que esta não é mais suportada.

ACERVOS ONLINE DE PESQUISAS

O Jornal de Pesquisas em Aprendizado de Máquinas, do inglês *Journal of Machine Learning Research* (JMLR) é um periódico eletrônico no campo de inteligência artificial, que não tem apenas acesso livre e gratuito para os leitores, mas também oferece gratuidade para os autores que publicam no veículo de informação. A missão do periódico é eliminar barreiras sócio-econômicas, na divulgação científica em sua área de atuação.

O JMLR também compartilha da convicção que somente publicações no campo computacional, não conseguem cumprir o propósito de transmitir de forma rápida e eficiente todos os aspectos necessários para incorporação de novas técnicas pelo público leitor. Dessa forma o jornal oferece armazenamento online de dados e códigos, que ficam disponíveis para *download* juntamente com a publicação.

Mendeley é uma aplicação de *desktop* e web com objetivo inicial de gerenciar documentos acadêmicos entre diferentes dispositivos e realizar referência de forma fácil e simples, que se expandiu para uma espécie de rede social científica agregando funcionalidades como compartilhamento de listas de leituras, grupos privados para escrita colaborativa, newsfeeds e mais importante para nosso contexto, armazenamento e compartilhamento de dados de pesquisas, banco de dados armazenados no site possuem um Digital Object Identifier(DOI), possibilitando referência do banco, e facilitando creditar os autores dos dados.

Similarmente aos exemplos anteriores a plataforma GigaDB é o repositório do Jornal GigaScience que atua nas áreas de biologia e biomedicina. O armazém virtual pode aceitar arquivos de dados, arquivos suportes, como até mesmo softwares, sem restrições de tamanho, abrigando inclusive dados de carcinoma hepatocelular compreendido em 15 Terabytes (SNEDDON; LI & EDMUNDS, 2012). O site pode aceitar hospedar dados de outros campos do conhecimento e de outros meios de veiculação de pesquisas mediante aprovação, levando em consideração que nem todos os veículos fornecem tal capacidade. Os conjuntos de dados anexados ao repositório também são indexados com a DOI único para referência em outras publicações.

CODE OCEAN

Code Ocean é uma start-up localizada em Nova Iorque e incubada no Jacobs Technion Cornell Institute, compreende um ambiente onde é possível armazenar a código dados encapsulada em um ambiente virtual que pode ser pertencente a 11 diferente linguagens, incluindo Python, R, Matlab e C++. Terceiros podem realizar o *download* dos arquivos, modificá-los ou re-executar na próprio site (PERKEL, 2018). Code Ocean oferece modalidade gratuita com tempo de execução e espaço limitados, uma hora por mês e cinco gigabytes respectivamente, e diferentes planos pagos, e ainda condições especiais para acadêmicos.

As capacidades da plataforma colaboram para a disseminação do conhecimento de forma simples e transparente de tal maneira que o Instituto de Engenheiros Eletricistas e Eletrônicos, do inglês *Institute of Electrical and Electronics Engineers* (IEEE) adotou a mesma como ferramenta de compartilhamento de pesquisas no portal de submissão de artigos de suas conferências.

BEAT PLATAFORM

Plataforma BEAT Platform (Plataforma de avaliação e teste biométricos, do inglês Biometrics Evaluation and Testing) (MARCEL, 2013), é uma plataforma desenvolvida pelo centro de pesquisa suíço Idiap e talvez seja o melhor exemplo de como pesquisas reprodutíveis devem ser realizadas. Com o intuito não somente de possibilitar um ambiente de pesquisa reprodutível no campo de inteligência artificial aplicada a segurança biométrica, mas concomitantemente fomentar desenvolvimentos de inovações e estabelecimentos de *benchmark*. Tal plataforma apresenta algumas características singulares em relação as apresentadas anteriormente.

Ela agrega mais de 60 bancos de dados consolidados, abrangendo diversos problemas da área, variando de reconhecimento de faces a identificação de identidade por padrão de digitação. A execução de algoritmos é realizada no servidor, apresentado resultados confiáveis e eliminação de problemas de compatibilidade entre maquinas e sistemas operacionais bem como a necessidade de instalação e download de softwares módulos e arquivos. Uma vez que os bancos de dados são incorporados pelos desenvolvedores da plataforma e a maioria deles, se não todos, possuem uma ou mais métricas obvias de avaliação, a comparação entre diversos métodos que compartilham o objetivo de mitigar o mesmo problema pode ser comparado de forma automática, se apoiando nessa capacidade da plataforma realizar competições entre pesquisadores, apresentando um

panorama da eficácia de diferenças abordagens.

Paralelamente, experimentos incorporados na plataforma são avaliados não somente se foram executados sem erros, mas são validados por revisores externos. Tal ferramenta, em conjunto com compartilhamento seletivo, o usuário pode conceder permissão de visualização para usuários específicos, oferece um recurso para técnicas serem atestadas pré-publicação por revisores de congressos e revistas.

2.4 DETECÇÃO DE NOVIDADE NO SINAL DO SEP

Para ilustrar a aplicação da plataforma desenvolvida nesse trabalho, algumas abordagens para detecção de novidade no sinal do SEP serão implementados e avaliados no Capítulo 4, configurando um dos objetivos específicos desse trabalho. Dessa forma, torna-se necessário apresentar o cenário atual sobre detecção de distúrbios e de novidades para o sinais do SEP e detalhar os dois métodos que serão comparados no trabalho.

O termo Detecção de Novidade, do inglês *Novelty Detection*, pode ser definido como a tarefa de se detectar que um determinado dado de teste é diferente dos dados disponíveis durante o treinamento. Métodos de detecção de novidade são geralmente empregados em situações onde muitas amostras de dados ditos normais estão disponíveis para o treinamento enquanto só existem poucas amostras disponíveis dos dados ditos anormais (PIMENTEL et al., 2014). O sistema então é treinado com esse conjunto de dados normais e aplicados para reconhecer os dados que são diferentes ou anormais.

A detecção de anormalidades é um importante passo dentro do contexto de processamento de sinais elétricos, pois de certa forma ela habilita o desenvolvimento de diversas áreas, uma vez que a mineração de eventos de interesse é necessária para os demais seguimentos. Abundantes estudos foram e são desenvolvidos na área de detecção de distúrbios elétricos. A identificação de variações na rede pode ser baseada na detecção de distúrbios elétricos ou pela detecção de novidade propriamente dita.

Em Meena, Rao e Ravishankar (2009) é proposto o uso de valores Raiz do Valor Quadrático Médio, do inglês *Root Mean Square* (RMS) para detecção de distúrbios com atualização de medição a cada amostra, para reduzido atraso de detecção. No entanto a abordagem é dedicada a identificação de distúrbios de níveis de tensão, como Sag Swell e interrupções. Se mostrando ineficiente para acusação de demais distúrbios. O uso do valor RMS é também explorado em (GONZALEZ; CARDENAS & ALVAREZ, 2006), aliado ao uso de filtro de Kalman para rápido diagnóstico da rede.

O uso de transformadas, i.e transformada S, Wavelet, Transformada de Fourier, do inglês *Fourier Transform* (FT), Hilbert Huang (HH), como técnica de pré-processamento de extração de características é largamente usada para essa aplicação. FT é uma das mais famosas técnicas de análise no domínio da frequência. Os autores Santos et al. (2000) usaram os coeficientes da FT como entrada do sistema especialista baseado em regras para a identificação de eventos. Em Heydt et al. (1999) uma FT janelada é usada para identificar segmentos de interesse dentre grandes registros temporais de sinais elétricos.

Um estudo comparativo entre *Wavelet Transform* (WT) e Transformada de Fourier de curta duração, do inglês *Short-Time Fourier Transform* (STFT) foi realizado por Gu e Bollen (2000) e apontou que para localização temporal de ocorrência de distúrbios pelo uso de WT apresenta resultados mais satisfatórios. WT apresenta representação local tanto no domínio do tempo, quanto no domínio da frequência, sendo mais adequada para análises de distúrbios transitórios do SEP (SANTOSO et al., 1996). Métodos baseados em WT para detecção de distúrbios podem ser observados em Gaouda et al. (1999), onde os autores usam WT e decomposição de sinais com múltiplas resolução. Em Xiangxun (2002) parâmetros de uma *Wavelet* bi-ortogonal complexa e parâmetros do domínio do tempo são usados para identificação de distúrbios de curta duração.

Em Lin e Tsao (2005), uma arquitetura em duas camadas, sendo a primeira transformação *Wavelet* e a segunda uma rede neural probabilística é apresentada. Apesar das grandes capacidades das transformadas *Wavelets*, seu desempenho pode ser degradado em cenários de baixos níveis de Relação Sinal Ruído, do inglês *Signal to Noise Ratio* (SNR) e em alta presença de componentes harmônicas (DUQUE et al., 2005). Em Yang e Liao (2001) e Dwivedi e Singh (2009) são usados artifícios de *de-noising* para mitigar a limitação da transformada em cenários ruidosos.

As abordagens mais clássicas apresentadas anteriormente oferecem bons resultados quanto ao monitoramento de distúrbios clássicos de qualidade de energia. Entretanto, sistemas de monitoramento baseados na detecção de novidades são mais adequados para o contexto atual de Redes Inteligentes. Tais sistemas de monitoramento devem identificar quaisquer mudanças de comportamento da rede, uma vez que as distorções já categorizadas podem não compreender todos os comportamentos atuais e futuros.

O monitoramento obedecendo esse conceito de inovação pode ser visto em Duque et al. (2005), onde o uso do princípio "dividir para conquistar" é usado para dividir o problema em subproblemas mais simples, neste caso, dividir o evento em sinais estacionários e não estacionários. Após a estimação de parâmetros do sinal como amplitude,

frequência e fase da componente fundamental, o conceito de novidade é usado para identificar variações entre janelas subsequentes do sinal.

Em Silva et al. (2016) é definido o conceito de detecção de novidades contextualizado no panorama das Redes Inteligentes, do inglês *Smart Grids* (SGs), os mesmos autores apresentaram em outros trabalhos (KAPISCH et al., 2014; SILVA et al., 2017) a idealização e implementação embarcada em plataforma Arranjo de Portas Programáveis em Campo, do inglês *Field Programmable Gate Array* (FPGA) de um sistema de monitoramento capaz de armazenar dados com alta taxa de amostragem por um longo período de tempo, através de uma taxa elevada de compressão de dados. A eficácia do sistema é atribuída, em grande parte, à gravação oscilográfica apenas de trechos que possuem novas informações, sendo o restante descomprimido com base em parâmetros atuais e informações passadas. Nesses trabalhos foi realizado ainda um estudo comparativo entre diferentes métricas de comparação entre trechos de sinal, como Morfologia Matemática (MM), energia da diferença ponto a ponto e diferença da energia dos trechos.

O reconhecimento de novidade pode ser realizado através do uso de Descrição de Dados por Vetores de Suporte, do inglês *Support Vector Data Description* (SVDD) com otimização de margem (LAZZARETTI; FERREIRA & NETO, 2016). Nesse caso, a performance do sistema é comparada com demais técnicas de inteligência artificial: SVDD clássico, clusterização *k-Means* e Estimador de Densidade Parzen, do inglês *Parzen Density Estimator* (PDE). Em Ribeiro et al. (2018), a detecção é realizada de maneira computacionalmente econômica através do valor quadrático da distância euclidiana de um vetor de alta dimensão, no qual cada dimensão corresponde a uma amostra do sinal. Em Alshahrani, Abbod e Taylor (2016) é usada transformada HH em conjunto com Redes Neurais, do inglês *Neural Network* (NN) para detecção e futura classificação de distúrbios. Em Chakravorti, Patnaik e Dash (2017) a etapa de detecção é realizada através filtro de media morfológica para monitoramento de uma micro-rede (arquitetura factível no contexto de SGs).

De e Debnath (2017) usaram os valores de correlação entre as medições aferidas e um sinal senoidal puro para acusar desvios na tensão. Tal método se mostrou efetivo tanto para sinais simulados quanto para reais em variados níveis de ruído. Em Nagata et al. (2018) o uso Análise de componente independente, do inglês *Independent Component Analysis* (ICA) de canal único é usado para separar componentes do sinal, se o valor de determinadas componentes excederem um limiar a variação é acusada. Já Monteiro, Zvietcovich e Braga (2018) determinou uma árvore de decisão cujo critério de decisão de cada nó é realizado por uma Máquina de Vetor Suporte, do inglês *Support Vector Ma-*

chine (SVM). Por sua vez os autores Rupal et al. (2017) utilizam decomposição de modo empírico para avaliação de novas informações no sinal.

2.4.1 TÉCNICAS PROMISSORAS QUE SERÃO AVALIADAS

Detecção de inovação no sinal do SEP é um assunto ainda atual na área como mostram as recentes publicações no tema, atraindo a atenção de pesquisadores da área.

As abordagens de detecção de novidade exploradas nesse trabalho são baseadas na identificação de inovação em um determinado segmento do sinal. O processo de segmentação de um sinal consiste em dividir o sinal em janelas consecutivas sem que haja sobreposição. O tamanho de cada segmento, ou janela, pode ser fixo ou variável (adaptativo), dependendo do comportamento do sinal e da técnica de processamento. É extraído então alguma característica do trecho do sinal compreendido na janela e essa característica é comparada com a característica da janela de referência. Se as janelas forem diferentes de acordo com alguma métrica, a inovação é acusada e a janela atual passa a ser a nova janela de referência.

Segmentação adaptativa é largamente utilizada no processamento de sinais biomédicos e áudio (AZAMI et al., 2015). O principal objetivo é determinar o comprimento do segmento em que o sinal possa ser representado pelo mesmo conjunto de parâmetros. Um conceito similar é introduzido por Bollen e Gu (2006) para analisar distúrbios em sistemas de potência.

Um estudo comparativo entre diferentes técnicas de detecção será realizado. As técnicas analisadas podem ser divididas em análises baseadas no domínio do tempo e baseadas no domínio da frequência. As abordagens referentes ao domínio do tempo estão descritas em Silva et al. (2016). No trabalho em questão foi desenvolvido um sistema de monitoramento capaz de monitorar e armazenar sinais do SEP por um grande período de tempo devido à uma alta taxa de compressão de dados obtidas em três etapas. Sendo a mais significativa baseada na identificação de novidade.

Os resultados obtidos nesse trabalho são comparados com a solução comercial Elspec G4000 BLACKBOX, que além de ser um analisador de Qualidade de Energia Classe A, é capaz de armazenar dados oscilográficos de maneira contínua, a uma taxa de amostragem de 1024 amostras por ciclo para os canais de tensão, e 256 amostras por ciclo para os canais de corrente, durante um ano em sua memória interna de 8 GB (ELSPEC, 2007). Apesar do funcionamento interno do medidor não ser conhecido em detalhes, sabe-se que a etapa de detecção de inovação é baseada na Transformada de Fourier Rápida, do

inglês *Fast Fourier Transform* (FFT) (COOLEY & TUKEY, 1965). Portanto, neste trabalho, também serão avaliados métodos de detecção de inovação no sinal do SEP baseado na FFT através da plataforma de pesquisa reprodutível.

MÉTODO BASEADO NO DOMÍNIO DO TEMPO (SILVA, 2016)

A fluxograma expresso na Figura 5 exemplifica o funcionamento do método de detecção de novidade baseado no domínio tempo. O processo de filtragem aplicada antes do cálculo da métrica de comparação, pode ou não ser empregado. Quando ao cálculo da métrica, duas variantes foram adotadas a Diferença de Energia Entre Frames (DEF) e a Energia do Resíduo (ER).

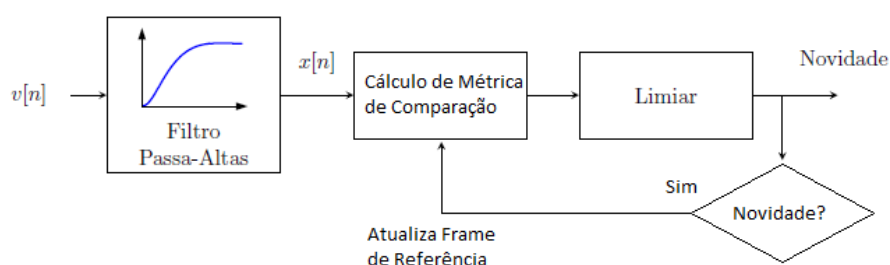


Figura 5: Fluxograma de detecção de novidade baseada no domínio do tempo.

Onde $v[n]$ é referente a n -ésima amostra da medição do sinal, e $x[n]$ é a n -ésima amostra do sinal filtrado. Quando o processo de filtragem não é utilizado, $x[n] = v[n]$. O cálculo da DEF ou ER é realizado, e comparado com o valor obtido pelo *frame* de referência. Se os valores em questão divergirem mais do que o estabelecido, o *flag* de inovação é disparado.

Como o conceito de inovação pretende acusar somente frames que possuem novas informações. O *frame* de referência é atualizado para o último *frame* onde ocorreu uma novidade. Para sistemas com o propósito de detecção de distúrbios, o *frame* de referência deverá conter informações do comportamento normal do sistema, e não deverá ser atualizado a cada detecção.

A DEF é calculada como o valor absoluto da subtração da energia da janela atual pela energia da janela de referência, como mostrado em (2.1).

$$DEF = |EF_A - EF_R| \quad (2.1)$$

Onde a energia da janela é calculada por (2.2).

$$EF = \sum_{n=0}^{M-1} |x[n]|^2 \quad (2.2)$$

em que, $x[n]$ é a n -ésima amostra do frame, e M é o número de amostras contidas em um frame.

O resíduo consiste no sinal resultante da subtração, amostra a amostra, entre a janela atual e janela de referência. A ER então, é calculada como mostrado em 2.2.

$$ER = \sum_{n=0}^{M-1} |x_A[n] - x_R[n]|^2 \quad (2.3)$$

em que, $x_A[n]$ é a n -ésima amostra da janela atual e $x_R[n]$ é a n -ésima amostra da janela de referência.

