MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

sid.inpe.br/mtc-m21d/2023/01.02.20.29-TDI

APLICAÇÃO DE TÉCNICAS DE MACHINE LEARNING NO ESTUDO DE TRANSIENTES DOS DETECTORES ADVANCED LIGO

Tábata Aira Ferreira

Tese de Doutorado do Curso de Pós-Graduação em Astrofísica, orientada pelo Dr. César Augusto Costa, aprovada em 09 de dezembro de 2022.

URL do documento original:
<http://urlib.net/8JMKD3MGP3W34T/48AFMK5>

INPE
São José dos Campos
2022

PUBLICADO POR:

Instituto Nacional de Pesquisas Espaciais - INPE
Coordenação de Ensino, Pesquisa e Extensão (COEPE)
Divisão de Biblioteca (DIBIB)
CEP 12.227-010
São José dos Campos - SP - Brasil
Tel.:(012) 3208-6923/7348
E-mail: pubtc@inpe.br

CONSELHO DE EDITORAÇÃO E PRESERVAÇÃO DA PRODUÇÃO INTELECTUAL DO INPE - CEPPII (PORTARIA Nº 176/2018/SEI-INPE):

Presidente:

Dra. Marley Cavalcante de Lima Moscati - Coordenação-Geral de Ciências da Terra (CGCT)

Membros:

Dra. Ieda Del Arco Sanches - Conselho de Pós-Graduação (CPG)

Dr. Evandro Marconi Rocco - Coordenação-Geral de Engenharia, Tecnologia e Ciência Espaciais (CGCE)

Dr. Rafael Duarte Coelho dos Santos - Coordenação-Geral de Infraestrutura e Pesquisas Aplicadas (CGIP)

Simone Angélica Del Ducca Barbedo - Divisão de Biblioteca (DIBIB)

BIBLIOTECA DIGITAL:

Dr. Gerald Jean Francis Banon

Clayton Martins Pereira - Divisão de Biblioteca (DIBIB)

REVISÃO E NORMALIZAÇÃO DOCUMENTÁRIA:

Simone Angélica Del Ducca Barbedo - Divisão de Biblioteca (DIBIB)

André Luis Dias Fernandes - Divisão de Biblioteca (DIBIB)

EDITORAÇÃO ELETRÔNICA:

Ivone Martins - Divisão de Biblioteca (DIBIB)

André Luis Dias Fernandes - Divisão de Biblioteca (DIBIB)

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

sid.inpe.br/mtc-m21d/2023/01.02.20.29-TDI

APLICAÇÃO DE TÉCNICAS DE MACHINE LEARNING NO ESTUDO DE TRANSIENTES DOS DETECTORES ADVANCED LIGO

Tábata Aira Ferreira

Tese de Doutorado do Curso de Pós-Graduação em Astrofísica, orientada pelo Dr. César Augusto Costa, aprovada em 09 de dezembro de 2022.

URL do documento original:
<http://urlib.net/8JMKD3MGP3W34T/48AFMK5>

INPE
São José dos Campos
2022

Dados Internacionais de Catalogação na Publicação (CIP)

Ferreira, Tábata Aira.

F413a Aplicação de técnicas de Machine Learning no estudo de transientes dos detectores Advanced LIGO / Tábata Aira Ferreira. – São José dos Campos : INPE, 2022.

xxiv + 122 p. ; (sid.inpe.br/mtc-m21d/2023/01.02.20.29-TDI)

Tese (Doutorado em Astrofísica) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2022.

Orientador : Dr. César Augusto Costa.

1. Glitches. 2. LIGO. 3. Aprendizado de Máquina. 4. Análise de redes. 5. Ondas gravitacionais. I.Título.

CDU 530.12

Esta obra foi licenciada sob uma Licença Creative Commons Atribuição-NãoComercial 3.0 Não Adaptada.

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported License.

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÕES

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS
Serviço de Pós-Graduação - SPGR

DEFESA FINAL DE TESE DE TÁBATA AIRA FERREIRA
REG. 135984/2018, BANCA Nº 328/2022

No dia 09 de dezembro de 2022, por teleconferência, o(a) aluno(a) mencionado(a) acima defendeu seu trabalho final (apresentação oral seguida de arguição) perante uma Banca Examinadora, cujos membros estão listados abaixo. O(A) aluno(a) foi APROVADO(A) pela Banca Examinadora, por unanimidade, em cumprimento ao requisito exigido para obtenção do Título de Doutora em Astrofísica. O trabalho precisa da incorporação das correções sugeridas pela Banca Examinadora e revisão final pelo(s) orientador(es).

Título: "Aplicação de técnicas de Machine Learning no estudo de transientes dos detectores Advanced LIGO"

Membros da banca:

Dr. Odylio Denys de Aguiar – Presidente - INPE
Dr. César Augusto Costa – Orientador - INPE
Dr. Massimo Tinto – Membro Interno - INPE
Dr. Francisco José Jablonski - Membro Interno - INPE
Dr. César Henrique Lenzi - Membro Externo - ITA
Dra. Iara Tosta e Melo - Membro Externo - INFN/LNS

Declaração de aprovação de Massimo Tinto anexa ao processo.

Documento assinado eletronicamente por **Odylio Denys de Aguiar**, **Pesquisador**, em 15/12/2022, às 15:44 (horário oficial de Brasília), com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Francisco Jose Jablonksi**, **Pesquisador**, em 15/12/2022, às 16:08 (horário oficial de Brasília), com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Cesar Augusto Costa (E)**, **Usuário Externo**, em 16/12/2022, às 09:20 (horário oficial de Brasília), com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Iara Tosta e Melo (E)**, **Usuário Externo**, em 16/12/2022, às 10:11 (horário oficial de Brasília), com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **César henrique lenzi (E)**, **Usuário Externo**, em 16/12/2022, às 13:05 (horário oficial de Brasília), com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.mcti.gov.br/verifica.html, informando o código verificador **10658508** e o código CRC **1D0E3DC7**.

Referência: Processo nº 01340.010021/2022-69 SEI nº 10658508

"Um livro é a prova de que os homens são capazes de fazer magia."

Carl Sagan

A meus pais ***Beatriz*** *e* ***Adilson*** *(in memoriam)*

AGRADECIMENTOS

A Deus por me dar força e saúde para lutar. À minha família, em especial minha mãe, quem está presente em todos os momentos da minha vida; por ser quem me apoia e me inspira a lutar pelos meus sonhos. Ao Delço, pela incrível companhia, pelas risadas e por todos os momentos juntos. Aos meus sobrinhos Luís Gustavo e Lucca que iluminaram muitos dos meus dias. A meus irmãos Maiara, Hudson, Sheila e Thiago por fazerem parte de todo esse processo.

Ao meu querido orientador César Augusto Costa pelos ensinamentos, amizade e parceria ao longo desses anos. Espero que ainda estejamos juntos por muito tempo.

Ao Odylio Aguiar por todos os conselhos, incentivos e amizade. Por ser uma grande inspiração para todos estudantes e por tornar possível meu doutorado sanduíche. Agradeço também à Professora Cláudia Vilega que me auxiliou com todos documentos e dúvidas do PrInt. Por acompanhar, sempre disposta, cada passo do meu processo. Aproveito para agradecer à Pricilla e ao Brian pelos auxílios no programa fellows do LIGO.

À Gabriela González (Gaby) por me receber com tanto carinho, me orientar, me dar apoio e permitir que eu fizesse parte do grupo de pesquisa da LSU. Agradeço também a todos os membros do grupo pela recepção.

Ao Massimo Tinto pela amizade, risadas, grandes ensinamentos e oportunidade para trabalharmos juntos no projeto do IMAGES. Também, agradeço a grande e indispensável parceria do estudante Manoel Felipe.

Ao Francisco Jablonski pelas partidas de xadrez e aulas em Séries Temporais. Ao Rafael Nunes pelas discussões e aulas que me abriram muito a mente. Ao Zé Carlos pela amizade e palavras de incentivo. Ao Flavio D'Amico pelos conselhos e pelas conversas na disciplina de Problemas Atuais da Astrofísica. Ao Braga, pela amizade.

A todos da astrofísica, em especial ao grupo GWINPE, pelo acolhimento. Pelas poucas discussões, mas super proveitosas com Júlio, Juliedson e Gabriela. Agradeço a todos os amigos que estiveram presentes nesse período. Em especial à Hemily por deixar os momentos no trabalho mais leves; ao Bruno por ser especial na minha vida e à Edi por me ensinar jogos sem dados.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

RESUMO

As detecções diretas de ondas gravitacionais não apenas trouxeram mais uma evidência da Teoria da Relatividade Geral de Einstein, mas inauguraram uma nova astronomia. Os observatórios LIGO foram os pioneiros na detecção desse tipo de sinal e dezenas de eventos já foram catalogados. O número progressivo de detecções fomenta a área e gera expectativas em diferentes observatórios e cientistas ao redor do mundo. Entretanto, não são apenas os sinais de eventos astrofísicos que aparecem nos dados destes detectores, mas diferentes ruídos transientes oriundos de diversos fatores ambientais, instrumentais ou antropogênicos. Estudar esses invasores locais, usualmente denominados *glitches*, é sempre um desafio para a colaboração científica, pois alguns deles têm alta taxa de ocorrência, podem mimetizar ondas gravitacionais, poluir os dados e diminuir a significância estatística de um sinal astrofísico real. Infelizmente, alguns desses transientes não têm causas identificadas ou definidas, e a tentativa de buscar tais indícios incentivou este trabalho. Esta tese apresenta uma forma alternativa para caracterizar e encontrar classes de *glitches* no canal gravitacional, a partir dos denominados *glitchgramas*. Duas técnicas computacionais foram utilizadas para avaliar a eficiência dessa caracterização proposta. A primeira aplicou ferramentas de Análise de Redes e a segunda, de Aprendizado de Máquina; ambos resultados foram comparados com as classificações prévias do Gravity Spy, ferramenta utilizada pela colaboração para classificar transientes. A análise de redes obteve resultados excelentes para determinadas classes, mas nem tanto para outras e, portanto, limitações no uso dessa técnica a partir de glitchgramas foram encontradas. No geral, o método teve $75,03\%$ de concordância com Gravity Spy e, com o cosseno de similaridade, apresentado na técnica, foi possível atribuir classes a *glitches* desconhecidos. O segundo método foi efetivo na busca de todas as classes investigadas. Com a aplicação de uma ferramenta do aprendizado de máquina supervisionado, uma validação cruzada foi realizada e o método concordou em $94,70\%$ com o Gravity Spy. Valor que poderia ter sido maior, pois o método apontou erros de classificação do atual modo de análise do LIGO. O aprendizado de máquina ainda mostrou-se independente, pôde ser aplicado em análises diárias de funcionamento do LIGO, em busca da presença de classes em canais auxiliares, abriu campos para diferentes aplicações e permitiu concluir que os glitchgramas caracterizam bem os *glitches*. Para exemplificar isso, esta tese também apresenta um estudo das duas classes mais presentes durante a terceira corrida observacional do LIGO: a Scattered Light e a Fast Scattering. Para esta última, uma investigação sobre sua relação com movimentos microssísmicos e antropológicos foi realizada.

Palavras-chave: Glitches. LIGO. Aprendizado de Máquina. Análise de redes. Ondas Gravitacionais.

APPLICATION OF MACHINE LEARNING TECHNIQUES IN THE STUDY OF TRANSIENTS FROM ADVANCED LIGO DETECTORS

ABSTRACT

The gravitational wave detections not only provided further evidence for Einstein's Theory of General Relativity but also inaugurated a new astronomy. LIGO observatories were pioneers in detecting such types of signals, and dozens of events have already been cataloged. The progressive number of detections has been promoting the area and generating expectations in scientists worldwide. However, it is not only the signals of astrophysical events that appear in the data of these detectors, different transient noise sources arise from various environmental, instrumental, or anthropogenic factors. Studying these local invaders, generally called *glitches*, is always a challenge faced by the scientific collaboration, as some of them have a high occurrence rate, may mimic gravitational waves, pollute the data and decrease the overall statistical significance of a real astrophysical signal. Unfortunately, some of these transients do not have identified or present well-defined reasons, and the attempt to look for such indications encouraged this work. This dissertation presents an alternative way to characterize and find classes of *glitches* in the gravitational channel, based on the so-called *glitchgrams*. Two computational techniques were used to evaluate the efficiency of this proposed characterization. The first applied Network Analysis tools and the second Machine Learning tools; both results were compared with previous Gravity Spy classifications, the tool used by the collaboration to categorize transients. Network Science obtained excellent results for some classes but not so much for others; therefore, limitations in using this technique from glitchgrams were found. Overall, the method had 75.03% of agreement with Gravity Spy, and, with the cosine of similarity, presented in the method, it was possible to assign classes to unknown *glitches*. The second method was effective in searching all investigated categories. With a supervised machine learning tool, cross-validation was performed, and the technique agreed at 94.70% with Gravity Spy. The value could have been higher, as the method pointed out classification errors in the current LIGO analysis mode. Machine learning still proved to be independent, could be applied in daily analyzes of LIGO's operation, in search of the presence of classes in auxiliary channels, opened fields for different applications, and allowed concluding that glitchgrams characterize well *glitches*. In order to exemplify this, this dissertation also studies the two most common classes during the third LIGO observational run: the Scattered Light and the Fast Scattering. For the latter, an investigation into its relationship with microseismic and anthropological motions was performed.

Keywords: Glitches. LIGO. Machine Learning. Network Science. Gravitational Waves.

LISTA DE FIGURAS

Pág.

LISTA DE TABELAS

Pág.

SUMÁRIO

Pág.

1 INTRODUÇÃO

A Teoria da Relatividade Geral de Albert Einstein propôs que a gravidade era causada por efeitos geométricos locais no espaço-tempo devido à presença de massa ou energia (KENYON, 1990). Em outras palavras, um corpo deforma o espaço-tempo ao seu redor e objetos próximos sentem a deformação e em torno dele gravitam. A Terra, por exemplo, sente a deformação que o Sol causa no espaço-tempo e, em consequência, o orbita. Essa ideia contestou o conceito desenvolvido por Isaac Newton de que a gravidade era uma força atuante em distâncias em um espaço infinitamente rígido, e de que a informação de qualquer variação da posição de uma massa era instantaneamente recebida; teoria matematizada na Lei da Gravitação Universal, publicada no livro *Principia*, que também pode ser acessada na versão de Newton (1999).

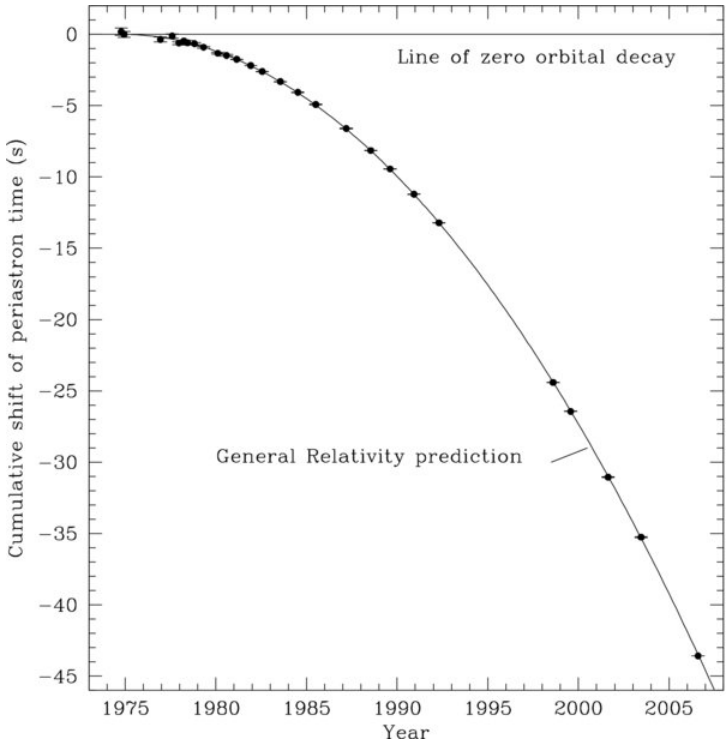
No Universo, eventos astrofísicos catastróficos podem gerar deformações que propagam-se através do espaço-tempo. Einstein pôde, com esse conceito, prever a existência dessas perturbações propagantes denominadas ondas gravitacionais. Elas são deduzidas a partir das Equações de Campo de Einstein e, de acordo com a Teoria da Relatividade geral, viajam na velocidade da luz (EINSTEIN, 1916; EINSTEIN, 1918).

A primeira evidência empírica da existência das ondas gravitacionais veio da observação de um sistema binário composto por um pulsar e uma estrela de nêutrons (PSR 1913+16). Esse sistema foi estudado por Hulse e Taylor e as observações mostraram um decaimento do seu período orbital (vide Figura 1.1); queda que só poderia ser justificada pela emissão de radiação gravitacional, prevista pela Teoria da Relatividade Geral (HULSE; TAYLOR, 1975; TAYLOR; WEISBERG, 1982; WEISBERG et al., 2010).

Joseph Weber foi o pioneiro no desenvolvimento de detectores de ondas gravitacionais na década de 60 (WEBER, 1960). No entanto, foi em torno de cem anos depois da previsão de Einstein, especificamente em 14 de setembro de 2015, que o primeiro sinal de ondas gravitacionais foi diretamente detectado nos dois observatórios LIGO, *Laser Interferometer Gravitational-Wave Observatory* (ABBOTT et al., 2016). Esse sinal era proveniente da coalescência de dois buracos negros de $36^{+5}_{-4}\,M_\odot$ e $29^{+4}_{-4}\,M_\odot$, que teve como remanescente um outro buraco negro de $62^{+4}_{-4}\,M_\odot$ e emitiu $3^{+0.5}_{-0.5}\,M_\odot c^2$ na forma de ondas gravitacionais (ABBOTT et al., 2016b). Tal evento, nomeado GW150914, inaugurou a astronomia de ondas gravitacionais e foi agraciado pelo Prêmio Nobel de Física para Reiner Weiss, Barry Barish e Kip Thorne em 2017

pela contribuição dos observatórios interferométricos na detecção (NOBEL, 2017).

Figura 1.1 - Medidas do pulsar PSR1913 + 16 até o ano de 2007. O eixo vertical apresenta a variação cumulativa do período orbital do seu sistema que concorda com a Teoria da Relatividade Geral, descrita pela curva. Caso o decaimento não existisse, os pontos deveriam ser constantes na linha horizontal.



Fonte: Weisberg et al. (2010).

Outras coalescências de buracos negros foram observadas. Algumas, inclusive, juntamente a um terceiro interferômetro localizado na Itália, Virgo (VIRGO, 2019), que possibilita prever, com melhor precisão, a localização da fonte emissora. Um outro importante evento que vale ser ressaltado aconteceu em 2017 e foi proveniente da coalescência de duas estrelas de nêutrons, o que permitiu uma contrapartida eletromagnética, e pôde ser observada por diferentes observatórios do mundo (ABBOTT et al., 2017b). Esse evento, GW170817, também inaugurou a astronomia multimenssageira com a presença de sinais de ondas gravitacionais. De acordo com os catálogos de ondas gravitacionais, cerca de noventa detecções de ondas gravitacionais já foram observadas até o momento (ABBOTT et al., 2019; ABBOTT et al., 2021a; ABBOTT et al., 2021b; ABBOTT et al., 2021c); e outras são esperadas.

O LIGO é composto por dois observatórios nos Estados Unidos. Um deles está localizado em Livingston, LA, e o outro em Hanford, WA (LIGO SCIENTIFIC COLLABORATION, 2022). Ambos têm funcionamento baseado no interferômetro de Michelson aliado à cavidades Fabry-Perot. Dois espelhos a uma distância de 4 km formam cada braço do interferômetro e a passagem de ondas gravitacionais causa diferença de fase do laser e permite a evidência da presença do sinal no fotodetector. Em virtude das polarizações das ondas gravitacionais, os detectores são mais sensíveis quando essas se propagam perpendicularmente ao plano do interferômetro (REITZE et al., 2019).

Os observatórios passam periodicamente por melhorias instrumentais para o aumento de sensibilidade. Atualmente, eles estão na versão aLIGO (*Advanced* LIGO), mas, aqui, serão referenciados apenas como LIGO. Essa sensibilidade é limitada por ruídos fundamentais como ruídos térmico, quântico e sísmico (AASI et al., 2015). Além dos ruídos fundamentais, aproximadamente estacionários (constantes no tempo), há também os ruídos transientes que podem estar relacionados a atividades antropológicas próximas aos detectores, fenômenos sísmicos, espalhamento do laser, variações ambientais e por muitos outros fatores. O objetivo principal desta tese é aplicar técnicas de Machine Learning para estudar e classificar esses ruídos transientes que, genericamente, são denominados *glitches*.

Glitches apresentam distribuições que diferem do comportamento gaussiano dos ruídos estacionários, sendo classificado, portanto, como não-gaussianos. Entender suas origens é fundamental para a busca dos sinais de ondas gravitacionais, pois eles afetam os detectores a todo instante, poluindo os dados, podendo diminuir a significância estatística e mimetizando sinais gravitacionais reais. Eles também podem acontecer durante um sinal gravitacional ou até mesmo vir de forma de parecida com o gorjeio (*chirp*) característico de um sinal de coalescência de objetos compactos, tornando-se falsos candidatos a ondas gravitacionais. Infelizmente, alguns *glitches* presentes no LIGO não têm causa física constatada e esse desafio motiva esta tese.

A Colaboração Científica do LIGO, LSC (do inglês *LIGO Scientific Collaboration*), tem um subgrupo, *DetChar* (de *detector characterization*), que trabalha na melhoria da qualidade dos dados e do qual esta tese e o grupo de ondas gravitacionais do INPE, GWINPE, fazem parte. O objetivo é analisar cuidadosamente os dados para permitir confiança e melhorias nas detecções de ondas gravitacionais (DETCHAR, 2021). Identificação, classificação e mitigação de ruídos (sejam estacionários ou não) fazem parte dessa linha de pesquisa; também monitorar (em tempo real) os interferômetros e seus canais auxiliares. Canais auxiliares são sensores que monitoram

os observatórios. Um terremoto é o exemplo simples para entender como um fator aleatório pode afetar o detector, fazendo com os espelhos se movam, indicando a passagem de uma possível falsa onda gravitacional. Um sismômetro (genericamente chamado de canal auxiliar) verifica essas coincidências e elimina falsos candidatos a sinais reais.

A maneira usual para caracterizar um *glitch* é através de sua morfologia no plano tempo-frequência. Nesse espaço de parâmetros, é possível atribuir uma classe a um determinado transiente de acordo com sua aparência. Inclusive, existe o projeto *Gravity Spy* (GRAVITYSPY, 2022) que junta aplicações de Aprendizado de Máquina com classificações de *glitches* feitas por cientistas da LSC e por usuários voluntários. O Gravity Spy foi desenvolvido através da plataforma *Zooniverse* (ZOONIVERSE, 2022) e é aberta a qualquer cidadão interessado em ajudar.

Ferramentas computacionais, a cada dia que passa, fazem mais parte do processo de pesquisa e extração de informações dos dados. Suas aplicações não ficam de lado na pesquisa de ondas gravitacionais e muito menos no estudo dos *glitches*. Foram desenvolvidas, por exemplo, algumas para classificações de transientes que podem ser encontradas em Bahaadini et al. (2018), Coughlin et al. (2019), Mukund et al. (2017), Biswas et al. (2013) e Powell et al. (2017). Ainda há aplicações para melhorar a significância de sinais de coalescências de binárias compactas (JADHAV et al., 2021), a qualidade dos dados (CUOCO et al., 2020) e para a busca das causas dos transientes (CAVAGLIA et al., 2018).

Não diferentemente, esta tese apresenta formas de classificação de *glitches* através do uso de técnicas de Aprendizado de Máquina e de Análise de redes. Ambas as técnicas são aplicadas em *glitchgramas*, uma representação alternativa (se comparada com a atual do LIGO através do Gravity Spy) no espaço de parâmetros de tempo, de frequência e de razão sinal-ruído. Os capítulos para essas abordagens e desenvolvimento dos conceitos estão organizados da seguinte forma:

- Capítulo 2: aborda o desenvolvimento simplificado da matemática envolvida para encontrar a equação de onda gravitacional a partir das Equações de Einstein. A Seção 2.1 apresenta as fontes geradoras de radiação gravitacional. Também, o princípio por trás do funcionamento dos interferômetros, os principais detectores terrestres e espaciais (existentes ou planejados para o futuro). Há breves comentários sobre outras técnicas de detecção e informações sobre próximas corridas observacionais envolvendo o LIGO;

- Capítulo 3: mostra as principais características do LIGO e os ruídos estacionários que compõem sua curva de sensibilidade que incluem: ruído quântico, ruído sísmico e ruído térmico. Também apresenta a taxa de ocorrência de ruídos transientes durante um dia, como eles podem afetar a sensibilidade do detector e como os canais auxiliares presentes nos detectores podem auxiliar na busca de causas desses *glitches*. Há um exemplo de como a eliminação dos *glitches* ajuda na significância de um sinal real de ondas gravitacionais e, por fim, há a apresentação do algoritmo Omicron que busca por transientes nos detectores e é o guia para o início deste estudo de *glitches*;

- Capítulo 4: descreve o principal modo de investigação de *glitches* por morfologia. Introduz os conceitos da ferramenta Gravity Spy utilizada atualmente pelo LIGO para classificação desses transientes ruidosos. Mostra a forma alternativa desenvolvida e utilizada na tese para caracterização do *glitch*. Há também as principais características das nove classes selecionadas para serem estudadas: Blip, Extremely Loud, Koi Fish, Tomte, Low Frequency Burst, Low Frequency Lines, Power Line, Scattered Light e Whistle. Por fim, mostra os passos para criação dos dados a serem analisados;

- Capítulo 5: uma vez que os dados a serem analisados são criados (a partir dos dados do LIGO), é possível aplicar métodos de análises. A princípio, os métodos testados seriam apenas de Machine Learning, mas para efeito de comparação, este capítulo apresenta a análise de transientes a partir do estudo de redes (ou *networks*). Para isso, são apresentados: a visualização dos dados em forma de grafo, o conceito de similaridade, o método de cosseno de similaridade, classificações de *glitches* a partir de buscas de comunidades e discussões gerais;

- Capítulo 6: este capítulo apresenta os conceitos gerais de Aprendizado de Máquina e como eles são aplicados aos mesmos dados do capítulo anterior. Como primeiro passo, o aprendizado não-supervisionado é utilizado para reduzir as dimensões dos dados através da ferramenta t-SNE; em seguida, a técnica supervisionada é aplicada para classificar e oferecer confiança numérica por meio de uma matriz de confusão e validação cruzada. No fim, o capítulo também discute os resultados, explica como essa técnica seria útil caso o Gravity Spy não existisse e ainda dá exemplos de aplicações para uma análise diária de dados do LIGO e para canais auxiliares;

- Capítulo 7: apresenta resultados do método de Machine Leaning não-supervisionado aplicado para o estudo de duas classes de transientes causadas por espalhamento do laser: a Scattered Light e a Fast Scattering. Essas duas classes foram as mais comuns durante a terceira corrida observacional do LIGO e, por isso, merecem atenção. Este capítulo também iniciou um estudo para buscar o motivo que faz o laser se espalhar e causar o Fast Scattering;

- Capítulo 8: Apresenta as discussões, conclusões e perspectivas futuras deste trabalho.

2 RADIAÇÃO GRAVITACIONAL

As ondas gravitacionais podem ser deduzidas a partir das Equações de Campo de Einstein que relacionam o tensor curvatura de Einstein, $G_{\mu\nu}$, e o Tensor momento-energia, $T_{\mu\nu}$, (MAGGIORE, 2007),

$$G_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}, \tag{2.1}$$

onde G representa a constante universal da gravitação e c, a velocidade da luz.

Essas equações vêm da Teoria da Relatividade Geral, TRG, e podem ser interpretadas pela famosa frase de Wheeler (FORD; WHEELER, 1999): *Spacetime tells matter how to move, matter tells spacetime how to curve* que diz, de forma simplificada, que a distribuição de massa no espaço-tempo ou a fonte do campo gravitacional ($T_{\mu\nu}$) define como vai ser a geometria ou a deformação do espaço-tempo local ($G_{\mu\nu}$). Se há massas aceleradas no Universo, tais deformações no espaço-tempo podem propagar-se; essas propagações são chamadas *ondas gravitacionais*.

Em um campo gravitacional fraco, as Equações de Einstein podem ser linearizadas. A linearização é feita assumindo que a métrica local ($g_{\mu\nu}$) é uma métrica plana $\eta_{\mu\nu}$ (também chamada métrica de Minkowski) mais uma pequena perturbação $h_{\mu\nu}$ ($|h_{\mu\nu}| \ll 1$). A métrica de Minkowski aqui tem a assinatura $diag(-1, 1, 1, 1)$ e o termo $h_{\mu\nu}$ pode ser interpretado como a perturbação métrica devido a presença da onda gravitacional. O sistema de coordenada, x^α, que satisfaz a condição de linearização ($g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$) é, algumas vezes, chamado sistema de coordenada Lorentziano (JARANOWSKI; KRÓLAK, 2009).

Com a linearização, a aplicação do *gauge* harmônico ($\partial^\mu \bar{h}_{\mu\nu} = 0$, também chamado de gauge de Lorentz, de De Donder ou de Hilbert (MAGGIORE, 2008)) e a desconsideração de termos de segunda ordem, as Equações de Einstein no vácuo ($T_{\mu\nu} = 0$) tornam-se:

$$\left(\nabla^2 - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \right) \bar{h}_{\mu\nu} = 0, \tag{2.2}$$

com $\bar{h}_{\mu\nu} = h_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu} h$, onde h representa o traço de $h_{\mu\nu}$. Essas são equações de onda para a métrica perturbativa $\bar{h}_{\mu\nu}$, equações de onda gravitacional em um espaço-tempo plano que propaga-se com velocidade da luz, c. Como solução, pode-

se escolher ondas planas do tipo

$$\bar{h}_{\mu\nu} = A_{\mu\nu} e^{ik_\alpha x^\alpha}, \tag{2.3}$$

na qual k_α representa o vetor de onda quadridimensional e $A_{\mu\nu}$ está relacionado a amplitude da onda e seu tensor de polarização $\epsilon_{\mu\nu}$.

Para estudar como é a interação das ondas gravitacionais com a matéria, é interessante escolher um sistema de coordenadas no qual o traço de h seja nulo ($h = 0$), o que implica que $\bar{h}_{\mu\nu} = h_{\mu\nu}$. Também que $h_{0\mu}$ seja zero, tornando a perturbação métrica puramente espacial, o que faz com que o gauge de Lorentz resulte numa perturbação métrica transversa, ou seja, $\partial^i h_{ij} = 0$ [1] (FLANAGAN; HUGHES, 2005). A onda ser transversa implica que a deformação no espaço-tempo será no plano perpendicular à sua direção de propagação.

Essa escolha de *gauge* simplifica e facilita o estudo das ondas gravitacionais. Ele também é chamado de *gauge* TT (do inglês *transverse-traceless*) e evidencia o fato das ondas gravitacionais terem duas polarizações. A partir do momento em que é considerado o *gauge* TT, muitas referências passam a chamar o tensor perturbação de h_{ij}^{TT}.

Das condições acima, $h_{xx} = h_{yy} \equiv h_+$ e $h_{xy} = h_{yx} \equiv h_\times$; assumindo que a onda viaja na direção $\hat{z}$, a solução reduz-se a

$$h_{ij}^{TT} = \begin{pmatrix} h_{xx} & h_{xy} & 0 \\ h_{yx} & -h_{yy} & 0 \\ 0 & 0 & 0 \end{pmatrix} cos(w(t - z/c)), \tag{2.4}$$

que também pode ser escrita como $h_{\mu\nu}^{TT} = h_+ \epsilon_{\mu\nu}^+ cos(w(t - z/c)) + h_\times \epsilon_{\mu\nu}^\times cos(w(t - z/c))$, onde $\epsilon_{xx}^+ = -\epsilon_{yy}^+ = 1, \epsilon_{xy}^\times = \epsilon_{yx}^\times = 1$ (com todos outros componentes nulos) equivalem aos dois tensores de polarização fundamentais das ondas gravitacionais. Usualmente, os dois recebem o nome de polarizações "mais" ou *plus* e "cruzada" ou *cross*.

Uma vez que a distância infinitesimal entre dois pontos é dada por (D'INVERNO, 1992)

[1] As letras gregas (μ, ν) têm valores de 0 a 3, enquanto as latinas (i,j) são os índices espaciais de 1 a 3.

$$ds^2 = g_{\mu\nu}dx^\mu dx^\nu, \tag{2.5}$$

a linearização pode ser aplicada e a equação para o elemento de linha ds^2 pode ser reescrita como

$$ds^2 = (\eta_{\mu\nu} + h_{\mu\nu})dx^\mu dx^\nu. \tag{2.6}$$

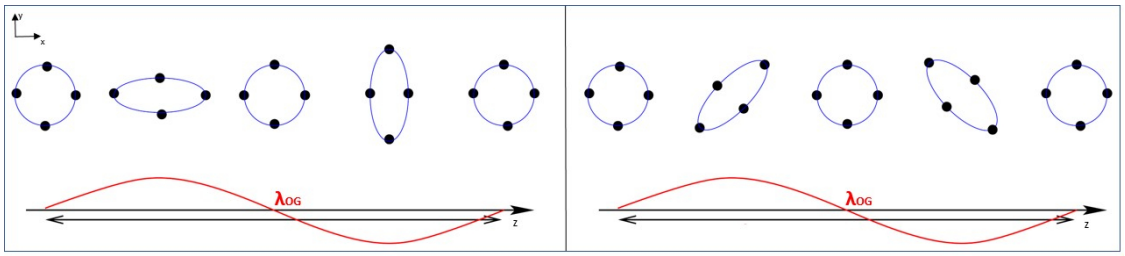
Se $\epsilon^{xy} = 0$, pode-se dizer que a polarização da onda é puramente "mais" e a Equação 2.6 torna-se:

$$ds^2 = -dt^2 + (1 + h_+ cos(\omega t - \omega z/c))dx^2 + (1 - h_+ cos(\omega t - \omega z/c))dy^2 + dz^2; \tag{2.7}$$

note que a métrica de Minkowski já foi substituída seguindo a assinatura mencionada anteriormente ($\eta_{00} = -1$ e $\eta_{11} = \eta_{22} = \eta_{33} = 1$).

A Equação 2.7 mostra que a perturbação métrica aparece apenas no plano $x - y$, que é o plano espacial perpendicular à propagação da onda. No entanto, ela aparece com efeitos contrários na distância própria em cada eixo, enquanto diminui em um, soma em outro. Para entender o efeito dessa polarização na prática, a Figura 2.1 (à esquerda) apresenta o efeito da onda gravitacional sobre um anel de massas testes. Considera-se que o plano da folha seja o plano $x - y$ e que a onda ainda viaja na direção z (perpendicular a esse plano). Inicialmente, as massas testes estão em repouso e, conforme a onda passa, algumas massas teste se afastam enquanto outras se aproximam. Em outras palavras, o anel se deforma, contraindo em uma direção e esticando-se em outra. Esse efeito se alterna até o fim da passagem da onda.

Figura 2.1 - Como massas testes reagem à passagem da onda gravitacional com uma polarização puramente "mais" (à esquerda) e puramente "cruzada" (à direita).



Fonte: Adaptado de Buonanno (2007).

O mesmo pode ser feito se apenas a polarização cruzada é considerada. Seus efeitos são iguais aos da polarização "mais", mas rotacionados por 45° (vide lado direito da Figura 2.1).

A polarização está diretamente relacionada às características inerentes da distribuição de massa e orientação da linha de visada entre a fonte de ondas gravitacionais e o detector. A combinação das duas polarizações fornece importante informação sobre a natureza da fonte e é a base fundamental para o princípio de detecções de ondas gravitacionais. Mas antes de entrar nesse detalhe, é interessante conhecer alguns tipos de fontes e princípios geradores de radiação gravitacional.

2.1 Fontes astrofísicas

As ondas gravitacionais acoplam-se fracamente com os detectores e, por isso, suas fontes detectáveis são eventos muito energéticos, até mesmo catastróficos (SATHYAPRAKASH; SCHUTZ, 2009). De forma análoga ao eletromagnetismo, onde cargas aceleradas emitem radiação eletromagnética, a radiação gravitacional resulta de aceleração de massas; estudo desenvolvido através do formalismo de quadrupolo.

Uma vez que as propriedades da fonte estão no tensor $T_{\mu\nu}$, é preciso voltar nas Equações de Einstein linearizadas fora do vácuo (Equações 2.8) para obter soluções para métrica perturbativa causada pela fonte,

$$\Box^2 \bar{h}_{\mu\nu} = -\frac{16\pi G}{c^4} T_{\mu\nu}. \tag{2.8}$$

Tais soluções não são triviais e, aproximações pós-Newtonianas podem ser aplicadas (SATHYAPRAKASH; SCHUTZ, 2009). Assim, no *gauge* TT, é possível encontrar que a deformação causada pela onda a uma distância r da fonte é,

$$h_{\mu\nu}(t) = \frac{2G}{rc^4}\frac{\partial^2 I_{\mu\nu}}{\partial t^2} = \frac{2G}{rc^4}\ddot{I}_{\mu\nu}(t - r/c). \tag{2.9}$$

Em outras palavras, a deformação dependente do tempo está relacionada à segunda derivada temporal do momento de quadrupolo reduzido do sistema, $I_{\mu\nu}$.

Há diferentes fontes astrofísicas com variações temporais no momento de quadrupolo que geram sinal gravitacional. Geralmente, elas são divididas em quatro tipos: espiralação, *burst*, contínuas e estocástica. As duas primeiras são de curta duração,

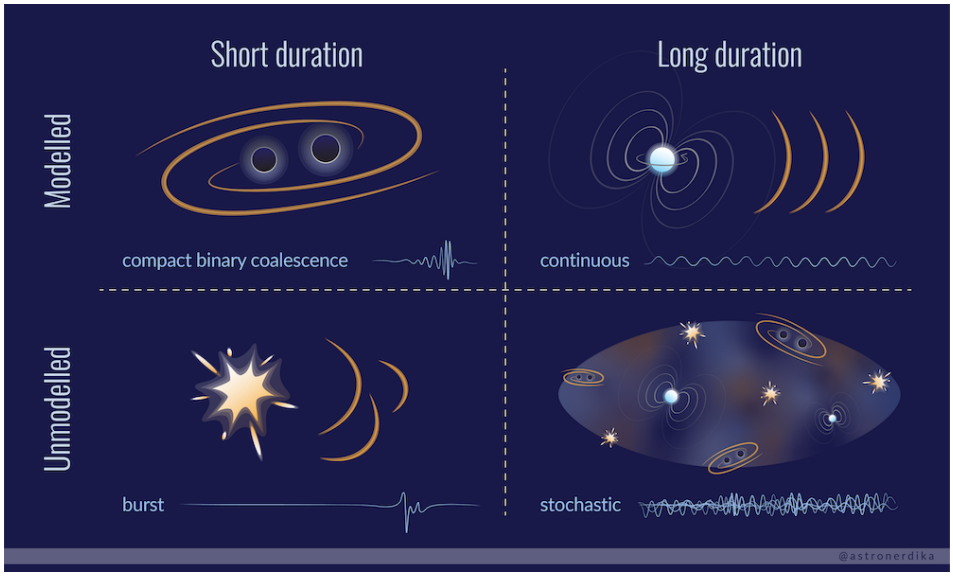
enquanto as duas últimas, não. A Figura 2.2 esquematiza os quatro tipos de fontes de acordo com a duração. À esquerda estão as fontes de curta duração e à direita, de longa duração; na parte superior estão as fontes que têm sinais bem modelados e na parte inferior, fontes de sinais não-modelados ou não conhecidos.

A fonte de longa duração e sem modelagem de sinal bem conhecida compõe o chamado fundo estocástico. Este é criado por superposição de sinais de ondas gravitacionais incoerentes, pode envolver diversos tipos de sinais astrofísicos de fontes independentes e também, de forma análoga à radiação cósmica de fundo, conter informações cosmológicas, como, por exemplo, sobre a evolução do Universo (CHRISTENSEN, 2018).

Também, em grande parte sem uma boa modelagem de sinal, está a fonte do tipo *burst* (ou impulsivas). São sinais muito curtos (da ordem de ms) que aparecem como transientes nos detectores e podem ser provenientes de eventos como supernovas e coalescência de buracos negros supermassivos (YAKUNIN et al., 2010).

As fontes contínuas têm sinais bem modelados e com longa duração. Binárias longe da coalescência e estrelas de nêutrons com irregularidades nas superfícies (sem simetria espacial axial) são exemplos de fontes geradoras desse tipo sinal. Vale ressaltar que, uma vez que os sinais são longos, o movimento do detector causa modulações em fase e amplitude no sinal (SCHUTZ; RICCI, 2010).

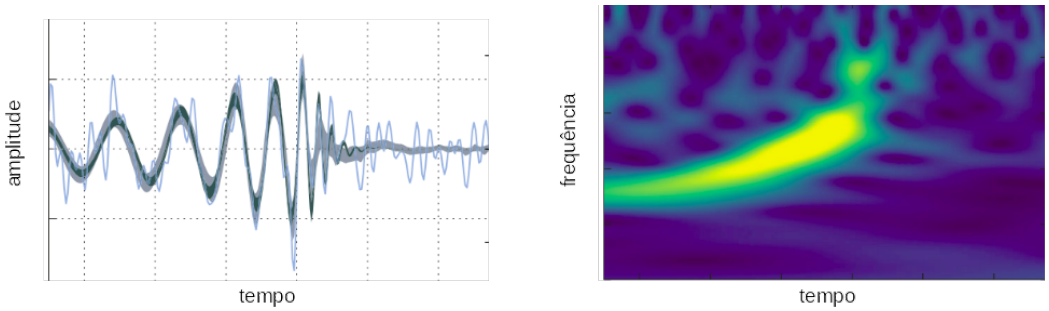
Figura 2.2 - Tipos de fontes de ondas gravitacionais.



Fonte: GALAUDAGE (2022).

Por fim, as fontes de espiralação são as que envolvem coalescências de objetos compactos como binárias de estrelas de nêutrons, binárias de buracos negros e binárias de estrela de nêutrons e buraco negro (LIGO, 2022). As formas de onda dessas fontes são bem modeladas e têm características próprias. Por exemplo, durante a coalescência de dois objetos compactos, a amplitude do sinal aumenta com a frequência (vide lado esquerdo da Figura 2.3). Também é possível visualizar, ao lado direito da imagem, esse sinal no plano tempo-frequência através de um espectrograma. A cor amarela pode ser interpretada como a razão sinal-ruído e é visível o aumento de frequência e de amplitude com o tempo; para tal representação atribui-se o nome *chirp*. Nesse instante, não é preciso se preocupar com as escalas, isso será tratado com detalhes mais adiante.

Figura 2.3 - A forma de onda de uma coalescência de objetos compactos, à esquerda, e sua representação em um espectrograma.



Fonte: Adaptada de Abbott et al. (2017a).

Todas as detecções, até o momento, foram de objetos compactos coalescentes. Haverá, na Seção 2.3, uma breve apresentação dos principais eventos de ondas gravitacionais já observados. No entanto, uma pergunta que pode aparecer nesse momento é: como as ondas gravitacionais são detectadas?