Se a DEF ou ER for superior a um limiar λ a janela contém uma inovação, conforme ilustrado em (2.4).

$$FLAG_{\text{inovação}} = \begin{cases} 1, & \text{Métrica} \geq \lambda \\ 0, & \text{Métrica} < \lambda \end{cases} \quad (2.4)$$

MÉTODO BASEADO NO DOMÍNIO DA FREQUÊNCIA (HEYDT ET AL., 1999;ELSPEC, 2007)

Variações temporais ocorridas no domínio do tempo são refletidas também no domínio da frequência e a estimação das componentes em frequência podem ser obtidas de maneira computacionalmente eficiente pelo algoritmo FFT. O fluxograma da identificação da inovação baseadas em frequência, se difere ligeiramente do que apresentado anteriormente e está exposta na Figura 6.

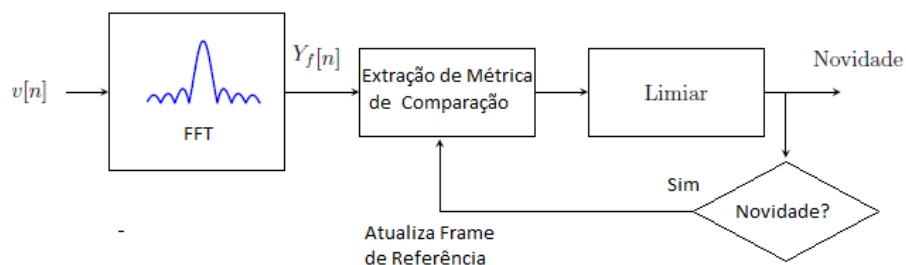


Figura 6: Fluxograma de detecção de novidade baseada no domínio da frequência.

Onde $v[n]$ continua representando a n -ésima amostra do sinal e $Y_f[n]$ corresponde a componente complexa no domínio discreto da frequência. Diferentes características

serão extraídas para serem usadas na comparação entre *frames* de acordo com as considerações realizadas a seguir. Quanto a comparação dos valores obtidos com um limiar e a atualização do *frame* de referência, as mesmas prerrogativas feitas para o domínio do tempo são válidas.

A detecção da novidade pode ser realizada através da Variação Máxima na Magnitude das Componentes (VMMC), conforme a Equação (2.5), semelhante ao explorado em Heydt et al. (1999). Onde a operação $|\cdot|$ se refere operação de módulo, ou distância do ponto até a origem de seu sistema de coordenadas.

$$VMMC = \max(|Y_{f[n]A}| - |Y_{f[n]R}|) \quad (2.5)$$

Como o monitoramento da magnitude espectral não contempla variações em fase, a comparação entre janelas também pode ser realizada segundo a maior diferença ponto a ponto entre os valores reais ou a maior diferença ponto a ponto entre os valores imaginários das frequências da janela de referência e da janela atual (2.6) (ELSPEC, 2007), denominada neste trabalho de Variação Máxima em Componentes de Frequência (VMCF).

$$VMCF = \max(\max(|REAL(Y_{f[n]A}) - REAL(Y_{f[n]R})|), \max(|IMG(Y_{f[n]A}) - IMG(Y_{f[n]R})|))) \quad (2.6)$$

As operações $REAL(\cdot)$ e $IMG(\cdot)$, se refere a componente real e imaginária de um número complexo respectivamente.

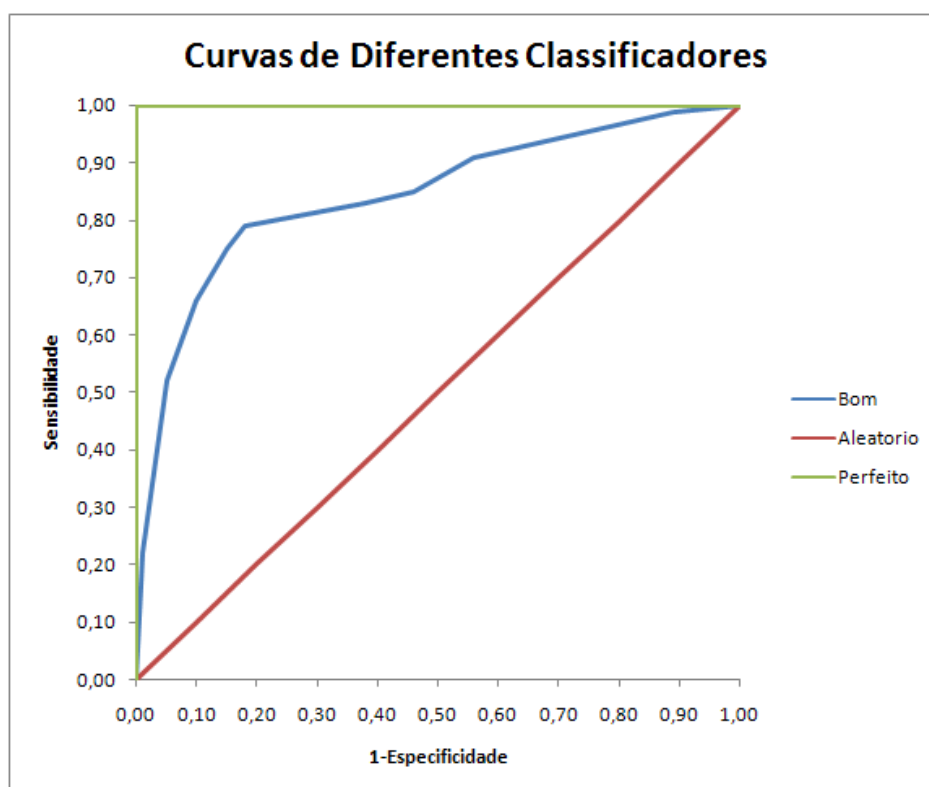
ESTRATÉGIAS DE COMPARAÇÃO E DEFINIÇÃO DE LIMIAR ÓTIMO EM DETECÇÃO

O problema de detecção de inovação de um sinal segmentado pode ser interpretado como um problema de classificação binário (se o trecho possui ou não uma novidade). Os métodos que serão comparados baseiam-se na diferença entre janelas do sinal, comparada com um certo limiar. Isso significa que o desempenho do sistema é dependente da escolha desse parâmetro. O uso de análises gráficas como a curva ROC (TREES, 2004) e a curva Precisão e revocação, do inglês *Precision Recall* (PR) (RIJSBERGEN, 1979), apontam para o comportamento do sistema de detecção com respeito a variação do patamar de decisão, possibilitando uma comparação entre métodos independente dessa parametrização.

A curva ROC mostra como o número de exemplos positivos corretamente identifica-

dos (probabilidade de detecção) varia de acordo com o número de exemplos negativos identificados como corretos (falsos positivos ou falso alarme) ao passo que o limiar de decisão é alterado. A Figura 7 mostra exemplos de curvas ROC para três diferentes classificadores, um bom um perfeito e um sem capacidade de discriminação. Em linhas gerais, o quanto mais próxima a curva se encontrar das bordas esquerda e superior, melhor é sua capacidade de seleção. Por esse motivo, uma métrica aceita para qualificação de curvas ROC é a Área Embaixo do Gráfico, do inglês *Area Under the Curve* (AUC).

Figura 7: Exemplo de curvas ROC.



A curva PR se diferencia da curva ROC uma vez que evidencia a relação entre precisão (2.7) e revocação (2.8). Os gráficos de análise PR vão apresentar o índice de precisão (P) e revocação (R) médio, que como o nome evidencia, é a média aritmética dos parâmetros através de diferentes limites de decisão.

$$P = \frac{T_p}{T_p + F_p} \quad (2.7)$$

$$R = \frac{T_p}{T_p + F_n} \quad (2.8)$$

onde, T_p é a porcentagem de positivos verdadeiros, F_p é a porcentagem de falsos positivos e F_n é a porcentagem de falsos negativos, todos em relação ao número total de

eventos. As duas análises serão utilizadas neste trabalho, pois análises com a ROC podem apresentar resultados otimistas, ainda mais quando realizada em problemas com classes desbalanceadas, em número de eventos ou em relevância de classificação (DRUMMOND & HOLTE, 2000).

O desempenho dos sistemas está intimamente ligado a escolha do limiar de decisão. A escolha de um patamar muito estreito acarreta em maior taxa de detecção de distúrbios. No entanto, um maior número de falsos positivos será acusado em contrapartida. A escolha de um patamar muito brando evita acusação de falsos inovações, porém a capacidade detectar eventos de interesse é comprometida. A escolha do limiar ótimo para cada algoritmo será realizada pela maximização de duas métricas: i) o índice Soma-Produto (SP) 2.9 (FILHO; SEIXAS & CALÔBA, 2008) e ii) o índice F_{score} 2.10 (SOKOLOVA; JAPKOWICZ & SZPAKOWICZ, 2006).

$$SP = \sqrt{\sqrt{S \times E} \times \frac{S + E}{2}} \quad (2.9)$$

Onde S se refere a sensibilidade do detector, definida pela probabilidade de detecção T_p e E é a especificidade do detector definida como sendo o complemento da probabilidade de falso alarme ($1 - F_p$).

$$F_{score} = 2 \times \frac{P \times R}{P + R} \quad (2.10)$$

2.5 CONCLUSÕES PARCIAIS

O processamento de sinais e técnicas de inteligência computacional têm cada vez aumentando sua importância para Sistema Elétrico de Potência. Por tais áreas apoiarem-se amplamente em pesquisa computacional, torna-se relevante e importante a possibilidade de realização de tais pesquisas de forma reprodutível. Adicionalmente, nesse capítulo, foram apresentados alguns avanços realizados na área de pesquisa reprodutível que possibilitam a integração e disseminação de conhecimento de forma mais rápida e fácil, demonstrando a atualidade do tema e como diversas áreas da ciência têm disponibilizado ferramentas de RR para a comunidade científica.

3 PLATAFORMA DE PESQUISA REPRODUTÍVEL PARA PROCESSAMENTO DE SINAIS APLICADO AO SEP

A proposta da plataforma é a criação de ambiente que possibilite o compartilhamento de pesquisas na área de processamento de sinais aplicado ao SEP de forma reprodutível. Para esse fim, existe a necessidade de implementação de alguns requisitos, cuja união resulte em ambiente virtual capaz de abrigar e expor os componentes essenciais de uma pesquisa. A primeira parte desse capítulo apresentará cada funcionalidade de maneira individual. E a segunda parte do capítulo será dedicada à descrição resumida das tecnologias utilizadas para construção do site e a explicação dos motivos de suas escolhas.

3.1 COMPONENTES DA PLATAFORMA

De forma a atender a necessidade de facilidade de acesso, a plataforma será dimensionada no formato de um site. A Figura 8 representa um mapa geral do site em questão. Os dois componentes essenciais para possibilitar a reutilização de pesquisas é a disponibilização de códigos e dados. Diversas ferramentas e plataformas expostas no capítulo 2 possibilitam de forma direta ou indireta a exposição desses elementos, porém com exceção da plataforma BEAT, estes dois elementos são armazenados de maneira conjunta, dificultando a reutilização deles de forma independente. A proposta da presente plataforma, consiste no armazenamento desses elementos de maneira independente entre si. Dessa forma o uso de um mesmo banco de dados pode ser acessado por diversos algoritmos e um mesmo código pode ser utilizado para processar diferentes bancos de dados de maneira mais fácil.

A associação de algoritmos, banco de dados e parametrização compõem um **experimento**, que após ser executado pelo servidor apresentará os resultados gerados. Resultados de um ou mais experimento podem ser analisados pela criação de instancia de **comparação**.

No primeiro acesso à plataforma, o usuário será direcionado à página de login ou



Figura 8: Visão geral do plataforma. Fonte: Elaborada pelo Autor.

cadastro de novo usuário, conforme pode ser observado pela Figura 9. Após o cadastro e efetivo login o usuário será direcionado à página inicial, local onde é possível acessar as quatro principais funcionalidades, que são elas: repositório de algoritmos, banco de dados, experimentos e comparação. A Figura 10 ilustra a tela com as opções citadas. Nas próximas subseções serão explicadas as características de cada uma das funcionalidades da plataforma desenvolvida.

A imagem mostra a interface de login da plataforma. No topo, há links para 'PQ', 'Register' e 'Log In'. O formulário principal contém:

- Título: Log In
- Campos de entrada: Username e Password.
- Botão de ação: Submit (em verde).
- Link de recuperação: "Don't have an account? [Click here to register.](#)"

Figura 9: Página de Login.

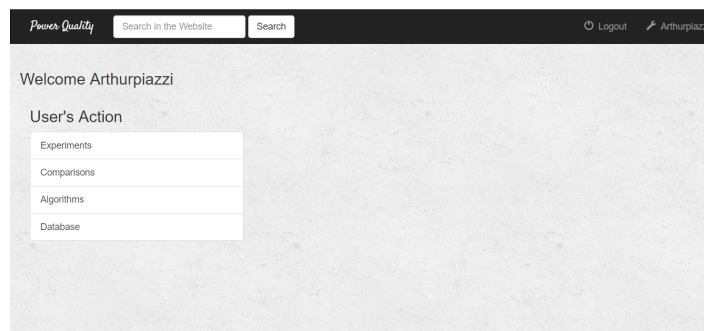


Figura 10: Visualização da página inicial.

3.1.1 REPOSITÓRIO DE ALGORITMOS

A disponibilização do código fonte e do conjunto de dados pode ser intercambiável de acordo com o área de conhecimento. Para o escopo do presente trabalho e até mesmo para o campo de processamento de sinais aplicado ao SEP, os códigos que compõem as análises de sinais elétricos podem ser considerados como primordiais. Dessa forma, a mais essencial funcionalidade oferecida pela plataforma é o repositório de rotinas¹.

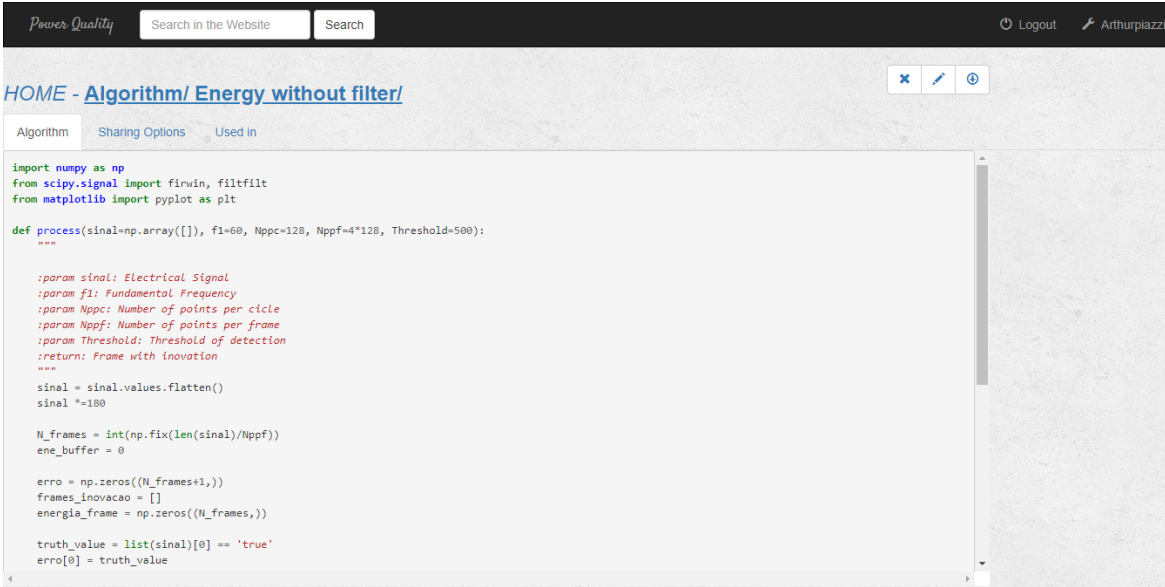
A Figura 11 mostra a página de apresentação com a listagem de algoritmos incorporados ao repositório. Além da exibição da listagem, o usuário pode adicionar um novo algoritmo à plataforma clicando no botão *New* localizado no campo superior direito da tela. Existe ainda a possibilidade de buscar algum algoritmo específico através do nome do código, autor ou descrição.

Updated	Name	Actions
July 25, 2018	Residue without filter - arthurpiazzi	Edit Delete Download
July 25, 2018	Energy without filter - arthurpiazzi	Edit Delete Download
July 25, 2018	FFT_abs - arthurpiazzi	Edit Delete Download
July 17, 2018	Inovation_Residuo - arthurpiazzi	Edit Delete Download
July 17, 2018	Inovation_FFT - arthurpiazzi	Edit Delete Download
July 17, 2018	Inovation_Energia - arthurpiazzi	Edit Delete Download
Oct. 26, 2017	Decton - arthurpiazzi	Edit Delete Download
Oct. 24, 2017	Detector de Inovação - arthurpiazzi	Edit Delete Download

Figura 11: Visualização da página de algoritmos.

¹Respeitando as diferenças entre as definições técnicas entre algoritmos, códigos, rotinas e *scripts*, neste texto todos eles se referem a um conjunto de instruções estruturadas com intuito de realizar certa tarefa ou problema

A página de visualização de algoritmos é ilustrada pela Figura 12. Além da exibição do código fonte, ações como download de arquivos, criação de experimentos a partir de tal rotina, visualização de experimentos que fazem uso de tal código estão disponíveis para o utilizador da plataforma. Para o autor do algoritmo outras operações são permitidas, como deleção e alteração das permissões de visualização do código. Apesar de ser contra o propósito da plataforma, algoritmos podem ser submetidos de maneira privada ou com visualização restrita a somente usuários específicos, sendo essa última opção especialmente útil para compartilhamento com revisores em período pré-publicação. De ambas formas, as rotinas ainda podem ser utilizadas para composição de experimentos como uma "caixa preta". Os algoritmos podem ser removidos da plataforma pelo autor, desde que ainda não tenham sido executados com sucesso em um experimento.



The screenshot shows a web interface for 'Power Quality'. At the top, there is a search bar and a user profile for 'Arthurpiuzzi'. The main content area displays a Python code routine titled 'HOME - Algorithm/ Energy without filter/'. The code is as follows:

```

import numpy as np
from scipy.signal import firwin, filtfilt
from matplotlib import pyplot as plt

def process(sinal=np.array([]), f1=60, Nppc=128, Nppf=4*128, Threshold=500):
    """
    :param sinal: Electrical Signal
    :param f1: Fundamental Frequency
    :param Nppc: Number of points per cycle
    :param Nppf: Number of points per frame
    :param Threshold: Threshold of detection
    :return: Frame with innovation
    """
    sinal = sinal.values.flatten()
    sinal *= 100

    N_frames = int(np.fix(len(sinal)/Nppf))
    ene_buffer = 0

    erro = np.zeros((N_frames+1,))
    frames_inovacao = []
    energia_frame = np.zeros((N_frames,))

    truth_value = list(sinal)[0] == 'true'
    erro[0] = truth_value
  
```

Figura 12: Visualização de uma rotina na página de algoritmos.