2.2 Detectores

Os dois principais tipos de detectores de ondas gravitacionais estão relacionados com massas ressonantes (ou antenas) e com sistemas interferométricos. Como mencionado anteriormente, a história experimental começou com Joseph Weber na Universidade de Maryland com a proposta de medir ondas gravitacionais utilizando barras ressonantes (WEBER, 1960). Apesar deste ter sido um experimento sem detecções confirmadas, serviu de impulso para diferentes desenvolvimentos experimentais no mundo.

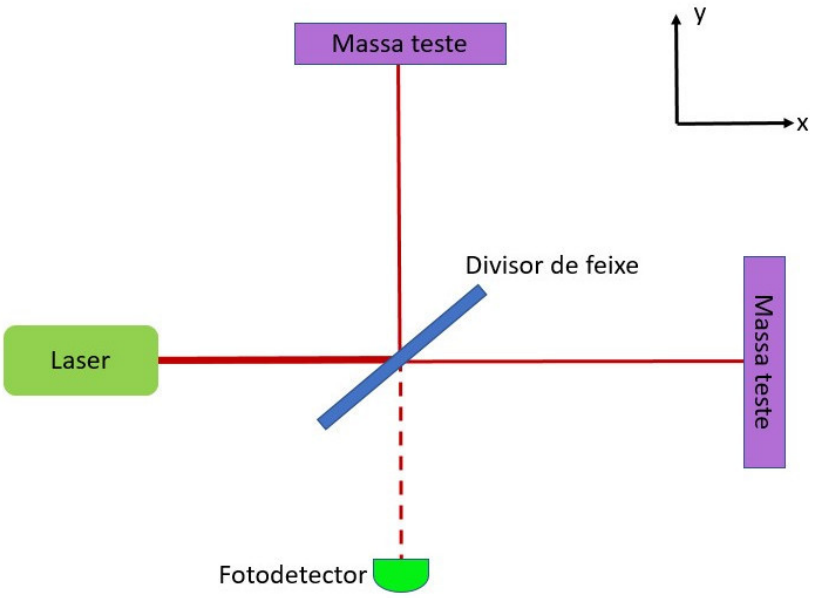
A ideia da massa ressonante é usar massas que oscilam após a passagem da onda gravitacional cuja frequência é a mesma de um dos modos de vibração da antena. Há detectores desse tipo em formatos cilíndricos e esféricos. Mais informações (incluindo referências sobre o detector brasileiro, Mario Schenberg) podem ser encontradas em Aguiar et al. (2005), Aguiar (2011), Oliveira e Aguiar (2016), Mauceli et al. (1996), Astone et al. (1999), Aufmuth e Danzmann (2005).

O foco deste trabalho, no entanto, está no detector LIGO, o responsável pela primeira detecção direta de ondas gravitacionais. A seção a seguir apresenta o princípio básico de seu funcionamento que faz parte dos sistemas interferométricos.

2.2.1 Interferômetros

Os dois detectores LIGO (ABBOTT et al., 2009) baseiam-se no interferômetro de Michelson que utiliza técnica de interferometria para medir deslocamentos. No caso do LIGO, tais informações são as indicações da possível presença de ondas gravitacionais. Um interferômetro desse tipo é composto por um laser, um divisor de feixe, espelhos (também chamados de massas testes) e um fotodetector. Há uma representação esquemática de um interferômetro desse tipo na Figura 2.4.

Figura 2.4 - Esquema de um interferômetro de Michelson.

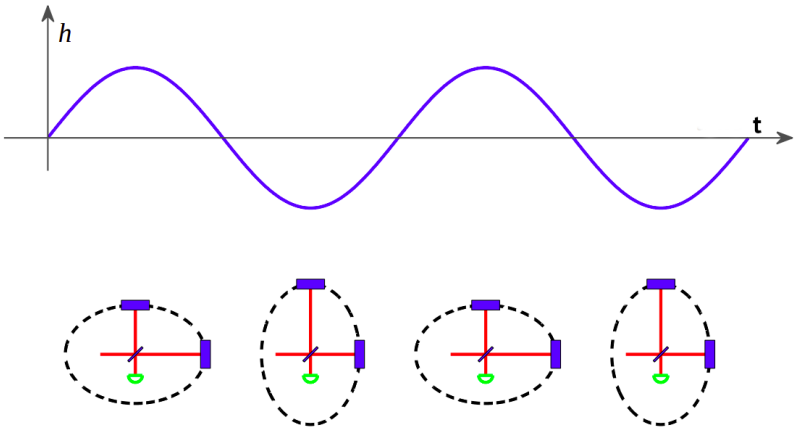


Fonte: Produção da autora.

O funcionamento para detecção de ondas gravitacionais é o seguinte: o laser (à esquerda) emite uma onda eletromagnética em direção ao centro que contém um divisor de feixe. Uma parte da luz é refletida em direção a um dos espelhos (direção y) e a outra é transmitida em direção ao outro (direção x). Cada feixe chega ao espelho, é refletido de volta e ambos são direcionados a um fotodetector. Esse direcionamento é feito de forma que haja uma interferência entre os dois feixes. Em um caso ideal, sem a presença de ondas gravitacionais, não haveria variações na medida da intensidade de luz no fotodetector.

Quando ondas gravitacionais passam pelo interferômetro, elas causam variações nos comprimentos dos braços relacionadas às deformações no espaço-tempo, e o padrão de interferência associado causa variações na intensidade de luz recebida no detector de fótons. Um exemplo é mostrado na Figura 2.5. Há na parte superior um sinal de onda gravitacional com amplitude h no tempo. Abaixo, há também como o interferômetro sente a presença de tais perturbações no espaço-tempo. Vale ressaltar que, nesse caso, a propagação da onda é perpendicular ao plano $x - y$ do interferômetro.

Figura 2.5 - Efeito sobre as massas testes de um interferômetro de Michelson devido à passagem de ondas gravitacionais propagando-se perpendicularmente ao plano da folha.



Fonte: Adaptada de Abbott et al. (2009).

Esse é o princípio básico do uso de interferômetros para detecção. Melhorias instrumentais devem ser aplicadas para tal alcance. No próximo capítulo, as principais modificações serão apresentadas, focando nos observatórios LIGO. Antes, a seção

seguinte apresenta os principais eventos já detectados por detectores interferométricos.

2.3 Detecções

Uma corrida observacional pode ser definida como o período em que os observatórios estão em funcionamento e medindo dados. O LIGO teve três corridas observacionais com detecções. A primeira, O1, foi de 12 de setembro de 2015 a 19 de janeiro de 2016. A O2 foi de 30 novembro de 2016 a 25 de agosto de 2017. E a terceira corrida foi dividida em duas partes: O3a (de 01 de abril de 2019 a 01 de outubro de 2019) e O3b (de 01 de novembro de 2019 a 27 de março de 2020). Até a presente data, antes da quarta corrida observacional, mais 90 sinais de ondas gravitacionais foram detectados. Todos eles podem ser encontrados nos catálogos de detecçoes: GWTC-1 (ABBOTT et al., 2019), GWTC-2 (ABBOTT et al., 2021a), GWTC-2.1 (ABBOTT et al., 2021b) e GWTC-3 (ABBOTT et al., 2021c).

O nome do evento detectado é dado pelas letras GW de *Gravitational Wave* seguidas pelo ano (aa), mês (mm) e dia (dd) da detecção do sinal. Dessa forma, o primeiro evento GW150914 aconteceu em 14 de setembro de 2015. Este, sem dúvida, é o evento extraordinário que inaugurou a astronomia de ondas gravitacionais. A lista a seguir apresenta esse e alguns outros eventos que merecem destaque:

→ **GW150914:** como mencionado anteriormente, esta é a primeira detecção direta (ABBOTT et al., 2016) que levou a contribuição do LIGO na observação de ondas gravitacionais ao prêmio Nobel em física (NOBEL, 2017). Esse evento aconteceu a uma distancia de 410 Mpc (ou $1,3.10^9$ anos-luz) da Terra (ABBOTT et al., 2016);

→ **GW170814:** O primeiro sinal medido por uma rede de três detectores. Este evento teve a participação do Virgo (ACCADIA et al., 2012) que permitiu uma melhor localização da fonte no céu e a primeira medida da polarização da onda (ABBOTT et al., 2017a);

→ **GW170817:** A primeira detecção de uma fusão de duas estrelas de nêutrons. Uma contrapartida eletromagnética pôde ser detectada por diferentes observatórios no mundo, inaugurando uma astronomia multimensageira que incluiu sinais de ondas gravitacionais (ABBOTT et al., 2017b; ABBOTT et al., 2017c);

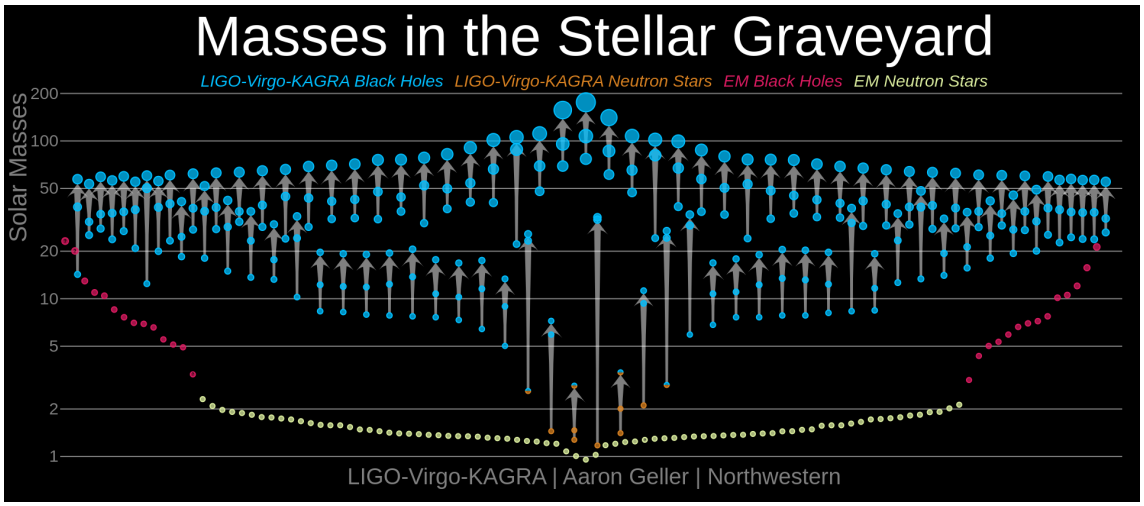
→ **GW190521:** primeira detecção com a evidência de um buraco negro de massa intermediária. Foi uma coalescência de dois buracos negros de $85M\odot$ e $66M\odot$ que resultou num remanescente de $142M\odot$ (ABBOTT et al., 2020a);

→ **GW190814**: evento com maior assimetria entre massas de objetos de mesma natureza. Foi a coalescência de dois buracos negros $23M\odot$ e $2.6M\odot$. Sendo o último um buraco negro menos massivo já observado ou a estrela de nêutrons mais pesada já descoberta num sistema duplo de objetos compactos (ABBOTT et al., 2020b);

→ **GW200105 e GW200115**: Primeiro e segundo eventos confirmados a serem uma coalescência de um buraco negro e uma estrela de nêutrons (COLLABORATION et al., 2021; BROEKGAARDEN; BERGER, 2021).

A Figura 2.6 apresenta a distribuição de massas de buracos negros e estrelas de nêutrons já observados. Em azul estão os buracos negros e em laranja as estrelas de nêutrons observados pelo LIGO. Em vermelho estão os buracos negros e em amarelo as estrelas de nêutrons observados por radiação eletromagnética. A distribuição das massas sugeria, anteriormente, a existência de uma lacuna de massas em torno de 2 a $5M\odot$, lacuna conhecida como *mass gap*, presente no intervalo de massa que fica entre a estrela de nêutrons mais pesada e o buraco negro mais leve conhecidos. No entanto, há evidências de que essa lacuna agora esteja sendo preenchida, inclusive pelas observações do LIGO (SHAO, 2022).

Figura 2.6 - Buracos negros e estrelas de nêutrons detectados. O ponto de metades laranja e azul simboliza que o objeto pode ser uma estrela de nêutrons ou um buraco negro.

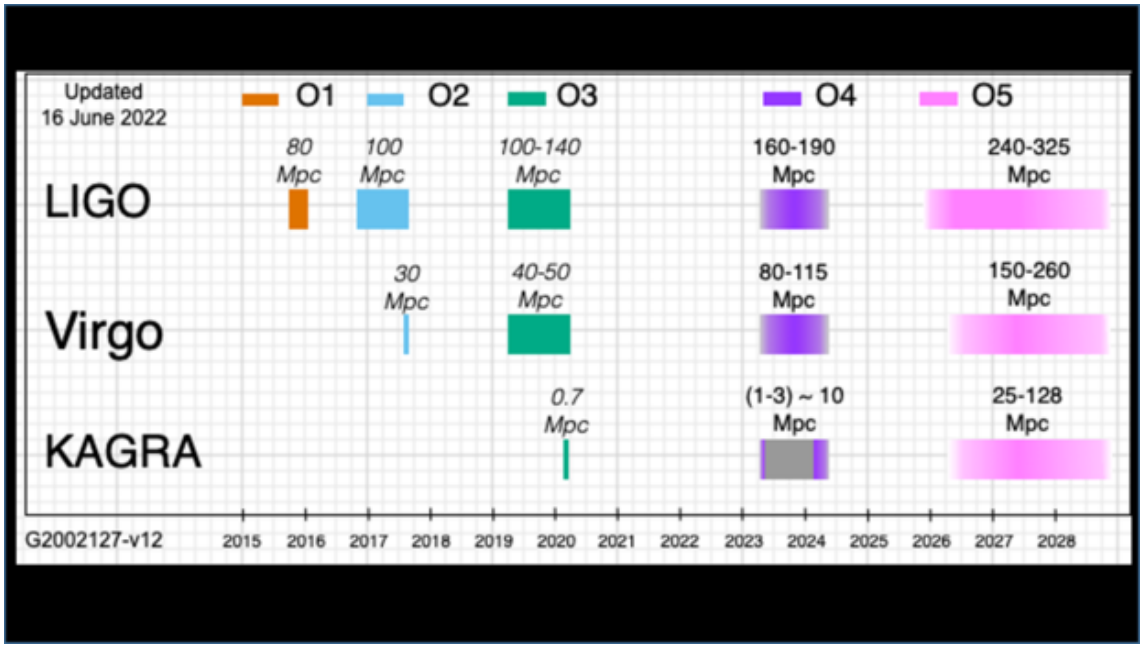


Fonte: Latest... (2022).

A Figura 2.6 mostra os nomes Virgo e KAGRA. Esses também são dois detectores interferométricos. O Virgo está localizado na Itália e seus braços interferométricos têm 3 km de comprimento (ACERNESE et al., 2014). Assim como o LIGO, ele também está na sua fase avançada que muitas vezes é referenciada como *advanced* Virgo ou aVirgo. O KAGRA é um interferômetro subterrâneo localizado no Japão, também com 3 km de braço (KAGRA..., 2019). Junto com LIGO e Virgo, formam a colaboração LVK (LIGO-Virgo-KAGRA).

A previsão para que o LIGO comece a corrida O4 é Março de 2023, juntamente com Virgo e KAGRA. Espera-se que o LIGO comece essa corrida com uma sensibilidade[2] de $160-190$ Mpc e o Virgo de $80-115$ Mpc (LATEST..., 2022). A Figura 2.7 apresenta esses planos observacionais para LVK, a sensibilidade esperada para cada corrida e informações sobre as corridas anteriores. O KAGRA irá começar a quarta corrida com um alcance observacional de $1-3$ Mpc, se afastará e voltará com uma sensibilidade maior.

Figura 2.7 - Linha do tempo das corridas observacionais.

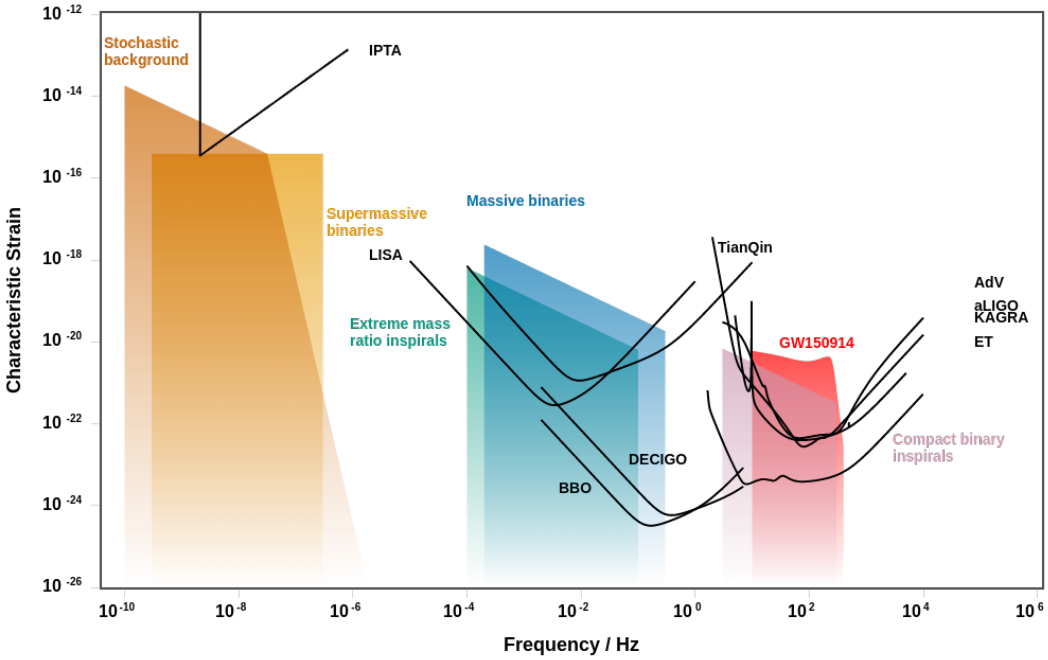


Fonte: LIGODCC (2022).

[2]essa sensibilidade também é conhecida como BNS *range*. Ela mede a distância na qual o observatório conseguiria detectar (com razão sinal-ruído igual a 8) a coalescência de duas estrelas de nêutrons de massas $1,4M\odot$.

Para o futuro, há projetos para a chamada próxima geração de detectores. Entre eles, há o ET (Einstein Telescope) e o CE (Cosmic Explorer). Além de projetos espaciais como LISA, TianQin, DECIGO, BBO e IMAGES. Estes diferem-se, principalmente, na alta sensibilidade em baixas frequências. A Figura 2.8 mostra a curva de sensibilidade de alguns deles e algumas da principais fontes astrofísicas que cada um é sensível. O interessante é que alguns eventos de coalescência também poderão ser detectados por observatórios espaciais muito antes da coalescência. Eles poderão estimar a posição no céu e quando a fusão irá acontecer.

Figura 2.8 - Fontes, detectores e curvas de sensibilidade.



Fonte: Gravitational... (2022).

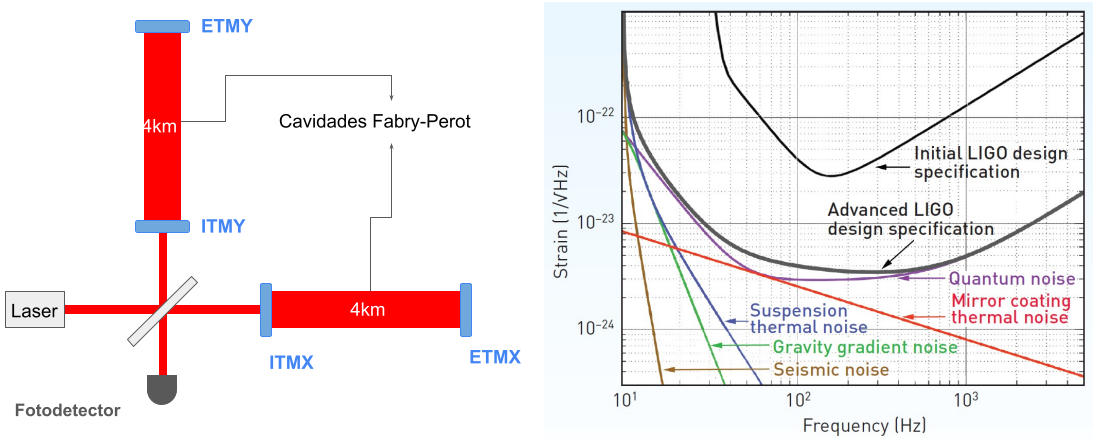
Por fim, apenas para complementação, existe uma outra técnica de detecção de ondas gravitacionais chamada PTA (Pulsar Timing Arrays). Ela usa radio telescópios que observam pulsares isolados e buscam por variações na chegadas dos pulsos que indicam a passagem de ondas gravitacionais. De acordo com NANOGrav (2022), o NANOGrav (North American Nanohertz Observatory for Gravitational Waves) é um consórcio de astrônomos que trabalham em experimentos desse tipo e, junto com projetos similares na Australia e Europa, formam o IPTA, International Pulsar Timing Array. A Figura 2.8 mostra quão baixas são as frequências das ondas gravitacionais esperadas por essa técnica.

3 CARACTERIZAÇÃO DO DETECTOR LIGO

O observatório de ondas gravitacionais por interferômetro laser, LIGO, foi fundado na década de 90 (ABRAMOVICI et al., 1992) e está tornando-se cada vez mais sensível. Seus observatórios em Livingston (LLO) e em Hanford (LHO) estão distanciados em aproximadamente 3.000 quilômetros. A distância entre eles implica que se um evento de onda gravitacional aparece em um dos detectores, o mesmo sinal deverá ser detectado no outro observatório em um intervalo de tempo máximo de 10 ms.

Cada observatório tem um interferômetro que segue a base de funcionamento explicada na Subseção 2.2.1. Uma importante modificação acrescentada, no entanto, é conhecida como cavidade *Fabry-Perot* que forma uma cavidade ressonante em cada braço do interferômetro, com um espelho depois do divisor de feixes (RAMIREZ, 2019). Esse espelho (chamado ITM$_i$ - *Input Test Mass*) possui uma alta transmitância e o espelho final (ETM$_i$ - *End Test Mass*) possui uma alta refletância. Assim, ao atingir o segundo espelho, o laser é refletido em direção ao primeiro novamente que também possui (nessa direção) alta refletância, de forma que a luz vá e volte várias vezes formando essa cavidade ressonante. Dessa forma, a luz fica confinada e, consequentemente, há um aumento significativo na sua potência (AASI et al., 2015). Além disso, há o aumento do tempo de exposição da luz com a onda gravitacional (RILES, 2013). Note que $i = x, y$, dependendo da direção escolhida. Aqui, x vai ser a mesma direção do laser. A Figura 3.1(a) esquematiza o LIGO com essa cavidade.

Figura 3.1 - À esquerda, há a representação das cavidades de Fabry-Perot; à direita, a curva de sensibilidade teórica do LIGO limitada por ruídos fundamentais e a diferença de sensibilidade entre o LIGO inicial e o aLIGO.



(a) (b)

Fonte: a) Produção da autora; b) OSA (2019).

Há também outras cavidades ópticas que compõem o detector. Entre elas, o IMC (*Input Mode Cleaner*) que é uma cavidade triangular antes da entrada para o divisor de feixe que filtra o feixe do laser de entrada (que no caso do LIGO tem comprimento de 1064 nm) e serve como referência para estabilização da frequência. Ainda antes do divisor de feixes, há também um PRM (*Power Recycling Mirror*). Quando a luz retorna dos braços para o divisor de feixes, parte dela também pode ser direcionada "de volta" ao laser. Para evitar isso, existe essa cavidade recicladora de potência (PRM) que reflete esse sinal recebido que, por sua vez, pode ser então reutilizado no interferômetro. Na saída do interferômetro, antes da leitura do sinal, há também o OMC (*Output Mode Cleaner*) que é similar ao IMC e limpa o feixe de luz. Essas informações e mais detalhes sobre a complexidade do LIGO podem ser encontrados em Abbott et al. (2009) e Aasi et al. (2015).

Quando uma onda gravitacional passa alterando os comprimentos dos braços do interferômetro, há no detector a projeção de sua amplitude relativa, chamada *strain amplitude* ou *strain* ou h. Se na passagem, o braço da direção $\hat{y}$ (de comprimento L) varia ΔL_y e o de $\hat{x}$ (também de comprimento L) varia ΔL_x, então (AASI et al., 2015),

$$h(t) = \frac{\Delta L_y - \Delta L_x}{L}. \tag{3.1}$$

A leitura dos fotodetectores compõe uma série temporal, onde o eixo vertical é amplitude strain. Ao representá-la no domínio da frequência, é possível obter a ASD (*amplitude spectral density*), também chamada curva de ruído instrumental (*strain noise*) ou de curva de sensibilidade do detector. Ela pode ser vista na Figura 3.1(b). O eixo vertical dessa imagem é a densidade espectral de amplitude que mostra a distribuição em frequência dos ruídos estacionários do detector.

Esses ruídos estão relacionados com as características físicas do interferômetro, local de instalação, instrumentação, etc. A Figura 3.1(b) também apresenta as curvas teóricas desses ruídos. Os principais, chamados ruídos fundamentais, são o térmico e o quântico. A seção a seguir explicará brevemente de onde eles vêm.

3.1 Ruídos estacionários

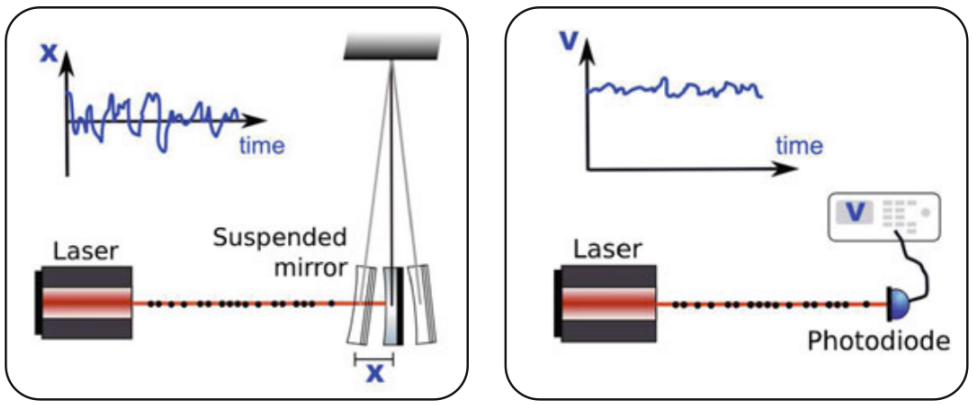
Ruído Quântico: O ruído quântico recebe esse nome porque está relacionado ao comportamento quântico da luz, à sua propriedade corpuscular. Esse ruído pode ser visto pela cor lilás na Figura 3.1(b), à direita, e é definido como a soma do ruído de Poisson (*Shot*

Noise) e do ruído de pressão de radiação. O *Shot Noise* aparece em frequências mais altas enquanto o ruído de pressão de radiação é dominante nas frequência menores.

A incidência de fótons no fotodetector é uma consequência básica para o funcionamento do interferômetro. Para determinadas medidas, é a quantidade de fótons que chega no detector que interessa. Contudo, os fótons no laser não estão igualmente espaçados no tempo (BASSAN, 2014), mas seguem uma distribuição de Poisson (SAULSON, 2017). Consequentemente, há uma flutuação na corrente elétrica devido à incidência do laser nos fotodetectores. Essa flutuação pode ser vista à direita da Figura 3.2 e representa o ruído de Poisson.

Além disso, fótons carregam momento, e quando o feixe de luz atinge a massa teste (espelho), ele exerce uma pressão sobre a superfície. Se essa pressão fosse constante, haveria uma correção simples na posição do espelho que recebe esses fótons. No entanto, como o número de fótons que chega no espelho varia (devido ao ruído de Poisson), a pressão sobre ele não é constante. Da segunda lei de Newton, uma variação de momento no tempo é uma força resultante. Logo, essa força move os espelhos (MAGGIORE, 2008) causando uma diferença de fase e indicando um sinal no detector. Esse é o chamado ruído de pressão de radiação. O lado esquerdo da Figura 3.2 ilustra esse efeito. Note que x é a posição relativa à sua posição inicial.

Figura 3.2 - Ilustração do ruído de pressão de radiação (à esquerda) e ruído de Poisson (à direita).



Fonte: Bassan (2014).

Ruído térmico de suspensão: O ruído térmico também é um ruído fundamental na sensibilidade de detectores de ondas gravitacionais. Ele está associado a fontes de dissi-

pação de energia, relacionado com o Teorema de flutuação-dissipação (GONZÁLEZ, 2000). Este teorema diz que o espectro de potência do movimento flutuante do sistema é dado por:

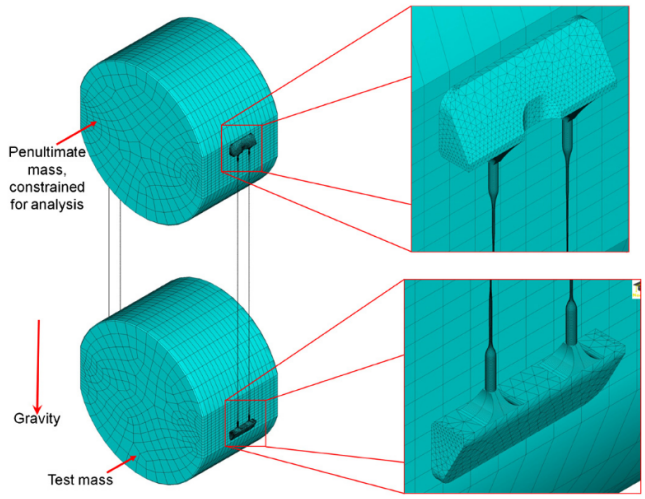
$$x^2_{term}(f) = \frac{k_B T}{\pi^2 f^2} R(Y(f)), \tag{3.2}$$

onde k_B é a constante de Boltzmann, T é a temperatura do sistema, $R(Y(f))$ é a parte real da admitância que é o inverso da impedância (SAULSON, 2017). A impedância, por sua vez, é definida como a razão entre a força aplicada ao sistema para movê-lo com velocidade v e o próprio valor de v ($Z = F/v$). Essa equação mostra que o deslocamento do espelho pode ser menor com a diminuição da temperatura. Por isso, gerações futuras do LIGO pretendem inserir criogenia.

Como o próprio nome diz, o ruído térmico de suspensão é proveniente do sistema de suspensão das massas testes e pode ser visto na cor azul da Figura 3.1(b). Vale informar aqui que as massas testes dos observatórios são suspensas para evitar o ruído sísmico (AASI et al., 2015). Na configuração do LIGO, esse ruído está presente em regiões de menores frequências (até mais ou menos 30 Hz), mas também é responsável por linhas evidentes em 510 Hz e harmônicos. A linha em 510 Hz acontece devido às fibras de sílica usadas no estágio final da suspensão das massas testes. Essas fibras têm ressonâncias chamadas modos violino e, devido às excitações térmicas desses modos nas fibras, há o movimento das massas testes nessas mesmas frequências. Então, esse ruído torna-se estacionário e aparece como uma linha na frequência de 510 Hz (HARRY et al., 2010) . A Figura 3.3 ilustra o início e o fim das fibras que sustentam a massa teste no sistema de suspensão.

Ruído revestimento Browniano: Há também um outro ruído térmico proveniente dos próprios espelhos (curva vermelha da Figura 3.1(b)). Esse ruído basicamente está relacionado com dissipações mecânicas nas camadas do revestimento dos espelhos. Flutuações de temperatura na cavidade do espelho implicam em mudança de fase do laser que pode ser devido à expansão térmica (ruído termoelástico) que faz com que a espessura do revestimento varie de acordo com a temperatura e por variação no índice de refração (efeito termo-refrativo). Se a fase varia, a interferência destrutiva é cancelada, causando um sinal no detector.

Figura 3.3 - Ilustração das fibras que sustentam as massas testes.



Fonte: Saulson (2017).

Ruído Sísmico: Um outro ruído também importante para ser mencionado é o ruído sísmico (cor marrom da Figura 3.1(b)). Ele está na região de baixas frequências na curva de sensibilidade do LIGO. O nome do ruído está relacionado com o fato do chão estar em constante movimento ou vibração. Na região de $1 - 10\,\text{Hz}$ ele é principalmente causado por atividades humanas incluindo carros, trens, indústrias, atividades agrícolas e também efeitos locais como ventos e raios (MAGGIORE, 2008). Além disso, ainda há um chamado fundo microssísmico (em frequência menores) que também afeta o detector podendo movimentar os mecanismos de suspensão e, consequentemente, o espelho. Acredita-se que este seja originado por ondas oceânicas. A Tabela 3.1 apresenta uma descrição sobre a banda de frequência desse ruído e suas fontes.

Tabela 3.1 - Tabela de algumas fontes de ruídos sísmicos, as frequências que elas atingem e a que distância elas ocorrem.

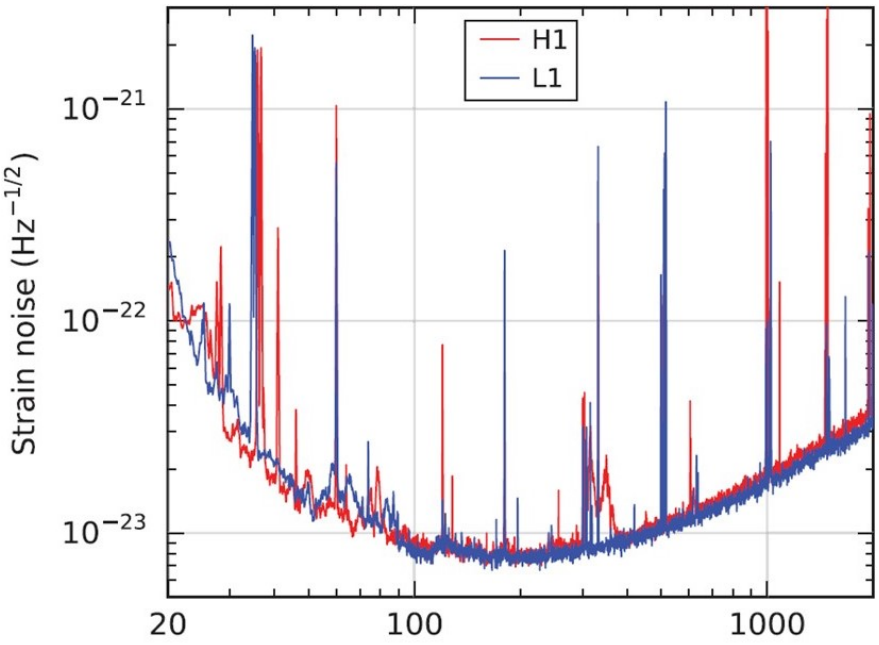
Frequência (Hz)	**Fonte**
0.01-1	Terremotos distantes, microssísmica
1-3	Ruídos antropogênicos distantes, terremotos próximos, ventos
3-10	Ruídos antropogênicos, ventos

Fonte: Adaptada de Macleod et al. (2012).

Ruído Newtoniano: Por fim, vale mencionar o ruído Newtoniano. Ele acontece em frequências menores e pode ser visto na cor verde da curva de sensibilidade do detector. Esse ruído também está relacionado com ondas sísmicas, pois a passagem delas próximas a um detector interferométrico cria variações na densidade da Terra que provocam forças gravitacionais variantes sobre os espelhos (HUGHES; THORNE, 1998). Há estudos que estimam que esse tipo de ruído aconteça também por flutuações atmosféricas, como variações na temperatura, por exemplo (CREIGHTON, 2008), e também por atividades humanas rotineiras (THORNE; WINSTEIN, 1999).

Além dos ruídos fundamentais mencionados acima, as fontes estocásticas também fazem parte do ruído estacionário. A Figura 3.4 mostra a curva de sensibilidade dos detectores LIGO durante a primeira corrida observacional. Se comparada com a curva teórica apresentada na Figura 3.1(b), ela tem o mesmo comportamento, porém com linhas verticais que chamam atenção; o que são elas e de onde vêm?

Figura 3.4 - Curva de sensibilidade dos detectores de Livingston (azul) e de Hanford (vermelho) durante a primeira detecção.



Fonte: Abbott et al. (2016).

3.2 Noise Budget

As linha verticais não estão na curvas teóricas, pois elas estão relacionadas com diferentes fatores instrumentais e/ou ambientais. Por exemplo, as linhas em 60 Hz e harmônicos (120 Hz, 180 Hz, etc.) surgem da corrente elétrica alternada nos Estados Unidos. Como mencionado anteriormente, as linhas em torno 510 Hz e harmônicas são os modos violinos provenientes de flutuações térmicas nas fibras de sílica usada na suspensão das massas testes. Em Hanford, há linhas verticais em torno de $10 - 30$ Hz devido à presença de neve no inverno. Há modos violinos causados na suspensão do divisor de feixes (em 300 Hz, 600 Hz, 900 Hz). Aparecem também linhas para calibração dos detectores. Há outras linhas devido às perturbações ambientais e outros fatores; todas essas que aparecem na curva de sensibilidade da Figura 3.4 podem ser encontradas em gwopenscience (2017). As linhas da terceira corrida observacional estão em gwopenscience (2020) [1].

3.3 Ruídos transientes

Além dos ruídos apresentados anteriormente, há também ruídos transientes nos detectores, os quais, usualmente, são chamados de *glitches* e compõem o foco de estudo desta tese. Transientes são eventos que aparecem como excesso de potência (em determinadas bandas de frequência) nas medidas do canal gravitacional e, tipicamente, possuem caráter não-gaussiano, não-estacionário e são de curta duração. Eles acontecem várias vezes ao dia de forma aleatória e independente de qual detector, poluindo e diminuindo a qualidade dos dados. Quando observados no espaço de tempo-frequência-potência, tais transientes apresentam características morfológicas distintas que possibilitam seu agrupamento e classificação por similaridade. Els podem aparecer por fatores ambientais como mudanças de temperatura, ventos, terremotos, tempestades; por atividades humanas como trabalhos em indústrias, passagens de trens e caminhões; e por problemas instrumentais como espalhamento do laser, rede elétrica, e até mesmo por uma partícula de poeira na câmara de vácuo (DAVIS et al., 2021).

É importante lembrar que ondas gravitacionais de coalescências de buracos negros e/ou estrelas de nêutrons também aparecem como transientes. Além do mais, apesar de não terem sido detectados ainda, sinais tipo *burst* podem ter essa característica. Um ruído transiente pode ser muito semelhante a um sinal astrofísico ou pode afetar o interferômetro de modo a diminuir a significância estatística de um sinal real. Por isso, evitá-los é fundamental. Um primeiro passo é recuperar informações do transiente para que seja possível atribuir a ele uma causa ruidosa ou defini-lo como um candidato a onda gravitacional. Para isso,

[1] informações e dados de ondas gravitacionais estão disponíveis em *GW Open Science*: https://www.gw-openscience.org/

o LIGO utiliza uma ferramenta chamada *Omicron* (ROBINET et al., 2020) que busca por sinais transientes no detector e os caracteriza.

3.3.1 Omicron

O algoritmo Omicron é um ETG (event-trigger-generator) que detecta eventos não-gaussianos com excesso de potência no sinal e o caracteriza com parâmetros como SNR, frequência e tempo. Para isso, o Omicron faz uma análise em tempo-frequência do sinal em diferentes resoluções. Além disso, salva tais informações em arquivos de dados que podem ser acessados pela LSC (ROBINET et al., 2020). Vale ressaltar que o Omicron não é um ETG de busca por ondas gravitacionais, mas por transientes em geral.

O algoritmo é baseado na transformada Q, Equação 3.3, que busca esses transientes projetando o sinal em bases seno-gaussianas no espaço tempo-frequência,

$$X(\tau, \phi, Q) = \int_{-\infty}^{\infty} x(t)\omega(t - \tau, \phi, Q)e^{-2i\pi\phi t}dt; \tag{3.3}$$

X mede a amplitude média do sinal $x(t)$ para uma janela seno-gaussiana ω. Essas janelas, por sua vez, dependem de um terceiro parâmetro conhecido como Q, também chamado de fator de qualidade, que é relacionado com a frequência central da seno-gaussiana, f_0, e sua largura de banda Δf,

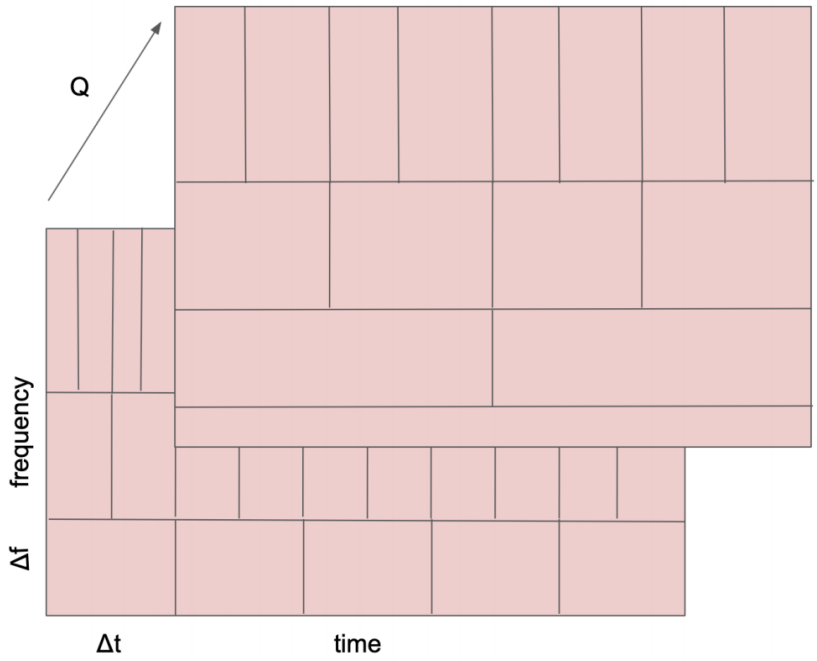
$$Q = \frac{f_0}{\Delta f}. \tag{3.4}$$

Dessa forma, para cada valor de Q, um plano tempo-frequência é construído. Cada pedaço do plano tem uma largura de banda (Δf) e uma duração (Δt) que formam um retângulo de área $\Delta f \Delta t$ chamado *tile*. Para manter o fator de qualidade constante no plano, a transformada altera a duração da janela. Pelo princípio da incerteza, quando a resolução em tempo diminui, a de frequência aumenta e vice-versa (GABOR, 1946). A Figura 3.5 mostra a construção desses *tiles* para diferentes valores de Q, como é feito pelo Omicron. O valor Q pode ser interpretado como o número de oscilações na função. Quanto maior o Q, para uma mesma frequência central, maior o número de oscilações da seno-gaussiana; isso implica uma maior resolução em frequência.

Para cada *tile*, o Omicron cria um ponto de dados representativo. Ele carrega informações como o valor da frequência central daquele *tile*, o Q, a SNR e o tempo. Esse ponto é chamado de *trigger*. O Omicron passa pelos dados do LIGO e, a cada transiente que encontra, cria uma tabela de dados com vários triggers e diferentes valores de Q. Essas informações são salvas em um arquivo no formato *.root* que é chamado de *unclustered file*.