Para realizar o armazenamento e o compartilhamento de algoritmos, poucas restrições são impostas ao usuário, exceto a formatação do código em algum formato de texto legível (i.e. .py, .m, .cpp) ao invés de executáveis. Porém, para tais arquivos, somente a funcionalidade de exibição do código é suportada. Para que o código possa ser executado na composição de experimentos, algumas restrições são impostas ao usuário.

Visto que a plataforma é aplicada a um campo muito amplo e diverso, é natural que as soluções associadas possuam diferentes requerimentos, quanto ao número e formato dos dados de saída e entrada. Para abranger o maior número de soluções possíveis mantendo a capacidade de execução no servidor, algumas limitações devem ser atendidas. A primeira e maior delas sendo referente à linguagem de programação. Devido ao fato de que RR está intimamente ligada a ideologia de código aberto e código livre, aliados as diver-

sas vantagens e popularidade da linguagem Python, esta foi escolhida como linguagem padrão do ambiente. Atualmente, somente *scripts* nessa linguagem possuem suporte de execução.

Além da linguagem, é necessário seguir um modelo conforme a Figura 13. Dentre os elementos presentes no exemplo, somente a função **process** possui a obrigatoriedade de implementação, pois esta função será chamada internamente pelo servidor na execução do experimento.

```
import module_1
from module_2 import Class_1

def process(data, *args, **kwargs):
    <statement 1>
    <statement 2>
    ...
    <statement n>

def Supporting_function(*args, **kwargs):
    < statement 1 >
    < statement 2 >
    ...
    < statement n >

class SupportingClass(object):
    def __init__(self, *args):
        < statement 1 >
        < statement 2 >
        ...
        < statement n >

    def foo(self):
        < statement 1 >
        < statement 2 >
        ...
        < statement n >
```

Figura 13: *Template* de um script incorporado à plataforma.

3.1.2 BANCO DE DADOS

Atualmente, existem sofisticadas e acessíveis ferramentas de compartilhamento de códigos, como GitHub e GitLab. Porém a reprodução de resultados pode ser comprometida quando não se tem acesso a mesma base de dados utilizadas pelos autores da publicação. Diferentemente do compartilhamento de códigos, ainda não existe uma consolidação de ferramentas para compartilhamento de dados, o que desencoraja a prática de ciência aberta. Atualmente existem notáveis repositórios de dados *online*, porém estes são específicas a um campo do conhecimento, como Mendeley e GigaDB citadas anteriormente no capítulo 2. O compartilhamento e padronização de bases de dados fomenta o avanço científico além do escopo da reprodutibilidade, a exemplo do repositório UCI (DHEERU & TANISKIDOU, 2017). Tal repositório que agrega diversos bancos de dados no campo de inteligência artificial e aprendizado de máquinas, representando uma importante contribuição para o campo.

Para explicar como tal repositório foi concebido, um paralelo entre as duas mais completas plataformas de pesquisa reprodutível se faz necessário, a plataforma BEAT e Code Ocean. A plataforma de segurança biométrica (BEAT) oferece importantes bancos de dados da área, mas não oferece a capacidade de um usuário comum adicionar novos. Essa limitação, apesar de não ser ideal, não invalida o propósito da plataforma, uma vez que é desejado a busca e consolidação de soluções para bases de dados específicas. O mesmo ainda não é verdade para o caso de processamento de sinais elétricos, e para cumprir a proposta da plataforma, os usuários devem ser capazes de incorporar novos dados à plataforma. Ao passo que a plataforma Code Ocean, que oferece a capacidade de inserção de banco de dados pelo usuário, encapsula os dados dentro do projeto, não existindo a possibilidade de reutilização em outros projetos ou a possibilidade de busca entre os conjuntos de dados arquivados. A Plataforma de Pesquisa Reprodutível para o Processamento de Sinais Aplicado ao SEP (PRESEP) se difere dessas plataformas. Oferecendo a capacidade do usuário adicionar novos dados, como é possível na plataforma Code Ocean, e mantendo os dados desvinculados de um experimento específico para facilitar o uso em diferentes experimentos, como a plataforma BEAT.

Sinais do SEP variam amplamente em sua formatação, podendo apresentar diferentes números de canais, tamanho de janelas temporais e taxa de amostragem, tendo a possibilidade de serem representados em diferentes domínios. Implementação da execução de diferentes rotinas, que por sua vez empregam dados com características diversas constitui um desafio. Dessa forma, para o estado atual da plataforma, foram adotadas as seguintes simplificações:

- Os dados serão contidos em uma instancia referida como ***Dataset***;
- Cada *Dataset* pode conter um ou mais arquivos, referidos como eventos elétricos;
- Um evento elétrico deve ser submetido com uma das seguintes formatações: Valores separados por virgula, do inglês *Comma Separated Values* (CSV) ".csv", arquivos de texto ".txt" e formato de dados salvos pelo *software* Matlab ".mat";
- Quanto à utilização dos dados em experimentos, diferentes formatos serão carregados de distintas maneiras. Arquivos com formatação CSV, serão carregados em objetos *Data Frame* (estrutura básica de dados da biblioteca **pandas**), ao passo que os demais serão carregados em *array* (estrutura de dados da biblioteca **numpy**).

Os arquivos de eventos elétricos apresentam peculiaridades, podendo até mesmo conter diversos sinais elétricos no mesmo arquivo. Dessa forma, é aconselhável o pro-

vimento de uma descrição sobre a utilização de dados salvos. Ações como download de todos os arquivos referentes a um *Dataset* e visualização online de eventos com dimensão inferior a dois são suportadas. As Figuras 14 e 15, mostram, respectivamente, as páginas de exibição de um *Dataset* e a visualização de um evento elétrico específico dentro de um conjunto de dados.

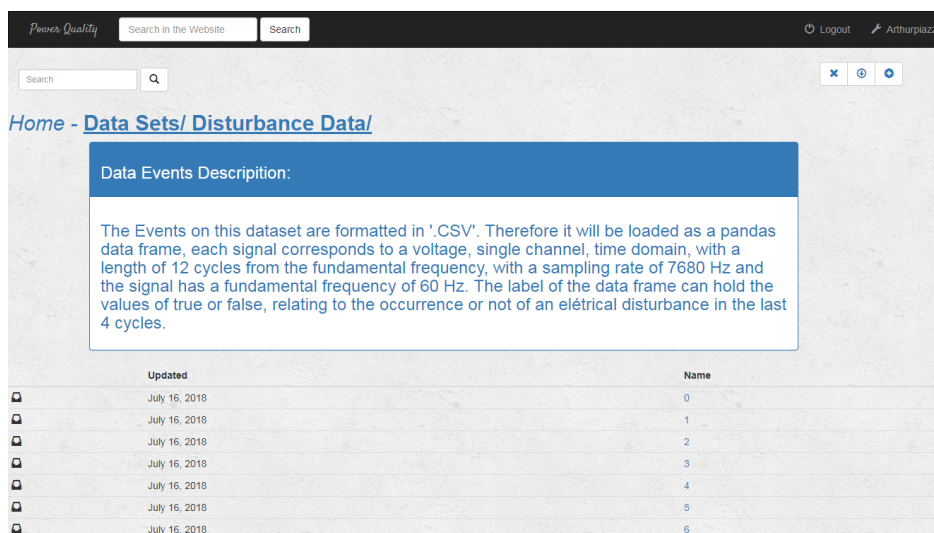


Figura 14: Visualização da página de exibição do *Dataset*.

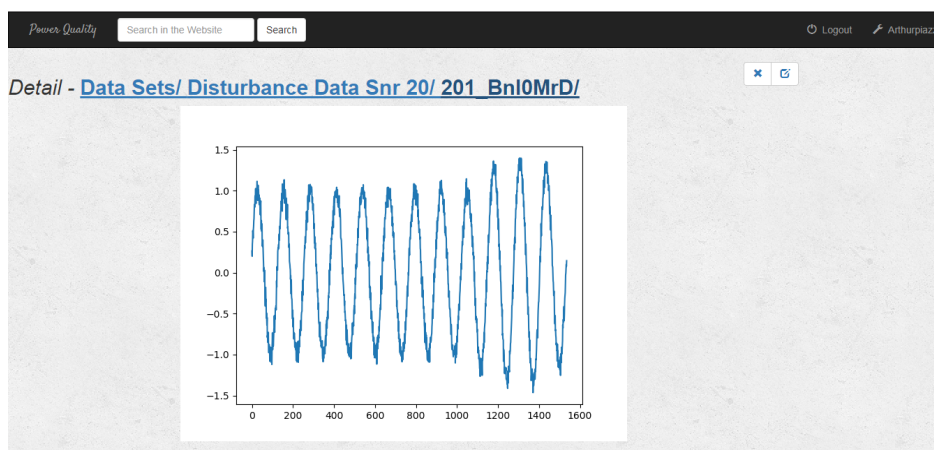


Figura 15: Visualização da página contendo um evento elétrico.

3.1.3 REPOSITÓRIO DE EXPERIMENTOS

A divulgação de código e dados quando realizada de forma conjunta, já significa um grandioso passo para pesquisa reprodutível. Contudo, somente tais elementos não garantem a reprodução de resultados. Apesar de não terem sido realizadas estatísticas sobre a facilidade ou factibilidade de reprodução de trabalhos publicados na área de análises de sinais elétricos, estudos realizados em outras áreas apoiadas na ciência computacional

evidenciam dificuldades na certificação de resultados. Em McCullough, McGeary e Harrison (2006) foi realizado um estudo sobre reprodução de publicações no *Journal of Money and Banking*, periódico da área de economia que exige submissão de códigos e dados para publicação. Levando em conta somente artigos que de fato obedeceram a política de transparência, somente 22% dos artigos tiveram os resultados atestados. Outro estudo sobre reprodutibilidade no campo da bioinformática (GILBERT et al., 2012), mesmo com acesso a código e dados, falharam em reproduzir 30% dos casos analisados.

Uma evidente causa da dificuldade de reprodução de resultados, quando códigos e dados estão disponíveis, pode ser atribuída a portabilidade. Diferentes sistemas operacionais, necessidade de softwares proprietários, quebra entre versões de módulos, bibliotecas, compiladores e interpretadores podem evitar a execução de rotinas em máquinas diferentes. A confiabilidade de resultados numéricos de um mesmo algoritmo executado em diferentes máquinas e implementado em diferentes linguagens foi investigada em (VITORINO et al., 2017), a fim de avaliar se erros decorrentes de diferentes métodos de arredondamento entre ambientes diferentes são capazes de impossibilitar a reprodutibilidade de um determinado algoritmo. A limitação atribuída a portabilidade pode ser mitigada através do uso de ambientes virtuais, e nesse contexto, a linguagem Python oferece uma conveniente ferramenta para criação de tal ambiente. Um ambiente virtual Python caracteriza-se pelo isolamento do interpretador Python, bibliotecas e scripts daqueles instalados em outros ambientes virtuais ou de quaisquer bibliotecas instaladas como parte do seu sistema operacional. Dessa forma, o usuário da PRESEP tem a capacidade de facilmente equiparar o ambiente virtual da plataforma e o ambiente virtual local, quer seja pra garantir compatibilidade antes de submeter rotinas a plataforma, ou posterior utilização das rotinas de demais pesquisadores salvas na plataforma.

Além da plataforma se basear em uma linguagem com alto grau de portabilidade, essa oferece a capacidade de execução no servidor. Como definido em (HURLIN et al., 2014), a execução de rotinas na nuvem é caracterizada como o próximo passo para pesquisa reprodutível. Dessa forma, resultados estão disponíveis de maneira cômoda, sem a necessidade de *download* de arquivos e instalação de *softwares*. A Figura 16 mostra a página de exibição de experimentos incorporados.

Para incorporar um novo experimento à plataforma, o usuário deve seguir os passos exemplificados no Fluxograma da Figura 17. A execução de cada passo pode ser observada na Figura 18. Após definir o algoritmo a ser empregado, o cliente da plataforma é interrogado sobre nome, banco de dados, descrição e link para publicação ou site relevante do experimento.

Uploaded	Name	Actions
July 26, 2018	arthurpiuzzi - Energy Freq	▶ ✕
July 26, 2018	arthurpiuzzi - Residue freq	▶ ✕
July 26, 2018	arthurpiuzzi - FFT Freq	▶ ✕
July 26, 2018	arthurpiuzzi - Energy without filter freq	▶ ✕
July 26, 2018	arthurpiuzzi - Residue without filter Freq	▶ ✕
July 26, 2018	arthurpiuzzi - Energy SNR 40	▶ ✕
July 26, 2018	arthurpiuzzi - Residue SNR 40	▶ ✕
July 26, 2018	arthurpiuzzi - Energy without filter SNR 40	▶ ✕
July 26, 2018	arthurpiuzzi - FFT SNR 40	▶ ✕
July 26, 2018	arthurpiuzzi - Residue without filter SNR 40	▶ ✕
July 26, 2018	arthurpiuzzi - Residue without filter SNR 20	▶ ✕

Figura 16: Página de exibição de experimentos.

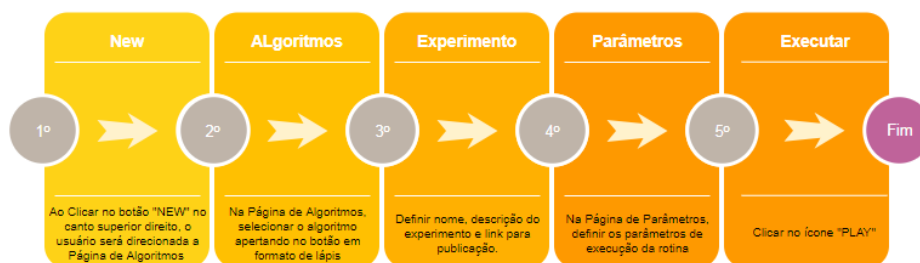


Figura 17: Página de exibição de experimentos.

No ato de submissão de um código na plataforma, o site faz um reconhecimento dos argumentos da função de **process**. No quarto passo do fluxograma, o usuário define os valores dos parâmetros de execução e é solicitado também identificar se algum dos argumentos se refere a dados de sinais elétricos. Caso haja alguma entrada referente a dados, o usuário deve excluí-lo da lista pois dados e parâmetros são salvos e carregados de forma diferente pela plataforma. O último passo a ser seguido é a solicitação de execução do experimento. A Figura 18 mostra a progressão das telas do site no processo de adição de um novo experimento. As caixas em vermelho evidenciam os ícones envolvidos nesse processo.

Uma vez que o tempo de execução dos experimentos é consideravelmente mais longo que os tempos de execução das outras tarefas do site. A execução das rotinas juntamente com as demais tarefas do site, pode repercutir em degradação no que se diz ao processamento do servidor. Dessa forma, foi necessário a implementação da execução de experimentos de forma paralela e independente das outras atividades do servidor.

Os resultados de um experimento podem ser salvos de duas maneiras. A primeira através da captura dos retornos da função **process** e, para tal, o retorno pode ser de qual-

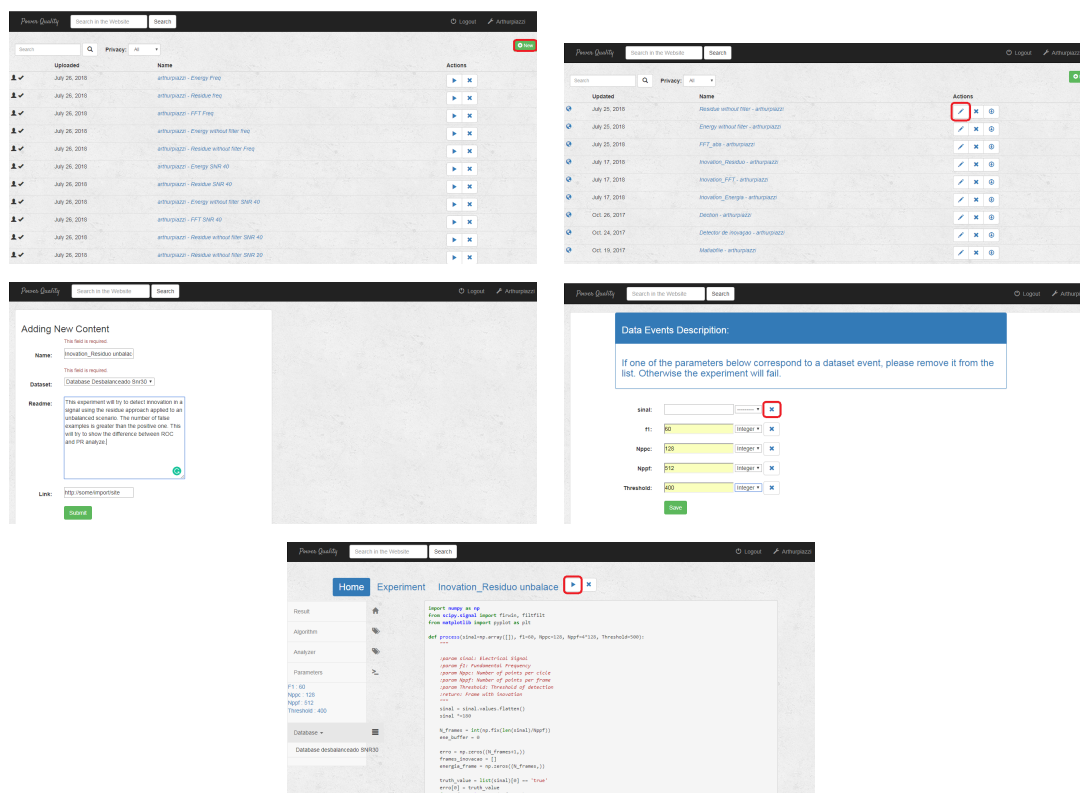


Figura 18: Passo a passo da adição de um novo experimento.

quer formato serializável (objetos que podem ser convertidos para Notação de Objetos *JavaScript*, do inglês *JavaScript Object Notation* (JSON), para ser compatível com o mecanismo de banco de dados), o que compreende todos os formatos nativos do python (i.e. listas, tuplas, *strings*). Sabendo que em diversos casos, somente a representação gráfica fornece compreensão clara dos resultados, a plataforma identifica se alguma figura foi criada durante a execução de função e a armazena para futura exibição. Devido a forma que o escopo de variáveis funciona na linguagem Python, a captura da figura é realizada sem a necessidade do retorno explícito do objeto.