Suponha intervalos de tempo e de frequência a serem analisados. Se diferentes planos Q forem sobrepostos, haverá diferentes *tiles* (ou triggers) para essa região e, por isso, o Omicron também cria um outro arquivo *.xml* com o trigger representante de todos os triggers; ele é escolhido por ter o maior valor de SNR. No fim, o novo arquivo carrega informações como tempo inicial e final das bordas dos *tiles* inseridos na região de interesse, a frequência de pico que é a frequência central do *tile* de maior significância e outros parâmetros. Esses parâmetros finais caracterizam um transiente encontrado nos dados. Esse novo arquivo é chamado de *clustered file*.

Figura 3.5 - Montagem de planos tempo-frequência a partir da transformada Q. Cada plano tem um valor de Q, uma resolução em tempo (Δt) e outra em frequência (Δf). O retângulo de área $\Delta f \Delta t$ forma um *tile*. Considerando apenas o plano superior (Q constante), para uma frequência f_0, Δf vai ser fixo e, portanto, para uma linha horizontal no plano, as resoluções em frequência e tempo são as mesmas. Se Q é constante, mas f_0 aumenta, então, de acordo com a Equação 3.4, Δf aumenta e Δt diminui. Por esse motivo, os *tiles* não são uniformes. Por outro lado, se são considerados vários planos para uma mesma frequência, conforme Q aumenta, Δf diminui (consequentemente Δt aumenta) e isso justifica diferentes larguras dos *tiles* conforme Q varia.



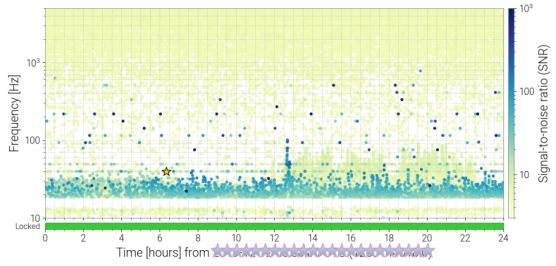
Fonte: Soni (2021).

A Figura 3.6(a) apresenta os transientes encontrados no canal gravitacional pelo Omicron durante um dia aleatório no LLO. A imagem mostra as frequências de pico desses sinais ao longo do dia; a cor de cada ponto representa os valores da SNR e o eixo horizontal apresenta o tempo em horas. Esse dia foi muito ruidoso, com altos valores de SNR na

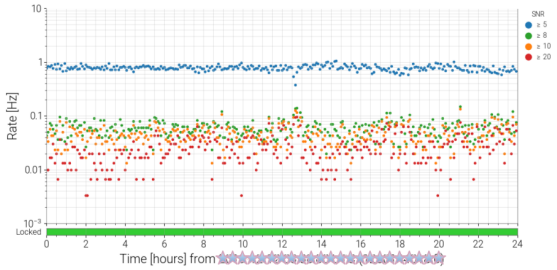
faixa de 20 a 30 Hz. Há também uma linha vertical proeminente um pouco depois das 12 h com transientes de SNRs consideráveis em diferentes frequências, atingindo até 100 Hz. A confirmação de que a quantidade de transientes aumentou nesse horário pode ser vista na Figura 3.6(b) que mostra a taxa de transientes durante o mesmo dia com seus respectivos SNRs. A taxa de ocorrência de transientes com menores valores de SNR é maior, enquanto que transientes com maiores valores de SNR acontecem com menos frequência.

A incidência de muitos transientes diminui a sensibilidade do detector. A Figura 3.6(c) mostra a BNS range desse mesmo dia no interferômetro de Livingston. Ela deveria ter uma tendência estacionária e, no entanto, em torno do mesmo horário, é visível como essa sensibilidade diminui.

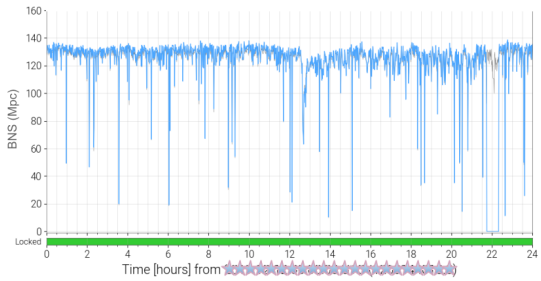
Figura 3.6 - Figura (a) mostra a quantidade de transientes encontrados pelo Omicron durante um dia aleatório no observatório de Livingston. A (b) mostra a taxa de ocorrência de transientes no detector durante esse mesmo dia bem como a sensibilidade do detector em (c). Na imagem (d), há vibrações terrestres na banda de $0,03$ a $0,1$ Hz que justificam o aumento de transientes e a queda de sensibilidade em torno das 12h.



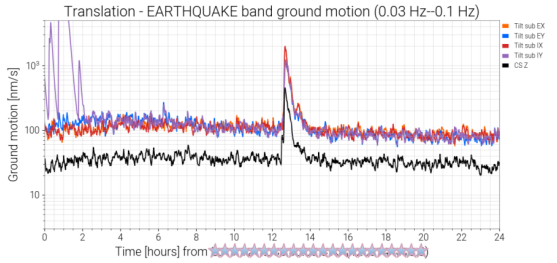
(a)



(b)



(c)



(d)

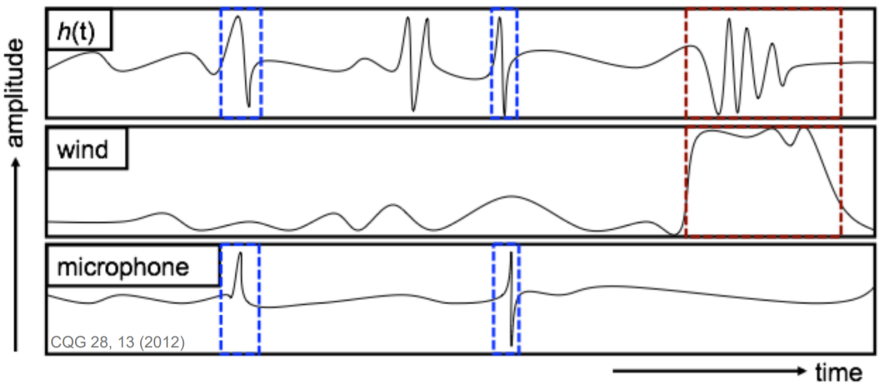
A remoção desses transientes está diretamente relacionada com a busca de coincidências entre sinais significantes do canal gravitacional e dos canais auxiliares. Canais auxiliares são os sensores que monitoram os observatórios como termômetros, sismômetros, acelerômetros, microfones, fotodiodos e tentam acompanhar tudo o que acontece em torno do interferômetro.

3.3.2 Canais auxiliares

A maneira usual de se encontrar a causa dos transientes no canal gravitacional é buscando coincidências temporais com canais auxiliares. Há mais de 200 mil canais auxiliares, divididos em várias bandas de frequências de interesse, monitorando comportamentos ambientais e instrumentais em cada observatório LIGO (NUTTALL, 2018) e alguns deles também são monitorados pelo Omicron. Como a maioria dos canais auxiliares não é sensível a ondas gravitacionais, as informações deles podem ser usadas para vetar sinais transientes (ruidosos) do canal gravitacional (SMITH et al., 2011).

Um esquema da busca pelas fontes de ruídos do tipo transiente é apresentado na Figura 3.7. O eixo vertical apresenta a amplitude do sinal e o horizontal, o tempo. A parte superior com $h(t)$ mostra quatro significantes transientes no canal gravitacional. O gráfico intermediário mostra a variação de velocidade no tempo para ventos e o inferior, para microfone. Quando faz-se a análise no tempo, pode-se se dizer, com uma alta probabilidade, que dois transientes do LIGO aconteceram por motivos de vibrações sonoras (em azul) e um por alguma ventania balançando as árvores ao redor sítio (em vermelho).

Figura 3.7 - Exemplo de busca da fonte de ruído transiente comparando sinais do canal gravitacional com dos canais auxiliares no mesmo intervalo de tempo.

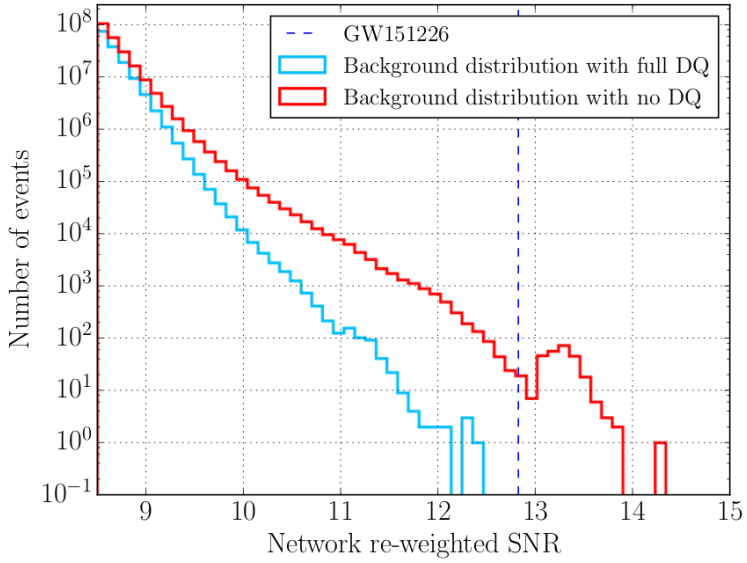


Fonte: Smith et al. (2011).

Um exemplo na prática pode ser visto na Figura 3.6(d). Ela mostra tremores terrestres (em nm/s) na banda de frequência de $0,03$ a $0,1$ Hz. É possível ver que logo depois das 12h há um pico de movimento terrestre; isso justifica o aumento dos transientes (ruidosos) e a queda da sensibilidade do detector, comentados anteriormente. Cada cor representa um canal auxiliar. Por exemplo, a cor vermelha mostra o sensor no espelho ITMX, ou, como descrito, IX.

Uma vez que é conhecida a causa do sinal, é possível descartá-lo como possível sinal proveniente de um evento astrofísico. Esse processo de excluir sinais com ajuda de canais auxiliares é chamado de *Data Quality vetoes* (DQ vetoes). A Figura 3.8 mostra como esses vetos melhoraram a significância para um evento evento de 2015, GW151226 (ABBOTT et al., 2016a). Tratam-se de distribuições de fundo (*backgroud*) de transientes. Todos sinais abaixo desse fundo podem ser considerados aleatórios.

Figura 3.8 - Histogramas de transientes com e sem aplicação de vetos provenientes de canais auxiliares.



Fonte: Abbott et al. (2018)

Em vermelho, o histograma mostra a quantidade de eventos por SNR (*background*) quando os DQ vetoes não são aplicados; em azul, por outro lado, quando os vetos são aplicados e os transientes detectados por outros canais são eliminados do canal gravitacional. Os tracejados verticais representam o evento GW151226 com uma SNR um pouco menor que 13. Note que se os vetos não são aplicados, GW151226 não seria mais significante que a distribuição de fundo. Mais uma vez, a melhoria da qualidade dos dados mostra-se indispensável.

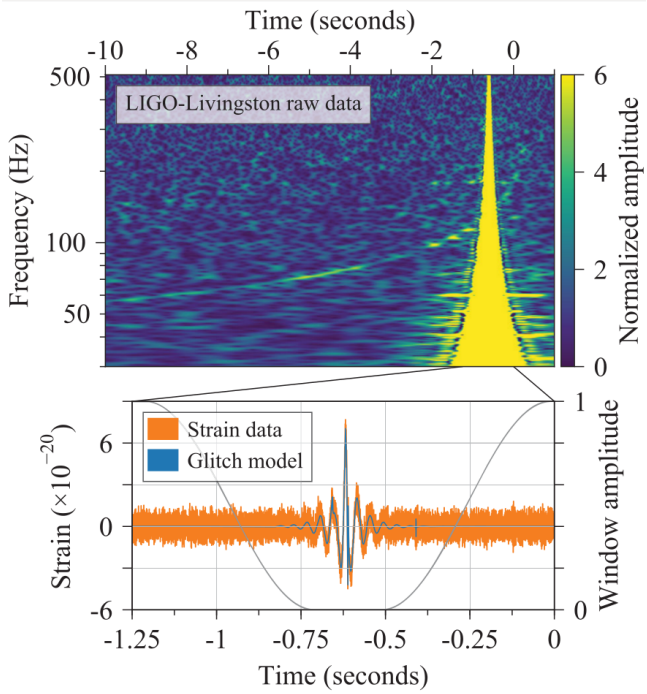
Mesmo que dois candidatos a ondas gravitacionais aconteçam sequencialmente nos dois observatórios LIGO, dentro do intervalo previsto pela relatividade geral, não há garantia de que sejam sinais reais. A taxa de ocorrência de transientes pode ser alta suficiente para que existam coincidências temporais entre dois ruídos. Por isso, estudá-los é essencial para ter uma boa qualidade de dados e buscar os tão esperados sinais gravitacionais. O próximo capítulo mostrará como os transientes ruidosos, também chamados de glitches a partir de agora, são normalmente estudados pela LSC e qual a forma proposta por esta tese.

4 GLITCHES E CRIAÇÃO DOS DADOS

Conhecer como os glitches se comportam é essencial, pois, infelizmente, ainda não é possível evitar que eles interfiram na busca de sinais de ondas gravitacionais. Eles podem mimetizar sinais reais, aumentar o background na busca de ondas e diminuir a significância de um sinal gravitacional. Além disso, eles reduzem o tamanho de dados a serem utilizados. Quando há vetos, dados são descartados e sinais astrofísicos podem ser perdidos.

Um caso real de impacto de glitches aconteceu durante o evento da primeira coalescência de uma binária de estrelas de nêutrons. A Figura 4.1 mostra o sinal tipo *chirp* característico de uma coalescência, porém com um glitch bem alto no final. Nesse caso, o glitch apareceu apenas em Livingston e não teve causa conhecida.

Figura 4.1 - Na parte superior há o sinal do evento GW170817 contaminado por um glitch que foi reconstruído via *wavelet* (parte inferior).



Fonte: Abbott et al. (2017b).

Felizmente, esse glitch também tinha acontecido poucas horas antes e membros da colaboração conseguiram reconstruí-lo via *wavelet* (parte inferior da figura) para retirá-lo do sinal de interesse (ABBOTT et al., 2017b).

Alguns glitches acontecem repetidamente, mas de forma independente. Por isso, uma forma para entender o comportamento deles é indispensável. Uma vez que o Omicron encontrou um transiente no detector, um estudo no plano tempo-frequência, através de um espectrograma, é realizado. Esse espectrograma também é criado a partir da transformada Q, a qual representa o sinal de interesse em combinações de seno-gaussianas.

O espectrograma normalmente usado como referência é o plano Q que contém o tile com maior valor de SNR. A parte superior da Figura 4.1 mostra um espectrograma via transformada Q. O eixo horizontal é o tempo, o vertical é a frequência de pico e a cor representa a energia normalizada, Z, defina para cada tile m (CHATTERJI et al., 2004),

$$Z = \frac{| X_m |^2}{<| X |^2>}; \tag{4.1}$$

o termo superior é o módulo ao quadrado do X (definido na Equação 3.3) para o tile m, e a parte inferior é referente à média de todos os tiles contidos no plano Q.

Nesse espaço de parâmetros, o glitch apresenta uma forma característica e, de acordo com sua morfologia, é nomeado. A Figura 4.2 mostra uma pequena demonstração visual de como é a "arte de nomear os glitches". O primeiro, à esquerda, por exemplo, parece um peixe e foi nomeado *Koi Fish*, o segundo lembra uma gota de chuva e foi nomeado *raindrop*, e assim por diante. Note que todos os espectrogramas apresentados aqui serão via transformada Q.

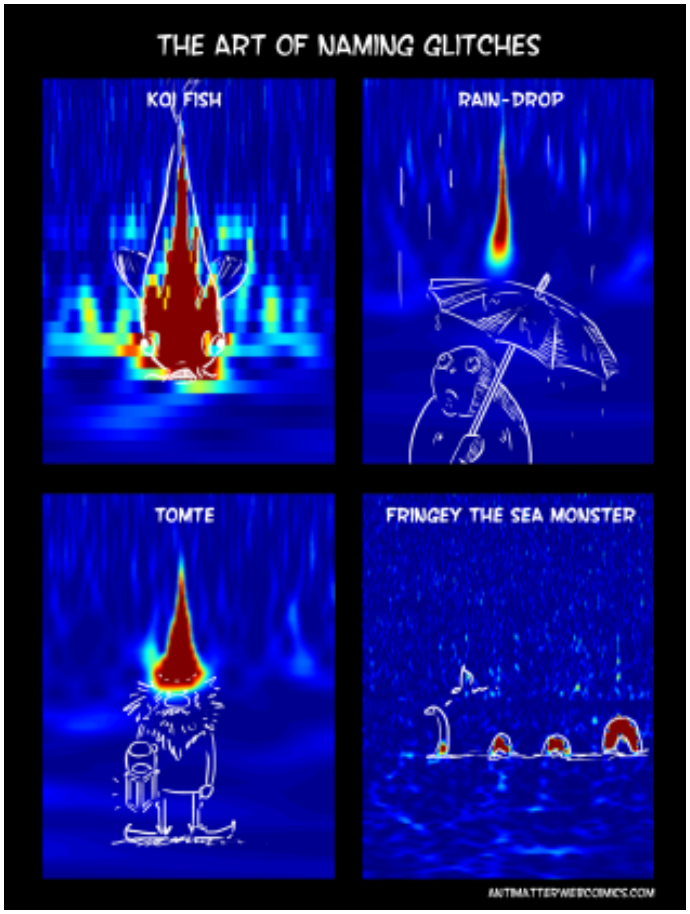
Uma vez que os glitches são identificados, é possível criar um banco de dados para tentar retirar informações importantes. É possível criar gráficos apresentando quais glitches são mais comuns em determinado observatório, quais se repetem, se uma estipulada classe aparece durante o ano todo ou em determinados períodos e tentar, de forma geral, buscar algum padrão no dados. Atualmente, o LIGO conta com o Gravity Spy para identificação e classificação de transientes.

4.1 Gravity Spy

O Gravity Spy, GS, é um projeto que combina ferramentas computacionais como machine learning e pessoas voluntárias de todo o mundo para alcançar um objetivo comum: classificar transientes. Isso acontece através da plataforma Zooniverse (ZOONIVERSE, 2022) que tem diversos projetos (em diferente áreas) e é aberta para quaisquer cidadãos interessados em aprender e auxiliar.

O GS utiliza redes neurais convolucionais (CNN, do inglês *convolutional neural network*) para classificar glitches a partir da imagem morfológica de seu espectrograma. Normal-

Figura 4.2 - A arte de nomear um glitch.

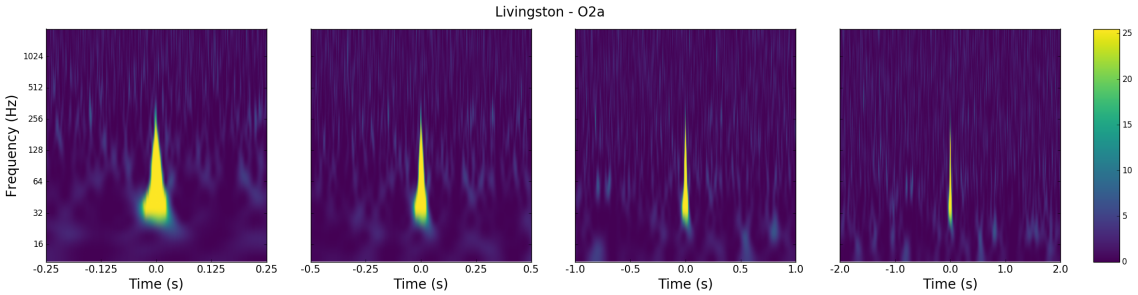


Fonte: Classical and Quantum Gravity (2016).

mente, ele utiliza um conjunto de dados já conhecido como entrada para processar o método CNN. Esse conjunto, também chamado de dados para treinamento da máquina, é composto por combinações de quatro espectrogramas centrados no transiente de interesse. Os espectrogramas de entrada variam em duração: 0,5 s, 1 s, 2 s e 4 s (ZEVIN et al., 2017). Um exemplo pode ser visto na Figura 4.3 para a classe de glitch conhecida como *Blip*.

Os espectrogramas podem ser gerados através do pacote *GWpy*. Ele contém diferentes ferramentas para o estudo (em Python) dos dados de ondas gravitacionais. Mais informações são encontradas em GWpy (2022). Existe também um curso aberto sobre como utilizar tais ferramentas e aprender a manipular os dados do LIGO em GWOSC (Gravitational Wave Open Science Center), acessível em GWOSC (2022).

Figura 4.3 - Espectrogramas de um glitch conhecido como Blip em quadro janelas de duração: 0,5s, 1,0s, 2,0s e 4,0s. Tais espectrogramas compõem a entrada para CNN e classficações via Gravity Spy.



Fonte: LIGO (2021).

As quatro janelas são feitas para tentar obter mais informações possíveis sobre o sinal. Assim, quando o algoritmo recebe a imagem de um glitch aleatório, sem classe, ele calcula a probabilidade dele pertencer a cada uma das classes de glitches já conhecidas. No fim, a classe atribuída a esse sinal vai ser a que tiver a probabilidade maior. Mais detalhes gerais sobre como machine learning funciona serão dados no capítulo 6.

Os voluntários são treinados com espectrogramas de glitches já classificados por cientistas do LIGO. No começo, eles passam a classificar transientes fáceis de serem distinguidos e, conforme acertam as classificações, evoluem para níveis seguintes. Nos níveis finais, o voluntário já é capaz de classificar glitches complexos e auxiliar nas classificações que o algoritmo ainda não tem tanta confiança (GRAVITYSPY, 2022).

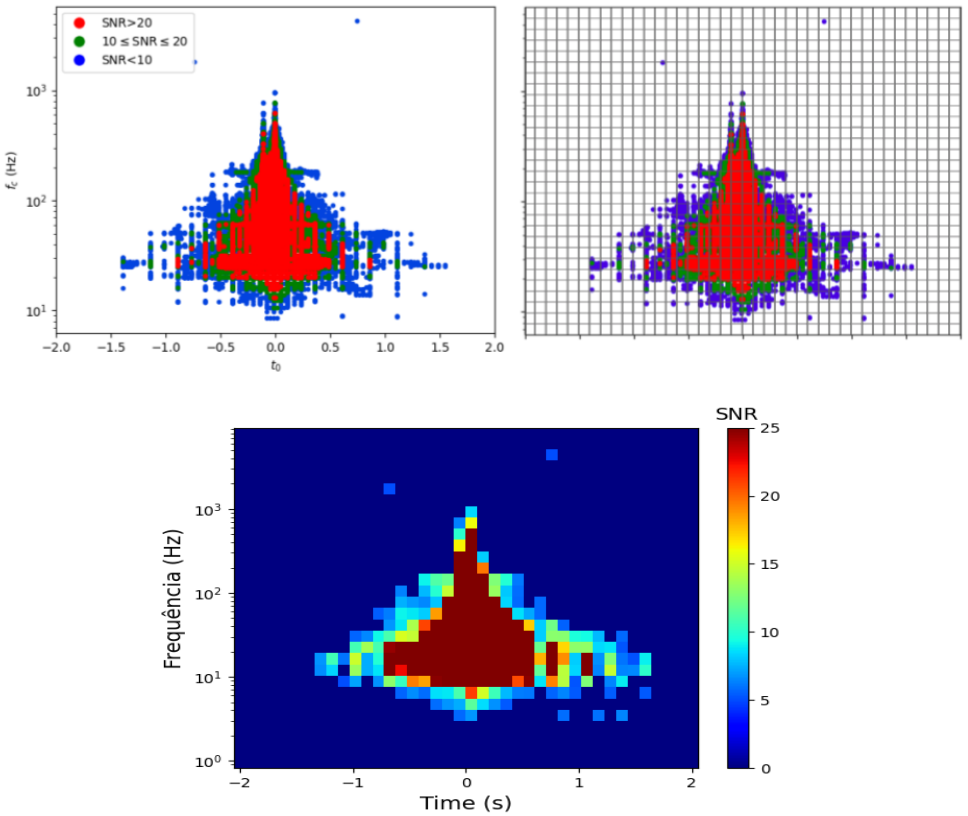
De forma resumida, o GS já tem uma base de glitches rotulados e, a partir dela, classifica novos transientes encontrados na lista do Omicron. Para cada transiente, ele gera o referente espectrograma e faz a análise de imagem via CNN. Quanto mais dados bem classificados, melhor o resultado final, melhor a predição. Para isso, o projeto conta com classificações de voluntários que ajudam a aumentar a base de referência e a auxiliar nas classificações menos confiantes do algoritmo. No fim, baseado na lista de transientes do Omicron, o Gravity Spy cria o próprio arquivo de dados com os tempos dos transientes encontrados, a classe de glitch que ele pertence e os parâmetros como tempo, SNR, amplitude, frequência de pico, largura de banda, duração, etc.

4.2 Glitchgramas - representação alternativa

O estudo desta tese está baseado no uso de *glitchgramas*. Essa caracterização alternativa iniciou-se durante o mestrado que, dentre os diferentes métodos testados, mostrou-se o

mais eficiente (FERREIRA, 2018). O glitchgrama é baseado no uso dos triggers encontrados pelo Omicron, salvos no unclustered file. Esses triggers são salvos com diferentes valores de Q que formam diferentes camadas de espectrogramas (*layers*), cada uma com um valor de Q. Quando visualizados no plano de tempo e frequência central, também desenham a morfologia característica do glitch. Um exemplo pode ser visto na Figura 4.4, à esquerda. Essa imagem mostra os triggers durante um glitch conhecido como *Extremely Loud*. Cada ponto da imagem representa um trigger e a cor corresponde ao intervalo do valor da SNR ao qual ele pertence que, por sua vez, representa um tile da transformada Q.

Figura 4.4 - Passos para construção do glitchgrama. Na parte superior, à esquerda, há a representação no plano tempo-frequência dos triggers encontrados pelo Omicron no instante em que um glitch conhecido como *Extremely Loud* foi classificado pelo Gravity Spy. À direita, há um exemplificação visual para mostrar o processo de construção do glitchgrama. A imagem inferior é o glitchgrama.



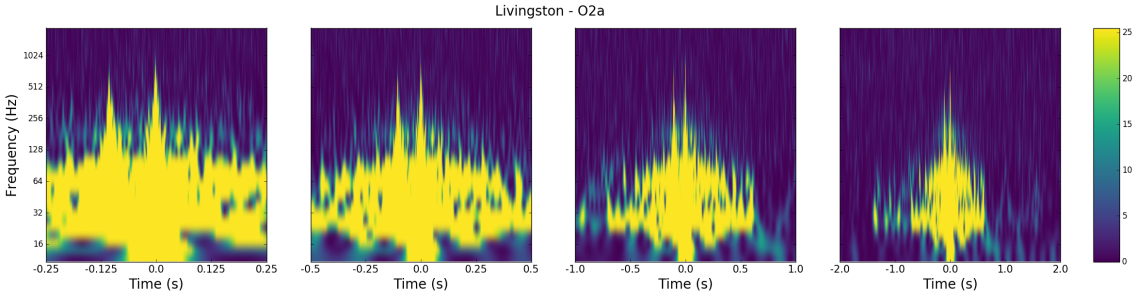
Fonte: Produção da autora.

É interessante ressaltar que a morfologia do glitch já é evidenciada nessa representação. Daí, a ideia de utilizá-la como um caracterizador de glitches, sem a necessidade da geração de espectrogramas.

Para criar a base de dados desta tese, a imagem é dividida em células (ou *bins*) (vide representação visual à direita da Figura 4.4) correspondentes aos tiles dos espectrogramas. Em outras palavras, é como se ela fosse repartida em vários retângulos, onde cada um tem uma duração e uma banda de frequência correspondente. Aqui, cada imagem dessa tem 4 segundos que foram divididos em 40 bins, ou seja, cada bin tem 0,1 segundo de duração. De forma análoga, o eixo vertical de frequências tem 30 divisões.

A partir dessa divisão foi criada uma matriz de dados para representar o glitch, denominada glitchgrama. A cada retângulo foi atribuído um valor de SNR. Os retângulos sem trigger (como nas bordas da Figura 4.4) receberam o valor zero. Os outros, receberam o valor do trigger com maior SNR. A parte inferior da Figura 4.4 mostra o glitchgrama final construído para esse glitch; ela nada mais é que uma matriz de dados (de 30×40 elementos). Essa matriz carrega a morfologia do sinal ruidoso e os valores de SNR em cada elemento, independente do valor de Q. Apenas para comparação, a Figura 4.5 mostra o espectrograma via transformada Q para esse mesmo glitch, o qual é a base de referência para o GS.

Figura 4.5 - Espectrogramas do glitch mencionado para efeitos de comparação com glitchgrama.



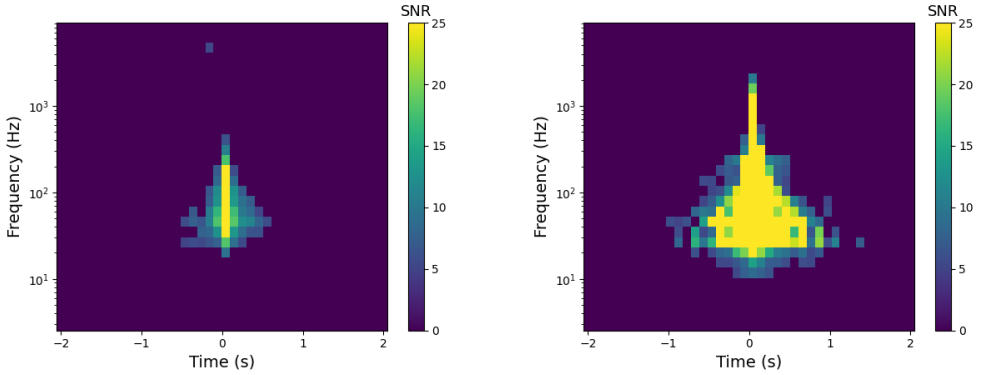
Fonte: LIGO (2021).

O glitchgrama é uma representação com menor resolução, principalmente, se comparado com os espectrogramas analisados pelo GS. No entanto, para criá-lo não há a necessidade de gerar a série temporal em torno do transiente. Isso é um ponto importante quando se pensa em rapidez na análise dos dados.

É importante destacar que o fato de dois glitches pertencerem à mesma classe não significa que sejam idênticos. Um exemplo pode ser visto na Figura 4.6 onde há glitchgramas para dois transientes. Eles aconteceram em tempos diferentes e são totalmente independentes, porém ambos foram classificados pelo Gravity Spy como *KoiFish*. É visível que, apesar

de serem do mesmo grupo, possuem diferenças tanto em duração quanto em SNR. Para facilitar e verificar se existe um padrão mais frequente, o glitchgrama de referência para determinada classe será a média de todos os glitchgramas dos grupos selecionados.

Figura 4.6 - Dois glitches classificados como KoiFish pelo GravitySpy. Eles são semelhantes, mas não idênticos.



Fonte: Produção da autora.

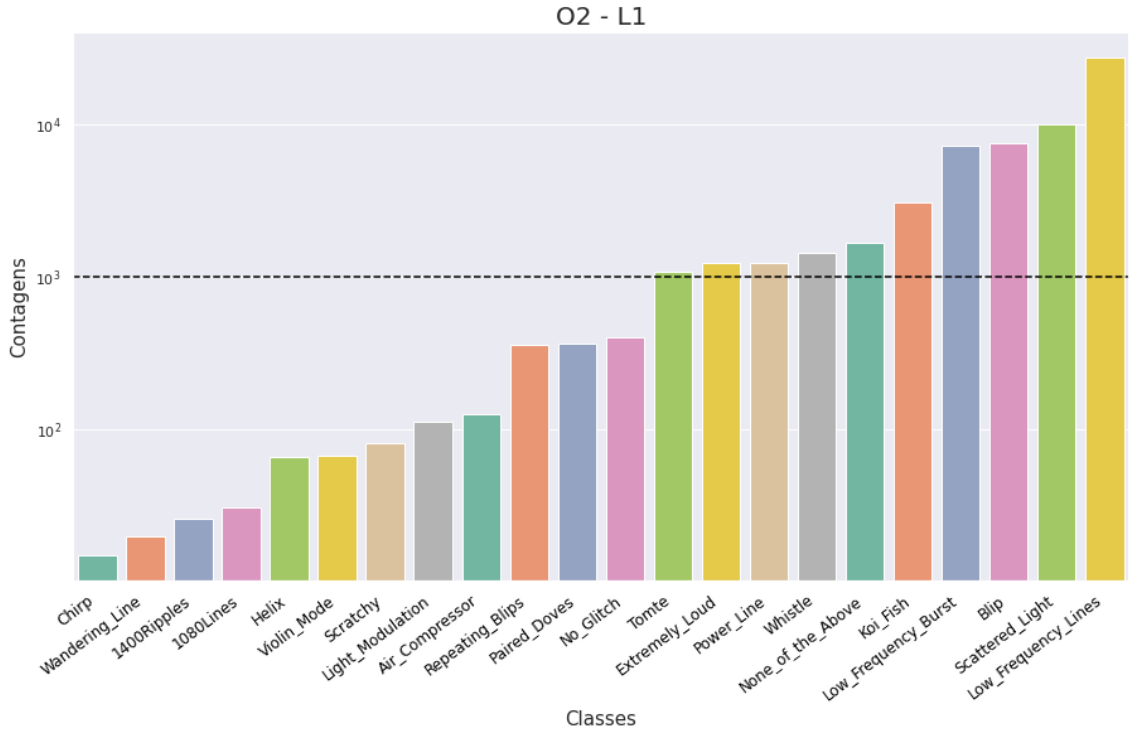
Antes da aplicação de técnicas computacionais para o estudo de glitches (a partir dos glitchgramas), as seções a seguir apresentarão um breve resumo das principais características do transientes mais comuns durante a O2.

4.3 Glitches durante a O2 em LLO

Milhares de transientes já foram encontrados e classificados no LIGO. A Figura 4.7, por exemplo, mostra a contagem deles (por classe) durante a segunda corrida observacional em Livingston. O transiente mais comum foi o chamado *Low Frequency Lines*, com mais de dez mil aparições. Dentre esses glitches, alguns têm causas conhecidas e outros, não. Esta tese pretende usar o glitchgrama para aplicar ferramentas computacionais e saber o quanto ele é bom para caracterizar um glitch e, dessa forma, verificar a possibilidade de utilizá-lo para buscar indícios de origens dos ruídos.

As ferramentas computacionais, em especial de aprendizado de máquina, precisam de dados conhecidos para uma boa criação de modelo de predição. Em geral, quanto mais dados, melhor. Por isso, esse estudo limitou-se aos glitches com no mínimo mil ocorrências no intervalo da O2, que estão demarcadas pela linha preta tracejada da Figura 4.7. Dessa forma, apenas os glitches *Tomte*, Extremely Loud, *Power Line*, *Whistle*, Koi Fish, *Low Frequency Burst*, Blip, *Scattered Light* e Low Frequency Lines estão sendo estudados aqui.

Figura 4.7 - Quantidade de glitches durante a O2 em Livingston.



Fonte: Produção da autora.

Note que a categoria *None of the above* não será utilizada na análise. Ela é uma categoria criada no GS para permitir que os voluntários atribuam uma classe para um glitch que não corresponda a nenhuma das outras, também para evitar erros por dúvidas no momento da atribuição. Assim, eles também contribuem para a descobertas de novas classes (SONI et al., 2021).

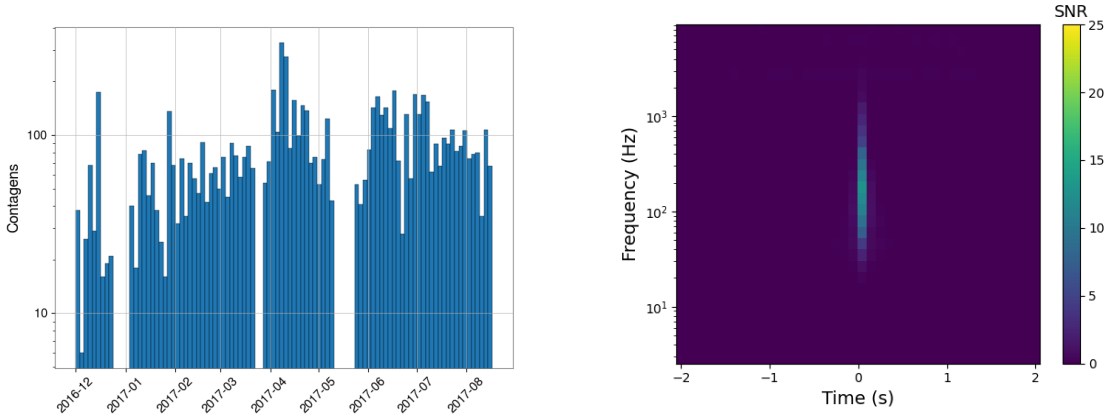
Cada uma das subseções a seguir vai apresentar algumas características de uma dessas nove classes. Haverá também o glitchgrama médio de cada categoria e alguns gráficos baseados nas informações da lista de glitches classificados pelo GS durante a O2 no LLO. Os glitchgramas médios são de exatamente mil glitches de cada uma das classes. Eles foram selecionados a partir da ordem de confiança na classificação do GS.

4.3.1 Blip

O Blip é uma das classes que mais incomoda os pesquisadores, pois além de estar presente nos dados (em alta quantidade), pode ser muito parecido com sinal tipo chirp (sinais indicativos de coalescências de binárias compactas), atingindo, inclusive, frequências si-

milares. Durante a O2, por exemplo, essa classe foi a terceira mais comum. A Figura 4.8 apresenta a quantidade de blip durante a O2 por mês. O pico foi em abril, mas há uma certa constância no tempo. À direita, há o glitchgrama médio dos mil glitches usados para estudo de comparação.

Figura 4.8 - A Figura (à esquerda) apresenta a quantidade de glitches da classe Blip por mês, durante a O2, e seu glitchgrama médio (à direita).

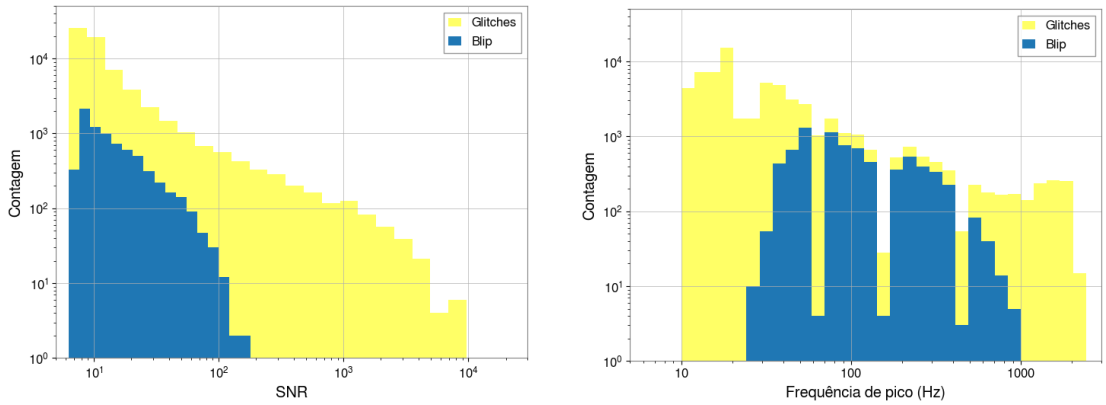


Fonte: Produção da autora.

Blips normalmente têm curta duração (em torno de 10 ms) e aparecem num formato de lágrima. Seu espectrograma também pode ser visto na Figura 4.3 em quatro diferentes janelas de tempo. A Figura 4.9 mostra que eles ocupam a região de valores de SNRs relativamente baixos (à esquerda, em azul), com valor máximo próximo a 200. A cor amarela representa a distribuição de SNR de todos as outras classes de glitches presentes na O2 (e mostradas no histograma da Figura 4.7). À direita dessa mesma imagem, é possível ver a distribuição da frequência de pico dessa classe. Ela aborda uma região considerável e aparenta ter quatro principais bandas.

Existe uma outra classe de glitches relacionada com Blip. É a chamada *Repeating Blips*. Ela é composta basicamente por múltiplos glitches tipo Blip que se repetem entre mais ou menos 0,25 a 0,50 s. Apesar do Blip ser um glitch comum nas corridas observacionais, ainda não é conhecida sua causa; também não há evidências que confirmam canais auxiliares envolvidos com ele (CABERO et al., 2019).

Figura 4.9 - Histogramas de SNR (esquerda) e de frequência de pico (direita) do Blip. Ele está em azul e é comparado com o histograma de todos os outros glitches em amarelo.

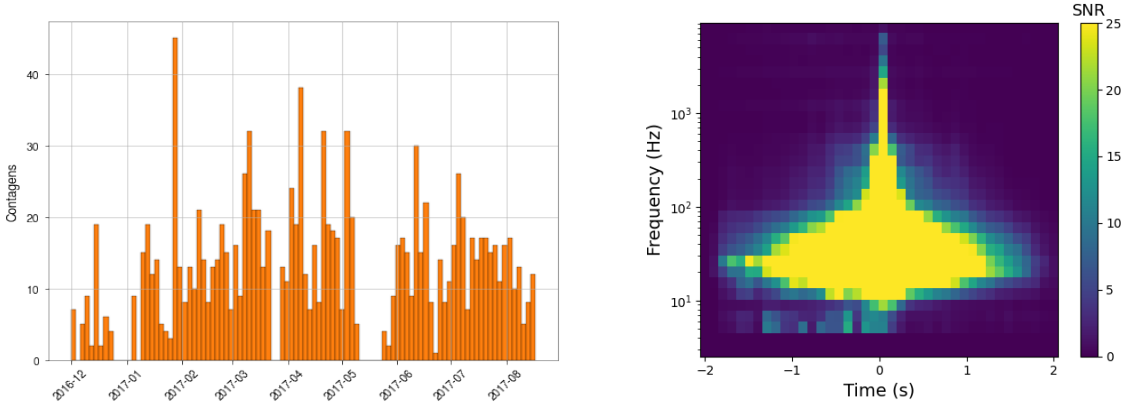


Fonte: Produção da autora.

4.3.2 Extremely Loud

A classe Extremely Loud, como o próprio nome sugere, é uma classe de ruído muito forte. Muito forte aqui significa com alto valor de SNR. Por tal característica, os espectrogramas (e glitchgramas) sempre aparecem saturados (vide glitchgrama médio à direita da Figura 4.10). Apesar de não ser tão comum quanto outras classes, tais transientes abrangem várias frequências e correspondem a fortes perturbações no detector que podem fazer com que haja queda de sensibilidade (como a mostrada na Figura 3.6(c)) (GLANZER et al., 2022).

Figura 4.10 - Quantidade de Extremely Loud durante a O2 e seu glitchgrama médio.

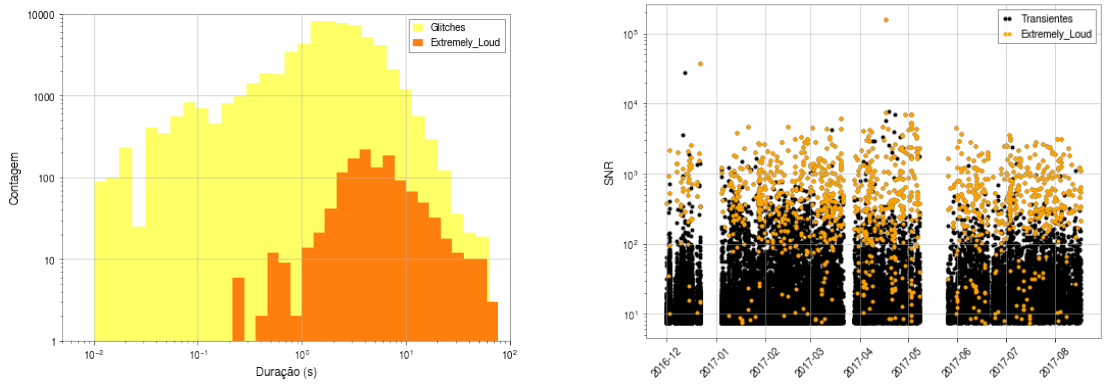


Fonte: Produção da autora.

Esses transientes compõem praticamente toda a região de glitches com alto SNR. A Figura 4.11, à direita, mostra um gráfico de dispersão de SNR por tempo dos transientes encontrados durante a O2 (em preto); os que foram classificados como Extremely Loud estão em laranja e é visível como eles estão praticamente em toda a região superior. Sua longa duração fica bem evidente no histograma à esquerda da mesma imagem.

Outros glitches, em especial o Koi Fish (mostrado a seguir), podem ter SNR suficientemente grande para também fazerem parte dessa categoria. Não há uma causa específica para esse tipo de transiente. Ele pode ser gerado por diferentes fatores que causem perturbações no posicionamento e alinhamento dos espelhos. O espectrograma de um Extremely Loud aleatório pode ser visto na Figura 4.5.

Figura 4.11 - Histograma de duração dos transientes Extremely Loud em laranja que é comparado com o histograma de todos os outros glitches em amarelo e gráfico de dispersão de SNR no tempo dele e outros transientes (em preto).



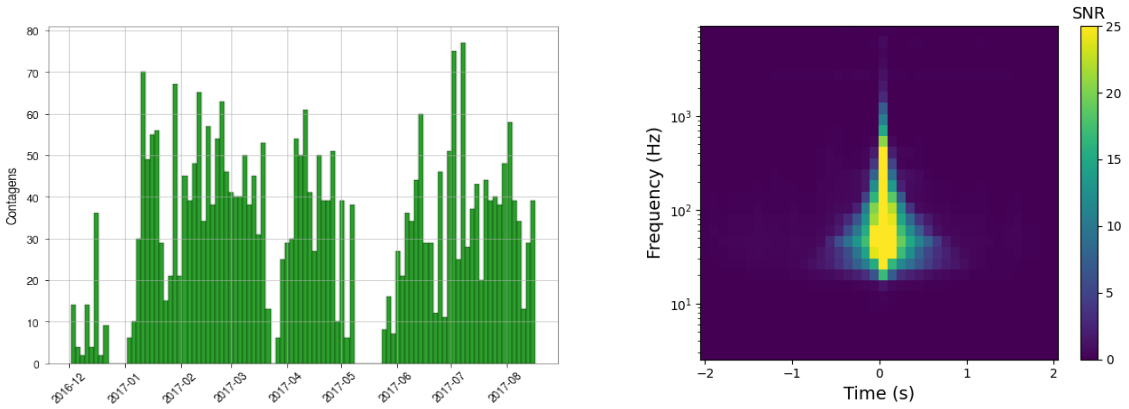
Fonte: Produção da autora.

4.3.3 Koi Fish e Tomte

Tanto o Koi Fish quanto o Tomte são glitches morfologicamente muito semelhantes ao Blip. No entanto, na parte inferior da morfologia do Koi Fish, em frequências mais baixas, há um achatamento que lembra a cabeça de um peixe. Além disso, normalmente, o Koi Fish tem valores de SNR e de duração mais altos. O campo de guia do site Gravity Spy diz que ele pode ser uma subclasse do Blip e suas origens físicas ainda não são conhecidas (GRAVITYSPY, 2022). A Figura 4.12 à esquerda mostra as aparições dessa classe durante a O2 e, à direita, seu glitchgrama médio.

Por outro lado, o Tomte difere-se do Blip por sua forma triangular, lembrando um cha-

Figura 4.12 - Ocorrência de Koi Fish durante a O2 (à esquerda) e seu glitchgrama médio (à direita).



Fonte: Produção da autora.

péu usado por gnomos. Ele pode ser visualizado na parte inferior, à esquerda, da Figura 4.2. Uma outra diferença entre eles é que o Tomte se apresenta em frequências menores, chegando até em torno de 100 Hz.

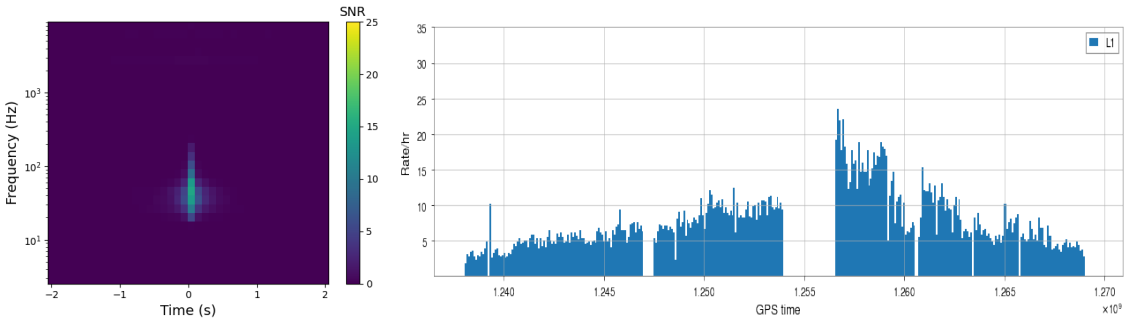
O Tomte teve apenas 1086 aparições e ficou próximo do limite mínimo imposto para estudos desta tese. No entanto, essa classe tem incomodado os pesquisadores, pois sua ocorrência aumentou significantemente na terceira corrida observacional. Na O3a, por exemplo, ele foi o segundo glitch mais comum, aparecendo cerca de 24 mil vezes. E, apesar de não ter sido o mais comum da O3b, teve mais de 26 mil aparições. A Figura 4.13 mostra (à esquerda) o glitchgrama médio do Tomte. Se comparado com Koi Fish, é visível como atinge frequências e SNR menores.

A Figura 4.13 (à direita) também mostra a taxa de ocorrência do Tomte por hora durante a O3. O espaço em branco da imagem é o intervalo entre a O3a e a O3b. É curioso que seu comportamento aparenta ser sazonal, com um aumento significante em torno de novembro (início da O3b). Um estudo mais específico precisa ser feito para afirmar isso e relacioná-lo com alguma causa específica. Por enquanto, há também suspeitas de que Tomte seja uma subclasse do Blip. https://pt.overleaf.com/project/61747c59f108284e694600ec

4.3.4 Low Frequency Burst e Low Frequency Lines

As classes Low Frequency Burst (LFB) e Low Frequency Lines (LFL) acontecem em baixas frequências. A LFB fica na região em torno de 10 a 20 Hz, tem duração curta e, devido à

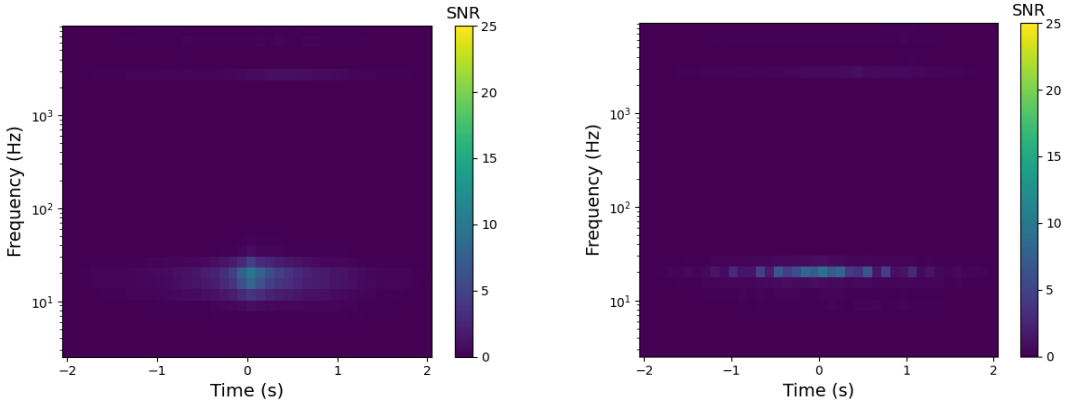
Figura 4.13 - Glitchgrama médio do Tomte e a sua taxa de ocorrência por hora durante a O3.



Fonte: Produção da autora.

sua morfologia, também é chamada de corcova. Por outro lado, a LFL pode ter duração maior (até 4 segundos) e aparece normalmente como uma linha horizontal em torno de 20 Hz. Ambas têm SNR relativamente pequenos e a Figura 4.14 apresenta os glitchgramas médios da classe LFB, à esquerda, e da LFL, à direita.

Figura 4.14 - Glitchgramas médios do Low Frequency Burst (esquerda) e Low Frequency Line (direita).



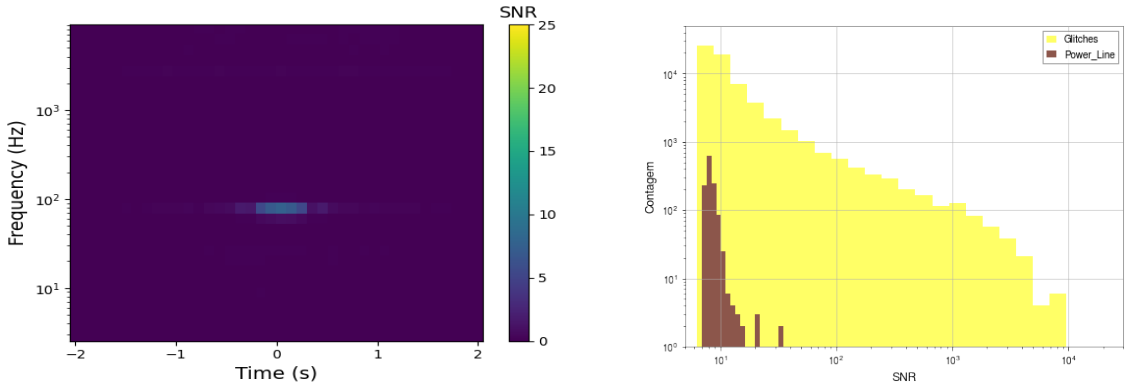
Fonte: Produção da autora.

Há evidências que o LFB seja causado por espalhamento da luz por movimento dos espelhos que apontam o feixe para o OMC. No entanto, ainda não se conhece o motivo do movimento desses espelhos (GRAVITYSPY, 2022).

4.3.5 Power Line

Nos Estados Unidos, a principal alimentação dos equipamentos é pela corrente alternada na frequência de 60 Hz. Como mencionado na Seção 3.2, tal corrente gera linhas verticais na sensibilidade do detector. Isso ocorre na frequência de 60 Hz. No entanto, às vezes, alguns equipamentos podem sofrer falhas, ligar e desligar, causando também um glitch em 60 Hz (ou harmônicos); esse é classificado como Power Line. A Figura 4.15 mostra o glitchgrama médio para o Power Line e evidencia o quanto a SNR dessa classe é pequena, se comparada com outros glitches.

Figura 4.15 - Glitchgrama do Power Line (à esquerda) e seu histograma de SNR.

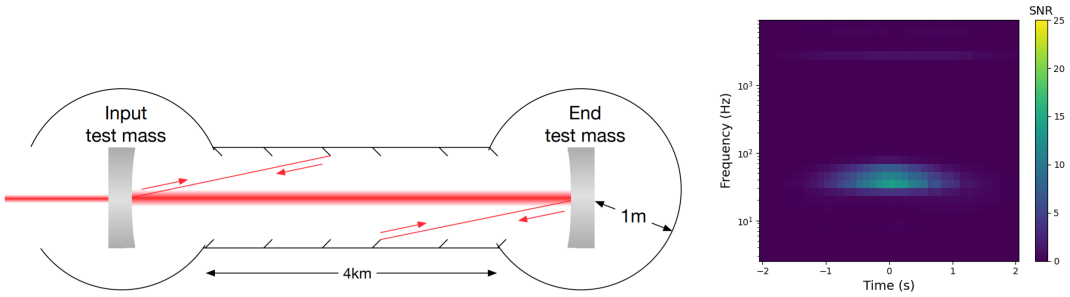


Fonte: Produção da autora.

4.3.6 Scattered Light

O Scattered Light, como o nome também sugere, acontece por espalhamento da luz do laser. Quando o feixe atinge o espelho, o fotodetector e o divisor de feixes, parte da luz pode ser refletida em direções aleatórias e direcionada às paredes da câmara de vácuo. Essas, por sua vez, podem redirecionar a luz ao feixe principal com uma diferença de fase (SENGUPTA, 2016), criando o glitch. Esse transiente acontece em ambos observatórios e normalmente em frequências em torno de 30 a 40 Hz. A Figura 4.16(a) ilustra esse efeito de espalhamento e a Figura 4.16(b) apresenta o glitchgrama médio do Scattered Light.

Figura 4.16 - Efeito por espalhamento do laser que causa transientes nos dados do LIGO (esquerda) e glitchgrama médio do Scattered Light (direita).



(a) (b)

Fonte: a) Sengupta (2016); b) Produção da autora.

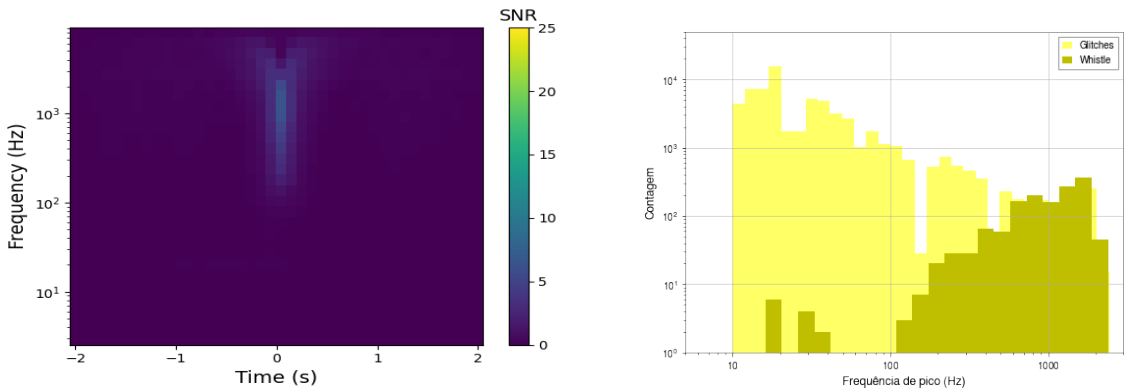
O espalhamento do laser também pode ser causado por movimentos das massas testes ou do divisor de feixes. Os glitches devido a tal espalhamento estão sendo os mais intrigantes atualmente, pois foram os mais presentes durante a terceira corrida observacional. Além disso, surgiu até uma nova classe de glitches por espalhamento durante a O3; ela é chamada de *Fast Scattering*. Para ter uma ideia, só na O3b, mais de cem mil transientes devido ao espalhamento da luz foram encontrados. Há um capítulo (Capítulo 7) com mais informações e estudos só sobre esses dois ruídos.

4.3.7 Whistle

Finalmente, o Whistle é o último glitch a ser estudado aqui. Essa classe também esteve presente na O3a e O3b e, comumente, aparece em forma de V ou W no espectrograma. Ele também é conhecido como notas de batimento de rádio frequência e é causado por sinais de rádio (em MHz) dos osciladores controlados por voltagem no LIGO (NUTTALL et al., 2015).

A Figura 4.17 mostra seu glitchgrama médio. Apesar de ter valores de SNR baixos, o formato em 'V' é visível. Além disso, ele praticamente é o responsável por toda a região de altas frequências, passando até de 2000 Hz.

Figura 4.17 - Glitchgrama representativo da classe Whistle (à esquerda) e seu histograma da frequência de pico.



Fonte: Produção da autora.

Com as informações dos glitches selecionados e a apresentação dos conceitos envolvidos, é possível, finalmente, criar o banco de dados a ser analisado.

4.4 Um resumo da criação dos dados a serem analisados

Conhecendo as principais características das nove classes selecionadas para estudo, é interessante aplicar técnicas computacionais para verificar o quão bom o glitchgrama é para caracterizar um glitch e buscar informações relevantes. A Figura 4.18 mostra o resumo do passo a passo para construção dos dados a serem analisados. Lembrando que foram escolhidos mil glitches de cada uma das nove classes apresentadas anteriormente. Eles foram selecionados a partir da lista gerada pelo Gravity Spy.

Nessa lista há todos os glitches classificados com o tempo em que eles aconteceram no observatório. Uma vez que cada glitch tem um tempo GPS correspondente, os passos seguintes são (note que cada passo está relacionado a um processo da Figura 4.18):

a) baseado no tempo GPS de um glitch secionado (t_{xxx}), os triggers do unclustered file são acessados. Todos os triggers entre (t_{xxx} - 2s) e (t_{xxx} + 2s) são salvos num outro arquivo de dados. A partir dele, é possível fazer um plot e ver sua morfologia em tempo e frequência central;

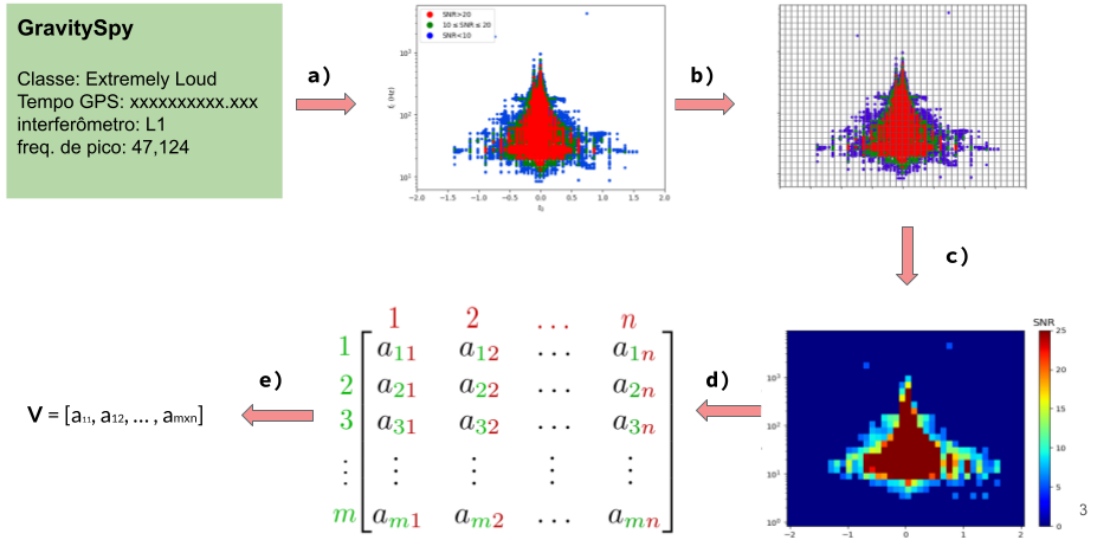
b) uma vez que os triggers formam a morfologia do glitch, essa imagem é dividida em 30×40 bins;

c) cada bin recebe apenas o valor do trigger de maior SNR, criando o chamado glitchgrama;

d) o glitchgrama é transformado numa matriz de dados, onde cada elemento carrega o valor da SNR correspondente;

e) a matriz de dados é transcrita num vetor final de 1200 dimensões.

Figura 4.18 - Passo a passo da criação dos dados a serem analisados.



Fonte: Produção da autora.

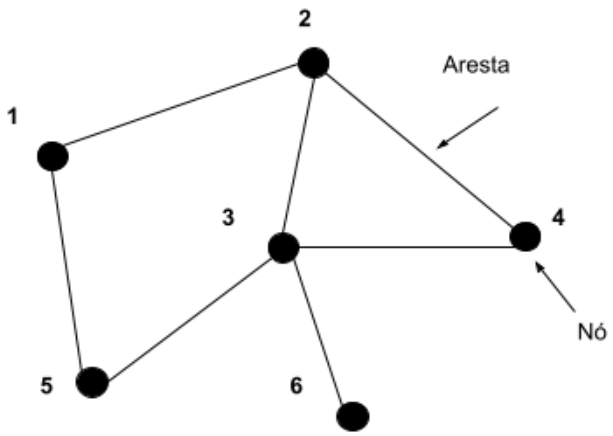
O processo listado acima é feito para todos os transientes ruidosos selecionados. Dessa forma, há um vetor representando cada glitch e nove mil vetores compõem a base de dados a ser analisada nos capítulos seguintes. Cada método irá validar se o glitchgrama é realmente um bom caracterizador de glitches.

5 ANÁLISE DE REDES E O ESTUDO DOS GLITCHES

A análise de redes é um estudo que busca informações e interações entre objetos. Está presente na biologia, física, matemática, internet e diversas áreas no mundo. De acordo com Newman (2018), uma rede (ou network) pode ser definida como a coleção de pontos unidos em pares por linhas. Os pontos, normalmente, são chamados de nós (ou vértices) e as linhas são as arestas (ou links).

Um exemplo simplificado de rede pode ser visto na Figura 5.1; tal representação também é chamada de grafo. Nesse caso esquematizado, há seis nós e sete arestas. Os nós são referenciados de 1 a 6 e as arestas podem ser especificadas por uma lista de pares: $\{(1,2),(1,5),(2,3),(2,4),(3,4),(3,5),(3,6)\}$; cada par é representado pelos números dos nós conectados.

Figura 5.1 - A representação de um grafo.



Fonte: Adaptada de Newman (2018).

Uma outra representação matemática e a mais comum para representar um grafo é a matriz de adjacência A_{ij}. Para o exemplo acima, ela poderia ser expressa pela Matriz 5.1. Cada linha i ou coluna j representa um nó e os valores diferentes de zero definem uma aresta. Note que, nesse caso, trata-se de uma matriz simétrica, pois a aresta que conecta os nós i e j é a mesma que conecta j e i.

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 2 & 2 & 0 & 0 \\ 0 & 2 & 0 & 1 & 3 & 1 \\ 0 & 2 & 1 & 0 & 0 & 0 \\ 1 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \tag{5.1}$$

Com esse tipo de estudo é possível criar um ambiente comum e fazer predições. Um exemplo é a famosa rede social. Cada nó pode ser uma pessoa e cada aresta uma amizade ou uma página a ser seguida/curtida. Um perfil social com muitos amigos terá várias arestas. Cada vez que uma amizade é conectada a alguém ou um assunto é sinalizado como de interesse, mais nós conectados vão se criando em torno da pessoa principal. A Figura 5.1 é só um exemplo de poucos nós, mas num caso real, pode haver centenas, milhares ou milhões deles.

Com os nós representando pessoas ou eventos de interesse, a rede começa a manifestar um certo padrão no perfil central. E com isso, passa a fazer sugestões de amizades, eventos, ofertas de emprego, páginas e etc. Se há muitas pessoas envolvidas, cada uma delas com seus correspondentes nós, aglomerados ou comunidades começam a surgir. Tais comunidades têm atividades, buscas e interesses similares.

Aqui entra um conceito importante: a similaridade. Ela mede o quão similar um objeto é do outro. Para o exemplo citado, suponha a existência de três pessoas: a A, a B e a C. Se A e B têm cem amigos em comum e, A e C têm cinco, então, diz-se que a similaridade entre os nós A e B é maior do que entre A e C. Na matriz de adjacência esse conceito pode vir em forma de pesos. Da Matriz 5.1, por exemplo, diz-se que a conexão entre 3 e 2 é duas vezes mais forte do que a 3 e 6. As arestas e seus pesos podem ter diferentes interpretações, dependendo do que a rede representa; as mais comuns são: amizade, fluxo, similaridade e distância (PLATT, 2019).

Esse conceito incentivou a aplicação no estudo de glitches. Se o glitchgrama é um bom caracterizador e cada glitch é representado por um nó, glitches similares vão fazer parte da mesma comunidade (classe) que, a princípio, deve estar pouco conectada a outro grupo que tem seus glitches mais conectados entre si e que, portanto, compõe outra classe. Se a rede é bem conhecida, é possível prever a qual classe um glitch desconhecido vai pertencer.

Há diferentes maneiras de medir similaridade entre dados. Como o glitchgrama forneceu vetores representativos para o glitch, então, o método utilizado aqui será o cosseno de similaridade.

5.1 Cosseno de similaridade

O cosseno de similaridade, como o nome sugere, mede a similaridade entre par de vetores, analisando a colinearidade e construindo uma rede de relação a partir do ângulo entre eles. Essa análise é utilizada frequentemente na busca de similaridade em textos e na detecção de plágios. O documento é convertido num vetor de n dimensões. Cada elemento é associado a uma palavra e o valor é o número de vezes em que a palavra aparece no documento (METCALF; CASEY, 2016). Além de identificação de plágio, buscas de artigos e documentos semelhantes na internet podem ser feitas com esse método.

Sejam dois vetores $\mathbf{x} = (x_1, x_2, ..., x_n)$ e $\mathbf{y} = (y_1, y_2, ..., y_n)$ separados por um ângulo θ; então, o cosseno entre eles é

$$\cos\theta = \frac{\mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}||\ \ ||\mathbf{y}||}, \tag{5.2}$$

onde, $\mathbf{x} \cdot \mathbf{y}$ é o produto interno entre os vetores, definido por

$$\mathbf{x} \cdot \mathbf{y} = \sum_{1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + ... + x_n y_n, \tag{5.3}$$

e $||x||$ é a norma do vetor $\mathbf{x}$, ou seja, a raiz quadrada da soma dos quadrados de cada elemento que também pode ser definida como o tamanho do vetor:

$$||\mathbf{x}|| = \sqrt{\mathbf{x} \cdot \mathbf{x}} = \sqrt{\sum_{i=1}^{n} x_i^2}. \tag{5.4}$$

O objetivo é encontrar um padrão ou grupos entre os vetores de dados que definam as classes de interesse apresentadas no capítulo anterior. Se o cosseno entre dois vetores (ou dois glitches) for 1, significa que eles são paralelos e proporcionais; se for zero, os vetores são ortogonais; se os vetores forem paralelos, mas com sentido contrário, terão um cosseno -1. Dessa forma, quanto mais o cosseno entre dois vetores for próximo de 1, maior a probabilidade deles pertencerem à mesma classe de glitch.

Sabendo disso, os vetores foram arranjados em uma matriz $\mathbf{M}$. Cada linha de $\mathbf{M}$ é um vetor e cada coluna corresponde ao elemento vetorial. Ou seja, se há 9000 glitches e cada glitch tem 1200 (30×40) dimensões, a matriz final $\mathbf{M}$ terá 9000×1200 elementos. O próximo passo foi multiplicar $\mathbf{M}$ pela sua transposta, obtendo o produto escalar entre todos os vetores. Por fim, cada linha foi dividida pelo produto das normas dos vetores multiplicados e assim, uma matriz de cossenos foi construída. Essa matriz corresponde a

nada mais que a matriz de adjacência $\mathbf{A}$ (de 9000×9000 elementos) para esse conjunto de dados.

Com a matriz de adjacência construída para os glitches (a partir dos glitchgramas), o *NetworkX* (NETWORKX, 2014-2022), pacote em Python gratuito e aberto, foi utilizado para analisar e visualizar as conexões entre os glitches. Ele é um pacote para criação, manipulação e estudo de redes complexas (PLATT, 2019). Além disso, retorna uma estrutura de grafos a partir da matriz $\mathbf{A}$ e permite sua visualização na forma da Figura 5.1. Vale ressaltar que a matriz $\mathbf{A}$ teve a diagonal zerada manualmente, pois não há interesses na similaridade entre um glitch com ele mesmo. Uma vantagem desse método é a possibilidade de trabalhar com vetores esparsos, o que torna muito mais interessante para tais aplicações, visto que os glitchgramas têm diversos *pixels* nulos.

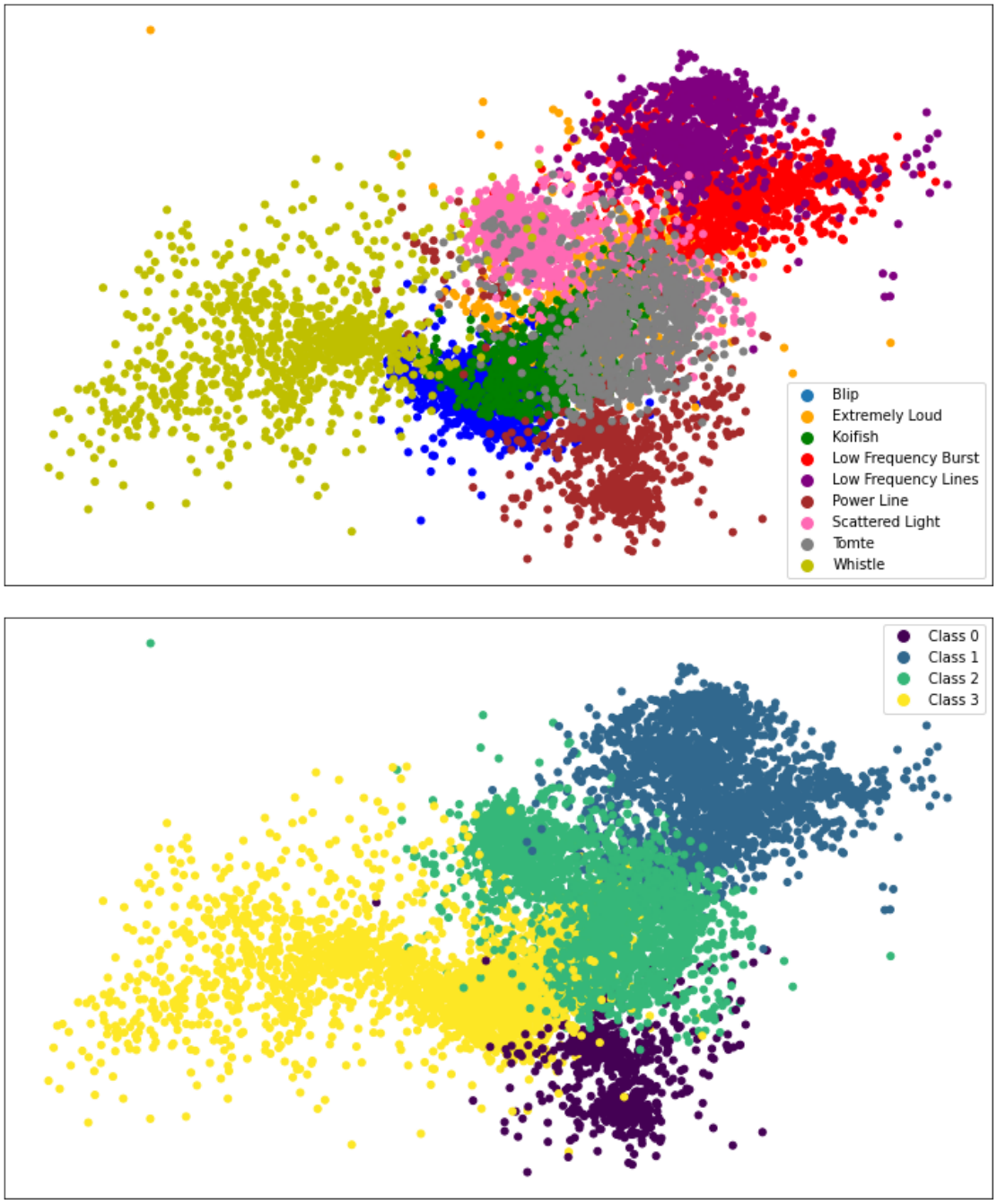
A parte superior da Figura 5.2 exibe o grafo a partir da matriz de adjacência desenhado pelo pacote NetworkX. Cada nó é um glitch e a cor representa a classe atribuída pelo Gravity Spy. Nesse caso, as arestas não foram desenhadas por efeitos de simplicidade.

O pacote oferece vários layouts para visualização dos dados. Pode ser em formato aleatório, circular, em três dimensões ou no formato de interesse do usuário. A escolhida para representar os dados dos glitches foi a *spring* que, inclusive, está na Figura 5.2 e é responsável pelas coordenadas das posições de cada glitch. Para este estudo, as posições (coordenadas x e y da Figura 5.2) são indiferentes; por isso, elas foram retiradas da imagem. Independente de onde os nós estejam, o que importa é retirar informações da rede; a identificação da presença de classes de glitches é uma delas.

O layout spring posiciona os nós a partir do algoritmo direcionado à força de Fruchterman-Reingold (FRUCHTERMAN; REINGOLD, 1991; KOBOUROV, 2012). Este trata cada aresta como uma mola que exerce uma força de atração entre os nós conectados; cada nó é considerado uma carga elétrica de forma que, se dois nós não são conectados pela mola, eles se repelem por uma força do tipo eletrostática. Como consequência disso, nós conectados (com similaridade) se atraem formando aglomerados que se repelem dos grupos não conectados. No fim, o algoritmo afasta e aproxima os nós para encontrar um equilíbrio que minimiza a energia do sistema.

Os grupos formados podem ser visualizados pelas cores na Figura 5.2. As classes Whistle e Power Line estão bem identificáveis, sobreposições entre LFL e LFB são notáveis e há uma região central que apresenta várias intersecções e atrapalham a visualização. No entanto, nesse caso, isso não implica se este é um método bom ou não. Para verificar estatisticamente a eficiência dele a partir do uso do glitchgrama e retirar informações importantes da rede, é necessário aplicar um localizador de comunidades, independente do Gravity Spy.

Figura 5.2 - A parte superior mostra o grafo criado a partir da matriz de adjacência (calculada através do cosseno de similaridade) desenhado pelo pacote NetworkX. Cada nó representa um glitch e cada cor é a classe atribuída pelo Gravity Spy. A parte inferior apresenta as quatro classes encontradas pelo Best Partition, independente das classificações do GS. As arestas entre os nós foram eliminadas para efeito de visualização.



Fonte: Produção da autora.

Localizar grupos ou comunidades significa, como mencionado no conceito da similaridade, encontrar um conjunto de nós densamente conectados que são pouco conectados com outro conjunto. A finalidade é obter sub-redes nessa rede geral de glitches que representem as nove classes escolhidas para estudo. Se isso é feito com sucesso, o método aplicado ao glitchgrama é capaz de analisar e classificar bem um transiente desconhecido.

Para esse tipo de análise, foi utilizado um outro pacote em Python: o *Best Partition* (AYNAUD, 2018). O Best Partition particiona o dados baseado no método *Louvain* e é um localizador de grupos muito popular por ser rápido e fácil de implementar em grandes networks (BLONDEL et al., 2008). Em suma, ele busca comunidades no grafo (criado a partir do NetworkX) e atribui classes a elas.

O algoritmo baseia-se na modularidade (BLONDEL et al., 2008) que é uma medida da qualidade da partição dos nós. Matematicamente é dada por

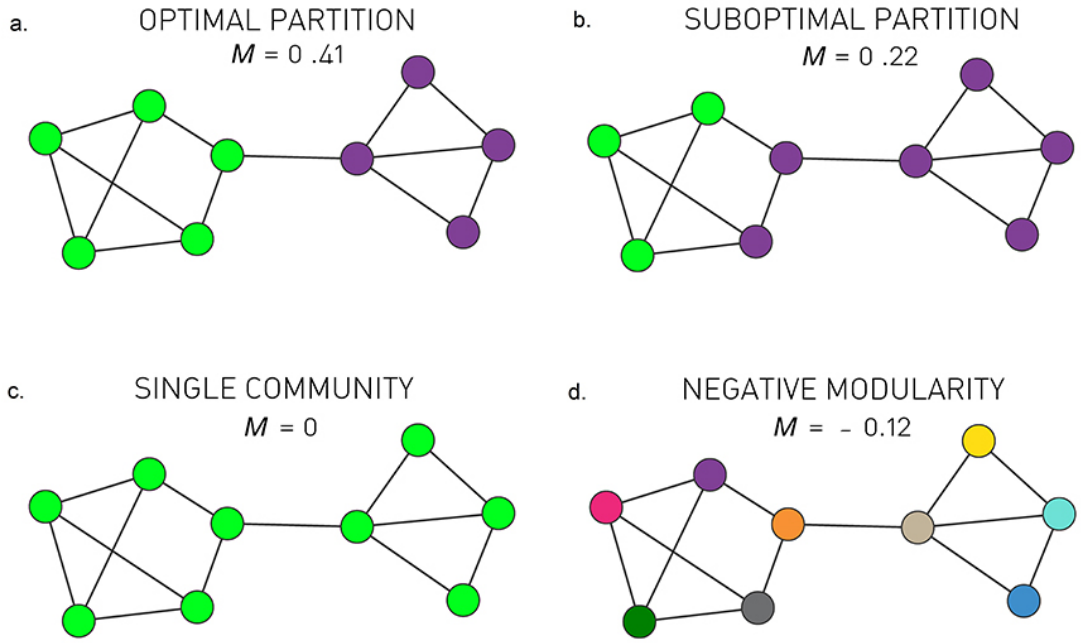
$$Mod = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j), \tag{5.5}$$

onde A_{ij} é peso da aresta entre i e j (da matriz de adjacência), k_i é a soma dos pesos das arestas anexadas ao nó i, c_i é a comunidade na qual o nó i é atribuído, $\delta(c_i, c_j)$ é o delta de Kronecker e m é a soma de todos os pesos $m = 1/2 \sum_{ij} A_{ij}$. O valor da modularidade varia de -1 a 1 e quanto maior o valor, melhor.

A Figura 5.3 mostra um exemplo de como seriam as partições de acordo com o valor da modularidade; cada cor indica uma uma comunidade encontrada. Na parte (a) há a modularidade ideal, onde existem duas classes e ambas são encontradas; na parte (b) uma modularidade abaixo do ideal, onde um grupo (em roxo) foi totalmente encontrado, mas dois nós da classe verde foram atribuídos à classe errada (roxa); na parte (c) é uma modularidade nula que indica que todos os nós pertencem à mesma comunidade; por fim, a de valor negativo (d) que atribui uma classe para cada nó.

O objetivo do Best Partition é procurar por uma modularidade ótima. Para isso, ele inicialmente atribui uma classe para cada nó. Feito isso, ele muda a classe do nó i para a classe do nó j. Se não houver ganho na modularidade, o nó i fica na mesma comunidade; caso haja, o nó i vai pertencer à classe do vizinho que ofereceu maior ganho de modularidade. Isso é feito em todos nós até que não haja mais possibilidade de ganho. Uma vez que esse processo é realizado, as comunidades são agregadas, formando uma nova rede. Cada comunidade construída é considerada um novo nó e todo processo é repetido até que não haja mais ganho.

Figura 5.3 - Exemplo para entendimento de como é a qualidade das partições (modularidade) encontradas por algoritmos.



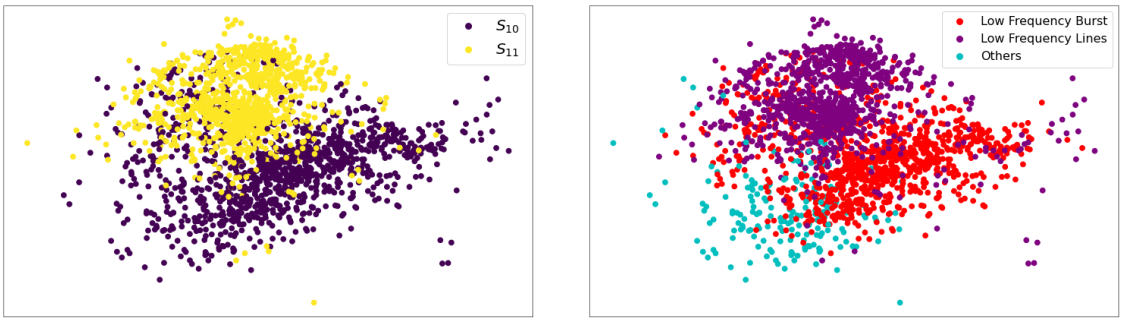
Fonte: Albert-László Barabási (2021).

A parte inferior da Figura 5.2 mostra as classes nos dados dos glitches encontradas pelo Best Partition. O algoritmo particionou os dados em apenas 4 classes das nove. Elas estão distribuídas em: Classe 0, Classe 1, Classe 2 e Classe 3. A Classe 0 tem um total de 923 glitches e 918 deles são Power Line. Dessa forma, pode-se dizer que 91,8% dos glitches Power Line (classificados pelo GS) estão na Classe 0. Ao comparar a imagem inferior com a da parte superior, é possível dizer (visualmente) que:

- A Classe 0 (em roxo) inclui somente Power Line;
- A Classe 1 (em azul) inclui Low Frequency Burst e Low Frequency Lines;
- A Classe 2 (em verde) inclui Tomte e Scattered Light;
- A Classe 3 (em amarelo) inclui Whistle, Blip e Koifish;
- Extremely Loud pode ser incluído na Classe 2 e ou na Classe 3.

Curiosamente, três partições encontradas pelo Best Partition têm mais classes envolvidas e, por isso, elas foram estudadas pelo mesmo algoritmo isoladamente. Por exemplo, a Figura 5.4 mostra os subgrupos encontrados pelo Best Partition quando somente os dados da Classe 1 foram analisados. O lado direito da figura mostra como o mesmo conjunto de dados foi classificado pelo Gravity Spy. Os pontos foram posicionados nas mesmas coordenadas da Figura 5.2 e, visualmente, há concordâncias entre as subclasses (S_{10} e S_{11}) e as classificações do GS (LFB e LFL). A cor ciano indica glitches de outras classes que foram incluídos na Classe 1 pelo algoritmo.

Figura 5.4 - Subgrupos encontrados com Best Partition quando a Classe 1 foi analisada isoladamente (à esquerda) e as correspondentes classificações atribuídas pelo Gravity Spy (à direita).



Fonte: Produção da autora.

Com esta nova análise, 942 de todos os 1000 Low Frequency Burst foram encontrados em S_{10}, e também 93,1% de Low Frequency Lines estavam em S_{11}. Este resultado está resumido na Tabela 5.1. Pode-se dizer que os subgrupos S_{10} e S_{11} têm grande equivalência com Low Frequency Burst e Low Frequency Lines, respectivamente

Tabela 5.1 - Concordância entre as subclasses encontradas na Classe 1 e as classificações do GS. A tabela deve ser lida da seguinte forma: $94,2\%$ dos Low Frequency Lines da lista do Gravity Spy são encontrados em S_{11}, $93,1\%$ dos Low Frequency Burst estão em S_{10}; esta última subclasse também contém $2,61\%$ de outras classes de glitches.