A Figura 19 mostra a página de resultados de uma rotina de detecção de novidade no sinal do SEP. O resultado para cada sinal elétrico pode ser observado em duas partes, uma parte gráfica contendo um *plot* do sinal de tensão segmentado em janelas e o resultado numérico composto pela indicação de quais janelas tiveram inovações identificadas.

Além dos resultados, a barra lateral expõe os demais elementos que constituem o experimento. Navegando pela barra lateral de cima para baixo, pode-se observar a aba de algoritmo, onde está presente a descrição do experimento, *link* para a publicação e exibição do código fonte. A aba de comparação, apresenta todas as comparações que fazem uso desse experimento. Na aba de parâmetros, é possível visualizar e alterar seus valores. E por fim, a indicação do banco de dados utilizados.

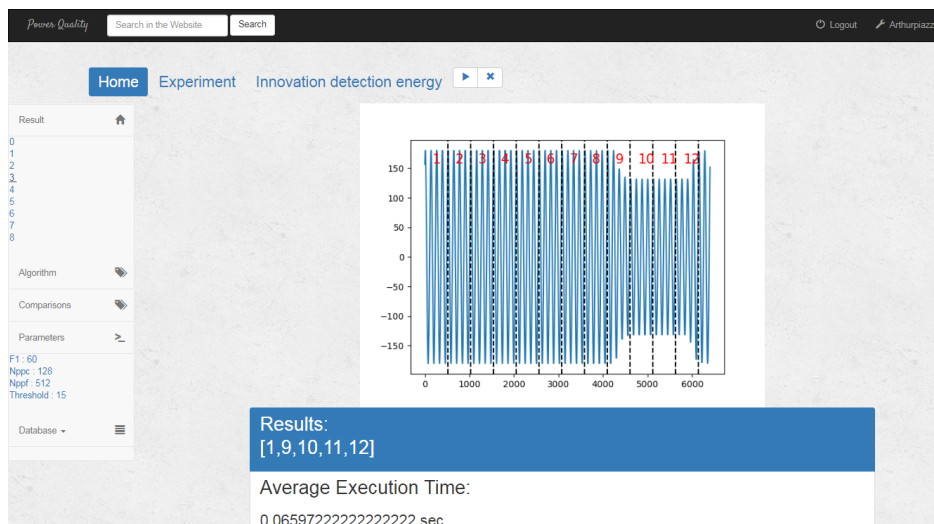


Figura 19: Página de exibição de resultados.

3.1.4 COMPARAÇÕES E ANÁLISES

O último componente da plataforma diz respeito a análise e comparação dos resultados gerados pelos experimentos. Comparações podem ser realizadas entre um ou mais experimentos, contando que todos eles usem a mesmo banco de dados. Uma vez que os resultados gerados pelos experimentos geralmente são relativos a sinais elétrico individuais. Comparações podem ser escritas referentes a somente um experimento, obtendo resultados referente a todo conjunto de dados.

Porém, o intuito original do componente de comparação é a comparação entre dois ou mais experimentos. Devido à grande multiplicidade de soluções para um mesmo problema, somente através da comparação de novas técnicas com demais já existentes é possível validar novas descobertas. A realização dessas comparações através do uso da plataforma elimina a necessidade de reimplementação de experimentos já submetidos a ela.

A possibilidade de simplificação do processo de comparação entre métodos impacta positivamente o campo de processamento de sinais aplicado ao SEP. Retomando o paralelo entre as demais plataformas de RR, a plataforma BEAT oferece a possibilidade da criação de *reports*, que comparam resultados de diferentes experimentos. Tal funcionalidade possibilitou uma estrutura que fornece critérios e métricas padronizados para avaliação de sistemas biométricos tanto para o meio acadêmico quanto para a indústria (MARCEL, 2013). A plataforma *Code Ocean* não contempla comparação entre resultados, pois sua proposta consiste em implementação de RR de forma ampla, não restrita a um campo do conhecimento. A generalidade da plataforma, abrangendo soluções tão distintas inviabi-

liza a possibilidade de comparação entre soluções.

Ainda justificado pela abrangência do campo, a comparação de resultados varia muito de aplicação a aplicação, contrariamente ao campo de segurança biométrica onde todos os problemas podem ser compreendidos em problemas de classificação. Por este motivo, na plataforma desenvolvida nesse trabalho, torna-se necessário fornecer o algoritmo de análise no ato de submissão. Algoritmos de comparação devem seguir as mesmas restrições que os algoritmos de processamento de sinais discutido na subseção 3.1.1.

Similar a adição de experimentos, o fornecimento de descrição da comparação e *link* para a publicação é aconselhado. A seleção dos experimentos que comporão a comparação deve ser realizada. Após a conclusão desses passos, a adição da comparação na fila de execução pode ser realizada. A Figura 20 mostra a progressão dos passos de adição de uma nova comparação.

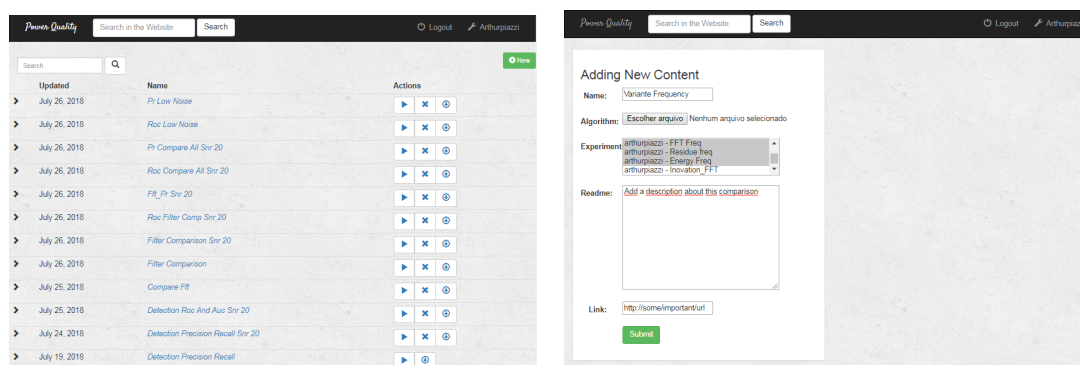


Figura 20: Passo a passo da adição de uma nova comparação.

Na página de visualização das comparações, exibida na Figura 21, é possível verificar os resultados, o código de análise e descrição da análise, os experimentos e o banco de dados utilizados na análise. O armazenamento dos resultados é realizado da mesma forma que os resultados gerados pelos experimentos, podendo ser expresso graficamente ou através de objetos serializáveis.

3.2 TECNOLOGIAS UTILIZADAS

O emprego de aplicações Web no cenário atual vem crescendo de forma acelerada. Vantagens como acessibilidade remota, não necessidade de instalação de softwares e compatibilidade entre diferentes sistemas operacionais justificam a migração de diversas soluções e produtos digitais das tradicionais aplicações de *Desktop* para recursos *online*. Devido à imensa demanda por serviços relacionados ao desenvolvimento Web, foram desenvolvidos nos últimos anos diversas linguagens, *frameworks* e recursos para garantir

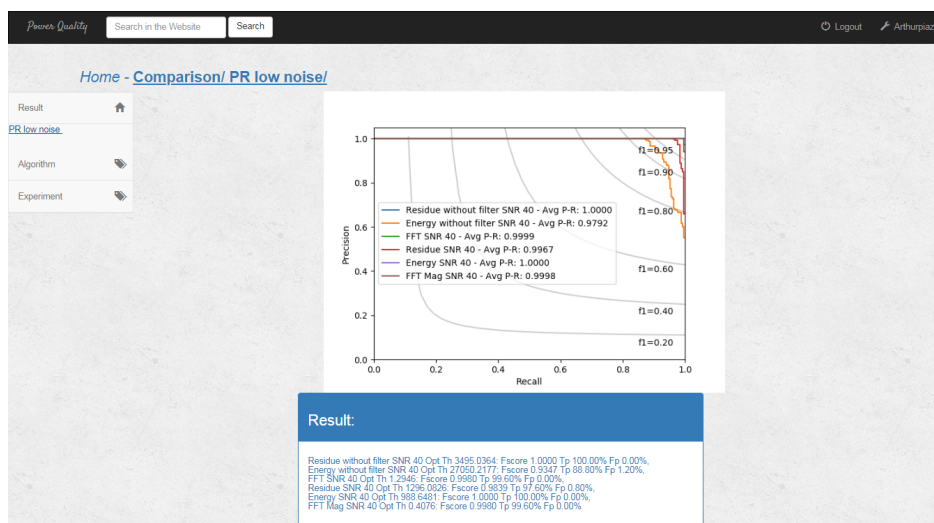


Figura 21: Página de exibição de comparação.

mais agilidade e eficiência ao campo, atenuando a lacuna entre a capacidade de produção e demanda do mercado.

As próximas subseções expõem as tecnologias escolhidas para o desenvolvimento do trabalho, bem como uma visão geral de algumas características de cada tecnologia que justificam o emprego das mesmas. Além das características específicas de cada tecnologia, um fator determinante para escolha foi a aceitação atual da tecnologia no mercado. Em linhas gerais, tecnologias mais populares possuem melhor documentação e menor latência na correção de limitações e erros.

Apesar de não existir um modo de mensurar com exatidão a popularidade das linguagens de programação empregadas no mundo, alguns indicadores podem ser utilizados para apontar essa relação, como estatísticas sobre o número de pesquisas sobre um determinado assunto, realizadas através da ferramenta de busca **Google** com o decorrer do tempo. Outra métrica que pode ser relacionada a popularidade de uma tecnologia de programação é o percentual de perguntas feitas no site **Stack Overflow**, maior site de pesquisas e repostas relacionado a ciência da computação.

3.2.1 PYTHON

Python é uma linguagem completa de Turing com multipropósito. Em outras palavras, é possível implementar qualquer aplicação através de sua utilização, desde simples automatização de tarefas até grandiosas aplicações de *Desktop*, sites e aplicativos para *SmartPhones*. Python é uma linguagem interpretada, o que acarreta desempenho inferior a linguagens compiladas como C e C++. Porém, sua sintaxe didática, código limpo com

obrigatoriedade de endentação e eliminação do uso de colchetes, aliados à existência de diversos módulos nativos ou facilmente instaláveis, conferiram à linguagem Python um lugar entre as linguagens mais populares do mundo e a mais proeminente no meio científico.

A popularidade na comunidade científica pode ser explicada por diversos motivos. Primeiramente por ser uma linguagem gratuita e *open source* e por seus poderosos módulos, como *numpy* para manipulação matemática, *scipy* para processamento de sinais e *tensorflow* e *keras* para o crescente campo do *Deep Learning*. Por esses motivos tal linguagem foi escolhida como base da plataforma, tanto em seu desenvolvimento, como para as rotinas que serão incorporadas a plataforma.

A aceitação de linguagem python pelo público pode ser observada nas Figuras 22 e 23, respectivamente, relacionadas a quantidade de pesquisas efetuados no Google e de perguntas realizadas no site Stack Overflow.

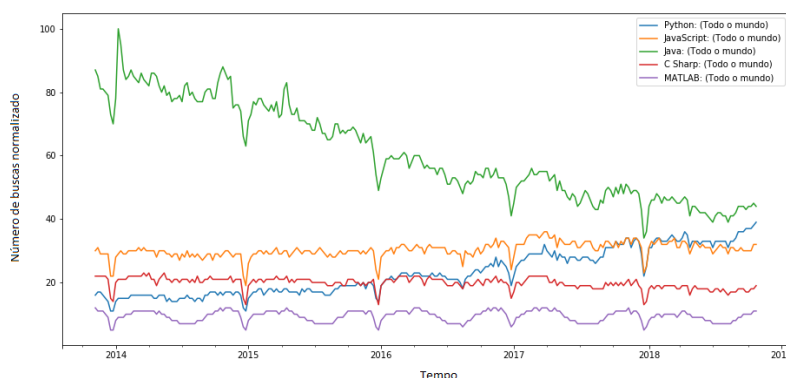


Figura 22: Busca realizada no Google sobre linguagens de programação. Fonte: <https://trends.google.com/>

Outro indicador que traduz o nível de penetração da linguagem na comunidade são as estatísticas sobre as linguagens incorporadas ao Website Github. A maior ferramenta de armazenamento e controle de versão do mundo. A Figura 24 mostra as estatísticas referente as linguagens dos projetos incorporados ao site no ano de 2017. Python ocupa a segunda posição no ranking, superando a consolidada linguagem Java pela primeira vez desde de sua criação.

3.2.2 DJANGO

Uma prática essencial do desenvolvimento web é o uso de *frameworks*. Todas as linguagens voltadas a esse destino possuem essas mini-linguagens (i.e.: C# possui o .NET,

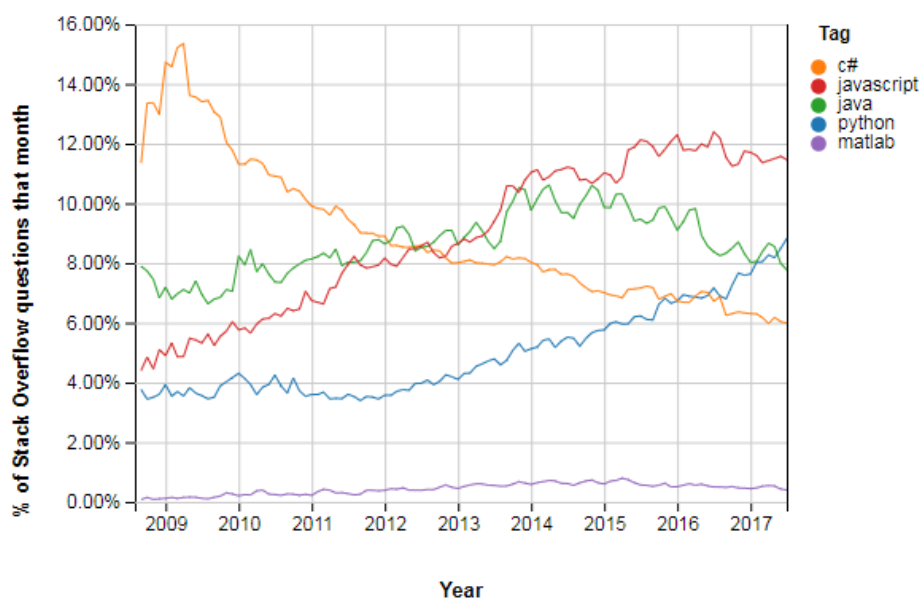


Figura 23: Perguntas realizadas no Stack OverFlow sobre linguagens de programação. Fonte: <https://insights.stackoverflow.com/trends>

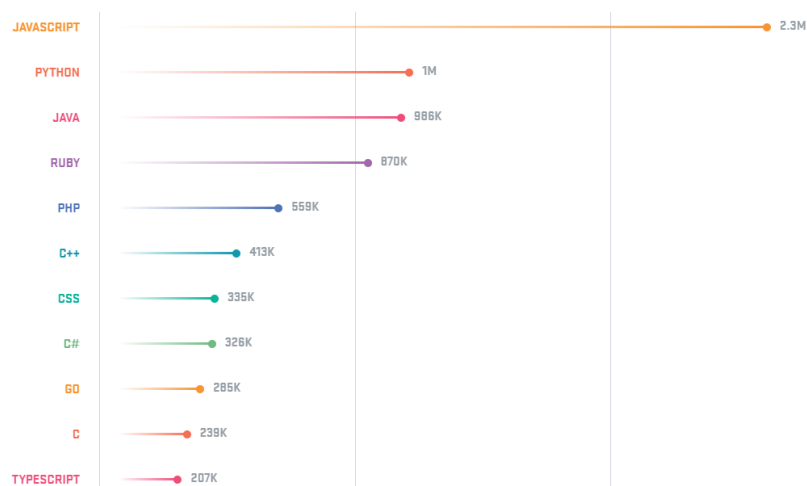


Figura 24: Estatísticas referente a linguagens utilizadas no site Github em 2017. Fonte: <https://octoverse.github.com/>

Ruby possui o Rails e o PHP o Laravel) que oferecem diversas abstrações e simplificações, diminuindo o tempo de programação e aumentando a qualidade dos recursos desenvolvidos. As Figuras 25 e 26 apresentam estatísticas referentes a alguns *Frameworks Backend*.

Dentre os *frameworks* da linguagem Python, o Django é o mais popular e mais completo. Tal framework faz uso da arquitetura Modelo Visão Controle, do inglês Model View Controller (MVC), que baseia-se na divisão de responsabilidades funcionais, facilitando para o desenvolvedor entender e modificar cada unidade em particular sem ter conhecimento e sem influenciar as demais (KRASNER; POPE et al., 1988). As responsabilidades

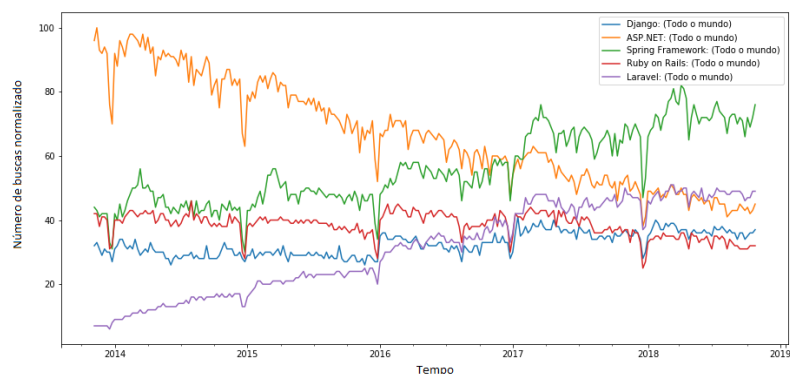


Figura 25: Busca realizadas no Google sobre *Frameworks Back-End*. Fonte: <https://trends.google.com/>

da arquitetura, como o próprio nome indica, são *Model*, referente aos dados necessários para a aplicação, *View*, referente a interface gráfica e *Controller*, referente as funções que geram e alteram as apresentações e os dados. A Figura 27 exemplifica o funcionamento da arquitetura.

Além da divisão de responsabilidade no nível funcional, o *framework* Django apresenta uma arquitetura de modularização em *Apps*, referente às diferentes finalidades que compõem o site. Tomando o presente trabalho como exemplo, o site necessita salvar informações do usuário e é necessário implementar um repositório de códigos, como esses dois recursos são bem distintos entre si, cada um representa um *App*. Essa segmentação por finalidade auxilia na reutilização do código e mantém a organização e escalabilidade do projeto.

A lista a seguir apresenta algumas das facilidades oferecidas pelo *framework*:

1. Abstração de banco de dados: o Django apresenta uma linguagem intermediária entre o servidor e bancos de dados, possibilitando o uso de diferentes softwares de banco de dados como SQL, PostgreSQL, Oracle, entre outros, através de uma sintaxe unificada.
2. Recursos de serialização de objetos para ágil criação de API (Interface de Programação de Aplicativos, do inglês Application Programming Interface).
3. Django *forms* que automatiza a criação de formulários e tratamentos de dados de entrada do usuário.
4. Tratamento de expressão regulares para roteamento de URL (Localizador Uniforme de Recursos, do inglês Uniform Resource Locator).

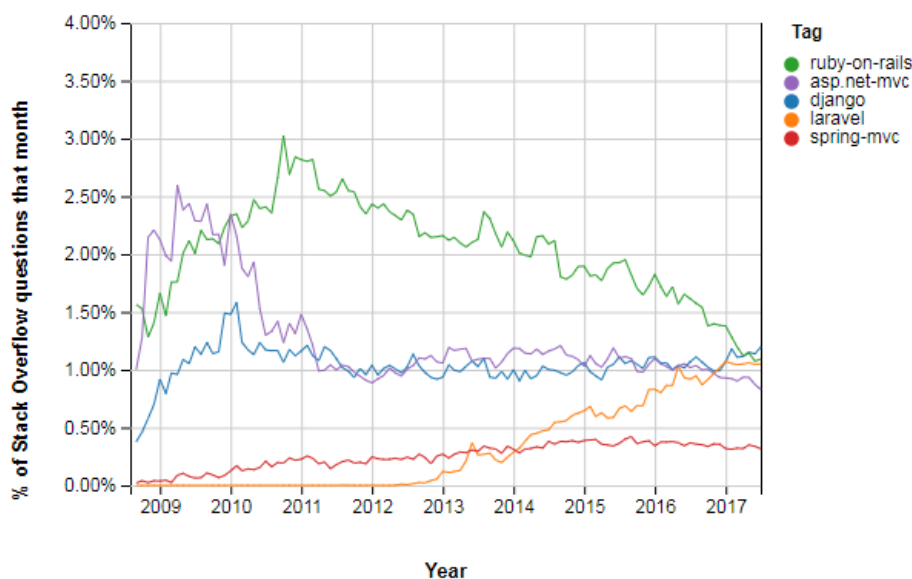


Figura 26: Perguntas realizadas no Stack OverFlow sobre *Frameworks Back-End*. Fonte: <https://insights.stackoverflow.com/trends>

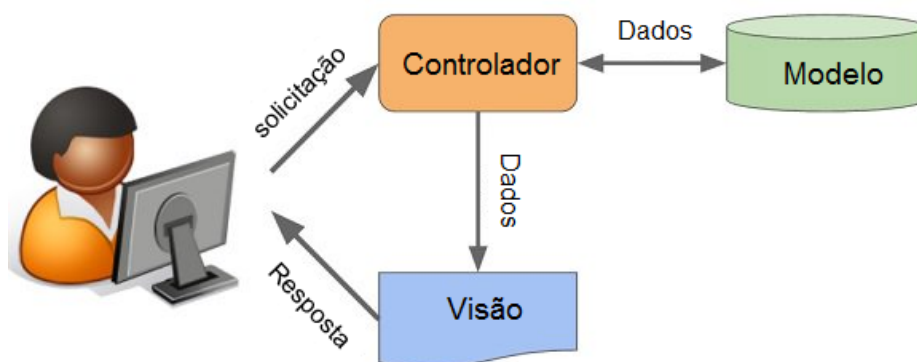


Figura 27: Arquitetura MVC.

3.2.3 JQUERY

Em ciência da computação, uma aplicação pode ser dividida em duas partes em relação aos seus conceitos de funcionamento, o *Back-end* concerne a camada de acesso de dados e o *Front-end* a camada de apresentação. Tal separação é mais importante e evidente no âmbito do desenvolvimento web, uma vez que tais camadas estão separadas fisicamente e é comum a denominação das mesmas como lado servidor e lado cliente.

O desenvolvimento do lado servidor apresenta uma multiplicidade enorme de linguagens e tecnologias que podem ser utilizadas, entretanto, no lado do cliente, devido à necessidade da interoperabilidade entre diferentes navegadores de internet, diferentes máquinas e sistemas operacionais, a consolidação do uso de algumas tecnologias se

tornou inevitável. Existem diferentes *frameworks*, mini-linguagens, recursos e protocolos relacionados ao lado cliente, porém sempre são baseados e continuarão sendo assim dentro de um futuro previsível, em HTML, Folhas de Estilo em Cascata, do inglês Cascading Style Sheets (CSS) e JavaScript (JS). HTML é um padrão de criação de páginas web baseado em elementos de marcação (*tags*), CSS em linhas gerais é um mecanismo de adição de estilo à página, e a linguagem JavaScript é utilizada no desenvolvimento das lógicas que controlam o navegador, realizando comunicação assíncrona e alterando o conteúdo do documento exibido.

Como aludido na subseção anterior, o desenvolvimento de *frameworks* e bibliotecas que agilizam o desenvolvimento é prática comum na ciência da computação. Existem algumas bibliotecas *Front-end* JavaScript como o jQuery, AngularJS e ReactJS. A mais completa e popular entre eles é o jQuery como evidenciado nas Figuras 28 e 29. Essa coleção de funcionalidades possibilita delegar funções para o lado servidor e criar interações na apresentação de forma mais dinâmica, de maneira mais simples e rápida do que o uso puro da linguagem JavaScript.

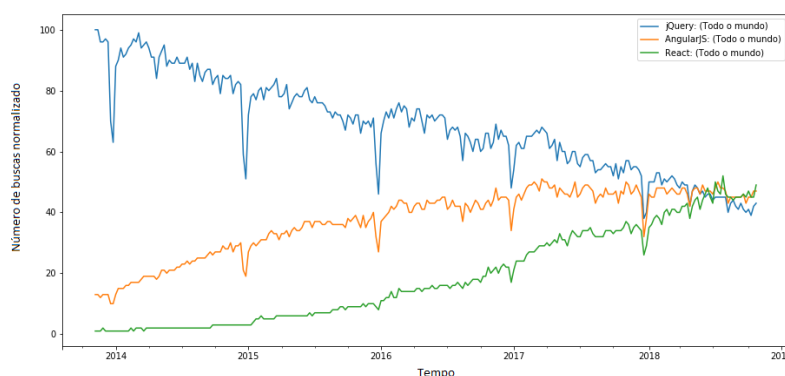


Figura 28: Busca realizadas no Google sobre *Frameworks Front-End*. Fonte: <https://trends.google.com/>

Dentre os recursos oferecidos pela biblioteca, o que merece mais destaque é a comodidade de implementação de comunicação AJAX (do inglês Asynchronous JavaScript + XML). O uso desse paradigma de comunicação permite que as páginas Web sejam totalmente carregadas uma vez ou poucas vezes e, posteriormente somente os dados são enviados entre cliente e servidor para atualização de forma assíncrona da página durante a navegação. O uso dessa técnica pode impactar grandemente na quantidade de dados enviados e requeridos e no tempo de vida da página (GARRETT et al., 2005). A Figura 30 exemplifica o funcionamento da ferramenta.

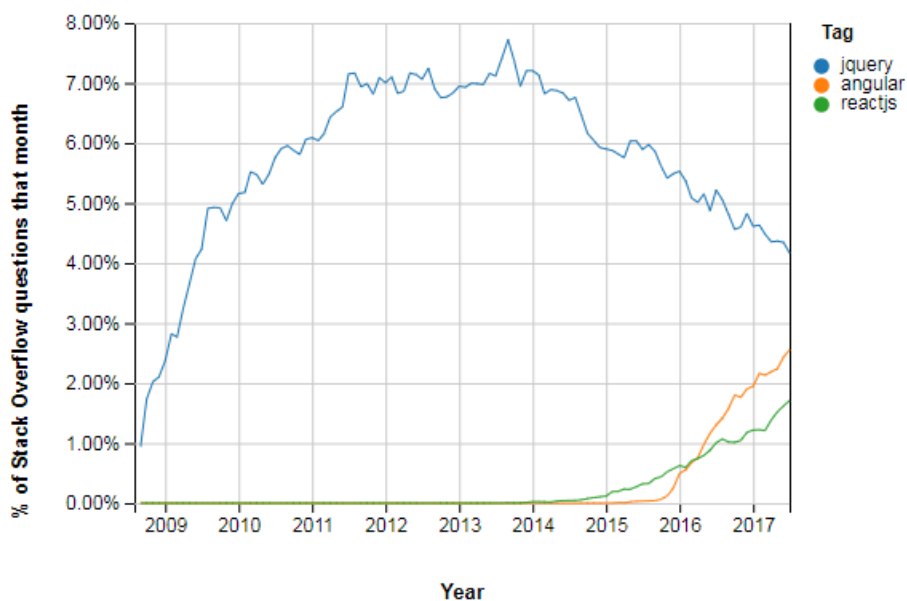


Figura 29: Perguntas realizadas no Stack OverFlow sobre *Frameworks Front-End*. Fonte: <https://insights.stackoverflow.com/trends>

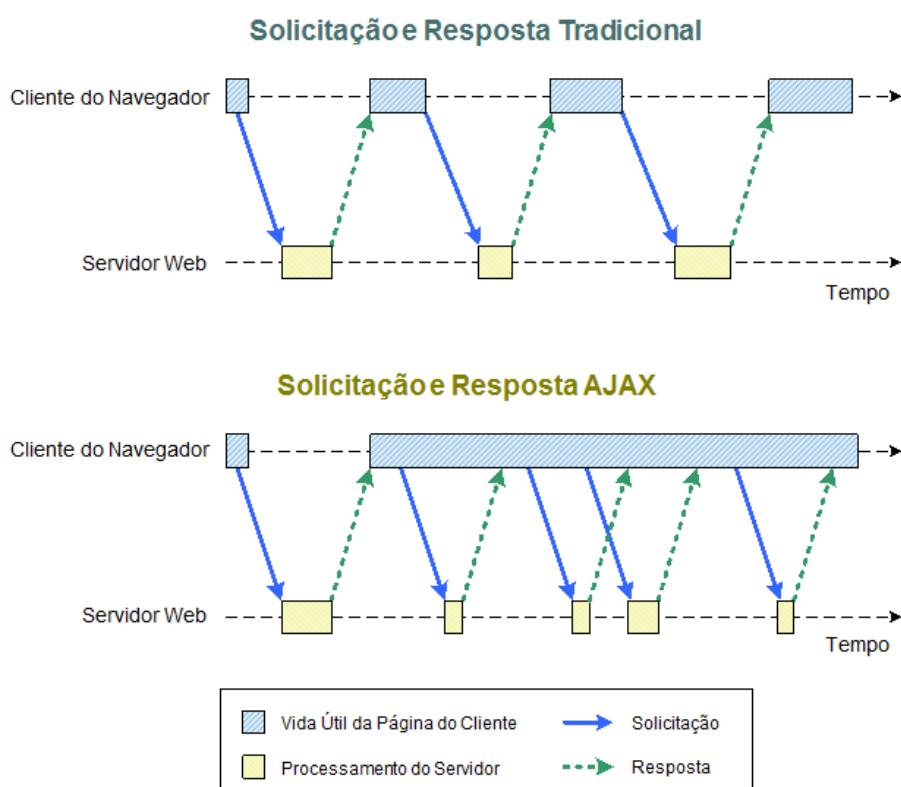


Figura 30: Explicação da ferramenta AJAX. Fonte: Adaptada de Garrett et al. (2005)

3.2.4 BOOTSTRAP

Construir um site ou aplicativo a partir do zero pode dispendar muito tempo e recursos financeiros e humanos. Um *framework* CSS é comumente utilizado por desenvol-

vedores e designers como uma ferramenta para acelerar o processo de criação da parte gráfica da aplicação e algumas funcionalidades a elas atribuídas. Um *framework* CSS é essencialmente um pacote composto de HTML, CSS e JavaScript predefinidos que são utilizados como elementos básicos na criação de sites. Dessa forma, elimina-se a necessidade de criação de milhares de elementos como formulários, barras de navegação, caixas de rolagem do princípio. Tais elementos são criados de forma eficiente e responsiva para se adequar a diferentes tipos de dispositivos, apresentando o conteúdo de forma correta e visualmente agradável para usuários em dispositivos móveis ou em computadores de mesa.

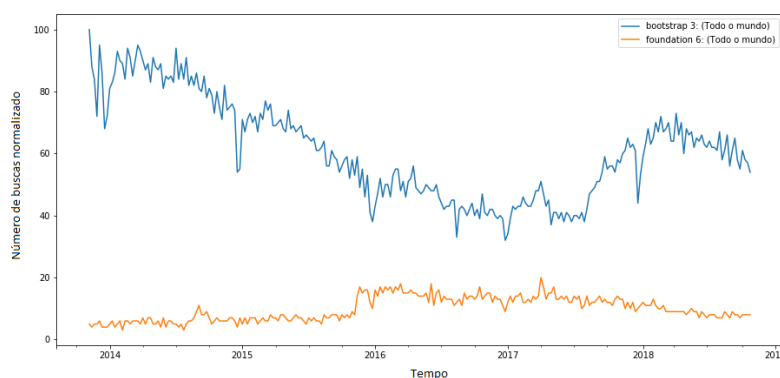


Figura 31: Bootstrap vs Foundation: Pesquisas no Google. Fonte: <https://trends.google.com/>

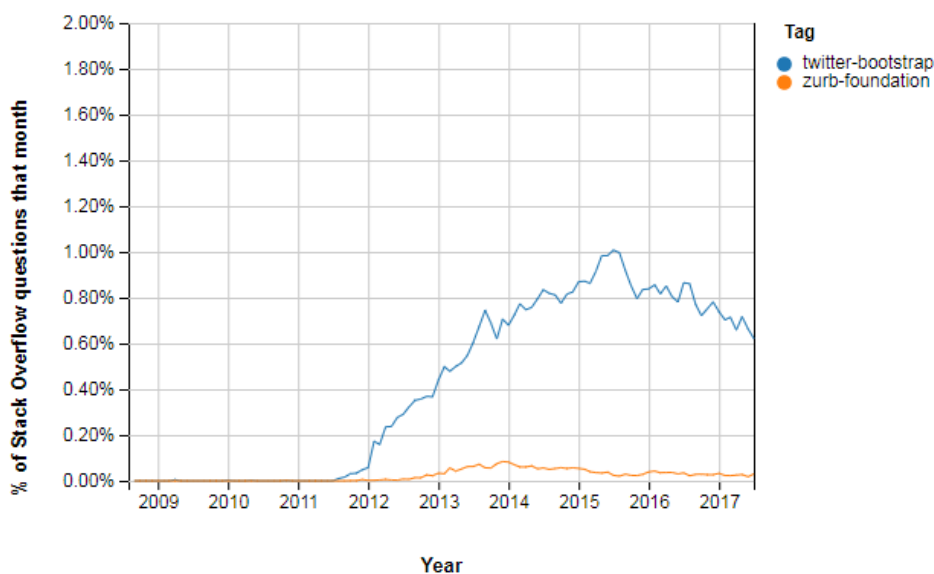


Figura 32: Bootstrap vs Foundation: Perguntas no Stack Overflow. Fonte: <https://insights.stackoverflow.com/trends>

Quanto ao número de opções publicamente disponíveis para auxiliar o desenvolvi-

mento gráfico, este é bem inferior ao observado nas outras aplicações (*Frameworks Backend* e *Front-end*). Atualmente somente dois *Frameworks*, possuem adoção expressiva na comunidade, o Bootstrap e o Foundation. Essa baixa quantidade de opções pode ser atribuída ao fato que *design* de aplicações sempre são baseados nas mesma linguagens (HTML, CSS, JS) e por possuírem uma implementação massante, criando alterações para uma imensa quantidade de elementos HTMLs.

Dentre as duas ferramentas disponíveis publicamente, Bootstrap possui maior aceitação, fato que pode ser constatado pelos gráficos exibidos na Figura 31 e 32. A maior aceitação da ferramenta, reflete em maior disponibilidade de modelos de soluções e páginas, melhor documentação e material suporte. O uso do Bootstrap foi preterível por esse critério.

3.3 CONCLUSÕES PARCIAIS

Nesse capítulo foi descrita a plataforma de pesquisa reproduzível para processamento de sinais aplicado ao SEP. A plataforma possibilita exposição de forma mais transparente de cada elemento relacionado à pesquisa, facilitando sua utilização por terceiros e aumentando seu impacto e utilidade. A integração de elementos como repositório de códigos, banco de dados e análises, permite a execução dos algoritmos na nuvem, a obtenção de resultados e análises facilmente atestadas.

Foram também apresentadas algumas das tecnologias utilizadas no desenvolvimento *Web* e no campo científico. A justificativa para utilização das diferentes tecnologias utilizadas na plataforma se apoiaram em geral na aceitação da linguagem pela comunidade científica e pelas características específicas das linguagens e *frameworks*. A Figura ?? mostra uma visão esquemática do site e também onde cada tecnologia é empregada.

A Tabela 2 mostra um quadro comparativo entre as funcionalidades oferecidas pela plataforma desenvolvida neste trabalho, e os demais recursos em prol da pesquisa reproduzível apresentados no Capítulo 2.

Tabela 2: Comparação entre as funcionalidades dos recursos em prol da RR.