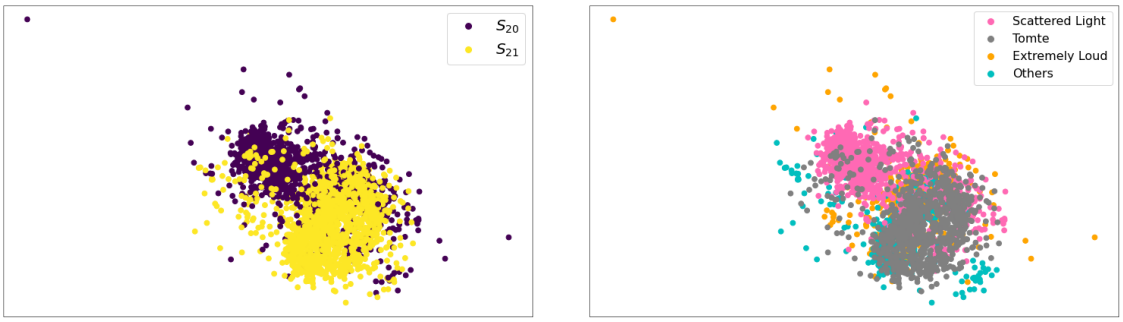
Classe	*Low Frequency Burst*	*Low Frequency Lines*	Outra
S_{10}	93,10%	5,70%	2,61%
S_{11}	6,40%	94,20%	0,13%

Note que cada subclasse será referenciada como S_{ijk}, onde i representa a primeira camada da aplicação do Best Partition, ou seja, a classe geral encontrada usando todos os dados, j a subclasse quando somente a classe i é analisada e k quando a subclasse j é examinada. As camadas de aplicação j e k estarão presentes apenas quando necessário. Nesse primeiro caso, por exemplo, como só os dados da Classe 1 estão sendo analisados mais uma vez, os subgrupos encontrados terão nomes do tipo S_{1j}. Não haverá o terceiro índice, pois nenhuma outra aplicação será feita.

A mesma análise foi aplicada às Classes 2 e 3, e os resultados estão nas Tabelas 5.2 e 5.3, respectivamente. Quando o conjunto de dados da Classe 2 foi a entrada para partição, dois subgrupos foram detectados: S_{20} e S_{21}. Quase noventa e nove por cento dos Scattered Light foram encontrados em S_{20} e 95,1% de Tomte em S_{21}. Embora os pontos estejam agrupados, é possível ver essas altas coincidências na Figura 5.5, que apresenta as partições definidas pelo Best Parition (à esquerda) e as classificações dos mesmos glitches pelo GS (à direita).

As aplicações poderiam ser finalizadas aqui, no entanto mais de cinquenta por cento dos Extremely Loud também estão em S_{21}; por isso, o localizador de comunidades foi aplicado novamente a S_{21}. Apesar de esperar por duas classes (uma para Tomte e outra para Extremely Loud), o conjunto foi subdividido em mais três. Os resultados estão abaixo da linha horizontal na Tabela 5.2. Esta nova aplicação reduziu a equivalência entre a classe S_{21} e Tomte. Agora, a classe encontrada pelo método com maior correspondência com Tomte é a S_{210}, com apenas 68,10% de concordância. O Extremely Loud teve uma correspondência com GS muito baixa; apenas 35,40% de concordância com S_{211}. A terceira classe (S_{212}) teve correspondências com Tomte, com Extremely Loud e com outras classes.

Figura 5.5 - Subgrupos encontrados com Best Partition quando a Classe 2 foi analisada isoladamente (à esquerda) e as correspondestes classificações atribuídas pelo Gravity Spy (à direita).



Fonte: Produção da autora.

Tabela 5.2 - Resultados da busca de subgrupos na análise da Classe 2.

Classe	*Scattered Light*	*Tomte*	*Extremely Loud*	Outra
S_{20}	98,90%	0,80%	13,90%	0,58%
S_{21}	0,00%	95,10%	51,50%	5,62%
S_{210}	0,00%	68,10%	0,00%	0,10%
S_{211}	0,00%	9,40%	35,40%	0,75%
S_{212}	0,00%	17,60%	16,10%	4,77%

De forma análoga, o Best Partition foi aplicado apenas para a Classe 3; nessa classe, o algoritmo particionou os dados em três novas partições: S_{30}, S_{31} e S_{32}. Os resultados podem ser vistos na Tabela 5.3. A subclasse S_{30} foi associada à classe Whistle com 99,4% de concordância, que, inclusive apresentou a maior equivalência entre um grupo do método e uma classe do Gravity Spy. Isso também pode ser observado na comparação das imagens da Figura 5.2. Por outro lado, a subclasse S_{31} não apresentou nenhuma conclusão interessante. Ela abordou 18,9% dos glitches Blip, 2,8% da classe Koi Fish e 0,5% dos glitches Extremely Loud.

Similarmente ao caso anterior, a subclasse S_{32} carrega mais de sessenta por cento de todos os Blip e Koi Fish, o que também induziu a reaplicar o localizador de comunidade. Após esta última aplicação em S_{32}, apenas 430 glitches da classe Blip foram encontrados em S_{321} e 51,4% dos glitches classificados como Koi Fish em S_{322}. Este último também incluiu mais 14% dos Extremely Loud. Houve uma outra subclasse (S_{320}) que teve um valor considerável de Blip (cerca de 22%). Os resultados desta última aplicação podem ser vistos depois da linha horizontal da Tabela 5.3. A representação gráfica está na Figura 5.6.

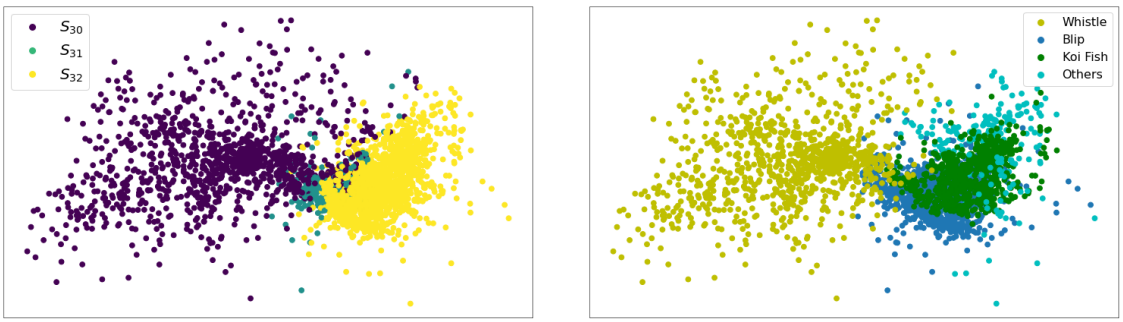
Tabela 5.3 - Resultados da busca de subgrupos na análise da Classe 3.

Classe	*Blip*	*Koi Fish*	*Whistle*	*Extremely Loud*	Outra
S_{30}	12,50%	2,60%	99,40%	3,00%	0,00%
S_{31}	18,90%	2,80%	0,00%	0,50%	0,00%
S_{32}	67,6%	65,60%	0,10%	14,50%	0,84%
S_{320}	21,90%	0,30%	0,00%	0,00%	0,26%
S_{321}	43,00%	13,90%	0,00%	0,20%	0,00%
S_{322}	2,70%	51,40%	0,10%	14,30%	0,58%

Um resumo das maiores equivalências entre uma classe encontrada pelo método e cada uma das nove classes selecionadas para estudo é mostrado na Tabela 5.4. Existem cinco equivalências com mais de 91% de concordância para Power Line, Low Frequency Lines,

Low Frequency Burst, Scattered Light e Whistle. Por outro lado, não há classes vinculadas a Blip e Extremely Loud com concordâncias altas. Elas atingiram 43% e 35,4%, respectivamente. Koi Fish e Tomte obtiveram 51,4% e 68,1%.

Figura 5.6 - Subgrupos encontrados com Best Partition quando a Classe 3 foi analisada isoladamente (à esquerda) e as correspondestes classificações atribuídas pelo Gravity Spy (à direita).



Fonte: Produção da autora.

É relevante lembrar que Koi Fish e Tomte são estudados como possíveis subclasses do Blip, o que pode ter colaborado para a mistura dos dados. A pior conexão, como comentado, foi entre uma classe do método e o Extremely Loud; a maior correspondência foi com o grupo S_{211}, com pontuação de 35,4%. Em geral, depois de estudar as três classes separadamente, há uma média calculada de 75,03% de coincidências entre o método com o Gravity Spy.

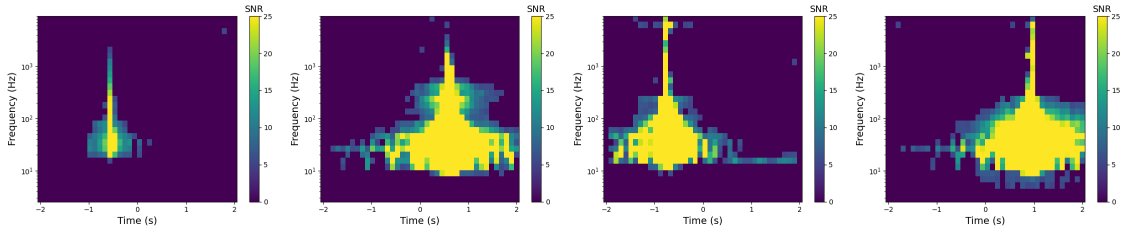
Tabela 5.4 - Resumo das classes encontradas pelo Best Partition com maiores equivalências com as classificações do Gravity Spy para as nove classes selecionadas. Por exemplo, 91,80% do glitches classificados com Power Line estão presentes na classe S_0.

Classe do Best Partition	Classe do Gravity Spy	Concordância
S_0	*Power Line*	91,80%
S_{10}	*Low Frequency Burst*	93,10%
S_{11}	*Low Frequency Lines*	94,20%
S_{20}	*Scattered Light*	98,90%
S_{210}	*Tomte*	68,10%
S_{211}	*Extremely Loud*	35,40%
S_{30}	*Whistle*	99,40%
S_{321}	*Blip*	43,00%
S_{322}	*Koi Fish*	51,40%

Com esse resultado final, surge uma questão intrigante: qual o motivo dos erros de equivalência entre as classificações do método de cosseno de similaridade e as do Gravity Spy?

Alguns transientes classificados como Extremely Loud pelo Gravity Spy, mas não encontrados em S_{211} pelo Best Partition, foram analisados. Os glitchgramas de quatro deles podem ser vistos na Figura 5.7. De fato, os glitches dos quatro têm uma morfologia semelhante a do Extremely Loud (veja seu glitchgrama médio na Figura 4.10). No entanto, todos eles estão deslocados do tempo central. Quando o cosseno entre um desses glitches é calculado com o vetor médio, muitos resultados do produto escalar serão nulos. Essa defasagem foi encontrada em vários glitches dessa classe, o que com certeza, colaborou para os erros. Além disso, a classe Koi Fish lembra muito a morfologia desse transiente, outro motivo que deve ser considerado.

Figura 5.7 - Glitchgramas referentes a glitches classificados como Extremely Loud pelo Gravity Spy, mas que estavam presentes na classe de maior equivalência com Scattered Light pelo Best Partition.



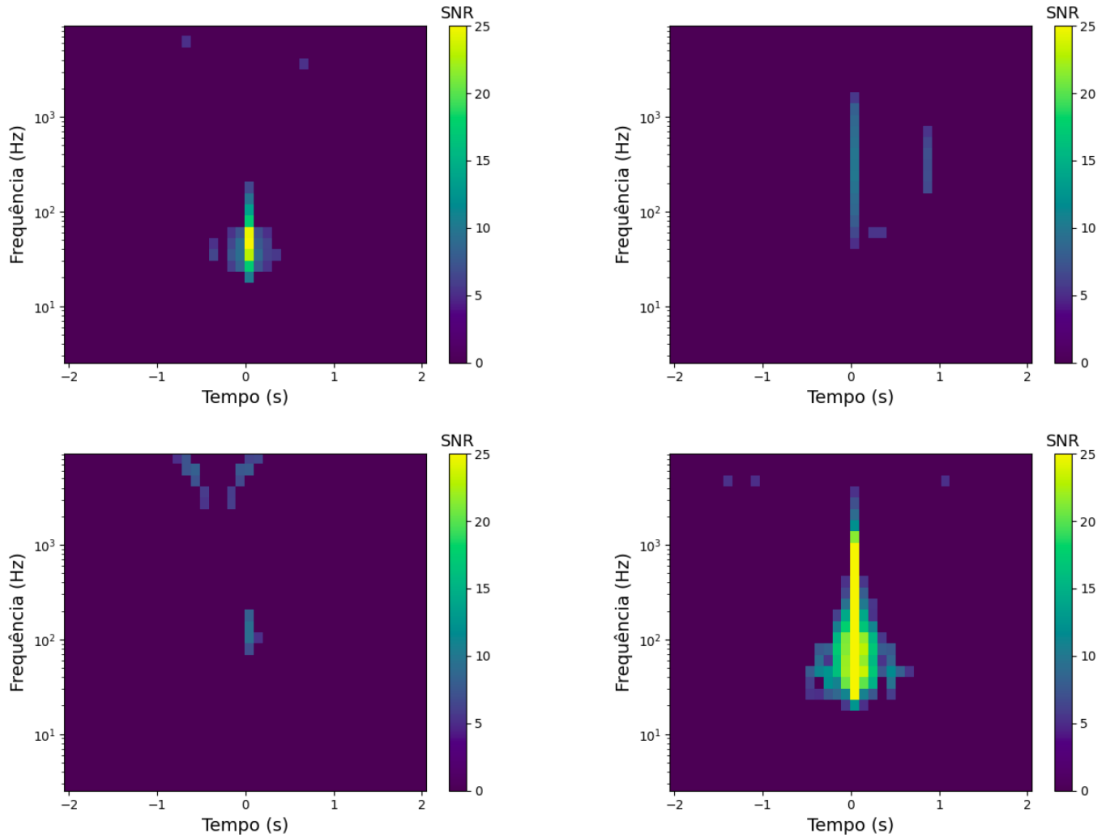
Fonte: Autoria própria.

O método do cosseno de similaridade e o glitchgrama carregam, claramente, seus próprios erros; não são métodos ideais e o resultado final não foi tão bom quanto se esperava (ele obteve apenas 75% de concordância com GS). No entanto, as classificações do Gravity Spy também podem estar equivocadas. Um exemplo é apresentado na Figura 5.8. Foram selecionados quatro glitches classificados como Blip pelo Gravity Spy que não estavam presentes na classe principal atribuída ao Blip pelo cosseno de similaridade (S_{321}). Seus glitchgramas foram comparados com os glitchgramas médios apresentados no Capítulo 4.

O primeiro glitchgrama da Figura 5.8, por exemplo, é muito mais parecido com o Tomte do que o Blip (como sugerido pelo GS); no glitch ao seu lado, à direita, há uma linha vertical lembrando o Blip, mas que não aparece sozinha, provavelmente, tratando-se do Repeating Blips. Na parte inferior da mesma imagem, à esquerda, há um outro glitchgrama que aparenta um Blip, mas um Blip que deve ter acontecido muito próximo a um Whistle. Seu formato em 'V' é bem evidente nas frequências altas. Por fim, à direita, na parte inferior da figura, há um outro glitchgrama que não é nem um pouco parecido com um Blip. Ele

aparenta um Koi Fish ou até mesmo um Extremely Loud. Todos esses exemplos mostram que o GS precisa de mais dados para melhorias na classificação e também justificam alguns dos erros de concordância com o método.

Figura 5.8 - Glitchgramas referentes a glitches classificados como Blip pelo Gravity Spy, mas que claramente não são.



Fonte: Produção da autora.

Discussões finais:

O estudo de glitches por análise de redes foi uma tentativa inovadora. Em geral, o método desse capítulo foi excelente para específicos grupos, mas ruins para outros. Comparando com o Gravity Spy, as classes encontradas pelo método coincidiram em 75%. Quando o algoritmo rodou nos dados, a princípio, encontrou apenas quatro classes. Depois, outras classes foram encontradas quando as principais foram analisadas isoladamente. Só foi possível relacionar um grupo do Best Partition para cada uma das nove escolhidas porque foram encontradas (visualmente) conexões com as classes do GS (Figura 5.2). Dessa forma, sem as classificações do Gravity Spy, não seria possível esse processo, pelo menos, não com

esse método aplicado aos glitchgramas. Se o GS não existisse, o Best Partition encontraria apenas as quatro classes iniciais. Não haveria possibilidade de reaplicar, pois não há uma condição limitante para cessar as aplicações. Isso, com certeza, pode ser um estudo futuro.

Por outro lado, o método foi capaz de encontrar erros de classificações do Gravity Spy. Também, mostrou-se sensível a glitches com deslocamentos em relação ao tempo central (que, nesse caso, foi escolhido como zero). Uma outra vantagem desse método é a busca da classe de um glitch aleatório. Se um vetor representativo pode ser feito a partir dos glitchgramas médios, haverá nove vetores característicos de cada classe. Agora, suponha um glitch de classe desconhecida; seu vetor representativo também será obtido pelo glitchgrama. O cosseno deste pode ser calculado com todos os outros nove vetores. O valor que for mais próximo de um, indicará a qual classe esse glitch pertencerá. Foram realizados vinte testes com esse tipo de classificação, nenhum cosseno teve valor igual a um, mas o valor mais alto indicou a classe correta para todos eles. Essa busca de classes é feita apenas multiplicando vetores e, portanto, quase não tem custo computacional. Vale lembrar, que isso não funciona para os glitches deslocados do centro, como no caso dos Extremely Loud apresentados na Figura 5.7. Para essa classe, a checagem morfológica é necessária antes da aplicação para não testar glitches deslocados.

Para saber se outros erros de classificação são provenientes do método ou dos glitchgramas, outra ferramenta computacional será aplicada. O capítulo seguinte mostrará os princípios básicos e os resultados obtidos para classificações de glitches (a partir do glitchgrama) utilizando técnicas de Aprendizado de Máquina.

6 MACHINE LEARNING E O ESTUDO DOS GLITCHES

Com o avanço da tecnologia e do acesso ao mundo digital, informações se propagam rapidamente e conectam o mundo a todo instante. Essas informações (ou dados) têm fluxo e volume altos; por isso, computadores para armazená-las são indispensáveis. Mas armazenar dados não é suficiente. É preciso também interpretá-los. E, novamente, os computadores não podem ser deixados de lado. Na busca de fazer com que as máquinas interprete dados como ser humano, técnicas de inteligência artificial, IA, têm sido constantemente desenvolvidas e aplicadas.

A inteligência artificial atua desde a identificação de imagens e ensinar um carro a dirigir sozinho, a conversar com você e programar agendas no seu celular. Ela também já é utilizada na verificação de *spam*, na medicina, na astronomia, nas empresas, na liberação de créditos, na previsão de falhas em equipamentos, e em diversas outras áreas. Há tantas aplicações de IA próximas que, às vezes, algumas são até ignoradas. Você, por exemplo, já colocou foto em alguma rede social e, automaticamente, houve o reconhecimento de rostos de pessoas que estavam nela? Isso é bem comum e é um exemplo de aplicação da IA. Tal reconhecimento acontece porque você certamente já colocou a foto desse rosto antes e a ligou com a pessoa correspondente. O algoritmo grava essas informações e busca por padrões entre as imagens novas e as anteriores para reconhecer pessoas. Esses padrões envolvem formato do rosto, modelo do nariz, tamanho do cabelo, cor dos olhos e outros parâmetros característicos.

O ato da máquina aprender a interpretar dados através de algoritmos é usualmente chamado de Aprendizado de Máquina, AM, do inglês *Machine Learning*, ML, que está dentro da IA. De acordo com Arthur Samuel, um dos pioneiros da IA, o AM pode ser definido como o campo de estudo que oferece aos computadores a habilidade de aprender sem ser explicitamente programado (SAMUEL, 1959). Aplicações nessa área incentivaram o uso no estudo de glitches. Afinal, a máquina reconhecer um sinal gravitacional no meio de tantos sinais ruidosos seria fantástico.

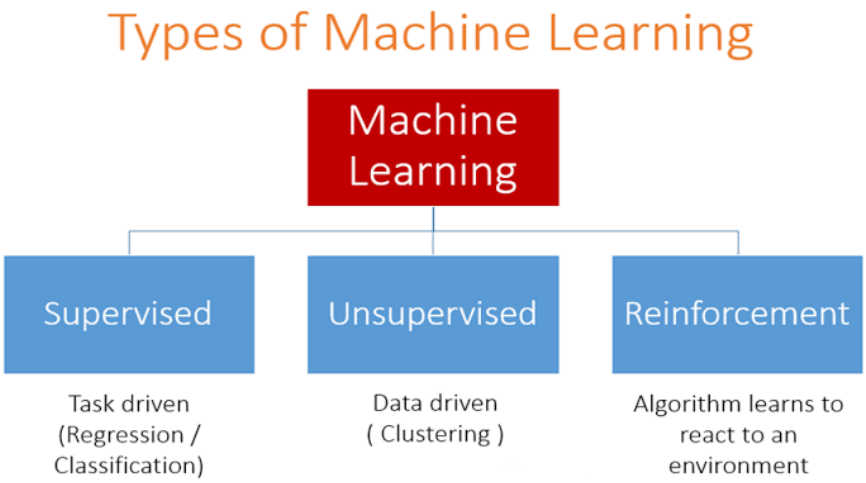
Usualmente, o AM é dividido em três partes principais: aprendizado supervisionado, aprendizado não-supervisionado e aprendizado de reforço (vide esquema na Figura 6.1). O AM supervisionado é baseado em dados já classificados (ou rotulados). Por exemplo, um conjunto de imagens pode ser inserido no computador com classes de frutas como laranja, banana e maçã. Essas imagens são utilizadas como base para treinamento do computador para que ele se torne capaz de classificar uma imagem aleatória. Dessa forma, ele cria um modelo a partir das características encontradas e busca semelhanças entre a nova imagem e a base que ele já conhece. Se o algoritmo, por exemplo, verificar a forma geométrica da imagem nova e ela for circular, a classe banana será facilmente descartada; se o segundo passo for verificar a cor e ela não for vermelha, a maçã também será uma opção eliminada

e, dessa forma, a predição poderá ser feita: a imagem desconhecida é uma laranja.

Por outro lado, o AM não-supervisionado não precisa de dados previamente classificados. Ele busca padrões em dados aparentemente aleatórios e agrupa por similaridade entre eles. Além disso, essa técnica também é utilizada para reduzir dimensões dos dados de interesse.

Por fim, existe o AM por reforço que não será utilizado neste trabalho. Este, basicamente, progride através do que chamam de punição e progresso. Cada vez que o algoritmo acerta uma previsão, ele é "recompensado" e toda vez que ele erra, ele é "punido". Para esse retorno, é preciso que uma pessoa avalie as respostas. Sempre que o algoritmo é punido, ele refaz a ação de forma a buscar um outro resultado que seja compensado. Essa técnica é frequentemente utilizada em jogos e aplicações com robôs; ela tem como objetivo aprender sozinha qual a melhor estratégia para obter o maior número de recompensas possível (GÉRON, 2019).

Figura 6.1 - Os principais tipos de Aprendizado de Máquina.



Fonte: Medium (2019).

6.1 AM não-supervisionado

Muitas vezes não é possível rotular os dados, principalmente se for uma grande quantidade ou algo custoso em tempo ou em dinheiro; outras vezes, como aqui, a classe pode ser omitida apenas para efeitos de verificação. Em todos esse casos, o AM não-supervisionado torna-se um método essencial. O princípio de funcionamento dele, como mencionado, é buscar padrões em dados misturados e sem classificações prévias. Ele busca por seme-

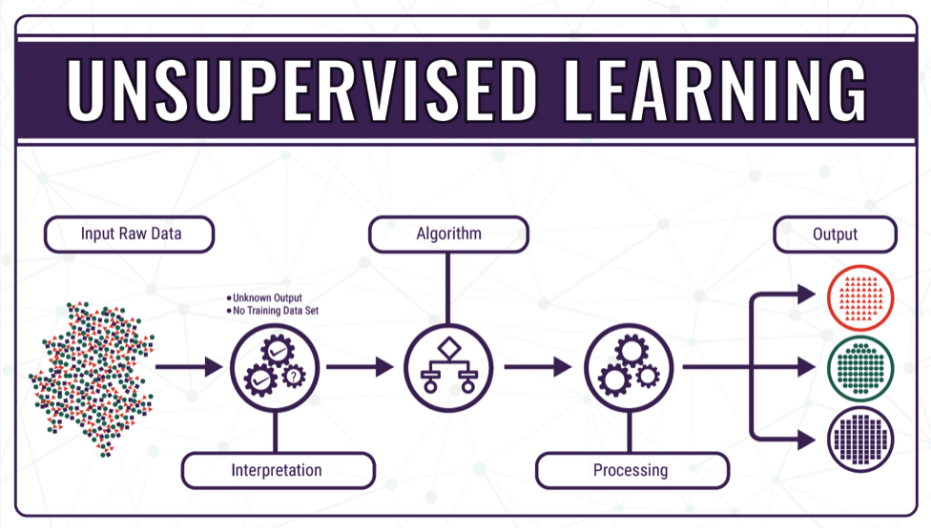
lhanças em dados que, a princípio, pareçam aleatórios, e podem servir como base para interpretações e obtenção de informações. Suas principais aplicações são: agrupamento de dados e redução de dimensões.

Agrupamento de dados

Essa técnica usualmente recebe um conjunto de dados "misturados" e busca padrões; como resposta, oferece grupos de dados semelhantes entre si. A Figura 6.2 esquematiza esse processo. Primeiramente, os dados sem classificação são inseridos no algoritmo que analisa os parâmetros, processa e separa em grupos similares. A resposta do algoritmo (*output*) na imagem está colorida apenas para indicar que, da mistura inicial, podem-se encontrar três grupos; mas vale ressaltar que nenhum dado sai classificado do algoritmo, isto é, com uma classe atribuída.

Em geral, a ideia aqui é muito parecida com a da aplicação do NetworkX do capítulo anterior, que agrupa dados semelhantes entre si e separa de outros distintos. Ao inserir os dados de glitches nesta técnica, as classes (Blip, Koi Fish, Power Line e etc.) são indiferentes para o algoritmo. Ele vai buscar os grupos através dos padrões, independentemente do conhecimento prévio. Isso é excelente para verificação do glitchgrama. Se ele é de fato um bom caracterizador do glitch, a presença das nove classes deverá ser evidente na saída do código. Em outras palavras, se a Figura 6.2 fosse referente aos dados dos glitches, nove grupos seriam apresentados no output.

Figura 6.2 - Esquema sobre o que é aprendizado de máquina não-supervisionado.



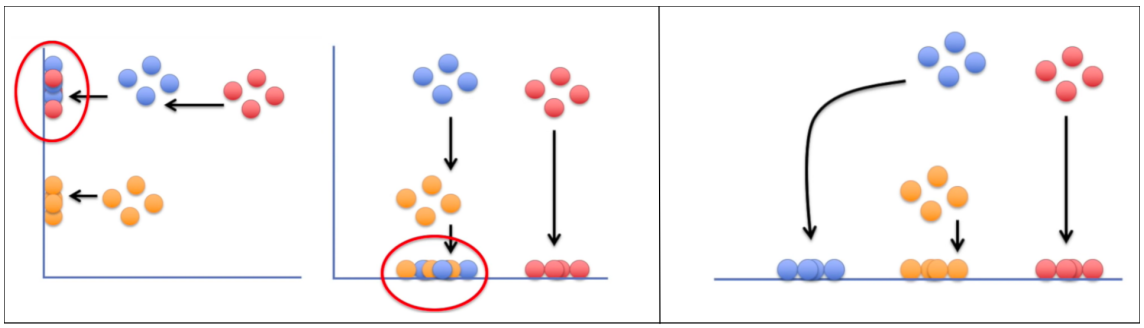
Fonte: NVIDIA (2018).

Redução das dimensões

Um computador pode buscar por padrões nos dados através de um algoritmo (uma sequência de instruções) que oferece uma resposta na saída do programa. Para isso, é preciso analisar dados. Esses dados podem conter muitos parâmetros (também chamados atributos) que os caracterizam. Cada parâmetro independente resulta em uma dimensão para o dado e visualizar conjuntos de dados multidimensionais não é uma tarefa fácil. No próprio caso do glitch, os dados são vetores criados a partir dos glitchgramas e, cada um, tem 1200 dimensões. Para resolver esse problema, o AM não-supervisionado também analisa os dados e reduz as dimensões para duas ou três, possibilitando a visualização em gráficos de dispersão.

Os dados poderiam simplesmente ser projetados em dimensões menores, mas isso não seria interessante, já que o objetivo é encontrar os grupos formados. Isso pode ser visto na Figura 6.3. Há uma simulação de como seria feita a projeção de 2D (duas dimensões) para 1D. Na dimensão maior, os dados são bem organizados em três grupos (vermelho, azul e laranja). Quando eles são projetados no eixo vertical (imagem da esquerda), o grupo laranja fica muito bem localizado, mas há uma grande mistura nos dados vermelhos e azuis. Lembrando que, por efeito de visualização, há cores, mas num caso real, pode não se conhecer nada sobre os dados e tal mistura resultaria em perda de informações ou conclusões falsas. O mesmo aconteceria no caso da projeção no eixo horizontal, onde há a mistura entre dados da classe laranja e azul, evidenciada pelo círculo vermelho. O objetivo, portanto, é encontrar uma técnica que faça isso e não destrua informações relevantes (como a quantidade de grupos).

Figura 6.3 - Exemplos de como seriam as projeções de dados em 2D para 1D nos eixos vertical e horizontal, respectivamente. Ambos têm perda de informações com a mistura de dados. À direita, há um exemplo de como tal redução seria feita pelo t-SNE, que preserva a existência de todos os grupos.



Fonte: Adaptado de StatQuest (2017).

Em Python, linguagem em que o programa deste trabalho está sendo desenvolvido, o algoritmo que está sendo utilizado para isso é o chamado t-SNE (t-Distributed Stochastic Neighbor Embedding). Trata-se de uma técnica de AM não-supervisionado e é um algoritmo peculiar, pois ao mesmo tempo que diminui as dimensões dos dados, ele agrupa por similaridade. O lado direito da Figura 6.3 mostra como seria feita a projeção dos dados utilizando o t-SNE. Note que os grupos estão bem separados em 1D, mas as distâncias não são preservadas (o que não importa para esta aplicação). A seguir haverá uma descrição dos principais conceitos do t-SNE baseada no artigo Maaten e Hinton (2008).

6.1.1 O t-SNE

O t-SNE pode ser entendido como uma melhoria no algoritmo SNE (Stochastic Neighbor Embedding) (HINTON; ROWEIS, 2002) que descreve distâncias Euclidianas multidimensionais em matrizes de probabilidades de similaridade. Dessa forma, a similaridade entre dois pontos x_i e x_j é dada pela probabilidade condicional $p_{j|i}$ (Equação 6.1) que compara a distância de x_j e outros pontos vizinhos na gaussiana centrada em x_i; σ é a variância dessa gaussiana.

$$p_{j|i} = \frac{\exp(-||\mathbf{x}_i - \mathbf{x}_j||^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-||\mathbf{x}_i - \mathbf{x}_k||^2/2\sigma_i^2)} \tag{6.1}$$

Se $p_{j|i}$ é relativamente alta, x_j é próximo de x_i; caso contrário, eles estão distantes. No caso de um *outlier*[1], os valores de $p_{j|i}$ serão extremamente pequenos para todos os j.

Também é possível calcular a probabilidade de similaridade entre os mesmos dois pontos na dimensão menor através da Equação 6.2. A ideia básica do SNE é obter essas probabilidades condicionais na maior dimensionalidade ($p_{i|j}$), na menor dimensionalidade ($q_{i|j}$) e fazer $q_{i|j}$ o mais próximo possível de $p_{i|j}$. Em um caso ideal, $p_{j|i}$ e $q_{j|i}$ seriam iguais, ou seja, a probabilidade de similaridade entre dois pontos em um espaço multidimensional deve ser a mesma em um espaço com menor dimensão,

$$q_{j|i} = \frac{\exp(-||\mathbf{y}_i - \mathbf{y}_j||)^2}{\sum_{k\neq i}\exp(-||\mathbf{y}_i - \mathbf{y}_k||^2)}. \tag{6.2}$$

Quando o t-SNE impõe que a similaridade entre dois pontos de dados seja equivalente tanto em alta quanto em baixa dimensionalidade, há, consequentemente, um agrupamento de dados semelhantes em baixa dimensão, tornando-se possível visualizar os grupos formados. Para isso, o método minimiza as divergências de Kullback-Leibler entre $p_{i|j}$ e $q_{i|j}$ usando um algoritmo para encontrar o mínimo local da função de custo (Equação 6.3). A divergência

[1] *outlier* é o nome atribuído a um ponto que se difere significantemente da maioria dos outros.

de Kullback-Leibler mede o quanto uma função de probabilidade difere-se de outra. Quanto menor seu valor, mais parecidas elas são.

$$C = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} \tag{6.3}$$

Para minimizar a função custo, o SNE usa o método do gradiente que busca pelo mínimo da função a cada iteração. O gradiente calculado pode ser visualizado pela Equação 6.4. O próprio artigo deste método diz que, fisicamente, essa equação pode ser interpretada como uma força resultante (na baixa dimensionalidade) criada por um conjunto de molas que conecta o ponto y_i e todos os outros pontos y_j; essa força vai repelir ou atrair se as distâncias entre os pontos forem pequenas ou grandes. O termo $(y_i - y_j)$ é interpretado como a direção da força, e $(p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})$ retrata a rigidez entre os pares de pontos obtidos pela diferença de probabilidades de similaridade em alta e em baixa dimensões.

$$\frac{\delta C}{\delta y_i} = 2 \sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j) \tag{6.4}$$

Uma diferença entre t-SNE e SNE é que t-SNE usa uma função de custo simetrizada, isto é, $p_{j|i}$ é trocado por $p_{ij} = (p_{j|i} + p_{i|j})/2n$, onde n é o número de dimensões. Além disso, ao invés de uma gaussiana centrada em x_i, o t-SNE usa uma distribuição t (de *t-student*) para obter a similaridade na baixa dimensão.

Em resumo, se o glitchgrama é um bom caracterizador, as mil e duzentas dimensões serão suficientes para caracterizar um glitch. O t-SNE vai calcular a p_{ij} entre dois glitches nessa alta dimensionalidade e criar uma matriz de similaridade entre todos os transientes. Ela segue o mesmo conceito da matriz de adjacência do capítulo anterior e terá 9000×9000 elementos. Feito isso, o algoritmo vai colocar os dados de forma aleatória na dimensão de interesse (neste caso, dois) e calcular uma nova matriz de similaridade através de $q_{j|i}$. A partir daí, ele vai alterar as posições dos pontos em 2D para fazer $q_{j|i}$ o mais próximo possível de $p_{j|i}$ e assim, os grupos formados na alta dimensão também emergirão em duas dimensões.

Neste trabalho, o t-SNE foi importado do pacote *scikit-learn*, no qual muitos algoritmos de Aprendizado de Máquina estão disponíveis (PEDREGOSA et al., 2011). Um parâmetro importante na aplicação da função do t-SNE no código é a perplexidade P, associada ao número de vizinhos efetivos. Ela pode ser definida matematicamente pela Equação 6.5 e é um valor inserido pelo programador no momento em que o algoritmo roda. Cada conjunto de dados tem seu melhor número de perplexidade e comumente é usado no intervalo de 5 a 50;

$$Perp(P_i) = 2^{H(P_i)}, \tag{6.5}$$

onde H_i é denominado entropia de Shannon (SHANNON, 1948),

$$H(P_i) = -\sum_j p_{j|i} \log_2 p_{j|i}. \tag{6.6}$$

A perplexidade também está relacionada com a variância σ da Equação 6.1, de forma que quanto maior for, mais vizinhos com similaridades não nulas entrarão em torno de x_i.

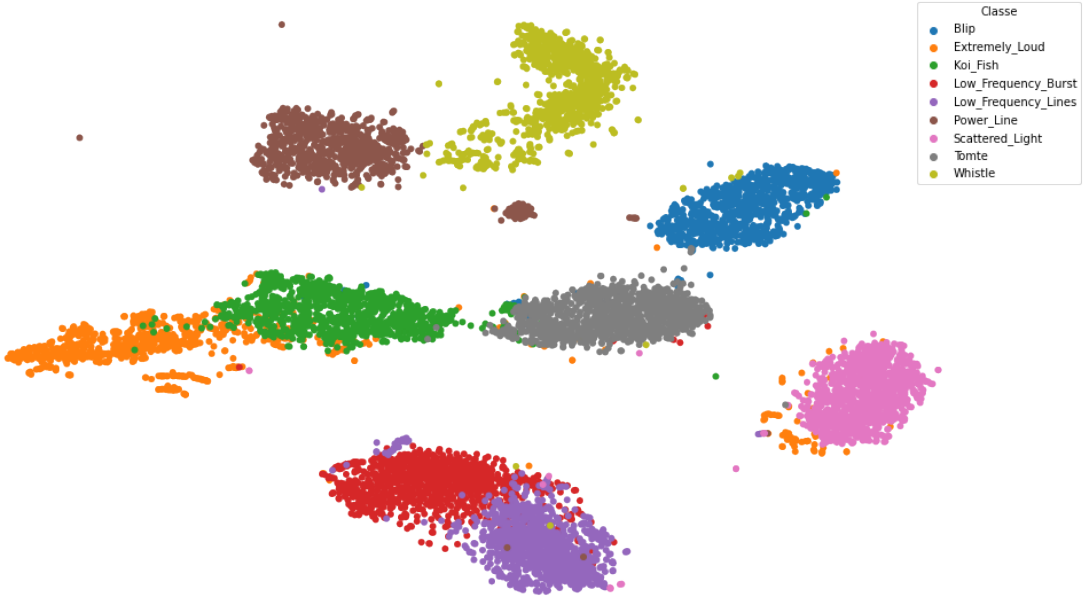
O resultado da aplicação do t-SNE para os dados dos glitches pode ser visualizado na Figura 6.4. Cada vetor tinha 1200 dimensões e o algoritmo reduziu para duas. A saída do código é, portanto, nove mil linhas com duas coordenadas cada. Essas duas coordenadas representam as posições x e y de cada vetor que foram usadas para montagem da Figura 6.4. Dessa forma, cada ponto representa um glitch e cada cor representa uma classe atribuída pelo Gravity Spy.

Novamente, as coordenadas horizontais e verticais não são de interesse deste trabalho e foram retiradas da imagem. Cada vez que o algoritmo roda, se a posição inicial não é imposta, os pontos estarão em lugares diferentes, prevalecendo apenas os grupos formados. Vale lembrar que o t-SNE não tem acesso às classes dos glitches e os pontos da Figura 6.4 foram coloridos (depois da aplicação do t-SNE) para verificar a eficiência do método com o uso dos glitchgramas.

A presença dos grupos na Figura 6.4 mostra que o método aplicado aos glitchgramas funciona; é visível a presença das nove classes. Há intersecções entre Low Frequency Burst (vermelho) e Low Frequency Lines (roxo), mas os principais aglomerados de cada classe são distinguíveis. Os grupos das classes Power Line (marrom), Whistle (verde claro), Tomte (cinza), Blip (azul) e Scattered Light (rosa) estão bem isolados e mostram poucas intersecções. Isso mostra que essas cinco classes são bem definidas pelos glitchgramas.

As classes Extremely Loud (laranja) e Koi Fish (verde escuro) estão próximas e com uma parte central bem conectada por uma mistura de ambas as classes; o grupo de Extremely Loud parece "invadir" uma parte da região de Koi Fish; aparentemente, é o grupo com mais intersecções e provavelmente terá maiores erros de classificações, assim como no método anterior. Como são glitches muito parecidos com Koi Fish, isso era esperado. Além disso, há também uma evidência de alguns glitches da classe Extremely Loud presentes no aglomerado principal do Scattered Light, alguns deles são os mesmos mostrados no capítulo anterior com deslocamentos no tempo central.

Figura 6.4 - Resposta do t-SNE (em 2D) para a análise dos nove mil glitchgramas de 1200 dimensões cada. O ponto representa um glitch e a cor, a classe. As cores foram aplicadas depois do algoritmo encontrar os grupos para conferir se eles foram bem determinados e alocados pelo método.

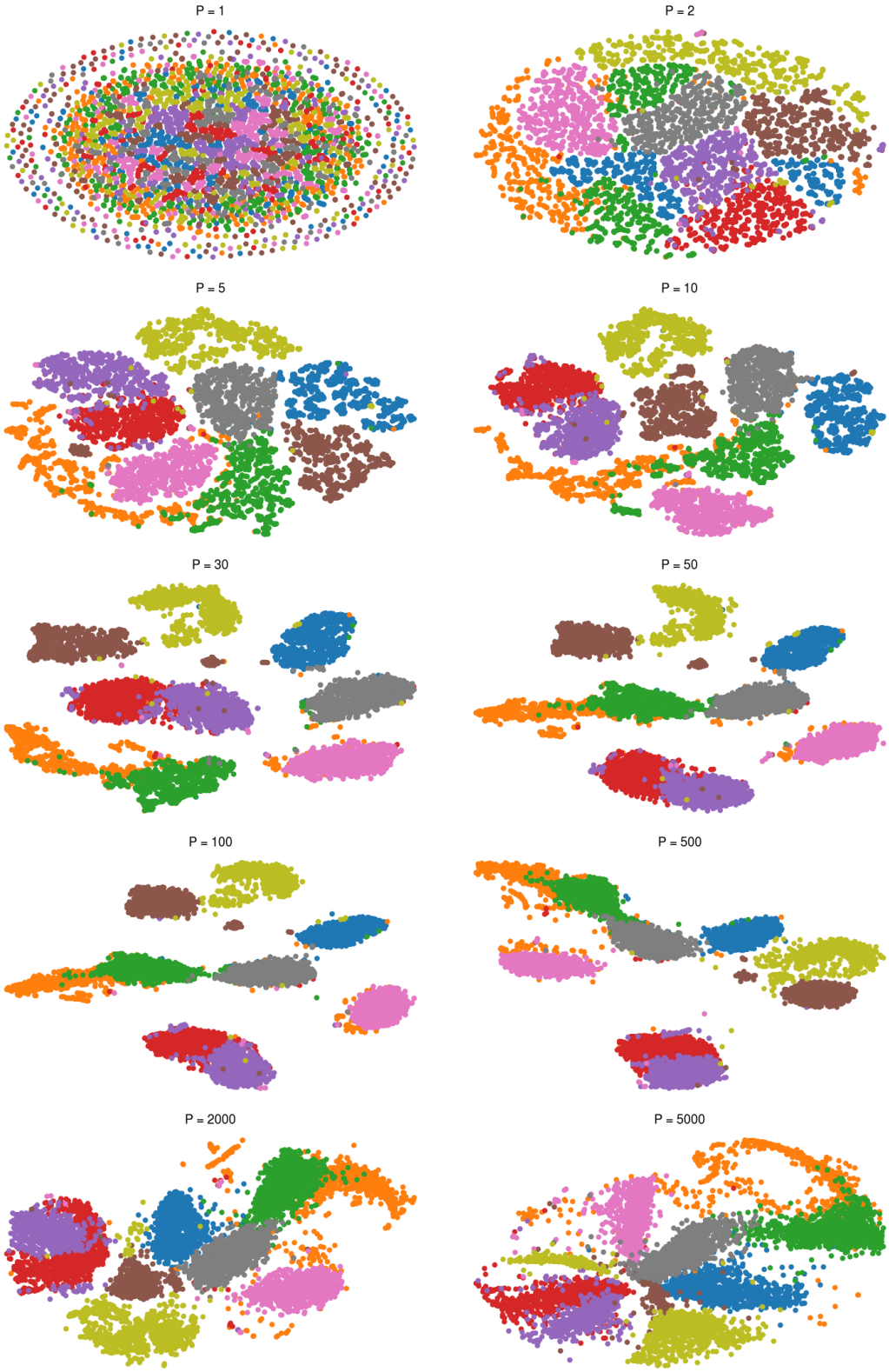


Fonte: Produção da autora.

O t-SNE aplicado aos glitchgramas funciona bem. Os grupos estão bem localizados na imagem e isso é importante para as classificações futuras, pois a classe de um glitch desconhecido vai depender da posição em que ele estiver na Figura 6.5. Se por exemplo, um transiente estiver perto dos pontos azuis, ele vai ter uma probabilidade maior de pertencer à classe Blip. Há dois outliers na parte superior, à esquerda, que foram classificados como Power Line pelo GS; foi verificado que são glitches Power Line, mas que aconteceram ao mesmo tempo que outro de classe diferente. Casos como esses são inevitáveis. Quanto menos outliers e menos intersecções melhor, mas isso só seria possível num caso ideal.

Para entender como a perplexidade interfere na prática, a Figura 6.5 mostra como seriam as respostas do t-SNE para perplexidades de valores 1, 2, 5, 10, 30, 50, 100, 500, 2000 e 5000. Na menor perplexidade, os pontos aparentam ser distribuídos de forma aleatória; para perplexidade igual a 2, já há indícios de formação de grupos de mesmas cores; para valor 5 (que é o valor mínimo sugerido no artigo da técnica), a presença dos grupos começa a ser mais evidente e isso vai melhorando conforme a perplexidade aumenta. A partir de 500, os grupos já ficam mais amontoados, dando a impressão que pertencem à mesma classe. Para facilitar essa visualização, pode-se imaginar todos os pontos em uma mesma cor e, se fossem, tanto para valores muito pequenos de perplexidade quanto para grandes, os grupos não seriam bem definidos.

Figura 6.5 - Resultados da aplicação do t-SNE para diferentes valores de perplexidade..



Fonte: Produção da autora.

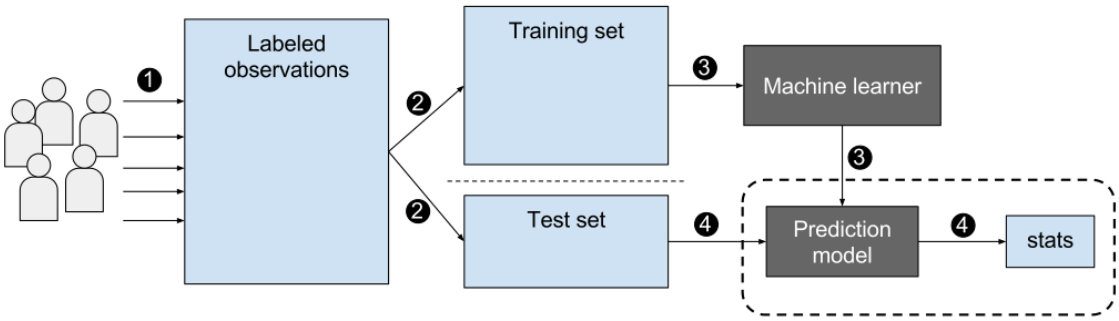
Todos os comentários e conclusões anteriores foram feitos por análise visual, mas para afirmações, é preciso apresentar um número de confiança para esses resultados. Por isso, um modelo preditivo de AM supervisionado será aplicado, usando como entrada a própria saída do t-SNE. Antes das análises, a próxima seção apresentará breves conceitos dessa área.

6.2 AM supervisionado

Chama-se *treinar* a máquina o ato de ensinar o computador com dados etiquetados, isto é, dados conhecidos e já classificados. Um dado contem atributos e aprender com eles é a base para o aprendizado de máquina. No aprendizado supervisionado, a máquina cria um modelo treinado a partir desses atributos e torna-se capaz de fazer predições de classificação para um dado aleatório; em outras palavras, a máquina é capaz de fazer uma inferência. Os atributos podem ser de dois tipos: quantitativo (idade, massa, altura, etc.) ou qualitativo (sexo, tamanho, grupo sanguíneo, classe de glitch, etc.) (FACELI et al., 2011). No caso do quantitativo, ainda uma subdivisão é feita em: contínuos (de infinitos valores) e discretos.

A Figura 6.6 apresenta um esquema sobre o que é o aprendizado de máquina supervisionado. Para aplicá-lo, o primeiro passo é obter um conjunto de dados já classificados. Esse conjunto é usualmente separado em duas partes: uma para treinamento da máquina e outra para teste. Apenas o conjunto de dados de treinamento é inserido para a máquina criar um modelo de predição. Depois que o modelo for criado, utilizam-se os dados de teste (como se fossem dados desconhecidos) para verificar a eficiência de predição e gerar estatísticas de acertos. É importante destacar que essa etapa é fundamental para uma boa aplicação do AM. Inserir os mesmos dados de treinamento para teste cria estatísticas falsas, já que há uma tendência maior de acerto.

Figura 6.6 - Esquema sobre o que é aprendizado de máquina supervisionado.



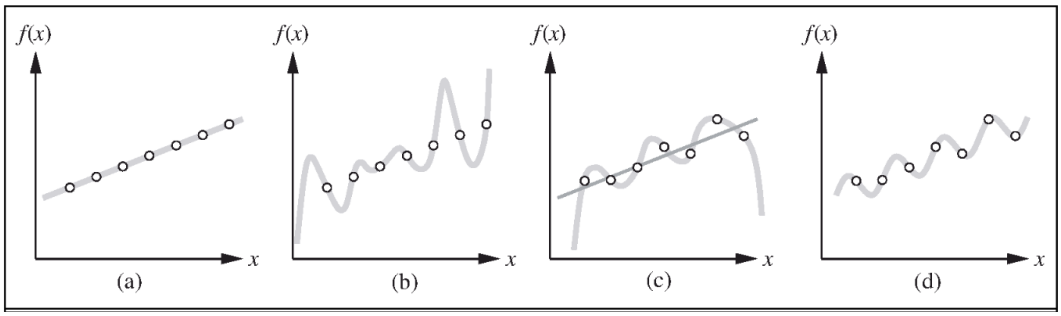
Fonte: NVIDIA (2018).

De acordo com Alpaydin (2009), tanto a classificação quanto a regressão são problemas de aprendizado de máquina supervisionado, onde há uma entrada X, uma saída Y e a tarefa é aprender a mapear a entrada até a saída. A diferença é que o modelo de regressão prevê valores contínuos, enquanto que um modelo de classificação prevê valores discretos (como definir se a imagem é uma laranja, uma maçã ou uma banana). Dessa forma, é possível dividir o AM supervisionado em dois principais casos: a busca de uma resposta para dados contínuos (regressão) e a busca de uma resposta para dados discretos (classificação).

Regressão

Se o objetivo é obter como resposta um valor numérico, o problema de aprendizagem é chamado de regressão. No aprendizado de máquina, a função numérica não é conhecida, mas existe um conjunto de dados para treinamento que torna a máquina capaz de prever valores (ALPAYDIN, 2009). Um exemplo de regressão está apresentado na Figura 6.7. Na parte a) da figura, há uma regressão linear que se ajusta perfeitamente ao conjunto de dados; tal ajuste usa como ferramenta o método de mínimos quadrados. Contudo, para esse mesmo conjunto, há uma outra possibilidade de ajuste polinomial de grau maior, que também se encaixa adequadamente aos dados e pode ser vista na parte b) da Figura 6.7.

Figura 6.7 - Exemplo de regressão.



Fonte: Russell e Norvig (2016).

Um outro conjunto de dados é mostrado na parte c). Há um ajuste polinomial e um ajuste linear que podem ser utilizados como modelo para a distribuição dos dados, além do outro polinômio na parte d). O próprio livro de Russell e Norvig (2016) discute essa imagem e faz uma questão relevante: dentre todas as possíveis hipóteses, como encontrar a melhor? A resposta mais coerente, quando não há condições específicas, é preferir a hipótese mais simples e consistente com os dados. Apesar do estudo de glitches não ser por regressão, essa ideia persiste (veja o exemplo a seguir).

Classificação

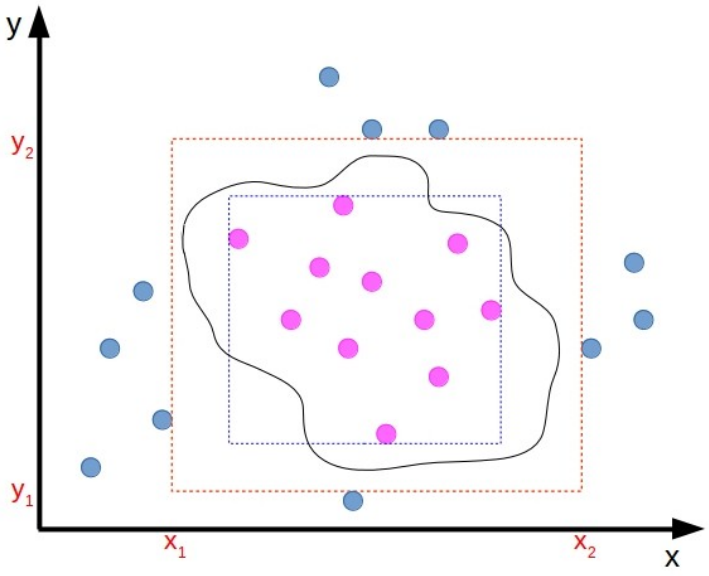
Quando não há interesse em obter um valor numérico, mas sim, uma classe, o problema de aprendizado é chamado de classificação; e de classificação Booleana ou binária se tem apenas duas respostas possíveis (RUSSELL; NORVIG, 2016). Por exemplo, uma pessoa pode ser classificada como doente ou saudável de acordo com seus atributos (sintomas), um animal como felino ou canino, um alimento como fruta ou legume, etc.

Conforme mencionado acima, para a aplicação do aprendizado de máquina supervisionado, é preciso dados já classificados para a máquina criar um modelo através deles. Quanto mais dados conhecidos, melhor. Por isso, o limite mínimo de mil contagens para cada classe de glitches foi imposto logo no início.

Na Figura 6.8 há o exemplo de um conjunto de vinte e dois dados já classificados entre azuis e rosas. Baseado nas classificações e na observância da imagem, um modelo que avalia um dado desconhecido com parâmetros x e y pode ser criado da seguinte forma:

- Se $y_1 \leq y \leq y_2$ e $x_1 \leq x \leq x_2$, então o dado de entrada é classificado como rosa. Caso contrário, azul. Esse modelo foi estabelecido usando o retângulo vermelho cujos lados tangenciam os círculos azuis.

Figura 6.8 - Distribuição de dados aleatórios com parâmetros x e y.



Fonte: Produção da autora.

Contudo, também é possível criar um modelo que tangencie os pontos rosas (retângulo azul) e ambos separam bem as duas classes. Note ainda que qualquer outro retângulo (entre esses dois) e qualquer outra função (como a desenhada em preto) podem ser usados como classificadores e, o que deve-se fazer é buscar por uma hipótese h do conjunto de hipóteses H que apresente menos erros. A escolha de um dos retângulos como modelo final ainda pode ser relevante já que trata-se de um modelo mais simples para explicar as distribuições dos padrões.

Esses exemplos de classificação e regressão foram apresentados para mostrar que dependendo do modelo escolhido, diferentes erros podem ser encontrados. Também, para confirmar que quanto mais dados, melhor o AM vai funcionar. Se neste último caso, mais dados fossem conhecidos, haveria menos opções, pois os modelos se encaixariam cada vez mais nos dados.

Erros podem vir de modelos, de dados e de classificações prévias; como visto no capítulo anterior, o Gravity Spy também tem erros e quando tais informações são usadas para treinar a máquina, classificações falsas vão surgir. O algoritmo escolhido aqui para classificar os glitches foi o SVM (Support Vector Machine). Ele é um dos métodos mais populares do AM e funciona tanto para classificação quanto para regressão (GÉRON, 2019); além disso, é robusto para outliers e pode ser usado para dados linearmente separados ou não.

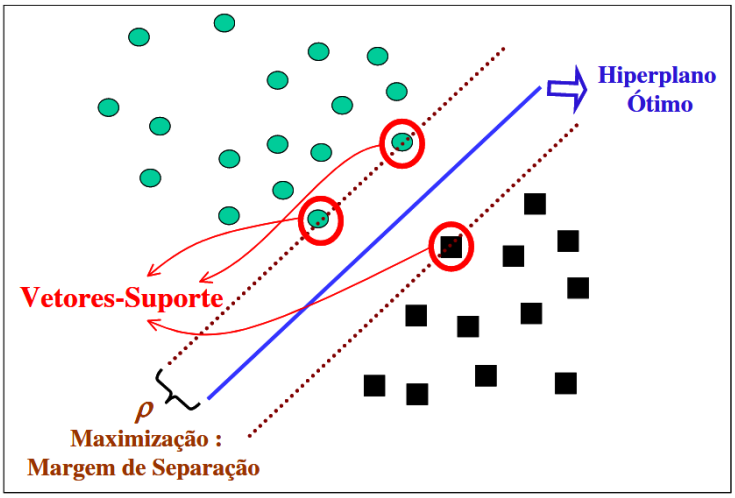
6.2.1 SVM

A ideia do SVM é criar hiperplanos entre dois vetores de suporte e encontrar um ótimo maximizando as distâncias entre o hiperplano e as linhas que passam pelos vetores de suporte, chamadas de margens (CORTES; VAPNIK, 1995). Para facilitar esse entendimento, a Figura 6.9 mostra como é esse processo. Novamente, há duas classes; como visto no exemplo anterior, diferentes modelos (ou classificadores) poderiam ser aplicados para criar funções capazes de separar uma classe da outra; tais linhas de separação são chamadas de limite ou fronteira de decisão e delimitam a região que cada classe é dominante.

Para fazer isso, o SVM seleciona dados da borda de cada classe vizinha (de forma que não haja outros dados entre eles) que são chamados de vetores de suporte e estão circulados em vermelho na Figura 6.9. Esses vetores de suporte são utilizados como referência para criar um hiperplano (no caso 2D, uma reta) que maximize a separação (ρ) entre ele próprio e cada vetor de suporte. Este é chamado de hiperplano ótimo e compõe o modelo a ser utilizado para classificação. Para esse exemplo de dados, o hiperplano ótimo está representado pela reta azul. Com tal modelo, pode-se dizer, portanto, se os dados estiverem à esquerda, acima do hiperplano ótimo, eles serão classificados como verdes, caso contrário, azuis.

Uma dúvida que pode surgir é que esse hiperplano poderia, por exemplo, estar rotacionado (de leve) no sentido anti-horário, usando apenas o vetor de suporte verde superior como margem dessa classe. Isso é verdade, no entanto, tal direção não geraria o máximo de ρ. Isso é feito para todos os pares de classes do conjunto de dados para que os limites de decisão sejam determinados. Nesse caso, as margens são denominadas rígidas, pois não há dados entre elas.

Figura 6.9 - Esquematização da construção de hiperplanos que delimitam as regiões de classificações, método do SVM.



Fonte: Horewicz et al. (2007).

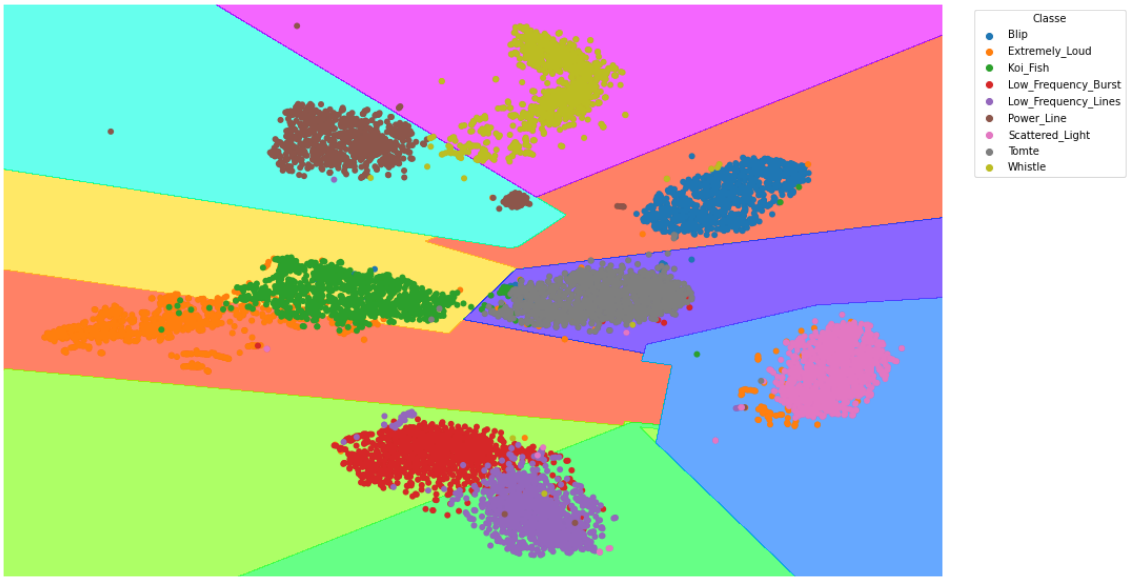
Criar regiões de classificação a partir da Figura 6.4 parece interessante. Pois além de classificar glitches, o SVM pode quantificar a eficiência do t-SNE. No entanto, num caso real como esse, há intersecções de dados e apenas cinco classes parecem linearmente bem separáveis. Como essa técnica funcionaria para as intersecções principalmente entre Extremely Loud e Koi Fish, e Low Frequency Lines e Low Frequency Burst? Não parece intuitivo encontrar vetores de suporte sem dados no meio deles.

Na verdade, o SVM também funciona nesses casos reais, o que foi mais um motivo para sua escolha. Nesse caso, o objetivo torna-se encontrar um hiperplano que tenha melhores classificações para maioria dos dados de treinamento, ou seja, tenha menos erros possível (JAMES et al., 2013). Algumas vezes, esse processo é denominado busca da margem suave, onde as margens são suavizadas permitindo que os dados também estejam presentes entre essas linhas impostas pelos vetores de suporte (USHIZIMA et al., 2005).

Essa característica é determinada pelo usuário através do chamado parâmetro de regularização, C. Quanto menor o C, maior a margem permitida para classificações erradas; quanto maior C, menor a margem e mais a função vai criando contornos para conseguir se adequar às classes corretas. Se esse parâmetro for muito pequeno, mais erros são permitidos, o que é ruim. Se for muito grande, todas as classes serão corretamente classificadas, levando ao conhecido *overfitting*. Em ambos casos, as classificações de dados serão ruins.

Dessa forma, para o caso dos glitches, essa técnica foi aplicada usando o valor de C padrão do algoritmo (C=1), que funcionou bem para esses dados. Para isso, as duas coordenadas de saída do t-SNE foram definidas como atributos para cada transiente e serviram de entrada para o SVM. Os contornos foram desenhados e as regiões foram coloridas para efeito de visualização; o resultado final está na Figura 6.10. Como esperado, há nove regiões determinando bem a dominância de cada classe. Se, por exemplo, um ruído aleatório cair na região rosa, ele deverá ser classificado como Whistle. Relembrando que, diferentemente do t-SNE, as classificações dos glitches (atribuídas pelo GS) também foram inseridas e foi, a partir delas, que o SVM criou as regiões de classificações apresentadas.

Figura 6.10 - Regiões de classificação para os glitches criadas a partir da técnica de AM supervisionado SVM. Esta só foi possível implementar depois da aplicação do t-SNE; caso contrário a visualização não seria possível. A entrada para o algoritmo foi composta pela classe de cada glitch e as duas coordenadas obtidas na redução de dimensões.



Fonte: Produção da autora.

Com classificador criado, o conjunto de teste pode ser utilizado. Esses dados entram no algoritmo sem categoria e são classificados pelo modelo. Os dados foram separados da seguinte forma: 70% dos glitches para treinar a máquina e 30% para testar. Cada um dos 2700 glitches recebeu uma classe pelo SVM que pôde ser comparada com a classe previamente concedida pelo Gravity Spy. Dessa forma, a acurácia do modelo foi obtida.

O modo mais utilizado para verificação de um método é chamado de validação cruzada, que é obtida através da matriz de confusão. A Tabela 6.1 apresenta essa matriz que relaciona os resultados preditos com os reais. O eixo vertical representa as classes verdadeiras, e o horizontal, as previstas pelo método. Assim, se o resultado real for positivo e o método der negativo, este será um dado falso negativo. No fim, os resultados principais estarão na diagonal da matriz (verdadeiro positivo e verdadeiro negativo), onde as classificações reais e do modelo concordam.

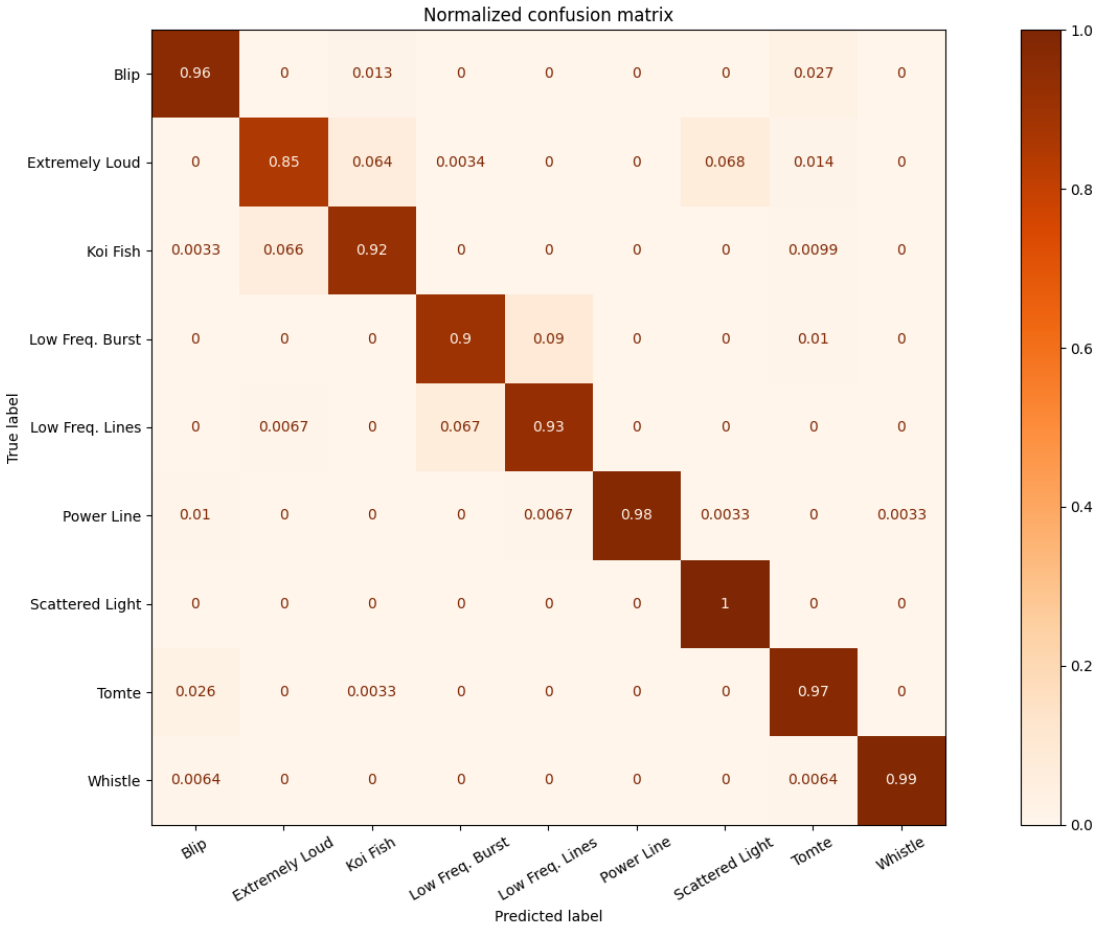
Tabela 6.1 - Montagem da matriz de confusão.

	Positivo	**Negativo**
Positivo	Verdadeiro positivo	Falso negativo
Negativo	Falso positivo	Verdadeiro negativo

A matriz de confusão para classificações dos glitches está na Figura 6.11; novamente, no eixo vertical está a verdadeira classe (*true label*), proveniente do GS, e no horizontal, o predito pelo modelo (do método SVM). O ideal seria que 100% dos transientes classificados pelo modelo fossem classificados corretamente e, nesse caso, a matriz de confusão (normalizada) teria uma diagonal unitária. No caso real, quanto mais próximo a 1, melhor.

Os resultados são bons, quanto mais escura a cor, mais próximo de um o valor está. Sete das nove classes tiveram mais de 90% de concordância com o GS. Os valores estão menores para Low Frequency Burst (90%) e para Extremely Loud (85%) que, inclusive, mais uma vez é o de menor acurácia. O maior erro cruzado foi de 9% para glitches da classe Low Frequency Burst que foram classificados como Low Frequency Lines pelo SVM. Também, há intersecções entre Extremely Loud e Koi Fish (como esperado pela distribuição de pontos do t-SNE), entre Extremely Loud e Scattered Light, poucas entre Tomte e Blip (que mais uma vez era esperado pela semelhança entre eles), mas no geral os resultados são interessantes.

Figura 6.11 - Matriz de confusão criada para os glitches a partir de classificações do SVM.



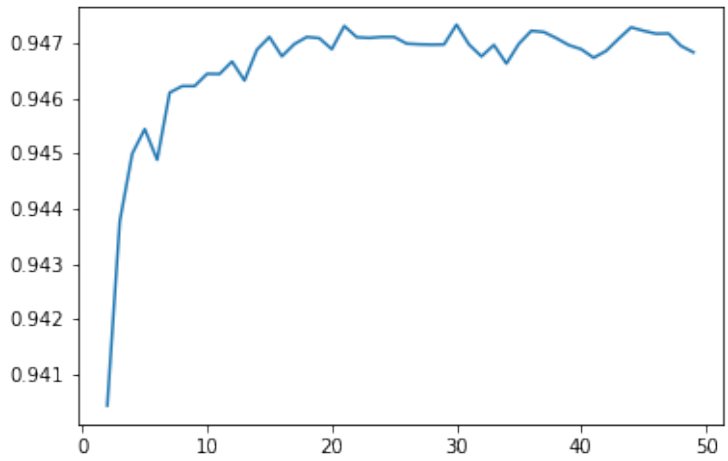
Fonte: Produção da autora.

A divisão do conjunto de dados entre teste e treinamento é feita de forma aleatória. Assim, cada vez que o modelo roda, pode haver mudanças na matriz de confusão. Uma maneira de contornar isso é fazendo a validação cruzada. Esta calcula a média dos acertos do modelo a partir de vários subconjuntos de teste e de treinamento. Ou seja, para o caso acima uma média para acertos pode ser feita. Caso outro conjunto de dados seja escolhido, ela pode variar. A validação cruzada pega diferentes subconjuntos e calcula a acurácia para obter um valor médio.

Para o caso dos glitches, foi calculada a acurácia média quando apenas um conjunto de dados é selecionado, depois quando dois conjuntos diferentes dos dados são selecionados, e assim por diante, até chegar em cinquenta. Todas tiveram dados distribuídos de formas diferentes, mas seguindo a razão de 0,7 para treinamento e de 0,3 para teste.

A precisão média final é indicada na estabilização da curva apresentada na Figura 6.12, que mostra como as acurácias médias (eixo vertical) variaram com a quantidade de subdivisões feitas (eixo horizontal). Pode-se dizer que, a partir de vinte subconjuntos, houve uma estabilidade relativa e uma acurácia de 94,7% foi obtida entre o método proposto (t-SNE + SVM) e as classificações do Gravity Spy.

Figura 6.12 - Como varia a acurácia das classificações dos glitches de acordo com a quantidade de vezes que os dados foram divididos (de formas diferentes) em treinamento e teste.



Fonte: Produção da autora.

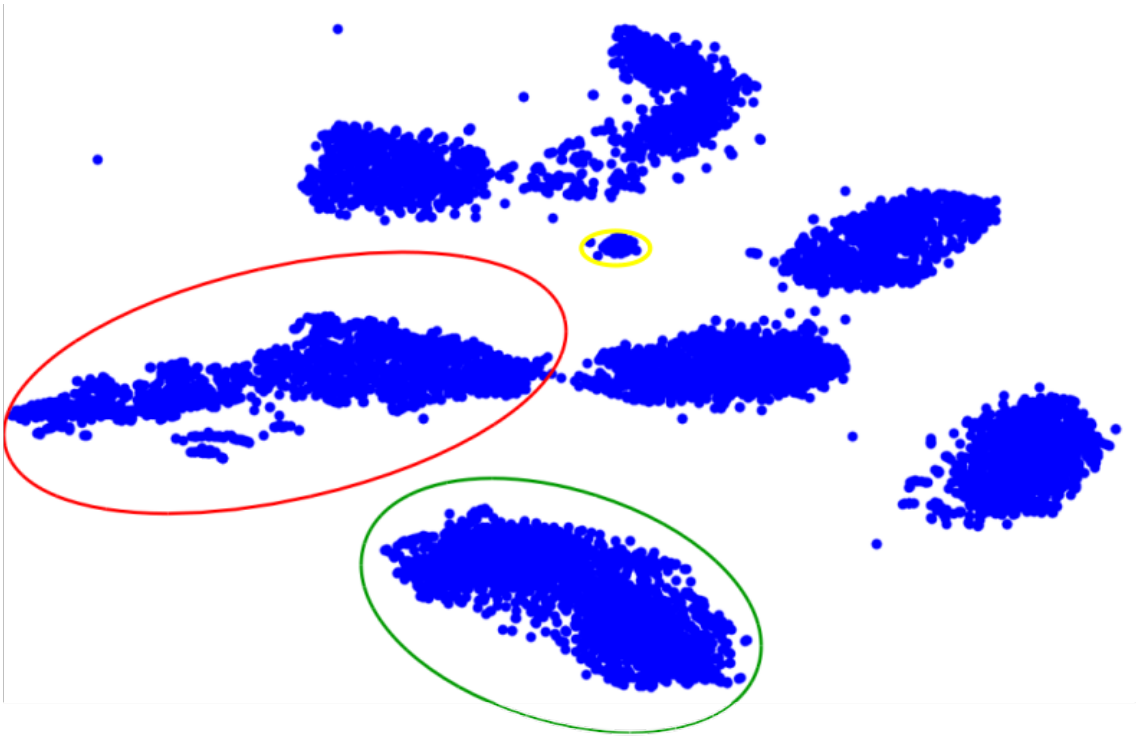
As técnicas de AM mostraram grande eficiência. No entanto, o SVM só foi possível de ser aplicado porque as classes dos glitches já eram conhecidas. A seção seguinte comenta como esse estudo poderia ser útil, mesmo sem a existência das classes fornecidas pelo Gravity Spy.

6.3 Como essa técnica poderia ser útil se o Gravity Spy não existisse?

Como foi apresentado, a análise de classificação do SVM só é possível por causa do GS. Além disso, a presença das nove classes foi inicialmente encontrada pelo t-SNE por análise visual (Figura 6.4). Para supor a inexistência do GS, é preciso voltar justamente na resposta do t-SNE. As cores não poderiam ter sido aplicadas, é como se a imagem tivesse apenas uma cor e só fosse evidente as presenças dos grupos formados. A Figura 6.13 representa como seria esse resultado.

Sem informações prévias, poderia-se dizer que há sete grupos. Talvez, seis bem definidos e um aparentando ter duas classes (o destacado em vermelho). Por outro lado, seria praticamente impossível dizer que o grupo destacado em verde tenha duas classes ou não. Apenas para lembrar, os grupos presentes no destaque vermelho são Extremely Loud e Koi Fish, e no verde, são *Low Frequency Lines* e *Low Frequency Burst*.

Figura 6.13 - Saída do t-SNE sem conhecimento prévio das classes fornecidas pelo GS.



Fonte: Produção da autora.

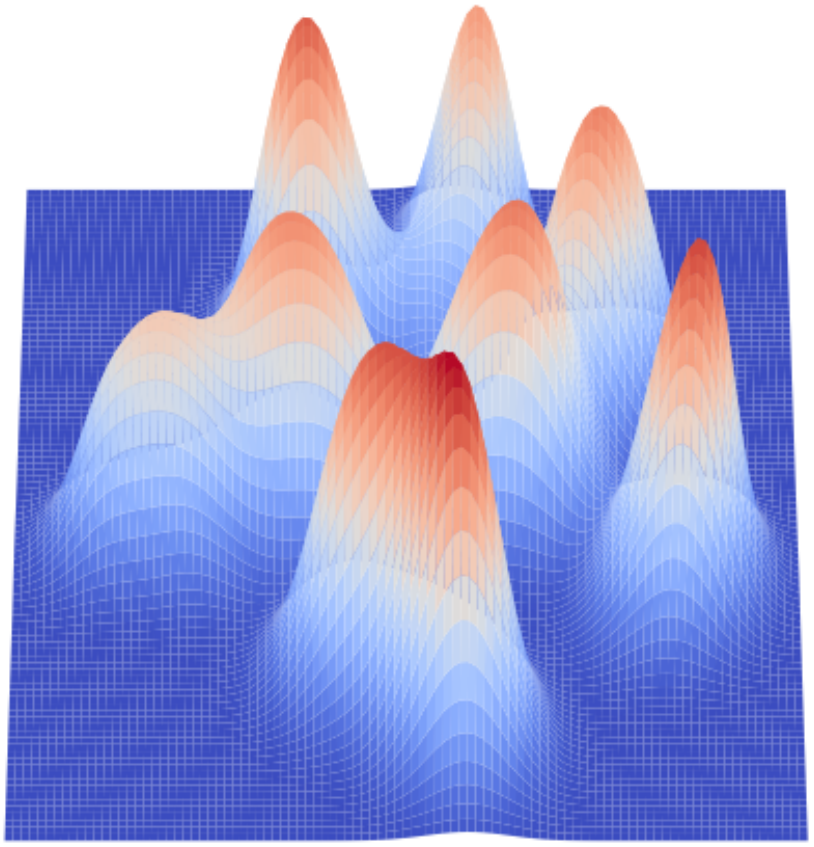
Um procedimento para verificar quantas classes realmente há neste conjunto de dados poderia ser feito desenhando um gráfico de densidade. A Figura 6.14 mostra como seria o gráfico de densidade 3D para este caso. A cor azul representa baixa quantidade de pontos, enquanto que a cor vermelha indica a região com alta densidade, que também é indicada pela presença dos picos. De fato, as cinco classes são fáceis de localizar e as outras são muito próximas. No entanto, há exatamente nove picos correspondentes às exatas nove classes estudadas. Ou seja, sem o GS, o t-SNE seria capaz de encontrar nove grupos de glitches nesse conjunto de dados.

Por outro lado, ele não é um classificador, e o SVM não poderia ser aplicado sem conhe-

cimento prévio. Seria preciso um classificador independente; daí, entra o Best Partition, discutido no capítulo anterior. Ele é capaz de encontrar comunidades e atribuir classes através da modularidade.

Sabendo disso, a técnica foi aplicada aos grupos do t-SNE que apresentaram picos próximos (e estão destacados em vermelho e verde). O método encontrou duas classes para o grupo destacado em verde: B_0 com 92,3% de concordância com Low Frequency Burst e B_1 com 94,9% com Low Frequency Lines. Para o outro grupo, uma classe teve equivalência de 66,4% com Extremely Loud e outra de 75,2% com Koi Fish. Claro que, novamente, as concordâncias foram apresentadas para avaliação do método, mas sem o GS, as classes poderiam ser chamadas de apenas B_i, com $i = 1, 2, ..., 9$. Caso houvesse interesse, a morfologia poderia ser analisada e um nome ser sugerido.

Figura 6.14 - Gráfico de densidade de pontos a partir da saída do t-SNE. A presença das nove classes de glitches é visível pela quantidade de picos. Caso o GS não existisse, ainda seria possível encontrar os nove grupo só com o uso do t-SNE.



Fonte: Produção da autora.

6.4 Discussões e outras possíveis aplicações do t-SNE

As técnicas de AM aplicadas aos glitches a partir dos glitchgramas foram muito boas. Em geral, tiveram quase 95% de concordância com o Gravity Spy e foram bem melhores se comparadas com o método de cosseno de similaridade (que teve uma média de coincidência de 75%). Um artigo comparando esses dois métodos aplicados aos glitchgramas foi publicado e pode ser acessado em Ferreira e Costa (2022).

Além dos erros desse método, o GS ainda tem erros próprios que, se não existissem, poderiam ter aumentado ainda mais a confiança. Um exemplo pode ser visto na Figura 6.15. Os quatro transientes foram classificados como Blip pelo GS. Os dois, da esquerda, são claramente Tomte e os dois da direita, Koi Fish; todos eles foram posicionados nos grupos corretos pelo t-SNE. A busca desses tipos de erros pode ser feita para todos os glitches nos quais as classificações entre o método e o GS divergem.

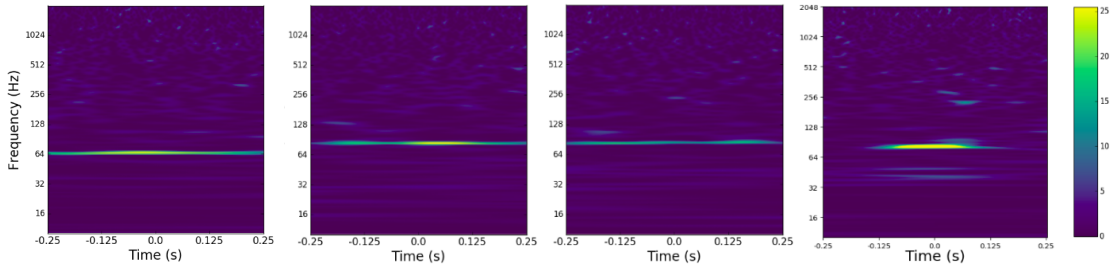
Figura 6.15 - Quatro glitches classificados com Blip pelo GS, mas que não são. O t-SNE colocou cada um no grupo correto.

Fonte: Produção da autora.

Mais um exemplo está um mini grupo na Figura 6.13 classificados como Power Line e destacados pelo contorno amarelo. Alguns desses glitches foram selecionados e os espectrogramas foram criados para investigar por que eles estão tão separados do grupo principal. Três deles estão à direita da Figura 6.16. Curiosamente, eles têm uma frequência de pico em torno de 83 Hz, enquanto que a frequência do grupo principal de Power Line é cerca de 60 Hz. À esquerda Figura 6.16, para comparação, há um espectrograma para um glitch desse grupo principal. Todos os glitches do grupo selecionado têm uma frequência de cerca de 80 Hz. Isso mostra que a técnica ainda é sensível a glitches com frequências diferentes, o que pode ser ótimo para aplicações na busca de novas classes de glitches.

Em geral, para classificar um glitch, o GS precisa obter a serie temporal em torno do tempo em que o Omicron encontrou um transiente, fazer a transformada Q, criar o espectrograma correspondente ao intervalo de tempo escolhido e só então, aplicar técnicas de AM. No

Figura 6.16 - À esquerda, há três espectrogramas do mini grupo isolado de Power Line. Eles têm frequência de pico em torno de 83 Hz, um pouco maior do que o usual encontrado no grupo principal que é em torno de 60 Hz. Um exemplo da classe principal está à direita.



Fonte: Criado a partir do GWpy (2022).

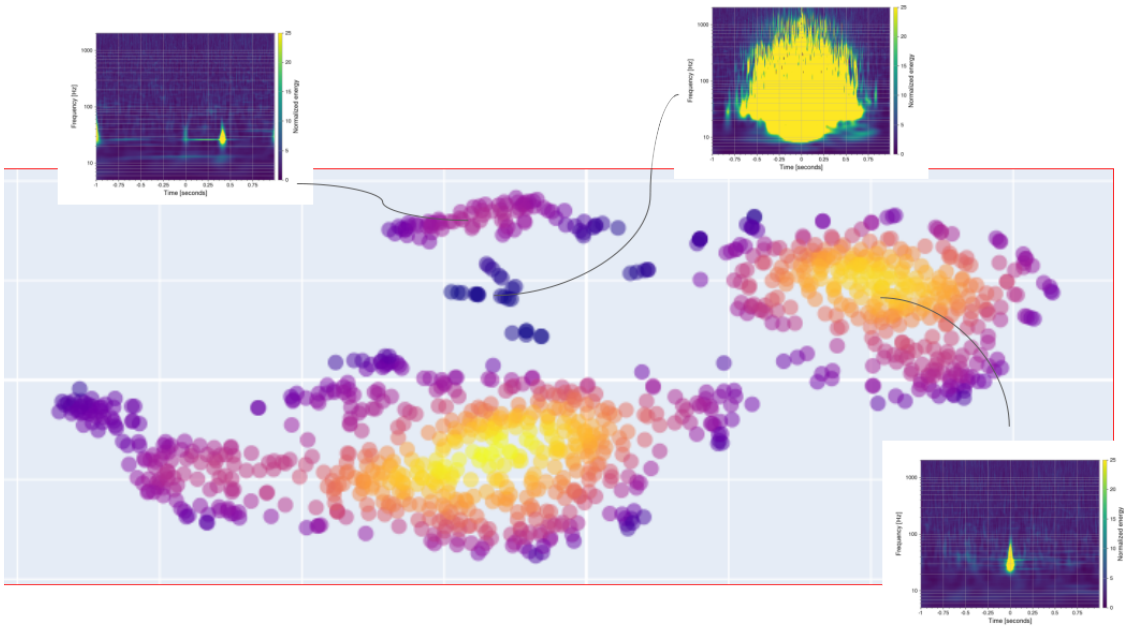
caso do GS, é utilizado um método de Deep Learning para análise da imagem formada no espectrograma, o que deixa o processo ainda mais demorado; além disso, ainda é preciso contar com as classificações de voluntários. Claro que a resolução e classificações desse modo são mais confiáveis, mas o método (t-SNE + SVM) através glitchgrama não ficou tão distante. Das duas ferramentas de AM apresentadas aqui, o processo de análise do t-SNE é mais rápido. Por outro lado, o uso do SVM é um pouco mais lento, principalmente, para desenhar as regiões de contorno de classificações. O uso deste foi essencial para verificar a eficácia do t-SNE.

Mesmo sem a existência do GS, o t-SNE mostrou-se efetivo. Através de gráficos de densidade de pontos, é possível encontrar as nove classes de glitches. A ideia deste projeto não é substituir o GS, mas apenas testar métodos alternativos e rápidos que possam colaborar com seus erros de classificação. Além disso, ter em mãos um método que seja capaz de encontrar indícios de glitches em canais auxiliares ou no próprio canal gravitacional. Por exemplo, há uma outra aplicação do t-SNE na Figura 6.17. Trata-se de um dia aleatório neste ano em que o LIGO estava fazendo testes. Os dados do Omicron para o canal gravitacional foram acessados e os glitchgramas criados para a aplicação do t-SNE. O resultado mostra a formação de diferentes grupos.

Há dois grupos formados com altas densidades de dados (em amarelo), um outro na parte superior de baixa densidade e alguns pontos agrupados entre esse grupo superior e o maior grupo na parte inferior. Os espectrogramas médios de quinze glitches aleatórios de cada grupo foram criados e estão na imagem conectados por uma linha para indicar qual grupo está sendo representado. O primeiro grupo de cima não trouxe nenhuma informação relevante, os glitches presentes nele são aleatórios e quatro podem ser vistos na primeira linha da Figura 6.18. Ao visualizá-los, é possível perceber que não há um padrão morfológico

nas imagens.