Funcionalidades	VisTrails	Knitr	Jupyter Notebook	RunMyCode	JMLR	Mendely	GigaDB	Code Ocean	Beat Platform	PRESEP
Armazenamento de códigos	x	x	x	✓	✓	△	△	✓	✓	✓
Armazenamento de dados	x	x	x	✓	✓	✓	✓	✓	△	✓
Execução na nuvem	x	x	✓	x	x	x	x	✓	✓	✓
Comparação entre métodos	x	x	x	x	x	x	x	x	✓	✓
Reutilização de BD entre métodos	x	x	x	x	x	x	x	x	✓	✓
Resultados rapidamente atestáveis	✓	✓	✓	x	x	x	x	✓	✓	✓
Acesso online	△	△	△	✓	✓	✓	✓	✓	✓	✓
Específico ao campo	x	x	x	x	✓	x	✓	x	✓	✓

Os símbolos ✓, △ e x significam respectivamente, funcionalidade suportada, funci-

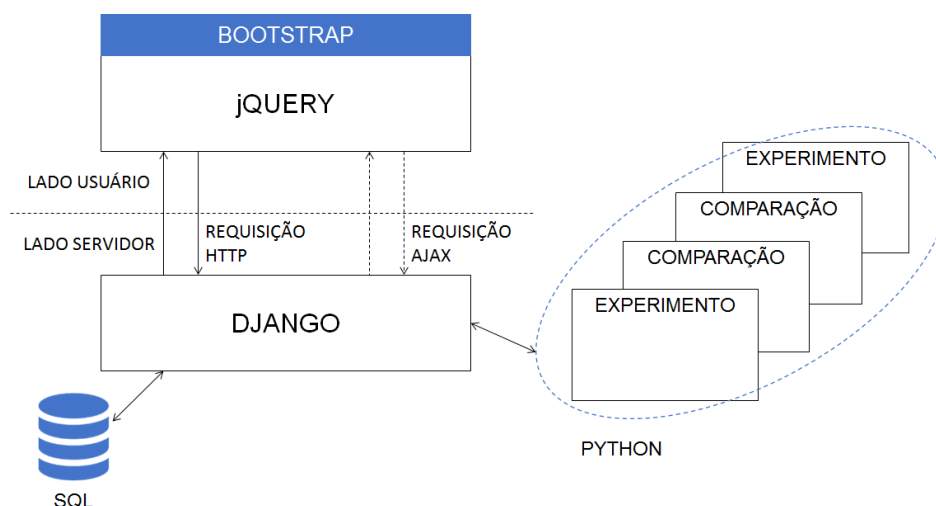


Figura 33: Visão esquemática do site evidenciando as tecnologias utilizadas.

onalidade suportada com ressalvas e não suportada ou não se aplica. Exemplo de funcionalidades suportadas com ressalvas, são as ferramentas Knitr e Jupyter Notebook. Sozinhas, não fornecem a possibilidade de compartilhamento online, as mesmas geram documentos com figuras, tabelas e resultados dinâmicos e atestáveis. Contudo a divulgação do conteúdo por terceiros não oferecida pela ferramenta em si, não cumprindo o propósito da RR. Por esse motivo foi criado domínios com o intuito de divulgação do conteúdo gerado por essas ferramentas. Para compartilhamento de relatórios desenvolvidos pela ferramenta Knitr, existem sites capazes de executarem os relatórios na nuvem como <http://shiny.rstudio.com/> e <https://www.opencpu.org/>. A exposição de resultados gerados através do Jupyter Notebook pode ser realizada através de sites com <https://nbviewer.jupyter.org/>, <https://github.com/> e <https://anaconda.org/>.

O desenvolvimento de uma ferramenta multidisciplinar para disseminação de trabalhos científicos configura uma necessidade atual no âmbito científico. A proposta da plataforma Code Ocean configura uma ótima opção para tal. Porém o desenvolvimento de um recurso que é específico de um campo do conhecimento traz benefícios que não podem ser alcançados de outra forma. Pode-se citar como exemplo a plataforma BEAT que é específica do campo de segurança biométrica, oferecendo banco de dados consolidados, comparação entre métodos, buscas de algoritmos similares, hospedagem, competições entre pesquisadores entre outros benefícios. Por este motivo, a característica de restrição a um campo do conhecimento foi assinalada como positiva.

4 UTILIZAÇÃO DA PRESEP PARA AVALIAÇÃO DE TÉCNICAS PARA DETECÇÃO DE NOVIDADES NO SINAL SEP

Neste capítulo, será utilizada a PRESEP aplicada à avaliação de técnicas para detecção de novidades no sinal do SEP. Um exemplo prático do uso dessa plataforma se faz necessário para melhor compreender as funcionalidades da mesma, adicionalmente, apresenta-se uma contribuição para a área de detecção de novidades, através da comparação de técnicas recentemente propostas na literatura.

4.1 DETECÇÃO DE NOVIDADES

Todos os resultados expostos nesta secção são capturas de telas da PRESEP, mostrando a viabilidade da realização e exposição de pesquisas apresentando resultados e comparações através do uso da plataforma. As análises expostas a seguir estarão divididas em dois tipos de telas com os seguintes conteúdos:

- Gráficos de curvas ROC, em que os valores de AUC são indicados na legenda. Logo abaixo são indicados os resultados referentes ao limiar ótimo com relação ao índice SP, o próprio índice P_{avg} , e as probabilidades de positivo verdadeiro (ou probabilidade de detecção) e falso alarme para o limiar ótimo.
- Gráficos de curvas PR com valores de precisão médios (P_{avg}) indicados na legenda, juntamente com os resultados de limiar, F_{score} , probabilidades de detecção e falso alarme para limiar otimizado em relação ao valores de F_{score} , é ainda apresentado curvas de nível pra diferentes valores de F_{score} .

4.1.1 EXPOSIÇÃO DE RESULTADOS ATRAVÉS DA PRESEP

Para avaliar o desempenho dos detectores, foram gerados mil e quinhentos sinais apresentando distúrbios de igualmente distribuídos entre as determinadas classes: *Sags*, *Swells*, distorções harmônicas, *Notches* e transitórios e igual quantidade de sinais com

ausência de distúrbios para compor balanceado número de contra exemplos para a detecção. Além da adição de distúrbios, é somado ruído gaussiano de fundo em um dos três possíveis níveis de relação sinal ruído, 40 dB, 30 dB e 20dB. A tabela 3 sintetiza as características do banco de dados gerado. Os sinais gerados possuem frequência fundamental de 60 Hz e taxa de amostragem de 7680 Hz. Foi adotada segmentação de tamanho fixo, comportando quatro ciclos da frequência fundamental por *frame*.

Tabela 3: Composição do banco de dados

	Sags	Swells	Distorção Hamônicas	Notches	Transitórios	Sem Distribuídos	Total
40 dB	100	100	100	100	100	500	1000
30 dB	100	100	100	100	100	500	1000
20 dB	100	100	100	100	100	500	1000
Total	300	300	300	300	300	1500	3000

É importante salientar que para cenários pouco ruidosos, segmentos sem distúrbios são praticamente idênticos. Logo, a diferença entre um segmento livre de distúrbios e um deturpado se torna evidente. Por consequência, é esperado que todos os detectores performem de maneira satisfatória. Como as condições de baixo ruído configuram situações mais simples que a demanda de uma aplicação real, todas as abordagens foram comparadas simultaneamente.

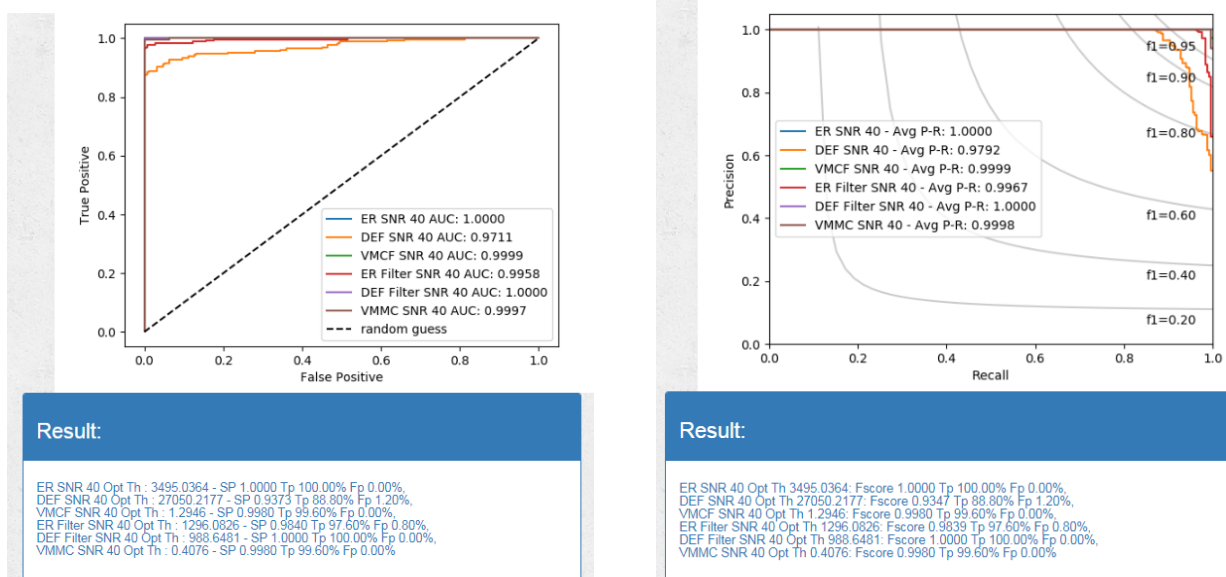
As Figuras 34(a) e 34(b) confirmam essa intuição, mostrando as respostas de todos os métodos de detecção explicados anteriormente e suas variantes que serão explicadas posteriormente. Dentre eles, dois métodos atingiram performance perfeita e somente um obteve performance abaixo de 97%. A tabela 4 recapitula o significado de cada sigla presente nas imagens e nas análises dos resultados a seguir.

Tabela 4: Recapitulação das siglas

ER	Energia do resíduo
DEF	Diferença de energia entre frames
ER Filter	Energia do resíduo após filtragem
DEF Filter	Diferença de energia entre frames após filtragem
VMCF	Varição máxima na magnitude das componentes de frequência
VMMC	Varição máxima em componentes de frequência

MÉTODOS NO DOMÍNIO DO TEMPO

Dentre as métricas de comparação, DEF e ER são baseadas nos valores no domínio temporal. Os resultados para o cenário de médio ruído são expostos nas Figuras 35(a) e

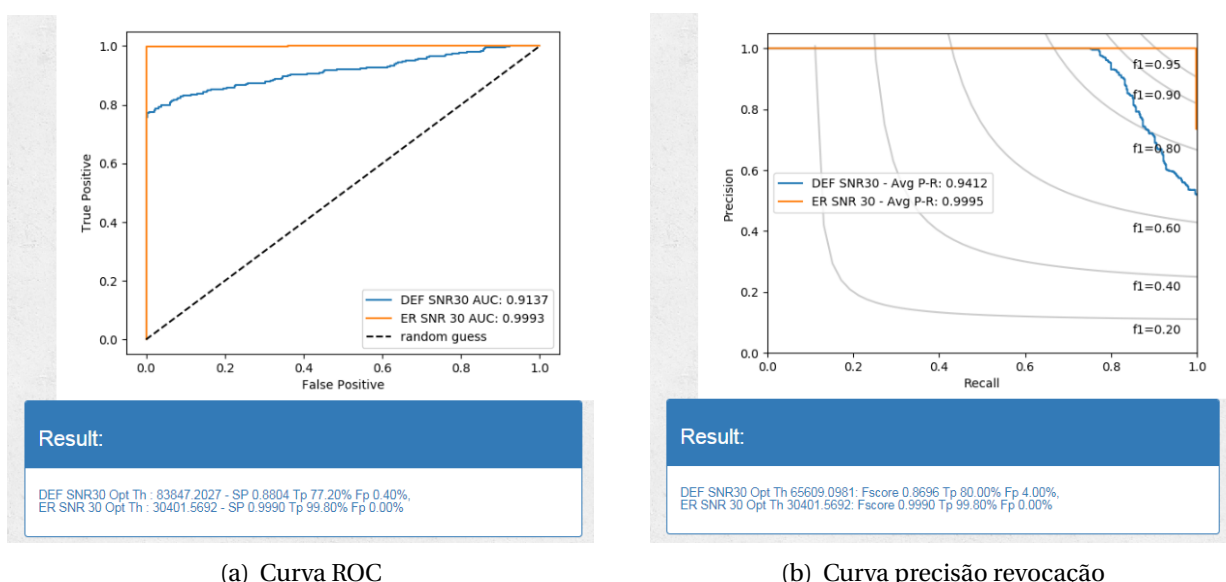


(a) Curva ROC

(b) Curva precisão revocação

Figura 34: Análise classificadores baixo ruído.

35(b). Para esse caso, o detector ER apresenta uma superioridade notável em relação ao DEF. Atingindo probabilidade de detecção de 99,8% sem probabilidade de falso alarme, independente do critério de escolha do patamar. Ao passo que o DEF só é capaz de identificar inovação em 77,2% com probabilidade de falso alarme de 0,4% ou 86,96% com falso alarme de 4%, para os critérios SP e F_{score} , respectivamente.



(a) Curva ROC

(b) Curva precisão revocação

Figura 35: Comparação entre DEF e ER SNR 30.

Em presença de ruídos maiores, a capacidade de detecção é deteriorada nos dois sistemas, como pode ser observado nas Figuras 36(a) e 36(b). Mas o ER ainda mantém ca-

pacidade de discernimento aceitável (93,6% de probabilidade de detecção, 1,6% de probabilidade de falso alarme), os dois critérios convergem na decisão do patamar ótimo. Ao passo que cenários ruidosos inviabilizam o uso da DEF, classificando positivamente inovações em somente 69,6% dos casos.

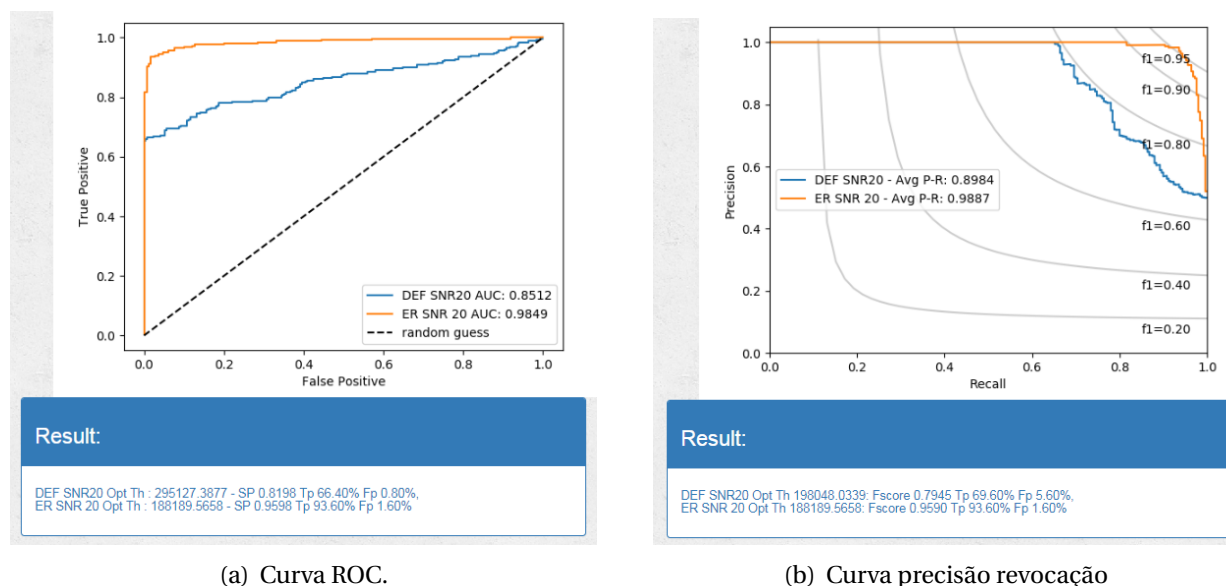
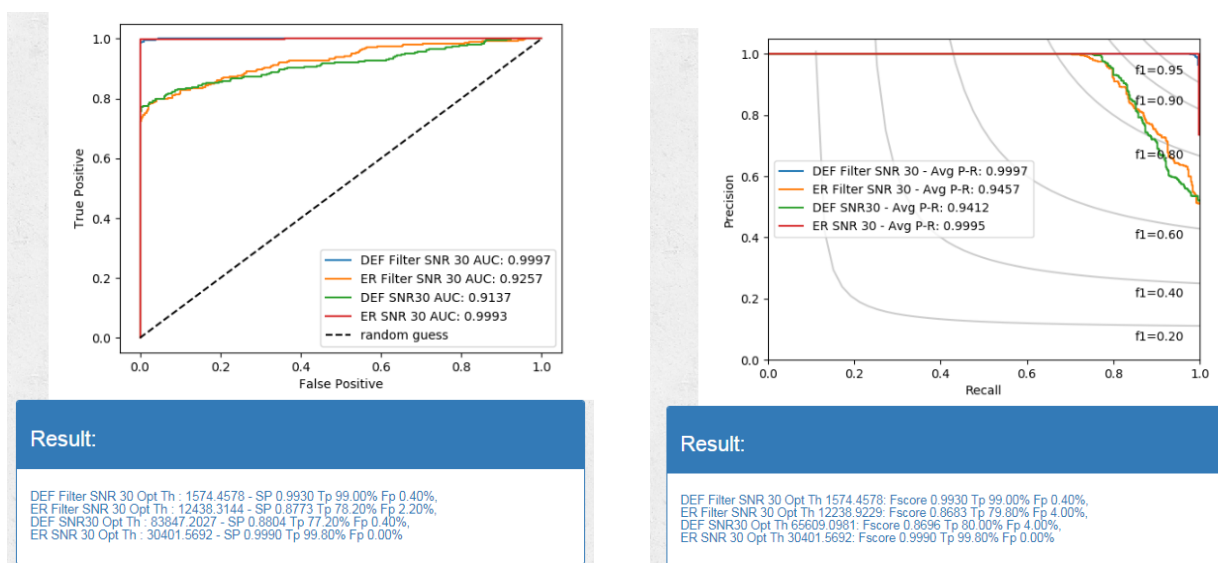


Figura 36: Comparação entre DEF e ER SNR 20.

Em (SILVA, 2016), é apontado que a detecção pode ser beneficiada através de uma filtragem passa-alta do sinal. O emprego de filtragem pode ser justificado pela eliminação de informação redundante (referente a componente de frequência fundamental), eliminação de detecção de pequenas variação na frequência fundamental e pelo entendimento que componentes de frequências mais baixas possuem maior energia do que as frequências mais altas. Identificação de distúrbios que ocorrem em diferentes faixas de frequência (por consequência diferentes valores de energia), através de um mesmo patamar de decisão pode constituir um desafio. Através de uma filtragem passa-alta, distúrbios de frequências baixas só afetam a energia do sinal através de amostras presentes em instantes transitórios do filtro, com energia compatível com as demais inovações.

A filtragem realizada nesse trabalho, segue o projeto usado em (SILVA, 2016). Essa consiste em um filtro Resposta Finita ao Impulso, do inglês *Finite Impulse Response* (FIR) passa-alta de décima ordem, com frequência de corte de 720 Hz. Os resultados dessa abordagem para ruídos SNR 30 dB podem ser observadas nas Figuras 37(a) e 37(b).

O processo de filtragem degrada o funcionamento do método baseado em ER, mas é muito benéfico pra a DEF. Superando sensivelmente o método ER sem filtragem em termos AUC (0,9997 contra 0,9993) e P_{avg} (0,9997 contra 0,9995). Com relação às taxas



(a) Curva ROC.

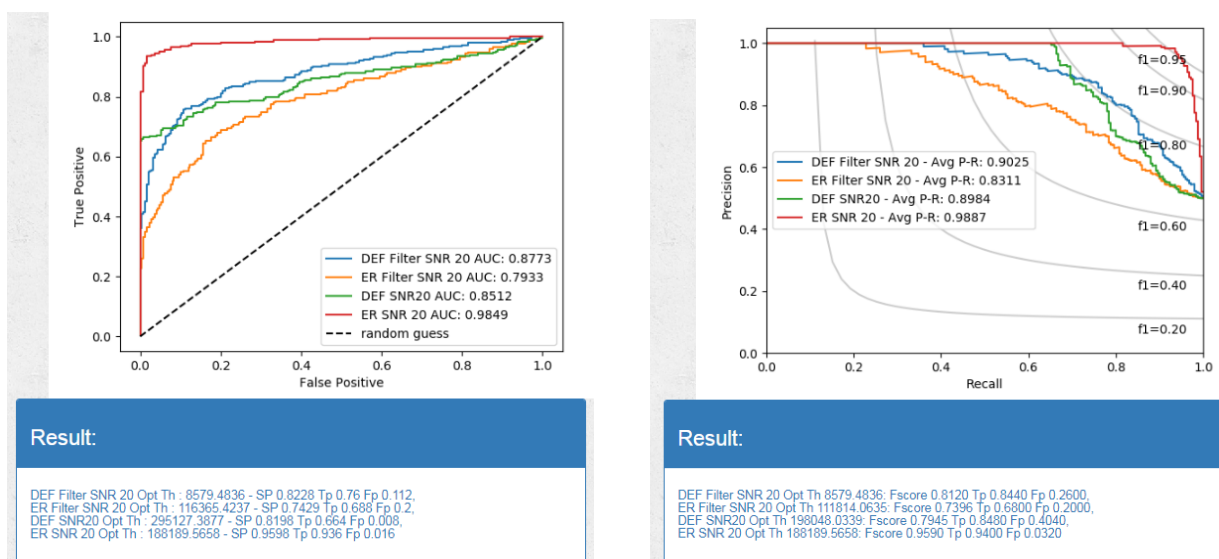
(b) Curva precisão revocação

Figura 37: Comparação entre métodos temporais com filtro SNR 30 dB.

de detecção e falso alarme no limiar ótimo, o método sem filtragem ainda se apresenta ligeiramente superior. Com probabilidades de 99,8% e 0% em oposição a 99% e 0,4%. Novamente os índices SP e F_{score} apontam para o mesmo limiar. Os detetores de menor desempenho (ER com filtro e DEF sem filtro) também apresentam resultados similares entre si. Esse fato pode ser percebido graficamente pela similaridade entre as curvas ROC e PR e pela confirmação quantitativa AUC, obtendo valores de 0,9257 e 0,9137, e P_{avg} obtendo 0,9457 e 0,9412.

Através da análise de ruído igual a 30 dB não é possível determinar qual abordagem temporal é preferível, pois as duas apresentam resultados semelhantes. Logo, a comparação em condições mais extremas se faz necessária. As Figuras 38(a) e 38(b) mostram os resultados em altíssima presença de ruídos (SNR 20 dB). Neste caso o desempenho dos detetores divergem, apontando para um sistema mais adequado.

Analisando a evolução do desempenho dos sistemas em relação à diminuição da SNR, pode ser constatado que o processo de filtragem de baixas frequências, introduz perda de robustez na capacidade de detecção, conforme pode ser atestado visualmente e pelos resultados numéricos. A capacidade de detecção DEF com filtro no patamar ótimo diminui de 99% para 76% e a probabilidade de falso alarme aumenta de 0,4% para 11,2%. A diferença no desempenho da detecção DEF sem filtro é menor, decaindo 10,4% na probabilidade de detecção e 1,6% na probabilidade de falso alarme, ao se usar o critério SP na escolha do patamar.



(a) Curva ROC.

(b) Curva precisão revocação

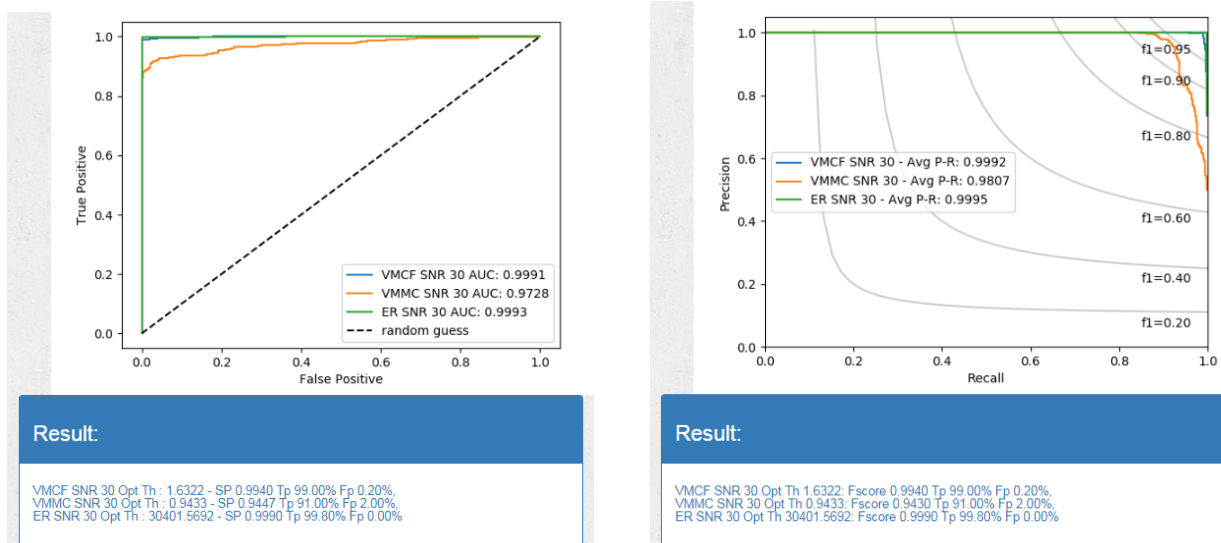
Figura 38: Comparação entre métodos temporais com filtro SNR 20.

A diferença é ainda maior ao comparar os métodos ER e ER Filter. Sem a presença do filtro o detector (ER) mantém um desempenho aceitável, com taxa de positivos verdadeiros de 93,6%, caindo somente 6,2%. O uso do filtro na ER (ER Filter), detector já com baixo desempenho em SNR 30 dB, não provoca muita alteração na taxa de detecção (decaimento de 79,8% para 78,4%), entretanto impacta de forma considerável na razão de falso alarme, elevando a taxa de 4% para 33,6%.

MÉTODOS NO DOMÍNIO DA FREQUÊNCIA

Além das métricas baseadas em valores temporais, a capacidade de detecção baseada no domínio da frequência também foi avaliada. A comparação entre os métodos espectrais será realizada em conjunto com o melhor classificador avaliado anteriormente, o ER sem filtro. As Figuras 39(a) e 39(b) expõem resultados referentes a sinais com SNR 30 dB.

A utilização da magnitude das frequências como métrica de comparação não apresenta resultados promissores. Apresentado correta classificação em somente 91% dos casos de médio ruído, onde a maioria dos detectores apresentam bom desempenho. No entanto, o uso dos valores complexos das componentes espectrais apresentam resultados comparáveis com o ER. O método ER supera apenas ligeiramente o método VMCF em relação às métricas gerais adotadas (AUC: 0,9993 contra 0,9991 e P_{avg} : 0,9992 contra 0,9995) e específicas (probabilidade de detecção 99,8% contra 99,8% e probabilidade de falso alarme 0,2% contra a inexistência de falso alarme).

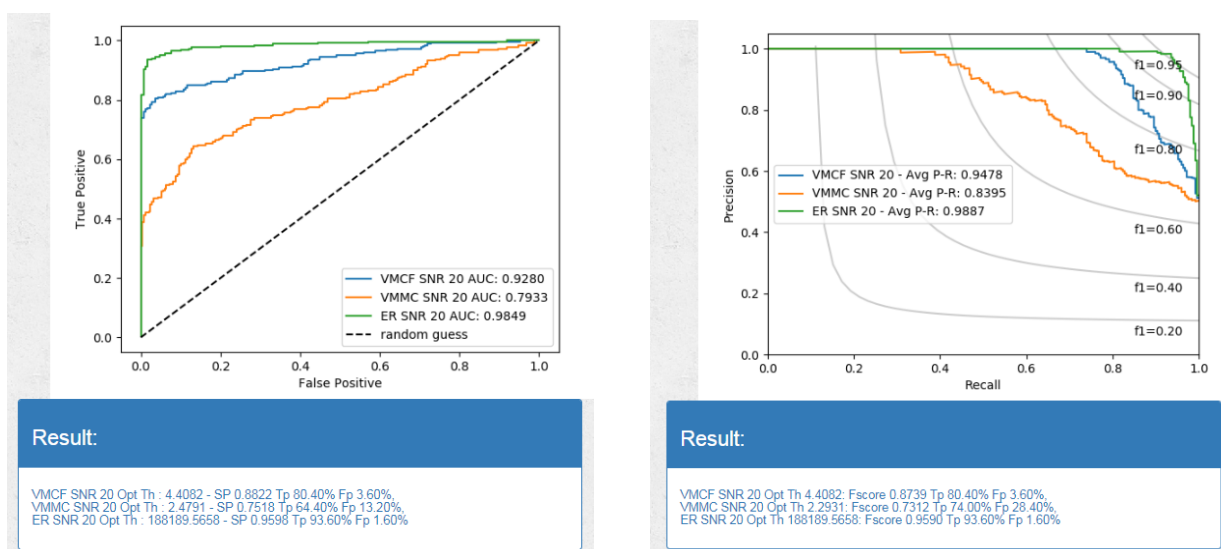


(a) Curva ROC.

(b) Curva precisão revocação

Figura 39: Comparação entre métodos espectrais SNR 30.

Na presença de altíssimo ruído no sinal (SNR 20dB), o desempenho da VMCF decai mais rapidamente que ER. Como pode ser observado nas Figuras 40(a) e 40(b). Não atingindo resultados satisfatórios, com apenas 84,4% de probabilidade de detecção e 3,6% de falso alarme. Enquanto, o melhor detector no domínio temporal (ER), atinge probabilidades de 93,6% e 1,6%.



(a) Curva ROC.

(b) Curva precisão revocação

Figura 40: Comparação entre métodos espectrais SNR 20 dB.

Já o uso de VMMC possui capacidade discriminativa comparáveis a decisão aleatória (50% para decisão binária com exemplos balanceados), quando o critério de escolha do limiar é o SP (64,4%). A capacidade de detecção é melhor quando é considerado critério

F_{score} , atingindo 74%. Em contra partida, a taxa de falso alarme aumenta de 13,2% para 28,4%. Isso reitera a hipótese de não adequação dos valores VMCM para identificação de novidades.

4.2 CONCLUSÕES PARCIAIS

A Tabela 5 sintetiza os resultados adquiridos pelos sistemas expostos acima. O uso de métricas temporais se mostrou vantajosa em relação às espectrais para o conjunto de dados analisado. Fato que corrobora com os resultados encontrados em (SILVA, 2016), onde o sistema de monitoramento baseado no domínio do tempo obtém resultados comparáveis ou superiores em comparação com sistemas baseados no domínio da frequência.

Tabela 5: Resultados de probabilidade de detecção e falso alarme.

	DEF		ER		DEF Filter		ER Filter		VMCF		VMCM	
	Pv(%)	Fa(%)	Pv(%)	Fa(%)	Pv(%)	Fa(%)	Pv(%)	Fa(%)	Pv(%)	Fa(%)	Pv(%)	Fa(%)
SNR 40 F_{score}	88,8	1,2	100,0	0,0	100,0	0,0	97,6	0,8	99,6	0,0	99,6	0,0
SNR 40 SP	88,8	1,2	100,0	0,0	100,0	0,0	97,6	0,8	99,6	0,0	99,6	0,0
SNR 30 F_{score}	80,0	4,0	99,8	0,0	99,0	0,4	79,8	4,0	99,0	0,2	91,0	2,0
SNR 30 SP	77,2	0,4	99,8	0,0	99,0	0,4	78,2	2,2	99,0	0,2	91,0	2,0
SNR 20 F_{score}	69,6	5,6	93,6	1,6	76,0	11,2	78,4	33,6	80,4	3,6	74,0	28,4
SNR 20 SP	66,4	8,0	93,6	1,6	76,0	11,2	68,8	20,0	80,4	3,6	64,4	13,2

No que concerne à escolha do limiar ótimo, na maioria dos casos, os dois índices apontam para um mesmo valor. Para casos onde há divergência nessa seleção, o uso do índice SP é preferível. Apesar de uma perda sensível na probabilidade de detecção, a probabilidade de falso alarme é reduzida quando se comparara com os resultados obtidos baseados na medida F_{score} . Isso é decorrente do fato da medida F_{score} buscar a melhor relação entre a classificação das duas classes envolvidas (acusar inovação, acusar normalidade), ao passo que o índice SP maximiza somente o desempenho de acusação de novidades.

Embora o desempenho do uso da energia do resíduo sem filtragem tenha sido igualado ou até mesmo superado sensivelmente quando consideradas métricas de comportamento geral do sistema (AUC e P_{avg}), sintetizados na Tabela 6. Esse apresentou resultados mais consistentes em diferentes situações de ruído, sendo o único capaz de atingir resultados superiores a 90% de probabilidade de detecção em presença de SNR 20 dB e valores de AUC e P_{avg} iguais a 0,9849 e 0,9887.

Os detectores foram testados para situações severas de ruídos (até SNR 20 dB). Entretanto, como especificado na norma (SMITH; HENSLEY & RAY, 1995), os níveis de ruído em sistemas elétricos de potência estão compreendidos na faixa de zero a um por cento da

Tabela 6: Resultados de área embaixo do gráfico e precisão média.

Funcionalidades	DEF		ER		DEF Filter		ER Filter		VMCF		VMMC	
	AUC	P_{avg}	AUC	P_{avg}	AUC	P_{avg}	AUC	P_{avg}	AUC	P_{avg}	AUC	P_{avg}
SNR 40	0,9711	0,9792	1,0000	1,0000	1,0000	1,0000	0,9958	0,9967	0,9999	0,9999	0,9997	0,9998
SNR 30	0,9137	0,9412	0,9993	0,9995	0,9997	0,9997	0,9257	0,9457	0,9991	0,9992	0,9728	0,9807
SNR 20	0,8512	0,8984	0,9849	0,9887	0,8773	0,9025	0,7933	0,8311	0,9280	0,9478	0,7933	0,8395

frequência fundamental, o que corresponde à SNR superior a 40 dB. Logo, para a maioria das aplicação reais os métodos DEF Filter e VMCF constituem soluções válidas.

É possível notar também que a adoção de um limiar de decisão estático não satisfaz as necessidades do sistema de decisão. Tomando os valores de patamar ótimo de melhor detector, o ER sem filtro. Os valores do patamar diferem muito para situações diferentes de ruídos (SNR 20 dB: 188189.56, SNR 30 dB: 30401.56, SNR 40dB: 3495.03). Uma seleção de patamar dinâmica baseada no nível estimado de ruído garantirá um melhor desempenho.

5 CONCLUSÕES

O presente trabalho apresentou um ambiente virtual que possibilita o armazenamento e compartilhamento de pesquisa de forma reprodutível. O conceito de pesquisa reprodutível ganhou atenção da comunidade científica nas últimas décadas, devido a observância das dificuldades inerentes à ciência computacional, quanto a disseminação de novas técnicas. Uma vez que os veículos de informação científicos tradicionais como artigos de jornais, revistas e congressos, nem sempre conseguem compreender todos os elementos necessários para o total entendimento por parte dos leitores e averiguação por parte de revisores.

Diversos jornais e revistas vêm estreitando seus requerimentos no que diz respeito a reprodutibilidade e divulgação de elementos essenciais da pesquisa, como códigos e dados. Diversas ferramentas gerais e específicas foram desenvolvidas ao longo dos anos para apoiar e facilitar uma mudança no paradigma tradicional de disseminação de conhecimento. Porém, o panorama atual de aplicações dedicadas ao desenvolvimento de pesquisas de forma reprodutível ainda é heterogêneo e limitado para diversas áreas do conhecimento, o que inabilita a cultura de reprodutibilidade ser tão forte quanto deveria.

Pode ser observado também, que a reunião de pesquisas de um campo específico em um único repositório apresenta um grandioso potencial didático para pesquisadores e estudiosos da área. Esse potencial é maximizado quando as pesquisas estão expostas seguindo os conceitos de reprodutibilidade. Observada a carência de um repositório de pesquisas na área de processamento de sinais para o sistema elétrico de potência, um campo imprescindível para manutenção e evolução dos demais segmentos da sociedade, a proposta do presente trabalho possui um grande potencial de fomento para a área.

A plataforma fruto desse trabalho possibilita a reunião e exposição de pesquisas relacionadas ao processamento de sinais para o sinal do SEP e também seus elementos constituintes de forma separada, através da associação de repositório de códigos com *links* vinculados a referente publicação, possibilitando a rápida utilização mesmo em cenários que a replicação por intermédio somente da leitura do artigo é possível. A plataforma

também oferece a possibilidade de armazenamento e exibição de conjuntos de dados, facilitando a realização e comparação de pesquisas usando uma base de dados comum.