Figura 6.17 - Resultado da aplicação do t-SNE para um dia aleatório do LIGO de Livingston. Há dois principais grupos (com altas densidades em amarelo), um subgrupo superior e alguns pontos entre eles. Cada linha liga um grupo ao espectrograma médio de quinze glitches aleatórios presentes nele.



Fonte: Produção da autora.

No mini grupo logo abaixo desse, há poucos glitches com altos SNRs que, provavelmente, fariam parte de uma classe como Extremely Loud. Quatro deles também podem ser vistos na segunda linha da Figura 6.18. Apesar dos dois da esquerda serem diferentes dos da direita, ambos são fortes e têm altas razões sinal-ruído.

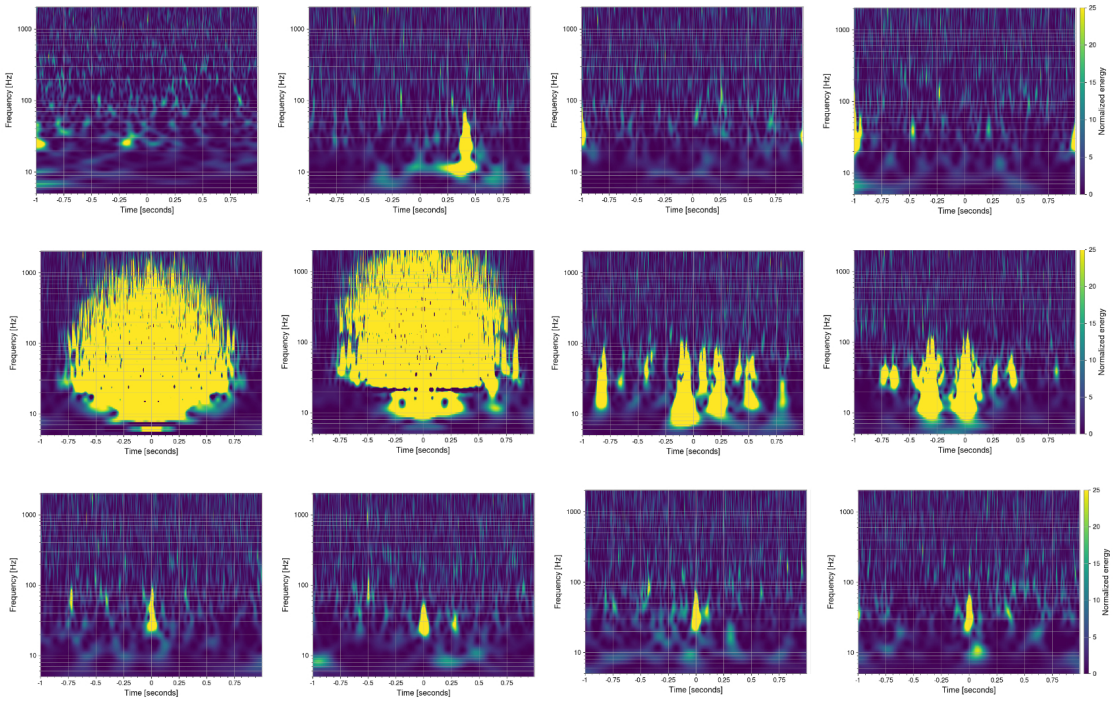
À direita, na parte superior, há um grupo grande de glitches parecidos com Tomte. A maior parte desse grupo só contém glitches com essa morfologia, quatro deles estão na terceira linha da Figura 6.18. Esse foi um grupo bem definido, isto é, ele agrupa bem e contém mesmos tipos de glitches, o que era esperado, pois o grupo tem uma alta densidade de pontos.

Isso significa que o t-SNE com glitchgrama podem ser aplicados durante um dia para saber quais classes de glitches foram predominantes. Inclusive, o outro grupo de alta densidade,

à esquerda, foi analisado e confirmou a presença de uma nova classe de glitch. Essa classe, infelizmente, não será apresentada por questões de sigilo, pois esses dados foram de apenas um dia de teste e o LIGO ainda não publicou informações sobre essa classe.

Outros dias foram avaliados e o t-SNE mostrou-se eficaz na busca dos glitches mais presentes e na evidência dessa nova classe. Esse tipo de estudo pode ser feito para um dia, uma semana ou um mês e, no fim, haverá a indicação da predominância de determinada classe por período. Por exemplo, se uma classe K foi dominante durante uma semana e, se nessa semana houve alguma situação diferente, a classe K poderá ser associada a essa situação. É possível fazer diferentes aplicações com essa técnica e todas elas poderão entrar como projetos futuros.

Figura 6.18 - Exemplos de quatro espectrogramas de cada um dos três grupos encontrados pelo t-SNE durante um dia. Todas as imagens foram geradas pelo pacote GWpy.



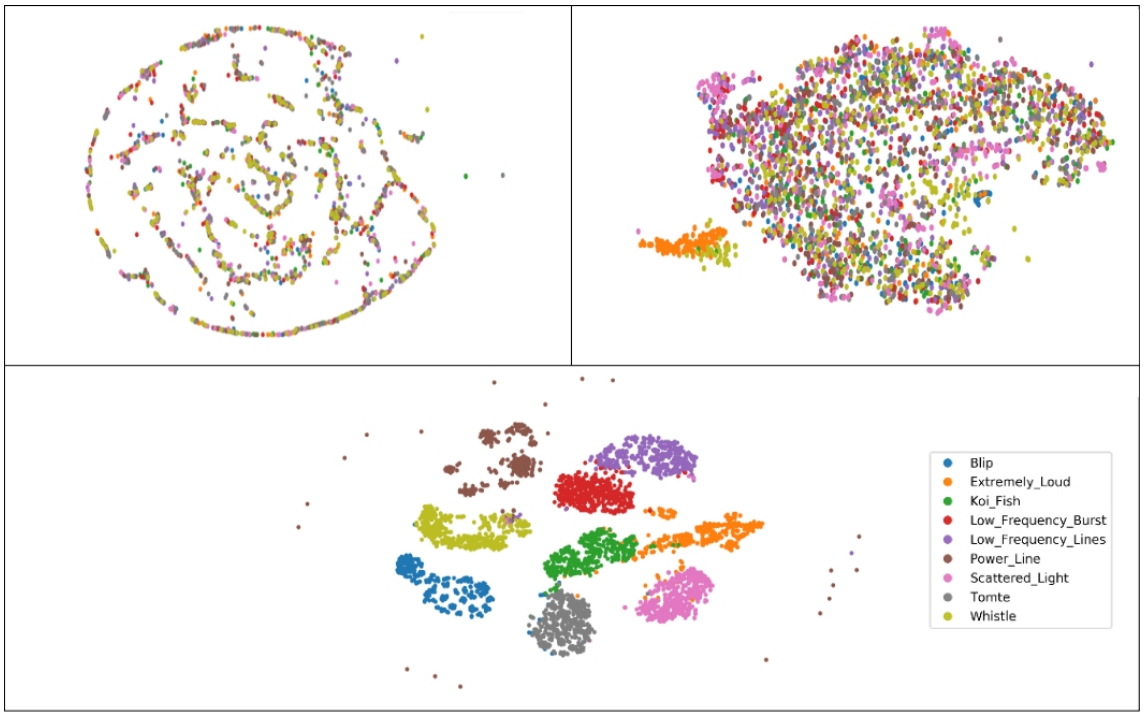
Fonte: Produção da autora.

Para finalizar, uma das intenções deste trabalho foi também criar uma ferramenta para buscar a presença de classes de glitches em canais auxiliares. Uma maneira de fazer isso é: uma vez que os tempos $\boldsymbol{t_g}$ em que os glitches aconteceram no canal gravitacional são

conhecidos, assume-se que esses glitches também aconteceram nos canais auxiliares num intervalo $t_g \pm 1\,\mathrm{s}$. A partir dessa imposição, os glitchgramas desses tempos são criados e o t-SNE aplicado. A presença de grupos no canal irá dar indicar se as classes escolhidas estão presentes nele ou não.

A Figura 6.19 apresenta o resultado dessa aplicação para três canais auxiliares aleatórios. A primeira imagem, à esquerda, trata-se de um canal auxiliar que não tem presença de classes evidentes. Existe um círculo principal com a mistura de todos os glitches selecionados para investigação; isso indica que não há padrões em subconjuntos dos dados, eles estão distribuídos de forma aleatória e, portanto, pode-se dizer que não há a presença dessas classes de glitches nesse canal.

Figura 6.19 - Exemplos do t-SNE aplicados a três canais auxiliares. O primeiro tem dados totalmente aleatórios; o segundo apresenta um grupo evidenciado, indicando a possível presença do Extremely Loud no canal; no terceiro, há a presença de quase todas classes, pois é um canal próximo ao gravitacional.



Fonte: Produção da autora.

A imagem à direita é a resposta do t-SNE para outro canal auxiliar; novamente, há uma mistura de dados num grande aglomerado à direita, indicando que nesses tempos aconteceram transientes nos canais, mas que não tiveram padrões repetitivos, apenas coincidiram

temporalmente com os glitches do canal gravitacional. No entanto, também há um mini grupo laranja no canto esquerdo. Isso significa que o algoritmo encontrou dados com morfologias similares que coincidem com os tempos de aparições do Extremely Loud no canal gravitacional, indicando a possível presença da classe nesse canal auxiliar.

Finalmente, na parte inferior, há um canal muito próximo ao canal gravitacional e, por esse motivo, contém a presença de todas as classes. Essa última imagem também diz que o intervalo de um segundo na busca de coincidências entre o canal gravitacional e canais auxiliares é interessante, mas ainda não é um valor definido.

Essa é apenas mais uma aplicação do t-SNE, mas outras podem ser feitas. Vale ressaltar que a busca de classes dessa forma ainda é um processo visual e que, no futuro, deve ser automatizada. A seguir, há um capítulo sobre como o t-SNE funciona no estudo de duas classes específicas de glitches.

7 UM ESTUDO MAIS PROFUNDO SOBRE SCATTERED LIGHT E FAST SCATTERING

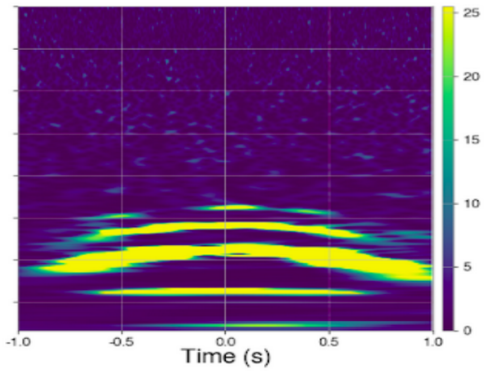
Este capítulo apresenta o trabalho desenvolvido durante o doutorado sanduíche na LSU (Louisiana State University), sob supervisão da Professora Dra. Gabriela González.

Dois transientes muito frequentes (que chamaram atenção durante a terceira corrida do LIGO) foram causados por espalhamento da luz do laser. Eles compõem as classes Scattered light e Fast Scattering. O Fast Scattering foi, inclusive, adicionado como nova classe durante a O3; além disso, foi o mais comum durante a O3a e o segundo com mais aparições durante O3b, perdendo apenas para o Scattered Light.

Aqui entra uma outra motivação para entender uma classe de glitch e automatizar seu processo de classificação (ou identificação) num conjunto de dados: novas classes de glitches poderão sempre surgir. Durante a O3, por exemplo, além do Fast Scatering, surgiu o *Blip Low Frequency* que é muito parecido com Blip, mas está presente em faixas de frequências menores. Esses surgimentos estão diretamente relacionados com as melhorias e mudanças instrumentais no observatório e, em geral, acompanham as faixas de frequências em que houve aumento de sensibilidade do detector.

O Scattered Light (também chamado de *Slow Scattering*) se apresenta em forma de arco, tem uma duração relativamente longa e pode acontecer em uma, duas ou mais faixas de frequências; seu espectrograma pode ser visto na Figura 7.1.

Figura 7.1 - O espectrograma de um Scattered Light.

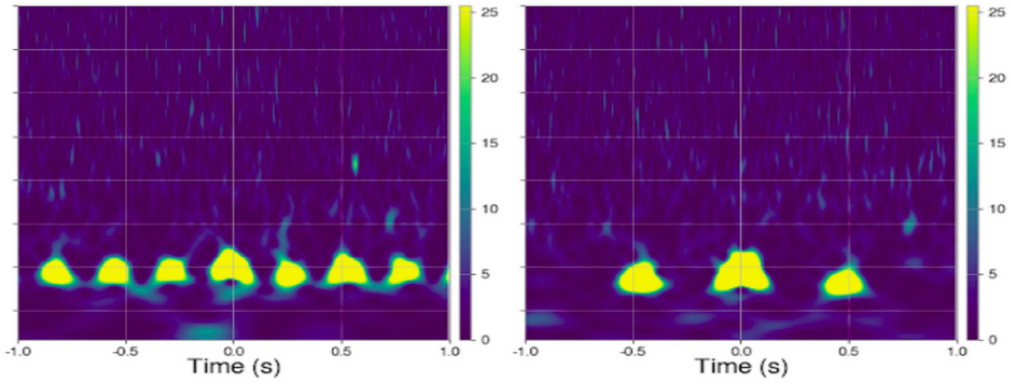


Fonte: Produção da autora.

Por outro lado, o grupo Fast Scattering, FS, é curto em duração e tem duas principais janelas de repetição: 0,25 segundo e 0,5 segundo. Quando um glitch FS tem intervalo

de repetição de 0,25 s, ele recebe o nome de FS de 4 Hz; o espectrograma de um glitch desse tipo pode ser visto à esquerda da Figura 7.2. Caso a repetição seja em 0,5 s, ele é denominado FS de 2 Hz (vide espectrograma à direita da Figura 7.2).

Figura 7.2 - Espectrogramas de dois glitches classificados com Fast Scattering. Eles são subdivididos em Fast Scattering de 4 Hz (à esquerda) e Fast Scattering de 2 Hz (à direita).

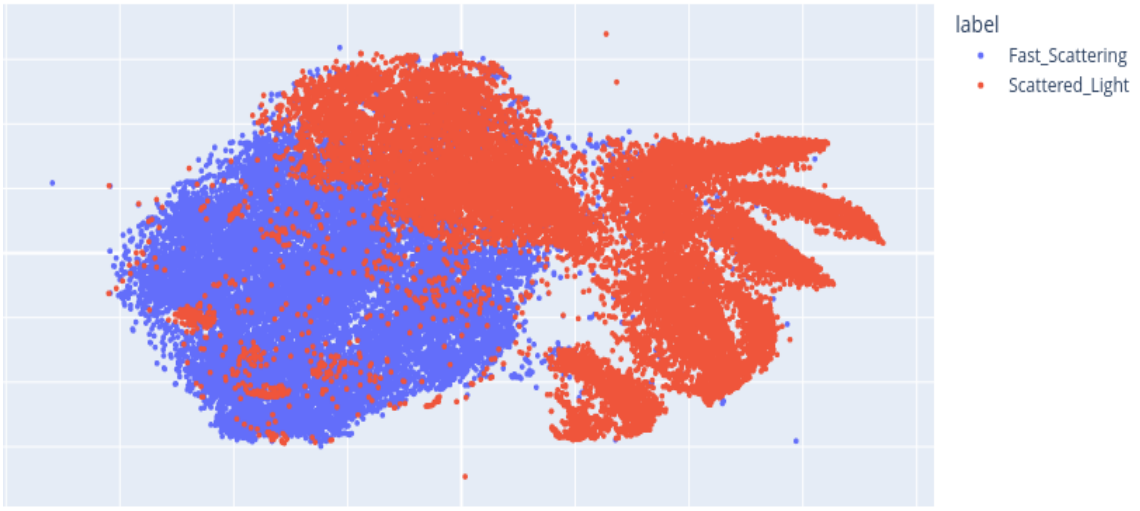


Fonte: Produção da autora.

Para estudar o comportamento desses dois tipos de ruídos e testar o método em casos específicos de glitches, o t-SNE também foi aplicado. Para isso, os glitchgramas de dez mil Scattered Light e dez mil Fast Scattering (da O3b) foram criados. Dessa forma, como no capítulo anterior, cada glitch teve seu corresponde vetor representativo de 1200 dimensões e o t-SNE reduziu cada vetor para 2D. O resultado pode ser visto na Figura 7.3; cada ponto representa um glitch, a cor vermelha indica que o glitch pertence à classe Scattered Light e a azul, à Fast Scattering.

Novamente, o algoritmo não sabe a qual classe corresponde cada vetor. Depois de agrupados, as cores foram aplicadas de acordo com as classes atribuídas pelo GS. Na Figura 7.3 é possível notar algumas intersecções entre as cores, mas em geral os glitches Scattered Light estão na parte superior e no lado direito, enquanto que os Fast Scattering dominam a região inferior à esquerda. Para ficar mais claro como estão as intersecções, as duas classes podem ser vistas na Figura 7.4 separadamente. A classe FS está à direita da imagem e a Scattered Light à esquerda.

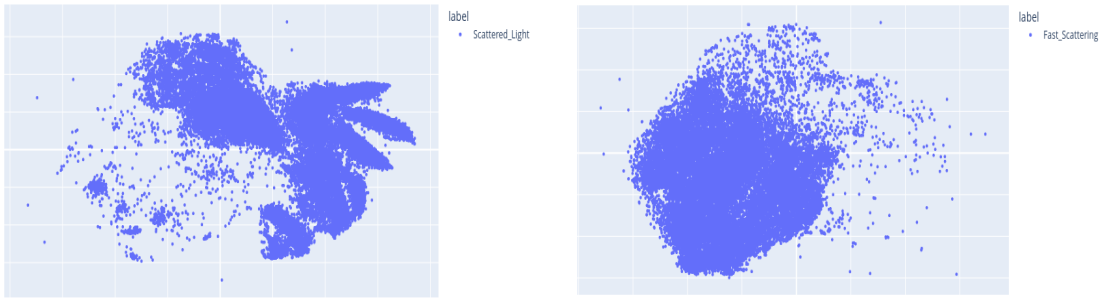
Figura 7.3 - Aplicação do t-SNE para Fast Scattering (azul) e Scattered Light (vermelho). Foram selecionados dez mil glitches de cada classe e cada um deles é representado por um ponto.



Fonte: Produção da autora.

O t-SNE separa bem as duas classes. No entanto, olhando para a Figura 7.4 (à direita), não é possível dizer se há duas subclasses do Fast Scattering ou não. O ideal seria que tivesse, pois uma deveria representar o grupo de FS de 4 Hz e a outra, o de FS de 2 Hz. Por outro lado, a distribuição dos pontos da classe Scattered Light chamou a atenção (lado esquerdo da Figura 7.4). Por que há tantos ramos/ilhas de pontos à direta? Será que são subclasses dessa classe? Se não, por que esses subgrupos existem?

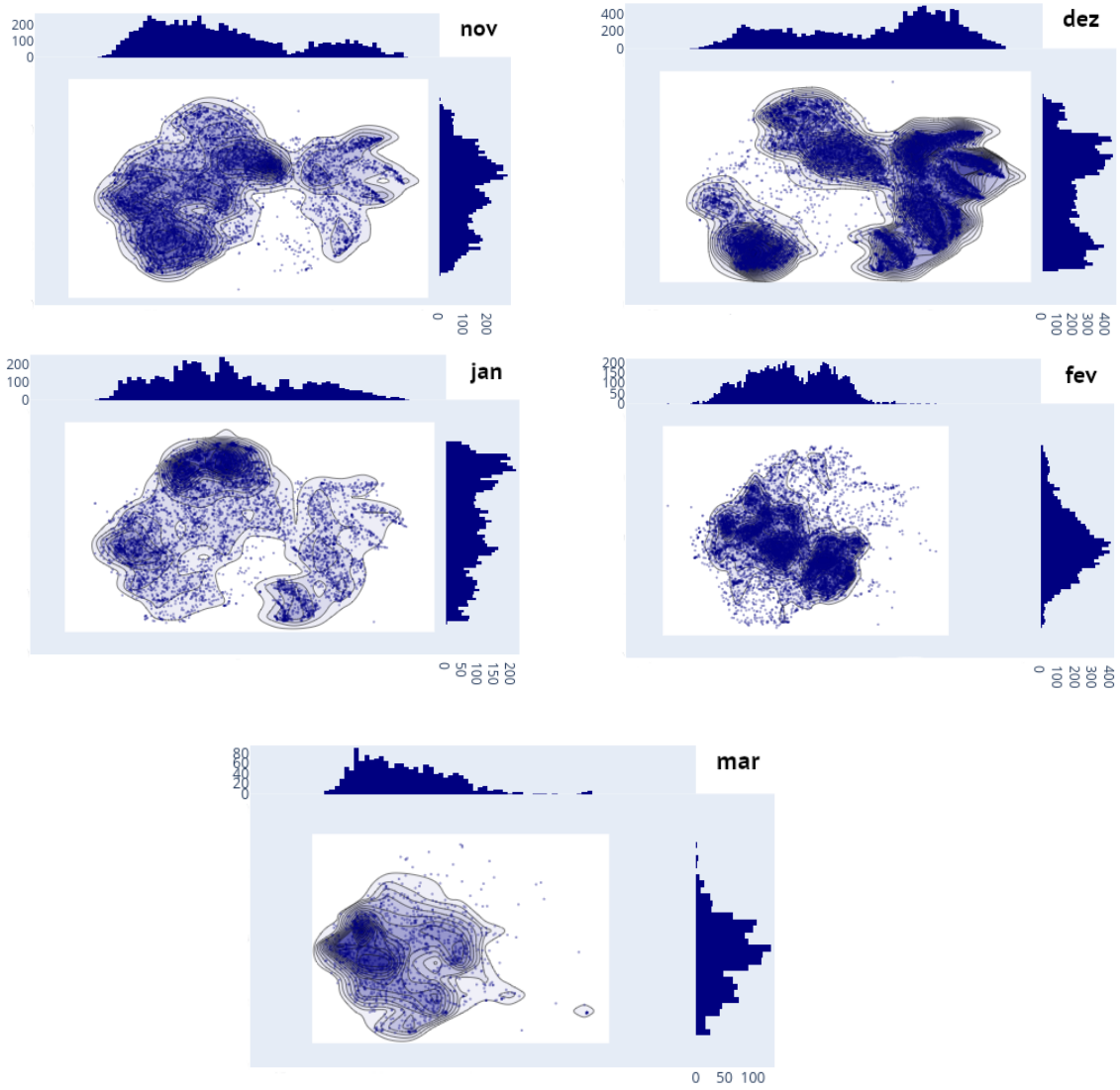
Figura 7.4 - Resultado da aplicação do t-sne para Scattered Light (à esquerda) e Fast Scattering (à direita).



Fonte: Autoria própria.

Antes dessa análise, as imagens da Figura 7.5 mostram o comportamento mensal desses dados da O3b (de novembro de 2019 a março de 2020). Em novembro, quase toda a região possui pontos, e as ilhas do Scattered Light são pouco povoadas. Em dezembro, a região central desaparece, restando apenas um subgrupo de FS. A partir de Janeiro de 2020 é visível o início do desaparecimento do Scattered Light. Em fevereiro quase não há a presença deles, sobressaindo apenas uma outra região de FS (a que não estava presente em dezembro). Em março, há uma densidade bem baixa de pontos em toda região de FS. Cada eixo tem um histograma da quantidade de pontos de dados.

Figura 7.5 - Resultado da aplicação do t-SNE para Fast Scattering e Scattered Light durante o mês de novembro de 2019, à esquerda, e durante o mês de dezembro de 2019, à direita.

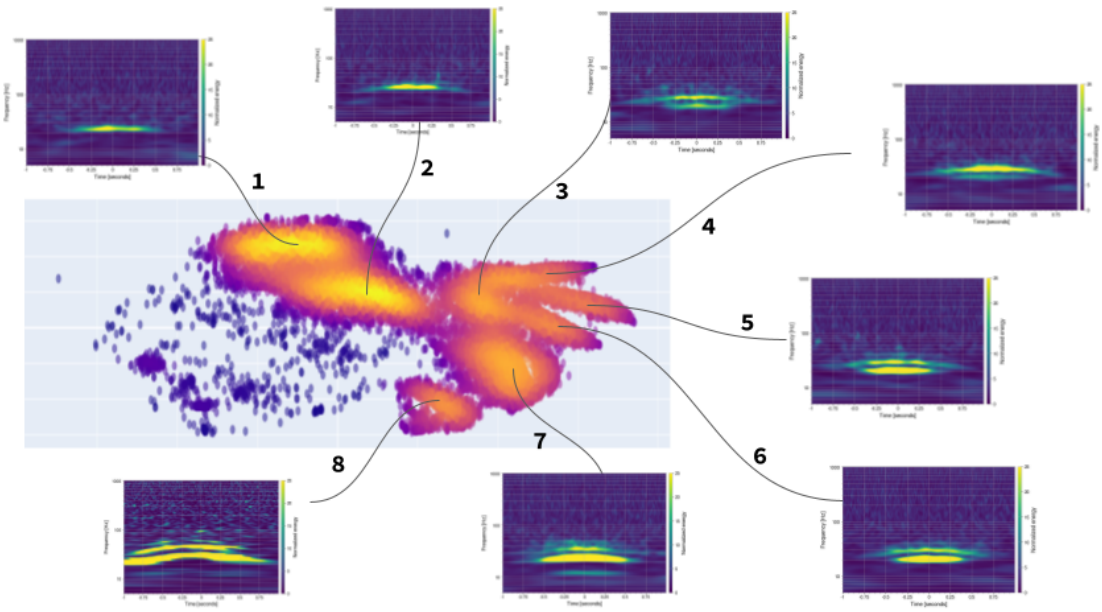


Fonte: Produção da autora.

Em geral, dependendo do mês, há regiões mais evidentes (com mais glitches) que outras. Curiosamente, fevereiro e dezembro parecem se complementar para formar a figura principal. O desaparecimento do Scattered Light está relacionado ao estudo que permitiu técnicas instrumentais para evitar esse ruído; tal pesquisa pode ser acessada no artigo Soni et al. (2020).

Voltando, para tentar responder às perguntas anteriores, foi construído um gráfico de densidade de pontos (em 2D) para cada uma das classes. As Figuras 7.6 e 7.7 são, respectivamente, os gráficos de densidade para Slow Scattering e Fast Scattering. As regiões amarelas apresentam alta densidade de glitches e as roxas, baixas densidades. De cada região amarela, quinze glitches aleatórios foram selecionados e tiveram seus espectrogramas criados para verificação.

Figura 7.6 - Densidade de Scattered Light a partir da saída do t-SNE.



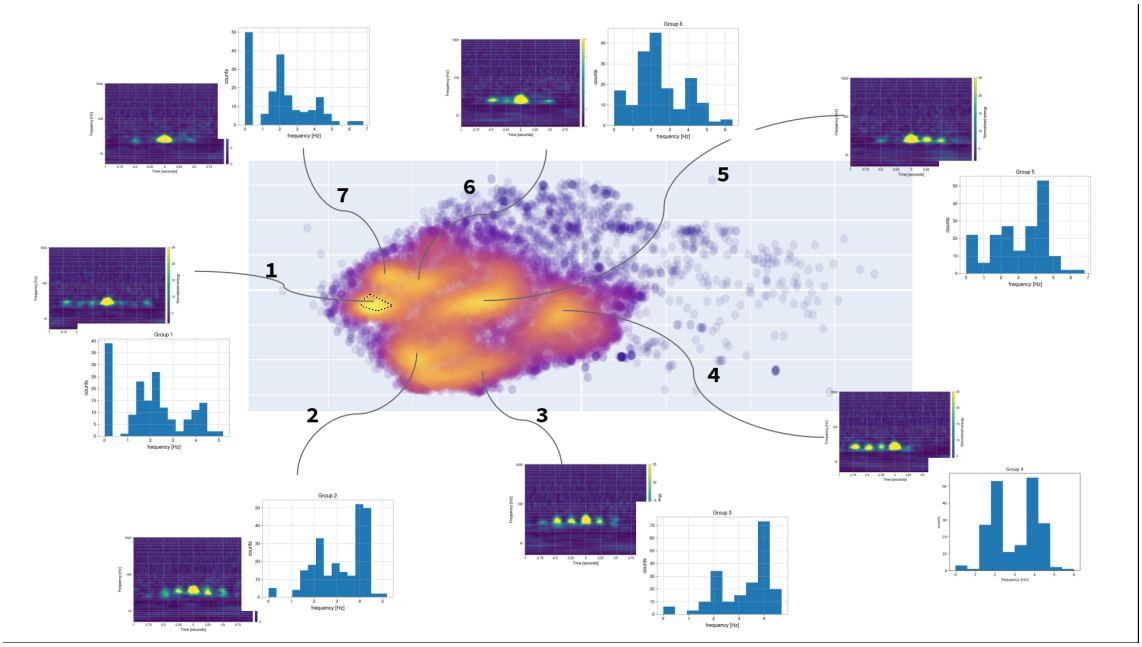
Fonte: Produção da autora.

A Figura 7.6 tem oito principais ilhas; cada uma delas está representada por um número e ligada à média desses quinze espectrogramas. Ao olhar atentamente, é perceptível que os espectrogramas médios têm diferenças significativas entre si; o da região 1, por exemplo, tem um arco em uma determinada banda de frequência e duração curta; o da 2 parece ser mais forte e mais curvado; na 3, já começa ter a presença de um outro arco em diferente faixa de frequência; na 4, o arco de cima é mais forte; os 5 e 6 são os mais similares e

parecem transições entre qual arco (superior ou inferior) é mais forte; o 7 tem uma faixa mais larga no arco de baixo que também é o mais dominante e da 8 tem dois principais arcos em duas (talvez três) largas bandas de frequência e ambos são fortes. No fim, cada grupo aparenta ter sua própria característica e os amontoados diferem um do outro em duração, bandas de frequência ou SNR.

A mesma análise pode ser realizada para a classe FS, observada na Figura 7.7. Foram selecionadas sete principais regiões de alta densidade. De forma análoga, cada uma delas tem exemplos da média de quinze espectrogramas aleatórios. Os grupos 1, 2, 3, 4 e 5 são claramente FS de 4 Hz, já os grupos 6 e 7 aparentam predominantemente FS de 2 Hz. Novamente, cada grupo difere entre si em detalhes. Por exemplo, os glitches do subgrupo 4 parecem estar deslocados para esquerda; da ilha 5, para direita; do grupo 2, os glitches parecem mais longos, e assim por diante.

Figura 7.7 - Densidade de Fast Scattering a partir da saída do t-SNE.



Fonte: Produção da autora.

Em geral, esses detalhes justificam os ramos encontrados e a presença de diferentes regiões com alta densidade de pontos. Aliás, esse método pode ser aplicado para esse tipo de pesquisa: busca por subclasses dentro uma classe de glitches; no entanto, mais confirmações e testes precisam ser feitos para verificar sua eficiência.

O Gravity Spy ainda não consegue identificar se um Fast Scattering é de 4 ou 2 Hz. Dessa forma, é preciso um método alternativo para verificar os resultados acima. Apesar das médias dos quinze espectrogramas terem sido calculadas diversas vezes, com diferentes subconjuntos de glitches aleatórios, e as características prevalecerem como as apresentadas nas Figuras 7.7 e 7.6, são poucas quantidades para fazer afirmações. Não é possível ainda garantir, por exemplo, que a maior parte dos FS de 2 Hz esteja nos grupos 6 e 7.

Para testar a eficiência das análises acima, também foi desenvolvido um código capaz de identificar se um Fast Scattering se repete a cada 0,25 ou a cada 0,5 segundos. Os passos para esse processo foram os seguintes:

- Criar um algoritmo capaz de selecionar dados da saída do t-SNE e retirar informações deles. Por exemplo, na Figura 7.7, há no grupo 1, um desenho tracejado. Esse desenho representa uma sub-região que foi escolhida. Ao fazer essa seleção, o algoritmo cria uma tabela com todos os dados que estão dentro desse desenho, carregando informações como classe e tempo de ocorrência do glitch; isso foi feito para as áreas mais amarelas de cada uma das sete regiões;

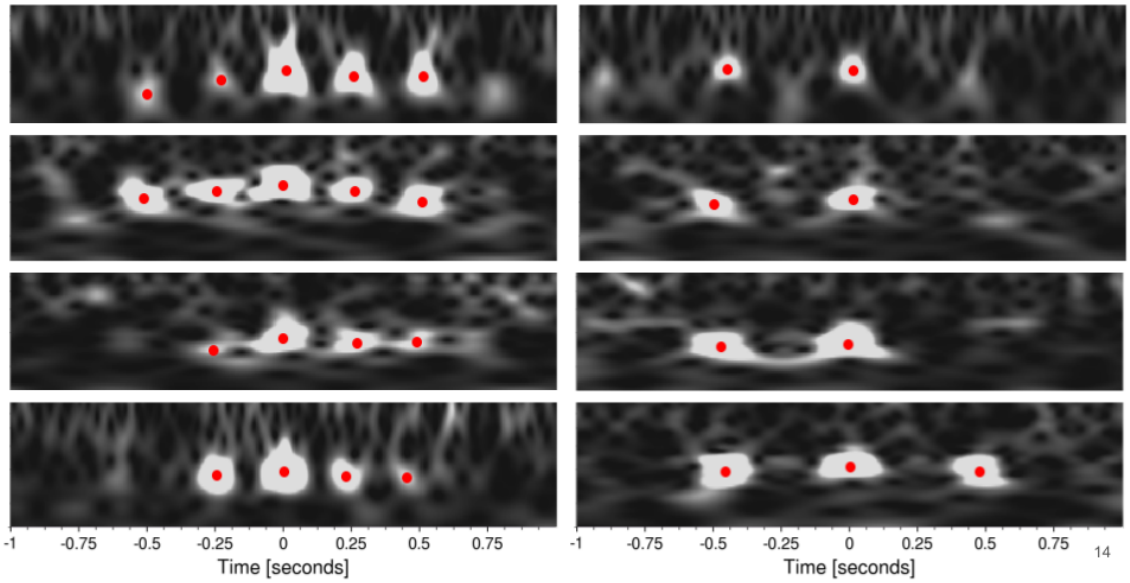
- Para cada tabela, os tempos foram lidos e seus correspondentes espectrogramas criados (através da tranformada Q). Cada espectrograma foi salvo num formato de imagem (.png);

- Um segundo algoritmo cortou cada uma dessas imagens, colocou numa escala de preto e branco e identificou os pixels mais brilhantes. Uma vez que os pixels mais brilhantes são encontrados, a distância entre eles é convertida em tempo e, dessa forma, é possível calcular o intervalo de repetição.

Com intervalo de repetição calculado para cada imagem selecionada, foi possível atribuir uma frequência correspondente. Dessa forma, no fim, cada glitch da tabela (de cada região da classe FS) recebeu uma subclassificação de 2 Hz ou 4 Hz.

A melhor maneira para checar a eficiência desse identificador de frequência é visualizando. A Figura 7.8 apresenta o output desse código para quatro transientes de 4 Hz (à esquerda) e quatro de 2 Hz (à direita). Os pontos vermelhos são atribuições do algoritmo para as regiões mais brilhantes e formam a base para o cálculo de distância entre os pixels (convertida em tempo). A identificação funcionou bem e, no geral, teve cerca de 82% de acurácia, que (por enquanto) é suficiente para ajudar a verificar o t-SNE, mas pode (e deve) ser aperfeiçoada como projeto futuro.

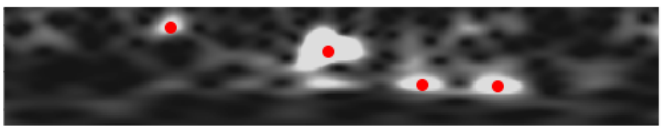
Figura 7.8 - Resposta do algoritmo criado para determinar se um Fast Scattering é de 2 ou 4 Hz.



Fonte: Produção da autora.

É importante destacar que nem todos espectrogramas são bem comportados (como os mostrados na Figura 7.8); há casos específicos que atrapalham a identificação. Para exemplificar, veja a Figura 7.9; se apenas o lado direito da imagem fosse analisado, o FS seria claramente de 4 Hz; no entanto, do lado esquerdo, há um outro ponto brilhante, cerca de 0,5 segundos antes do central. Isso atrapalha a determinação da frequência, pois a distância entre cada par de pontos vizinhos é calculada e, no fim, a distância final é dada pela média de todas as distâncias. Nessa caso específico, os dois primeiros pontos se distanciam em 0,5 s, o segundo com o terceiro em 0,25 s e o terceiro com o quarto em 0,25 s; no fim, a distância atribuída vai ser 0,33 s (e a frequência 3,03 Hz).

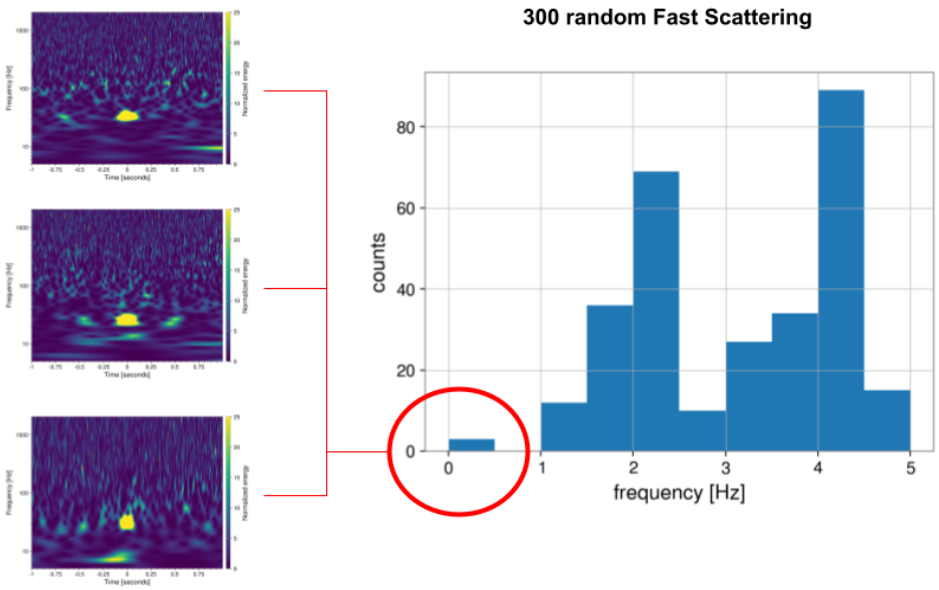
Figura 7.9 - Um exemplo de espectrograma de FS que não tem frequência bem determinada.



Fonte: Produção da autora.

Para ter uma ideia dos valores encontrados pelo algoritmo, foram selecionados 300 glitches aleatórios classificados como FS. O histograma da frequência atribuída a eles pode ser visto na Figura 7.10. Há dois principais picos em torno de 2 Hz e 4 Hz, mas há também valores entre eles que, provavelmente, correspondem a classes como da Figura 7.9. Dentre esses trezentos, três diferentes FS apareceram sem repetições de 0,25s ou 0,5s e estão destacados pelo círculo vermelho. Os espectrogramas deles também estão na imagem. É possível ver que eles têm apenas uma mancha central e, para diferenciar, eles serão denominados "0 Hz". Apenas para facilitar a escrita, glitches identificados como FS de 4 Hz serão chamados de FS4, os de 2 Hz serão FS2 e, por analogia, os de 0 Hz serão FS0.

Figura 7.10 - Histograma da frequência de 300 Fast Scattering selecionados aleatoriamente. É possível ver principais regiões em torno de 2 Hz e 4 Hz. Além disso, também há alguns sem repetição, denominado 0 Hz. Nesse exemplo, há três deles (circulados em vermelho; seus espectrogramas podem ser visto à esquerda.



Fonte: Produção da autora.

Esse algoritmo foi aplicado para alguns glitches de cada uma das sete principais regiões do t-SNE, e um histograma (como da figura anterior) foi criado para cada uma delas. O resultado pode ser visto na Figura 7.7 ao lado de cada espectrograma médio. De acordo com o identificador, os grupos 1 e 7 têm mais FS0 e FS2, a grande presença de FS0 faz com que os espectrogramas médios tenham a mancha central mais evidente; o grupo 6 tem mais

FS2, o que também é visível no espectrograma médio; o grupo 4 parece igualmente divido entre FS2 e FS4; o grupo 5 tem um pico grande de FS4 e talvez uma mesma quantidade dividida em FS2 e FS4; os grupos 2, 3 têm predominância de FS4.

O t-SNE parece separar os grupos com mais FS0 e FS2 (grupos 1, 6 e 7 na região superior à esquerda) dos grupos com mais evidências de FS4 (grupos 2, 3), mas não define uma região específica para cada uma das frequências. Há intersecções praticamente em todas as regiões e, em especial, no grupo 4 com a presença de quase a mesma quantidade de FS2 e FS4. Como projeto futuro, é interessante testar melhorias no método (incluindo no glitchgrama) para casos específicos como esses, em que há interesse na busca de subclasses de uma única categoria.

Para aproveitar o identificador criado, um outro estudo foi realizado: o início da busca da causa específica do Fast Scattering. A seção seguinte discute os conceitos e as conclusões até o momento.

7.1 A busca da origem de glitches da classe Fast Scattering

Existe uma hipótese de um pesquisador da colaboração LIGO (SONI, 2022) que diz que o espalhamento do laser que causa o glitch Fast Scattering possa ser proveniente de relações entre movimentos microssísmicos (MM) e antropogênicos (MA). A frequência de repetição das manchas dependeria da força relativa entre os movimentos nessas duas bandas. Ele diz que: altos MA e baixos MM resultam em FS4, e valores próximos de MM e MA resultam em FS2. Em outras palavras,

$$\text{Se} \quad R = \frac{MA}{MM} > T, \quad \text{há a presença de um FS de } 4\,\text{Hz no detector;} \tag{7.1}$$

$$\text{Se} \quad R = \frac{MA}{MM} < T, \quad \text{há a presença de um FS de } 2\,\text{Hz,} \tag{7.2}$$

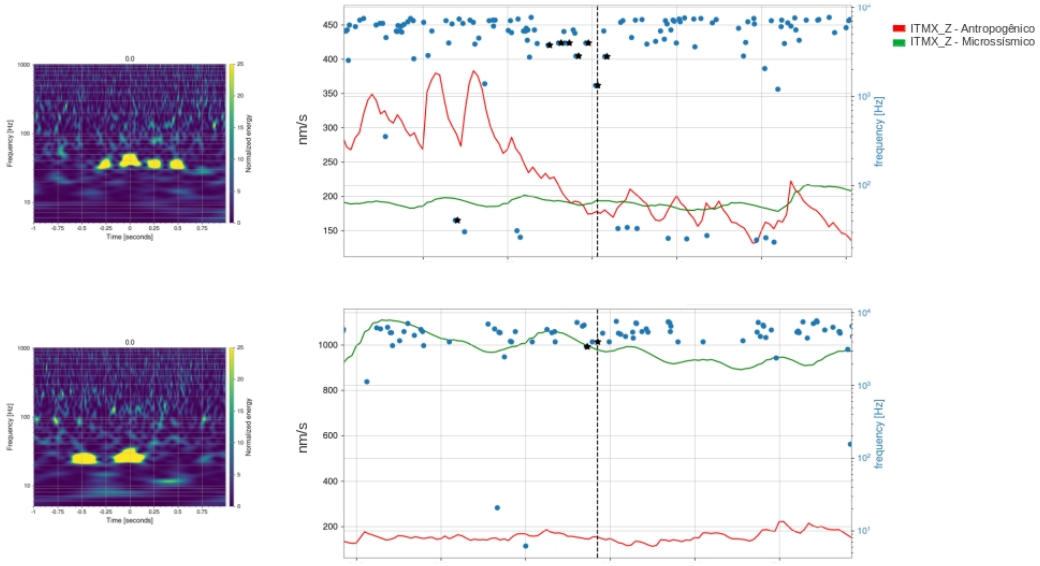
onde o valor T é um limite imposto que vai definir a frequência de repetição das manchas. Trata-se de um valor desconhecido que ainda precisa ser determinado.