Os elementos citados anteriormente podem ser associados para criação de experimentos que são executados no servidor, gerando resultados atestáveis para os autores sem a necessidade de *download* e instalação de arquivos e programas. Testar diferentes abordagens com o mesmo propósito é a maneira mais eficiente de consolidar e padronizar técnicas. Neste sentido, foi apresentada a capacidade de comparar diferentes experimentos incorporados a plataforma de forma fácil, rápida e confiável, auxiliando para que o processo de desenvolvimento científico seja realizado de forma mais coesa.

As possibilidades apontadas acima foram validadas pela apresentação de um estudo comparativo de diferentes técnicas de detecção de inovação em sinais do SEP. Duas técnicas utilizam o domínio do tempo e outras duas o domínio da frequência. A comparação realizada através da plataforma, se apoiou no mesmo banco de dados e utilizou as mesmas métricas para análise de desempenho. Foi concluído que para os distúrbios analisados e sinais com relação sinal ruído superior a 30 dB é possível extrair características capazes de identificar inovação tanto no domínio do tempo quanto na frequência. E foi evidenciado também que para maiores níveis de ruído a análise temporal se tornou vantajosa, através do uso da métrica da energia do resíduo entre *frames*.

5.1 TRABALHOS FUTUROS

Através de uma análise crítica sobre o estado atual da plataforma é possível apontar alguns pontos que podem abranger e facilitar o uso da mesma:

1. Adicionar suporte para execução de algoritmos em outras linguagens.
2. Atualmente, as análises são realizadas em no máximo dois níveis (Experimento e Comparação). Possibilitar a execução de experimentos em blocos, sem restrições quanto quantidade e maneira de conexão (série, paralelo). Permitirá melhor modularização das rotinas envolvidas, isolando tarefas como pre-processamento, extração de características, processamento, avaliação etc.
3. Melhorar a interface do usuário de forma a se tornar mais intuitiva ao utilizador da plataforma.
4. Adicionar suporte de execução de rotinas com multiprocessadas em Unidade de Processamento Gráfico, do inglês *Graphics Processing Unit* (GPU).

5. Consolidar o uso da plataforma internamente pelos pesquisadores de PSCOPE, como realização de teste *beta*, para posterior abrangência ao resto da comunidade científica.

REFERÊNCIAS

- ABBOTT, B. P. et al. Observation of gravitational waves from a binary black hole merger. *Physical review letters*, APS, v. 116, n. 6, p. 061102, 2016.
- ALSHAHRANI, S.; ABBOD, M.; TAYLOR, G. Detection and classification of power quality disturbances based on hilbert-huang transform and feed forward neural networks. In: IEEE. *Power Engineering Conference (UPEC), 2016 51st International Universities*. 2016. p. 1–6.
- AZAMI, H. et al. An intelligent approach for variable size segmentation of non-stationary signals. *Journal of advanced research*, Elsevier, v. 6, n. 5, p. 687–698, 2015.
- BAKER, M. Reproducibility crisis? *Nature*, v. 533, p. 26, 2016.
- BOLLEN, M. H.; GU, I. Y. *Signal processing of power quality disturbances*. : John Wiley & Sons, 2006.
- BUCKHEIT, J. B.; DONOHO, D. L. Wavelab and reproducible research. In: *Wavelets and statistics*. : Springer, 1995. p. 55–81.
- CHAKRAVORTI, T.; PATNAIK, R. K.; DASH, P. K. Advanced signal processing techniques for multiclass disturbance detection and classification in microgrids. *IET Science, Measurement & Technology*, IET, v. 11, n. 4, p. 504–515, 2017.
- CHAMBERS, C. *Registered reports: A step change in scientific publishing*. : Elsevier. Retrieved from <http://www.elsevier.com/reviewers-update/story/innovation-in-publishing/registered-reports-a-step-change-in-scientific-publishing>, 2014.
- COOLEY, J. W.; TUKEY, J. W. An algorithm for the machine calculation of complex fourier series. *Mathematics of computation*, JSTOR, v. 19, n. 90, p. 297–301, 1965.
- COPI, I.; COHEN, C.; FLAGE, D. *Essentials of logic*. : Taylor & Francis, 2016.
- DE, S.; DEBNATH, S. Real-time cross-correlation-based technique for detection and classification of power quality disturbances. *IET Generation, Transmission & Distribution*, IET, v. 12, n. 3, p. 688–695, 2017.
- DHEERU, D.; TANISKIDOU, E. K. *UCI Machine Learning Repository*. 2017. Disponível em: <<http://archive.ics.uci.edu/ml>>.
- DONOHO, D. L. et al. Reproducible research in computational harmonic analysis. *Computing in Science & Engineering*, IEEE, v. 11, n. 1, 2009.
- DRUMMOND, C.; HOLTE, R. C. Explicitly representing expected cost: An alternative to roc representation. In: ACM. *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2000. p. 198–207.

- DUQUE, C. A. et al. Power quality event detection based on the divide and conquer principle and innovation concept. *IEEE Transactions on Power Delivery*, IEEE, v. 20, n. 4, p. 2361–2369, 2005.
- DWIVEDI, U.; SINGH, S. Denoising techniques with change-point approach for wavelet-based power-quality monitoring. *IEEE transactions on power delivery*, IEEE, v. 24, n. 3, p. 1719–1727, 2009.
- ELSPEC, E. L. *Elspec G4000 BLACKBOX*. [S.l.]. 2007.
- FILHO, E. F. S.; SEIXAS, J. M.; CALÔBA, L. P. Optimized calorimeter signal compaction for an independent component based atlas electron/jet second-level trigger. *PoS*, p. 102, 2008.
- FREIRE, J.; BONNET, P.; SHASHA, D. Computational reproducibility: state-of-the-art, challenges, and database research opportunities. In: ACM. *Proceedings of the 2012 ACM SIGMOD international conference on management of data*. 2012. p. 593–596.
- FREIRE, J. et al. Reproducibility using vistrails. *Implementing Reproducible Computational Research*, Chapman & Hall/CRC Boca Raton, FL, v. 33, 2014.
- FREIRE, J. et al. Provenance for computational tasks: A survey. *Computing in Science & Engineering*, IEEE, v. 10, n. 3, 2008.
- GAOUDA, A. et al. Power quality detection and classification using wavelet-multiresolution signal decomposition. *IEEE Transactions on power delivery*, IEEE, v. 14, n. 4, p. 1469–1476, 1999.
- GARRETT, J. J. et al. Ajax: A new approach to web applications. 2005.
- GILBERT, K. J. et al. Recommendations for utilizing and reporting population genetic analyses: the reproducibility of genetic clustering using the program structure. *Molecular Ecology*, Wiley Online Library, v. 21, n. 20, p. 4925–4930, 2012.
- GONZALEZ, M.; CARDENAS, V.; ALVAREZ, R. Detection of sags, swells, and interruptions using the digital rms method and kalman filter with fast response. In: IEEE. *IEEE Industrial Electronics, IECON 2006-32nd Annual Conference on*. 2006. p. 2249–2254.
- GU, I. Y.; BOLLEN, M.; LE, C. D. Signal processing and classification tools for intelligent distributed monitoring and analysis of the smart grid. In: *IEEE PES Innovative Smart Grid Technologies Conference Europe. 2011 2nd IEEE PES International Conference and Exhibition on Innovative Smart Grid Technologies, ISGT Europe 2011, Manchester, 5-7 December 2011*. 2011.
- GU, Y. H.; BOLLEN, M. H. Time-frequency and time-scale domain analysis of voltage disturbances. *IEEE Transactions on Power Delivery*, IEEE, v. 15, n. 4, p. 1279–1284, 2000.
- HEYDT, G. et al. Applications of the windowed fft to electric power quality assessment. *IEEE Transactions on Power Delivery*, IEEE, v. 14, n. 4, p. 1411–1416, 1999.
- HURLIN, C. et al. Runmycode. org: A research-reproducibility tool for computational sciences. *Implementing reproducible research*. CRC Press, Boca Raton, FL, p. 367–381, 2014.

- IOANNIDIS, J. P. Why most published research findings are false. *PLoS medicine*, Public Library of Science, v. 2, n. 8, p. e124, 2005.
- JIANG, Z. et al. A vision of smart transmission grids. In: IEEE. *Power & Energy Society General Meeting, 2009. PES'09. IEEE*. 2009. p. 1–10.
- KAPISCH, E. B. et al. An electrical signal disturbance detector and compressor based on fpga platform. In: IEEE. *Harmonics and Quality of Power (ICHQP), 2014 IEEE 16th International Conference on*. 2014. p. 278–282.
- KENALL, A. et al. *Better reporting for better research: a checklist for reproducibility*. : BioMed Central, 2015.
- KERR, N. L. Harking: Hypothesizing after the results are known. *Personality and Social Psychology Review*, Sage Publications Sage CA: Los Angeles, CA, v. 2, n. 3, p. 196–217, 1998.
- KLUYVER, T. et al. Jupyter notebooks-a publishing format for reproducible computational workflows. In: *ELPUB*. 2016. p. 87–90.
- KNUTH, D. E. Literate programming. *The Computer Journal*, Oxford University Press, v. 27, n. 2, p. 97–111, 1984.
- KRASNER, G. E.; POPE, S. T. et al. A description of the model-view-controller user interface paradigm in the smalltalk-80 system. *Journal of object oriented programming*, v. 1, n. 3, p. 26–49, 1988.
- LAZZARETTI, A. E.; FERREIRA, V. H.; NETO, H. V. New trends in power quality event analysis: novelty detection and unsupervised classification. *Journal of Control, Automation and Electrical Systems*, Springer, v. 27, n. 6, p. 718–727, 2016.
- LEISCH, F. Sweave: Dynamic generation of statistical reports using literate data analysis. In: SPRINGER. *Compstat*. 2002. p. 575–580.
- LI, F. et al. Smart transmission grid: Vision and framework. *IEEE transactions on Smart Grid*, IEEE, v. 1, n. 2, p. 168–177, 2010.
- LIN, C.-H.; TSAO, M.-C. Power quality detection with classification enhancible wavelet-probabilistic network in a power system. *IEE Proceedings-Generation, Transmission and Distribution*, IET, v. 152, n. 6, p. 969–976, 2005.
- MARCEL, S. Beat-biometrics evaluation and testing. *Biometric technology today*, Elsevier, v. 2013, n. 1, p. 5–7, 2013.
- MCCULLOUGH, B. D.; MCGEARY, K. A.; HARRISON, T. D. Lessons from the jmcb archive. *Journal of Money, Credit and Banking*, JSTOR, p. 1093–1107, 2006.
- MCNUTT, M. *Journals unite for reproducibility*. : American Association for the Advancement of Science, 2014.
- MEENA, P.; RAO, K. U.; RAVISHANKAR, D. A modified simple algorithm for detection of voltage sags and swells in practical loads. In: IEEE. *Power Systems, 2009. ICPS'09. International Conference on*. 2009. p. 1–6.

- MONTEIRO, D. do A.; ZVIETCOVICH, W. G.; BRAGA, M. F. Detection and classification of power quality disturbances with wavelet transform, decision tree algorithm and support vector machines. In: IEEE. *2018 Simposio Brasileiro de Sistemas Eletricos (SBSE)*. 2018.
- MOXHAM, N.; FYFE, A. The royal society and the prehistory of peer review, 1665–1965. *The Historical Journal*, Cambridge University Press, p. 1–27, 2017.
- NAGATA, E. A. et al. Voltage sag and swell detection and segmentation based on independent component analysis. *Electric Power Systems Research*, Elsevier, v. 155, p. 274–280, 2018.
- PENG, R. D. Reproducible research in computational science. *Science*, American Association for the Advancement of Science, v. 334, n. 6060, p. 1226–1227, 2011.
- PENG, R. D.; ECKEL, S. P. Distributed reproducible research using cached computations. *Computing in Science & Engineering*, IEEE, v. 11, n. 1, p. 28–34, 2009.
- PÉREZ, F.; GRANGER, B. E. Ipython: a system for interactive scientific computing. *Computing in Science & Engineering*, IEEE, v. 9, n. 3, 2007.
- PERKEL, J. M. Data visualization tools drive interactivity and reproducibility in online publishing. *Nature*, v. 554, n. 7690, p. 133–134, 2018.
- PIMENTEL, M. A. et al. A review of novelty detection. *Signal Processing*, Elsevier, v. 99, p. 215–249, 2014.
- RIBEIRO, E. G. et al. Real-time system for automatic detection and classification of single and multiple power quality disturbances. *Measurement*, Elsevier, v. 128, p. 276–283, 2018.
- RIJSBERGEN, C. V. Information retrieval. dept. of computer science, university of glasgow. URL: citeseer.ist.psu.edu/vanrijsbergen79information.html, v. 14, 1979.
- RUPAL, H. S. et al. Detection and classification of power quality disturbances using signal processing techniques. In: IEEE. *Asia-Pacific Power and Energy Engineering Conference (APPEEC), 2017 IEEE PES*. 2017. p. 1–6.
- SANTOSO, S. et al. Characterization of distribution power quality events with fourier and wavelet transforms. *IEEE Transactions on Power Delivery*, IEEE, v. 15, n. 1, p. 247–254, 2000.
- SANTOSO, S. et al. Power quality assessment via wavelet transform analysis. *IEEE transactions on Power Delivery*, v. 11, n. 2, p. 924–930, 1996.
- SILVA, L. et al. Gapless power-quality disturbance recorder. *IEEE Transactions on Power Delivery*, IEEE, v. 32, n. 2, p. 862–871, 2017.
- SILVA, L. R. M. *Estudo e implementação de técnicas de detecção e compressão de distúrbios elétricos*. Tese (Doutorado) — Universidade Federal de Juiz de Fora, 2016.
- SILVA, L. R. M.; DUQUE, C. A.; RIBEIRO, P. F. Smart signal processing for an evolving electric grid. *EURASIP Journal on Advances in Signal Processing*, Nature Publishing Group, v. 2015, n. 1, p. 44, 2015.

- SILVA, L. R. M. et al. The concept of novelty detection applied to power quality. In: IEEE. *Power and Energy Society General Meeting (PESGM), 2016*. 2016. p. 1–5.
- SIMMONS, J. P.; NELSON, L. D.; SIMONSOHN, U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, Sage Publications Sage CA: Los Angeles, CA, v. 22, n. 11, p. 1359–1366, 2011.
- SMITH, J. C.; HENSLEY, G.; RAY, L. Ieee recommended practice for monitoring electric power quality. *IEEE Std*, p. 1159–1995, 1995.
- SNEDDON, T. P.; LI, P.; EDMUNDS, S. C. *GigaDB: announcing the GigaScience database*. : BioMed Central, 2012.
- SOKOLOVA, M.; JAPKOWICZ, N.; SZPAKOWICZ, S. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In: SPRINGER. *Australasian joint conference on artificial intelligence*. 2006. p. 1015–1021.
- STODDEN, V.; HURLIN, C.; PÉRIGNON, C. Runmycode. org: a novel dissemination and collaboration platform for executing published computational results. In: IEEE. *E-Science (e-Science), 2012 IEEE 8th International Conference on*. 2012. p. 1–8.
- TREES, H. L. V. *Detection, estimation, and modulation theory, part I: detection, estimation, and linear modulation theory*. : John Wiley & Sons, 2004.
- UDDIN, Z. et al. Recent advances of the signal processing techniques in future smart grids. *Human-centric Computing and Information Sciences*, SpringerOpen, v. 8, n. 1, p. 2, 2018.
- VANDEWALLE, P.; KOVACEVIC, J.; VETTERLI, M. Reproducible research in signal processing. *IEEE Signal Processing Magazine*, IEEE, v. 26, n. 3, 2009.
- VITORINO, A. et al. Evaluating the reproducibility of multiagent systems. In: IEEE. *World Engineering Education Conference (EDUNINE), IEEE*. 2017. p. 66–70.
- XIANGXUN, C. Wavelet-based detection, localization, quantification and classification of short duration power quality disturbances. In: IEEE. *Power Engineering Society Winter Meeting, 2002. IEEE*. 2002. v. 2, p. 931–936.
- XIE, Y. knitr: a comprehensive tool for reproducible research in r. *Implement Reprod Res*, v. 1, p. 20, 2014.
- YANG, H.-T.; LIAO, C.-C. A de-noising scheme for enhancing wavelet-based power quality monitoring system. *IEEE Transactions on Power Delivery*, IEEE, v. 16, n. 3, p. 353–360, 2001.

APÊNDICE A -- PRODUÇÃO CIENTÍFICA

“Plataforma de Pesquisa Reprodutível para Qualidade de Energia Elétrica”. PIAZZI, A. C.; CERQUEIRA, A. S.; SILVA, L. R. M.; DUQUE, T. F.; RIBEIRO, P. F, XII Conferência Brasileira sobre Qualidade de Energia Elétrica - CBQEE2017, Curitiba, 2017.

Resumo - O conceito de pesquisa reprodutível vem ganhando importância nos últimos anos. Pesquisas facilmente reproduzíveis podem ter resultados atestados e sua pesquisa avançada por outros pesquisadores. Para atingir tal objetivo, diversas etapas devem ser seguidas, como a divulgação do código, divulgação do banco de dados usado, parâmetros escolhidos, softwares e versões utilizados na pesquisa, entre outras. O presente trabalho, a fim de criar uma ferramenta para agilizar a difusão do conhecimento, propõem uma plataforma web para pesquisas reprodutível de rotinas para a Qualidade de Energia Elétrica (QEE).

"Reproducible research platform for electric power quality algorithms". PIAZZI, A. C., CERQUEIRA, A. S.; and MANSO, L. R; DUQUE, C. A, ICHQP - *18th International Conference on Harmonics and Quality of Power*. IEEE, 2018

Abstract - Scientific computation is emerging as the central pillar of science. However, loosed research practices and the own nature of the work is leading to a credibility crisis. Reproducible computational research is gaining attraction in the last decades not only because it benefits the validation and attestation of findings, making a more trustworthy publication, but also lowering the barrier of entry, making it more understandable for the target audience and, therefore, increasing its impact. In this article, an online platform where it is possible to store, share and execute linked code and data, from signal processing, applied to electric power quality, is presented. In such manner, this article presents an alternative for researchers in the field to conduct and share their work satisfying the requirements of reproducible research.