Para verificar essa hipótese, é preciso analisar como é o comportamento de canais auxiliares (sismômetros) que monitoram esses movimentos do solo. Os movimentos sísmicos são medidos em diferentes bandas de frequência. Quando o microssísmico é referenciado significa que a banda analisada é a de $0,1$ a $0,3$ Hz; no caso antropogênico, é a de 1 a 3 Hz.

A Figura 7.11 mostra o comportamento desses canais em dois períodos de tempo. A parte superior é em torno de um glitch FS de 4 Hz. Esse glitch está representado por uma estrela que é intersectada por uma linha vertical tracejada. O eixo horizontal mostra 120 segundos em torno desse glitch, e o eixo vertical indica as medidas dos canais em nm/s (nanômetros por segundo). Em vermelho está a curva dos MA e, em verde, dos MM. Cada ponto azul representa um transiente e cada estrela, um glitch classificado como FS. Pode-se dizer, portanto, que no instante em que esse FS apareceu nos dados, os MA estavam muito próximos aos MM. A razão média entre esses dois movimentos foi calculada (seguindo as Equações 7.1 e 7.2) e o valor obtido foi $0,910$.

De forma análoga, a imagem inferior da Figura 7.11 mostra as medições dos movimentos do solo em torno de um FS de 2 Hz. Os MM são mais fortes (em torno de 1000 nm/s) e os MA mais fracos (cerca de 200 nm/s). Se a imagem superior for comparada com a inferior, pode-se dizer que, no tempo do FS4, os MM foram mais fracos e os MA pouco mais fortes, enquanto que o grande diferencial do caso inferior seja altos MM. A razão R para esse caso foi 0,155. À esquerda de ambos os casos, está o espectrograma de cada glitch escolhido como referência.

Figura 7.11 - Movimentos (em nm/s) antropogênicos, em vermelho, e microssísmicos, em verde, em torno de dois glitches classificados como FS (representados pela estrela na linha vertical tracejada). Os dados superiores são referentes a um FS de 4 Hz e os inferiores a um FS de 2 Hz. Todas as estrelas representam FS e os pontos em azul outros tipos de transientes. À esquerda de cada serie temporal, há o espectrograma do FS selecionado.



Fonte: Produção da autora.

Esses são apenas dois exemplos que concordam com a hipótese proposta, mas para testar mais casos, um algoritmo foi criado para acessar os dados dos dois canais (em torno de glitches classificados como FS) e calcular as razões R. Eles foram divididos em três subclasses (utilizando o código identificador mencionado anteriormente): 0 Hz, 2 Hz e 4 Hz. Para cada uma dessas três, glitches foram selecionados e as razões foram calculadas.

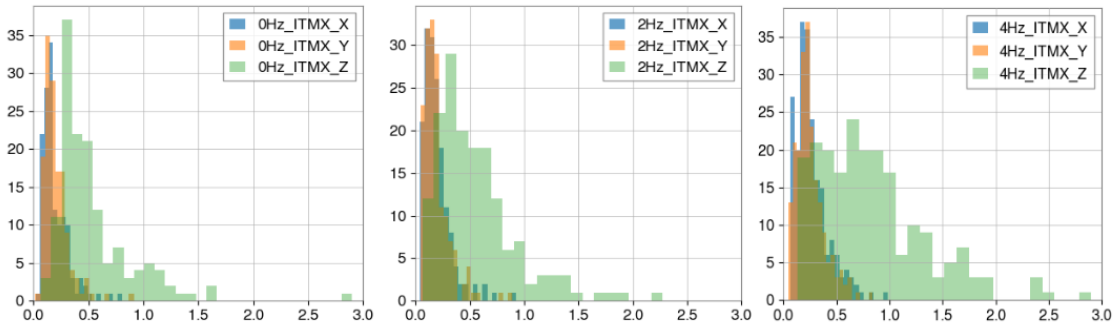
A hipótese é para movimentos nos canais que monitoram as regiões das massas testes iniciais, ITM**I**. Novamente, **I** indica se é o espelho da direção X ou Y. Além disso, mais uma outra informação será acrescentada. As posições dos canais auxiliares serão agora referenciadas por ITMI_**J**, com **J** = X, Y, Z indicando a orientação do movimento medido; X mede movimentos horizontais na direção do braço X, Y mede movimentos na direção do braço Y, e Z mede movimentos verticais em relação aos espelhos. Dessa forma ITMX_Z referencia-se aos movimentos do solo verticais na posição da massa teste de entrada do braço X.

Para cada uma das três subclasses de FS, um subconjunto foi escolhido (de forma aleatória) e em torno de cada glitch, a razão entre os movimentos antropogênicos e microssísmicos foi calculada. Suponha que t_{fs} seja o tempo em que um FS foi classificado; os dados dos canais auxiliares foram coletados de $t_{fs} - 1$s a $t_{fs} + 1$s, a média em nm/s para esse intervalo foi calculada para cada canal e, a partir dessa média, a razão pôde ser obtida.

Os histogramas das razões podem ser vistos na Figura 7.12. À esquerda, estão três histogramas referentes a FS0; a cor azul indica que as razões foram calculadas a partir de movimentos do solo na direção X, a cor laranja indica movimentos na direção Y e a verde, na direção Z. É possível notar que todos os maiores picos nesse caso estão antes de $R = 0{,}5$. A direção Z tem razões um pouco maiores que as outras duas, mas mesmo assim são (em maioria) menores que 0,5. De forma análoga, na imagem central, há os histogramas para os mesmos canais quando apenas FS2 são estudados. As razões no canal vertical também apresentam valores pouco maiores mas, novamente, os picos das três direções estão dentro ou muito próximos a $R = 0{,}5$. Por fim, na imagem à direita, estão os histogramas de FS4. As direções de X e Y têm razões pequenas, mas na direção Z, as razões são relativamente maiores.

Considerando esses três casos (0 Hz, 2 Hz e 4 Hz), seria impossível dizer que as razões são capazes de diferenciar uma frequência da outra olhando apenas ITMX_X e ITMX_Y. Isso pode ser feito imaginando os resultados sem a cor verde. Em outras palavras, apenas a orientação vertical é necessária para a análise da hipótese, e os canais ITMX_X e ITMX_Y podem ser dispensados. Essa informação é muito interessante, pois diz que se a hipótese estiver correta, o FS está relacionado com movimentos verticais dos espelhos de entrada no interferômetro.

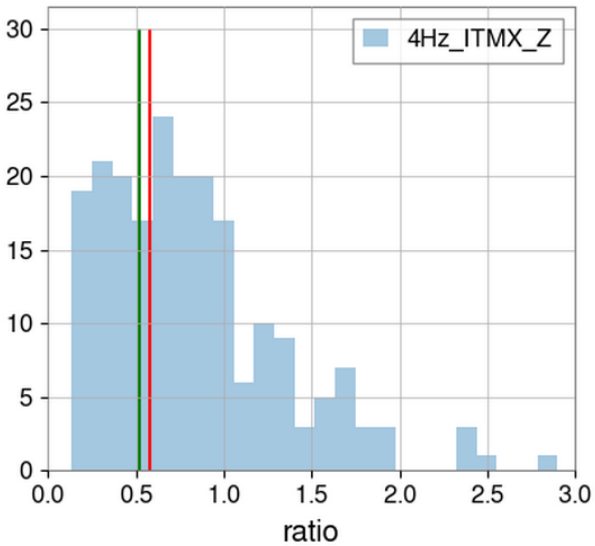
Figura 7.12 - Histogramas dos valores das razões $R = MA/MM$ para cada uma das três frequências de FS: 0 Hz, 2 Hz, 4 Hz. Para cada uma dessas frequência há três histogramas representando as direções (X, Y ou Z) nas quais os movimentos são medidos. Neste caso, são movimentos nas posições do espelho ITMX.



Fonte: Produção da autora.

Uma outra informação retirada é que, se apenas o histograma do FS de 4 Hz na direção Z for analisado, é possível encontrar dois picos principais: um com valor alto para R (como esperado pela hipótese) e outro com valor baixo e muito próximo do calculado para os FS de 2 Hz ou 0 Hz. A Figura 7.13 mostra apenas esse histograma.

Figura 7.13 - Histograma da razão entre movimentos antropogênicos e microssísmicos para FS de 4 Hz. As linhas verticais separam as regiões em altas (depois da linha vermelha) e baixas (antes da linha verde) razões.



Fonte: Produção da autora.

Esses dois picos indicam que, de fato, há FS4 com altos valores de R que estão de acordo com a hipótese; no entanto, também existe um grupo de FS de 4 Hz com razões baixas. As linhas verticais vermelha e verde na Figura 7.13 representam, respectivamente, as posições que delimitariam essas duas regiões de altas e baixas razões. Isso significa que existe um grupo de FS4 com valores de razões baixos como os de FS2. Com essa informação, uma outra hipótese surge: será que alguns FS de 4 Hz podem estar relacionados com FS de 2 Hz?

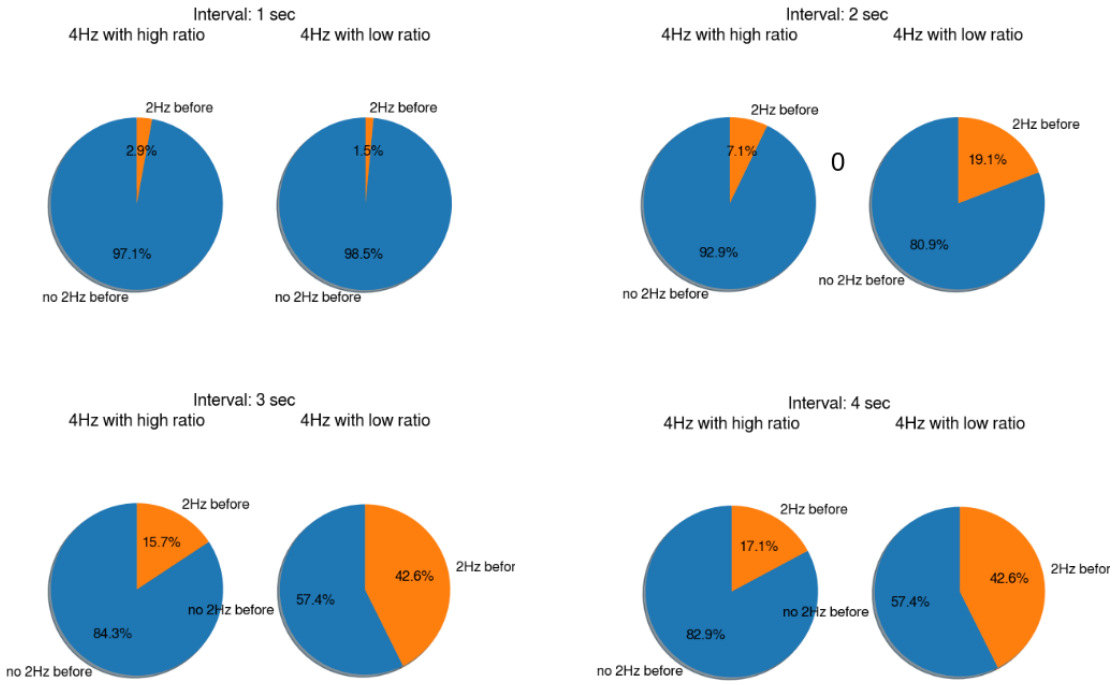
Para verificar, uma lista de FS de 4 Hz com razões baixas foi criada. Para cada glitch dessa lista, definido como glitch principal, foi analisada a presença de outros FS próximos em tempo. Todos eles tiveram a frequência determinada pelo identificador. A partir daí, foi verificado se existia algum FS de 2 Hz antes de cada FS4 da lista. Se sim, os FS4 de baixas razões R podem de fato estar relacionados com FS2.

O resultado final pode ser visto na Figura 7.14. Há quatro imagens e em cada uma delas, à direita, há a porcentagem dos FS4 que tiveram pelo menos um FS2 antes. Eles estão divididos da seguinte maneira: a primeira imagem considera FS de 2 Hz presentes em um segundo antes da identificação do de 4 Hz; a segunda imagem, da direita, busca a presença de pelo menos um FS2 em até dois segundos antes do glitch principal de 4 Hz, e assim por diante. À direita da terceira imagem, portanto, deve ser lida como: 42,6% dos FS de 4 Hz com baixa razão R tiveram pelo menos um FS de 2 Hz em até três segundos antes de suas ocorrências. A mesma análise foi feita para os FS4 com altas razões R e, para efeito de comparação, estão à esquerda de cada um dos quatro casos.

Comparando os quatro gráficos da direita, pode-se dizer que se só um segundo de dado (antes do glitch principal) é considerado, apenas 1,5% dos FS de 4 Hz têm a presença de FS2, o que é praticamente metade da porcentagem de evidência no caso de 4 Hz com altas razões. Conforme o tempo aumenta, a presença de 2 Hz vai se tornando mais significativa. Considerando tempos de 1 s, 2 s, 3 s e 4 s, as presenças de FS de 2 Hz antes de FS de 4 Hz (de baixas razões) foram de $1,5\%$, 19,1%, 42,6% e 42,6%, respectivamente. No caso de altas razões foram 2,9%, 7,1%, 15,7% e 17,1%.

A presença de FS2 antes de um FS4 aumenta significantemente no caso de baixas razões; isso é visível com o aumento da área laranja nas imagens. De três segundos para quatro, não houve diferença e, portanto, um limite em três segundos pode ser aplicado. Claro que quando há o aumento do tempo na busca de FS2 antes de FS4, a chance de coincidências é maior. Por isso, também há um aumento da presença de FS2 antes de FS4 de razões altas, mas o crescimento do aumento é menor que no caso de razões baixas. Por exemplo, para três segundos, apenas 15,7% dos FS4 de altos R têm pelo menos um FS2 antes; no caso de baixos valores de R, há 42,6%.

Figura 7.14 - Gráficos tipo pizza que indicam se há a presença de FS de 2 Hz antes de FS de 4 Hz. A busca da presença de 2 Hz foi feita para 1s, 2s, 3s e 4s antes do FS de 4 Hz. Há quatro imagens representando a análise para esses quatro intervalos de tempo. À direita de cada, há a porcentagem dos FS de 4 Hz de razões R baixas que tiveram 2 Hz antes; à esquerda, para comparação, há a porcentagem dos FS de 4 Hz de razões R altas que tiveram 2 Hz antes.



Fonte: Autoria própria.

A princípio, pode-se concluir que:

- FS de 2 Hz e de 0 Hz têm (em maior parte) razões R menor que 0,5;
- um FS que tenha razão alta R será provavelmente de 4 Hz, confirmando a hipótese sugerida inicialmente;
- um FS de 4 Hz não necessariamente aconteceu no momento em que houve uma razão R alta; foram encontrados dois grupos para FS de 4 Hz: um com alta razão R e outro com baixa;
- há mais indícios da presença de FS de 2 Hz antes de um FS de 4 Hz medidos em baixas razões R do que em altas razões;
- mesmo que um FS de 4 Hz seja uma consequência de um FS de 2 Hz, isso não

explicaria a presença dos 100% de casos de 4 Hz com razões baixas;

Esse estudo de transientes causados pelo espalhamento do laser ainda está em desenvolvimento e precisa de mais análises. Ele também faz parte de dos possíveis trabalhos futuros.

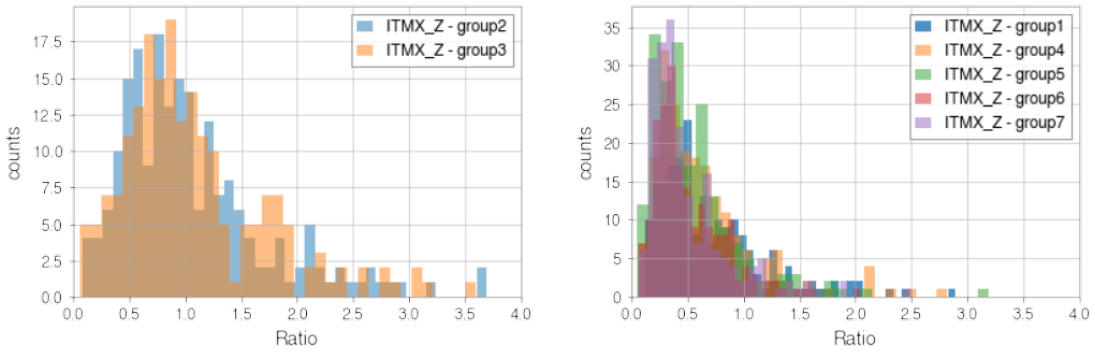
Discussões finais:

Este capítulo apresentou um estudo mais profundo dos glitches causados por espalhamento do laser. Um deles é o Scattered Light e o outro, o Fast Scattering; ambos formaram as classes mais presentes durante a O3. Os glitchgramas de cada uma delas foram criados e o t-SNE aplicado, possibilitando visualizar o comportamento mensal de cada categoria. Além disso, para verificar a eficiência do t-SNE para subclasses do Fast Scattering, um identificador foi criado, pois o Gravity Spy ainda não é capaz de diferenciar se um FS é de 4 Hz, 2 Hz ou 0 Hz. A partir dele, foi possível testar a hipótese de que o FS4 é causado quando há altos movimentos antropogênicos e baixos microssísmicos. De fato, há glitches FS4 classificados nos momentos em que os canais auxiliares mostraram altas razões (MA/MM); no entanto, existe um outro grupo de 4 Hz que tem razão baixa. Há indícios que este último possa ser explicado como uma consequência do FS de 2 Hz, mas ainda é preciso mais estudo. Além disso, são necessários mais testes e a inclusão das informações dos FS0.

Apenas por curiosidade, as razões R também foram calculadas para cada um dos sete grupos encontrados pelo t-SNE, presentes na Figura 7.7. Os histogramas dessas razões estão na Figura 7.16. Por motivos mostrados anteriormente, apenas os movimentos da direção Z estão sendo utilizados. Os histogramas foram separados em duas partes: à esquerda, estão os histogramas com maiores valores de R que foram encontrados nos grupos 2 e 3; à direita, estão os grupos 1, 4, 5, 6 e 7, cujos picos estão antes de $R = 0{,}5$.

As medianas das razões de cada grupo foram calculadas. O grupo 01 teve uma mediana igual a 0,575; o grupo 04: 0,511, o grupo 05: 0,436; o grupo 06: 0,415; o grupo 07: 0,370; o grupo 02: 0,913 e o grupo 03: 0,921. De fato, os grupos 2 e 3 foram os grupos nos quais o identificador encontrou mais FS de 4 Hz (e são os que têm maiores razões); logo, devem ser os FS4 pertencentes ao conjunto que concorda com a hipótese inicial. Também, os grupos 1, 6 e 7 foram os que tiveram mais FS de 2 Hz e 0 Hz (e são os que têm menores razões). O grupo 4 teve praticamente metade dos casos como 2 Hz e a outra como 4 Hz. Como a razão mediana do grupo 4 é baixa, é provável que nesse grupo estejam presentes os FS de 4 Hz com razões baixas. A princípio, essas separações e concordâncias com os grupos do t-SNE são apenas coincidências.

Figura 7.15 - Histogramas da razão R calculados para glitches de cada uma das sete regiões de alta densidade encontradas pelo t-SNE para a classe FS. Dois grupos tiveram picos com valores maiores que $R = 0,5$ e estão à esquerda da imagem; os outros cinco grupos tiveram picos menores e estão à direita.



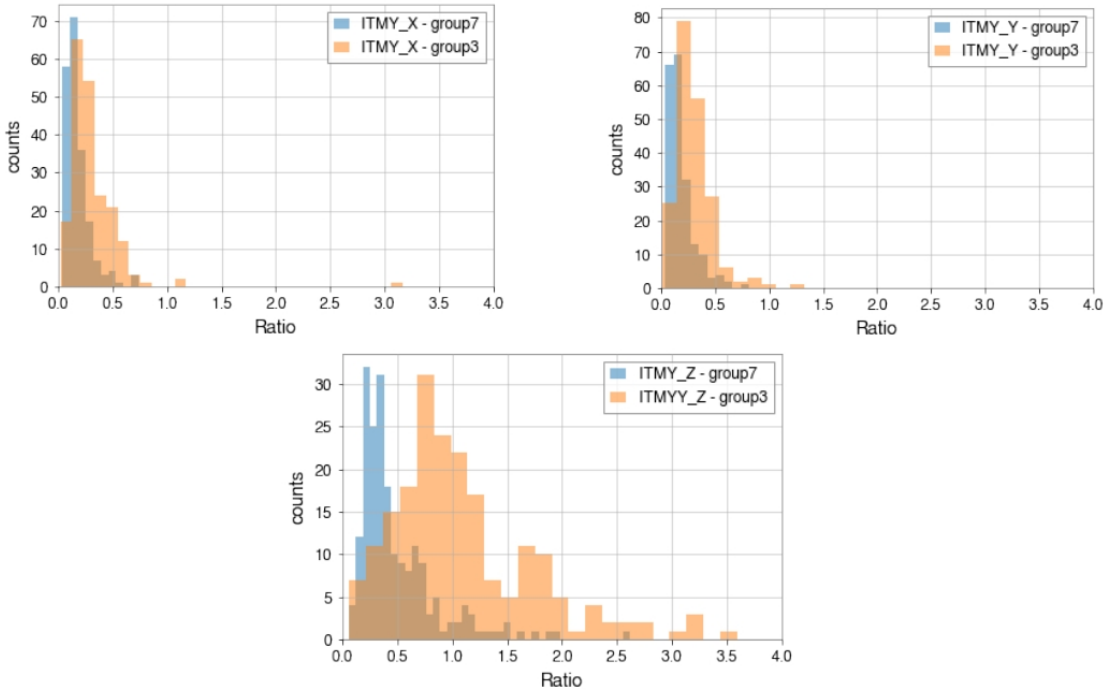
Fonte: Produção da autora.

A partir dos valores das medianas e dos histogramas da Figura 7.12, pode-se dizer que o valor de T, limite para definir se é um FS de 4 ou 2 Hz, esteja entre 0,5 e 0,6. Lembrando que esse valor é válido apenas para separar 2 Hz e 0 Hz de um dos grupos de 4 Hz. Ainda não há detalhes sobre como discernir um FS 4 Hz de baixas razões dos outros FS.

Para finalizar, a Figura 7.16 mostra os histogramas para os grupos 7 e 3 que tiveram menor e maior mediana, respectivamente. Novamente, nota-se na figura, que a diferença nos picos é predominante apenas no eixo vertical (direção Z). O grupo 3 é o que mais tem FS de 0 e 2 Hz e o grupo 7 o que mais tem FS de 4 Hz com razões altas. Essas medidas foram feitas nos espelhos ITMY_J. Os movimentos medidos nos espelhos ITMX_J e ITMY_J são muito similares ou idênticos, e não faz diferença escolher um ou outro para esse estudo.

Apesar da hipótese inicial não ter sido totalmente confirmada, foi possível encontrar indícios de que existam outro motivos para o FS de 4 Hz (com razões baixas). Além disso, verificou-se que os glitches FS4 de altas razões estão relacionados apenas com os movimentos verticais nas massas testes iniciais. Descobrir a origem específica de um glitch, principalmente de um do tipo FS, que foi o mais presente durante a O3, é fundamental. Continuidades nessa pesquisas devem ser feitas.

Figura 7.16 - Histogramas da razão R para os grupos 3 e 7 encontrados pelo t-SNE. Esses foram os grupos com maiores e menores valores de R (na direção Z), respectivamente.



Fonte: Produção da autora.

8 CONCLUSÕES

Este trabalho apresentou um estudo de caracterização de glitches presentes em um dos observatórios de ondas gravitacionais LIGO. Glitches são sinais ruidosos não-gaussianos que poluem os dados dos detectores a todo momento. Eles podem acontecer por diversos motivos, incluindo fatores ambientais, instrumentais e antropológicos; no entanto, muitos deles não têm causa específica determinada. Além de aparecerem em sinais gravitacionais, atrapalhando a reconstrução dos parâmetros da fonte, eles aumentam o background na busca de ondas e com isso, diminuem a significância de eventos candidatos. Ainda, reduzem o tamanho dos dados a serem utilizados. Por isso, identificar as causas para tentar eliminar glitches do canal gravitacional é fundamental.

Os glitches são usualmente analisados por morfologia no plano tempo-frequência. Dessa forma, é possível encontrar padrões nos dados que oferecem a possibilidade de atribuir classes a esses ruídos; as classes são nomeadas de acordo com a aparência do glitch nesse espaço de parâmetros. Atualmente, esse estudo de atribuir classes a glitches é feito pelo Gravity Spy, que é a combinação de técnicas de Machine Learning, classificações prévias de cientistas do LIGO e cidadãos voluntários que auxiliam na atribuição de classes.

Para atribuir classes, o Gravity Spy seleciona os transientes de interesse encontrados pelo algoritmo Omicron, analisa o tempo de ocorrência do sinal no canal gravitacional, gera a série temporal em torno desse tempo principal, faz a transformada Q e cria um espectrograma para representar o transiente. A partir disso, o Gravity Spy, baseado em informações prévias, aplica uma análise de imagens através de ferramentas de Deep Learning, e classifica os transientes ruidosos.

Apesar de ser um método muito efetivo, o processo é lento, e esse é um dos motivos pelo qual este trabalho foi proposto. Os estudos desta tese são construídos a partir de uma forma alternativa para desenhar a morfologia dos glitches. Ela é baseada no processo de criação de arquivos do Omicron, de onde é possível recriar a forma principal do transiente ruidoso no espaço de tempo, frequência e razão sinal-ruído. Depois de recriada, ela é transformada em uma matriz de dados, chamada glitchgrama, e então, em um vetor. Dessa forma, cada glitch tem um vetor de dados representando sua morfologia, onde cada elemento carrega o valor da razão sinal-ruído naquele pixel; neste caso, o glitchgrama foi construído por uma matriz de 30×40 elementos, o que gerou um vetor de 1200 dimensões. Com essa representação, não são realizados todos os passos para acessar a série temporal, fazer a transformada Q dos dados e criar os referentes espectrogramas.

Para este estudo foram escolhidas nove classes glitches (as que tiveram no mínimo mil aparições durante a segunda corrida observacional do LIGO, O2). Elas são: Blip, Koi Fish, Low Frequency Burst, Low Frequency Lines, Power Line, Scattered Light, Tomte e

Whistle. Com isso, nove mil vetores de mil e duzentas dimensões cada foram criados e métodos para verificar a eficiência de caracterização de um glitch a partir do glitchgrama foram aplicados. Verificar a eficiência significa confirmar se o glitchgrama é de fato capaz de diferenciar uma classe de glitch de outra.

Como os dados estão salvos em vetores, um método de análise provável de ser utilizado é algum que inclua combinações entre eles e aplicações de álgebra linear. Por isso, o método de cosseno de similaridade foi testado; ele está dentro da análise de redes que, por sua vez, busca informações de objetos e a similaridade entre eles. O método cosseno de similaridade entrou justamente na busca dessa similaridade entre os vetores (glitches). Ele calculou a similaridade através do cosseno entre eles e, a partir daí, foi possível construir uma matriz de adjacência para os glitches que permitiu visualizar os dados em forma de grafos (a partir do pacote NetworkX). Vale ressaltar que essa construção foi independente da classe do glitch, as classes não entraram como parâmetros, apenas cada elemento do vetor.

Para testar o método com os glitchgramas, foi preciso aplicar um localizador de comunidades, chamado Best Partition. Ele particionou os dados e buscou por grupos através da modularidade. Nesse caso, isso equivale a encontrar um grupo que represente cada uma das nove classes escolhidas. Se o método e o glitchgrama funcionam bem, então, nove grupos devem ser encontrados, pois glitches similares (da mesma classe) devem estar fortemente conectados entre si e fracamente ligados a outros grupos.

Após a aplicação do Best Partition nos dados, apenas quatro das nove classes foram encontradas. Elas foram chamadas de S_0, S_1, S_2 e S_3. Para comparação, o mesmo modo de visualização através dos grafos foi usado com as classificações do Gravity Spy aplicadas. Apenas uma das quatro classes encontradas (a S_0) teve uma alta equivalência com Power Line, cerca de 91,8%. Isso significa que $91,8\%$ dos glitches classificados como Power Line estavam presentes em S_0. No entanto, visualmente, foi possível observar que as outras três comunidades incluíam duas ou mais classes e, dessa forma, o mesmo algoritmo foi aplicado novamente para cada uma das três comunidades.

Com essa nova aplicação, mais dois subgrupos (S_{10} e S_{11}) tiveram concordâncias maiores que 93% para Low Frequency Lines e Low Frequency Burst, e outros dois (S_{20} e S_{30}) tiveram mais que 98,00% com Scattered Light e Whistle. Depois disso, o algoritmo teve que ser aplicado novamente. E, com a terceira aplicação, foram encontrados mais três subgrupos (S_{321}, S_{322}, e S_{210}) com equivalências de 43,00%, 51,40% e 68,10% com Blip, Koi Fish e Tomte, respectivamente; a classe de glitch menos encontrada nos dados pelo Best Partition foi a Extremely Loud. Apenas 35,40% de glitches dessa classe foram encontrados no subgrupo S_{211}. Uma das razões apresentadas para isso foi que alguns glitches classificados como Extremely Loud estão deslocados do tempo central. O cosseno de similaridade é sensível a glitches deslocados no tempo, o que é compreensível, pois o glitchgrama possui

muitos pixels com valores zero. O produto escalar entre um vetor com muitos zeros e um similar deslocado tende a ser afetado. Além disso, em um estudo posterior, há a intenção de determinar os efeitos da resolução dos glitchgramas no resultado final das classificações.

Ao final, foi encontrada uma média de 75,03% de concordância entre o método e o Gravity Spy. Pode-se dizer que ele foi excelente para algumas classes específicas e até encontrou erros de classificações do Gravity Spy, mas foi ruim para outras. Não seriam possíveis reaplicações do algoritmo se as classes do Gravity Spy não fossem previamente conhecidas. Nessa situação, o método encontraria apenas quatro das nove classes estudadas e não haveria, em princípio, condições de parada nas aplicações. Essas condições podem ser estudadas futuramente. Uma vantagem desse método é a busca na semelhança para um vetor de classe desconhecida. Se há um vetor (bem definido) representando cada classe, sua similaridade com um vetor de dados desconhecido poderá indicar a qual classe o sinal pertence. Isso não é válido para as classes mal definidas e glitches que aparentam deslocados no tempo.

O segundo método para verificar a eficiência dos glitchgramas foi com aplicações de técnicas de Machine Learning ou Aprendizado de Máquina. O Aprendizado de Máquina, como o próprio nome já diz, vem do ato de ensinar a máquina a "pensar", isto é, ensiná-la a tomar decisões a partir de um conjunto de decisões que já foram tomadas por humanos.

Se o glitchgrama caracteriza bem os glitches, então, em mil e duzentas dimensões as classes estariam agrupadas. Mas como visualizar mil e duzentas dimensões? Para isso, a primeira parte do processo foi reduzir as dimensões dos vetores para 2D (duas dimensões) de forma que a formação dos grupos presentes em 1200D ainda fossem evidentes em 2D. Isso foi feito através da ferramenta t-SNE (t-Distributed Stochastic Neighbor Embedding), que está presente no Aprendizado de Máquina não-supervisionado. Ele é um algoritmo que reduz dimensões, fazendo com que as similaridades entre pares de dados sejam as mesmas em alta e baixa dimensões. Isso significa que vetores próximos (glitches parecidos) em 1200D também estariam agrupados em 2D.

Com a aplicação, foi possível visualizar (em 2D) a formação dos nove grupos de glitches selecionados. Aqui também, o algoritmo não conhece as classificações corretas, elas foram atribuídas para verificação e, todos os grupos estavam bem posicionados. Para confirmar numericamente, uma técnica de Aprendizado Supervisionado foi aplicada, a SVM (Support Vector Machine). Nela, os dados de entrada foram separados em duas partes: 70% para treino e 30% para teste. Isso significa que 6300 glitches (com classificações) foram usados para treinar a máquina a criar um modelo preditivo; com o modelo criado, 30% dos dados foram usados para testar o quão bom o modelo classificaria os glitches. O algoritmo não sabia a classe dessa parte dos dados e atribuiu uma categoria de acordo com o modelo criado previamente. Assim, foi possível confrontar as classes atribuídas com as verdadeiras

(do Gravity Spy) para verificação.

Houve intersecções entre algumas classes, porém, em geral, o resultado foi muito bom. A menor acurácia (de acordo com a matriz de confusão gerada) foi para a classe de Extremely Loud com 85% que, coincidentemente, é a mesma classe que teve o pior resultado no primeiro método. Com a validação cruzada, encontrou-se 94,70% de equivalência entre as classificações desse método e do Gravity Spy. Esse valor, inclusive, é significantemente maior que o do método anterior. O método ainda foi capaz de encontrar erros de classificações do Gravity Spy; além disso, encontrou um subgrupo de Power Line com frequências de pico diferentes da frequência média do grupo principal. Se os erros no Gravity Spy não existissem, a acurácia poderia ter sido maior.

Mostrou-se ainda que, mesmo que o Gravity Spy não existisse, seria possível encontrar as nove classes de glitches selecionadas apenas com a aplicação do t-SNE, sem conhecimento prévio de alguma das classes. Isso é um ponto positivo, pois o t-SNE aplicado aos glitchgramas torna-se independente do Gravity Spy. Inclusive, foi apresentado o resultado de um estudo desse tipo. O t-SNE foi aplicado durante um dia aleatório de testes do LIGO e grupos de glitches foram encontrados. Em especial, dois deles com altas densidades foram analisados e um evidenciou a presença de uma nova classe de glitch no detector. Essas aplicações podem ser aplicadas durante períodos diferentes para análises estatísticas de glitches.

Sem o Gravity Spy, não seria possível aplicar o método SVM, pois ele depende do conhecimento prévio do número de classes para predições. No caso do Gravity Spy não existir, o t-SNE encontraria nove grupos, mas a morfologia de cada um teria que ser analisada para atribuição de nomes de forma similar a feita atualmente pelo LIGO. Ainda nesse caso, se grupos estivessem muito próximos, o Best Partition poderia ser utilizado para separá-los. Apesar das suposições, o Gravity Spy existe e o objetivo não é substitui-lo, mas colaborar com alertas de classificações erradas e ter uma ferramenta que possa analisar rapidamente classes presentes nos dados.

Dessa forma, conclui-se que os glitchgramas caracterizam bem os glitches. Há limitações quando ferramentas de Network Science são aplicadas a eles, mas funcionam muito bem quando acoplados com Machine Learning. Neste último caso, diferentes aplicações podem ainda ser feitas. Por exemplo, foi apresentado como o método poderia ser utilizado para buscar glitches em canais auxiliares, sensores que monitoram os observatórios. A presença de classes em determinados canais podem oferecer indícios de suas causas.

Como o uso dos glitchgramas no caso de classes diferentes foi promissor, houve também a aplicação para a análise de apenas duas classes muito comuns durante a terceira corrida observacional do LIGO: a Fast Scattering e a Scattered Light. No resultado, o t-SNE sepa-

rou bem as duas, e foi possível visualizar o comportamento mensal de ambas. O algoritmo criou mini grupos do Scattered Light que tinham diferenças nos arcos na representação morfológica. Dentre essas diferenças, estavam duração do glitch, quantidade de arcos e razão sinal-ruído. Também foi possível visualizar pequenos grupos de altas densidades para o Fast Scattering.

A classe Fast Scattering tem subclasses. Elas acontecem na mesma frequência de pico, porém se repetem de formas diferentes. O Gravity Spy ainda não é capaz de diferenciar essas formas de repetições e, por isso, um algoritmo identificador de frequências foi criado para verificar se o t-SNE teria separado bem essas subcategorias ou não. O identificador encontrou três subclasses: a que a mancha se repete a cada 0,25 segundos (Fast Scattering de 4 Hz), a que se repete a cada 0,5 segundos (Fast Scattering de 2 Hz) e a que não se repete (denominado Fast Scattering de 0 Hz).

Apesar do t-SNE encontrar grupos com alta quantidade de Fast Scattering de 2 Hz e 0 Hz e grupos com alta quantidade de Fast Scattering de 4 Hz, também existiram grupos da mistura de todas as subclasses. A princípio, pode-se dizer que o t-SNE é bom para classes bem diferentes; modificações devem ser feitas para estudos de subclasses de glitch. Essas modificações são muito interessantes para projetos futuros, pois a identificação de subclasses é fundamental para conhecimento da classe principal.

Com o algoritmo criado para identificação da frequência do Fast Scattering, foi possível verificar a hipótese de que as frequências 2 Hz e 4 Hz são definidas pela força relativa entre movimentos antropogênicos, MA, e microssísmicos, MM. A sugestão é que os glitches Fast Scattering de 4 Hz sejam causados quando há altos MA e baixos MM. Nesse caso, a razão $\mathrm{R} = \mathrm{MA}/\mathrm{MM}$ seria alta; se a razão fosse baixa, haveria a presença de um Fast Scattering de 2 Hz no detector. Foram encontrados glitches que concordam com essa hipótese, porém, também foi encontrado um outro grupo de 4 Hz em que essa afirmação não é válida. Nesse grupo, as razões dos Fast Scattering de 4 Hz são baixas. Há indícios de que eles possam ser consequências do de 2 Hz, mas mais estudos são necessários. Com os teste aplicados, estimou-se que T, valor que define se uma razão é alta ou baixa, está entre 0,5 e 0,6.

Este trabalho e as aplicações de técnicas computacionais aos glitchgramas podem ser continuados e aprimorados de diferentes formas. Identificar, conhecer e buscar fontes de ruídos transientes são de extrema importância para a limpeza dos dados; com as ideias e ferramentas apresentadas, esta tese espera poder colaborar e incentivar pesquisas para a astronomia de ondas gravitacionais.

REFERÊNCIAS BIBLIOGRÁFICAS

AASI, J. et al. Advanced ligo. **Classical and Quantum Gravity**, v. 32, n. 7, p. 074001, 2015. 3, 19, 20, 22

ABBOTT, B. et al. Ligo: the laser interferometer gravitational-wave observatory. **Reports on Progress in Physics**, v. 72, n. 7, p. 076901, 2009. 13, 14, 20

_____. Gw150914: First results from the search for binary black hole coalescence with advanced ligo. **Physical Review D**, v. 93, n. 12, p. 122003, 2016. 1, 15, 24

_____. Gwtc-1: a gravitational-wave transient catalog of compact binary mergers observed by ligo and virgo during the first and second observing runs. **Physical Review X**, v. 9, n. 3, p. 031040, 2019. 2, 15

ABBOTT, B. P. et al. Gw151226: observation of gravitational waves from a 22-solar-mass binary black hole coalescence. **Physical Review Letters**, v. 116, n. 24, p. 241103, 2016. 30

_____. Observation of gravitational waves from a binary black hole merger. **Physical Review Letters**, v. 116, n. 6, p. 061102, 2016. 1

_____. Gw170814: a three-detector observation of gravitational waves from a binary black hole coalescence. **Physical Review Letters**, v. 119, n. 14, p. 141101, 2017. 12, 15

_____. Gw170817: observation of gravitational waves from a binary neutron star inspiral. **Physical Review Letters**, v. 119, n. 16, p. 161101, 2017. 2, 15, 31

_____. Multi-messenger observations of a binary neutron star merger. **Astrophysical Journal Letters.**, v. 848, n. 2, p. L12, 2017. 15

_____. Effects of data quality vetoes on a search for compact binary coalescences in advanced ligo's first observing run. **Classical and Quantum Gravity**, v. 35, n. 6, p. 065010, 2018. 30

ABBOTT, R. et al. Gw190521: a binary black hole merger with a total mass of 150 m. **Physical Review Letters**, v. 125, n. 10, p. 101102, 2020. 15

_____. Gw190814: gravitational waves from the coalescence of a 23 solar mass black hole with a 2.6 solar mass compact object. **The Astrophysical Journal Letters**, v. 896, n. 2, p. L44, 2020. 16

_____. Gwtc-2: compact binary coalescences observed by ligo and virgo during the first half of the third observing run. **Physical Review X**, v. 11, n. 2, p. 021053, 2021. 2, 15

_____. Gwtc-2.1: deep extended catalog of compact binary coalescences observed by ligo and virgo during the first half of the third observing run. **arXiv preprint arXiv:2108.01045**, 2021. 2, 15

_____. Gwtc-3: Compact binary coalescences observed by ligo and virgo during the second part of the third observing run. **arXiv preprint arXiv:2111.03606**, 2021. 2, 15

ABRAMOVICI, A. et al. Ligo: the laser interferometer gravitational-wave observatory. **Science**, v. 256, n. 5055, p. 325–333, 1992. 19

ACCADIA, T. et al. Virgo: a laser interferometer to detect gravitational waves. **Journal of Instrumentation**, v. 7, n. 03, p. P03012, 2012. 15

ACERNESE, F. a. et al. Advanced virgo: a second-generation interferometric gravitational wave detector. **Classical and Quantum Gravity**, v. 32, n. 2, p. 024001, 2014. 17

AGUIAR, O. D. Past, present and future of the resonant-mass gravitational wave detectors. **Research in Astronomy and Astrophysics**, v. 11, n. 1, p. 1, 2011. 13

AGUIAR, O. D. et al. The brazilian gravitational wave detector mario schenberg: progress and plans. **Classical and Quantum Gravity**, v. 22, n. 10, p. S209, 2005. 13

ALBERT-LáSZLó BARABáSI. **Network science**. 2021. Disponível em: <https:http://networksciencebook.com/chapter/9#hierarchical>. Acesso em: 03 out. 2022. 55

ALPAYDIN, E. **Introduction to machine learning**. [S.l.]: MIT Press, 2009. 73

ASTONE, P. et al. Search for coincident excitation of the widely spaced resonant gravitational wave detectors explorer, nautilus and niobe. **Astroparticle Physics**, v. 10, n. 1, p. 83–92, 1999. 13

AUFMUTH, P.; DANZMANN, K. Gravitational wave detectors. **New Journal of Physics**, v. 7, n. 1, p. 202, 2005. 13

AYNAUD, T. **Community detection for NetworkX's documentation**. [S.l.]: Tech. Rep., 2018. Disponível em: https://media. readthedocs. org/pdf . . . , 2018. 54

BAHAADINI, S.; NOROOZI, V.; ROHANI, N.; COUGHLIN, S.; ZEVIN, M.; SMITH, J. R.; KALOGERA, V.; KATSAGGELOS, A. Machine learning for gravity spy: glitch classification and dataset. **Information Sciences**, v. 444, p. 172–186, 2018. 4

BASSAN, M. Advanced interferometers and the search for gravitational waves. **Astrophysics and Space Science Library**, v. 404, p. 275–290, 2014. 21

BISWAS, R. et al. Application of machine learning algorithms to the study of noise artifacts in gravitational-wave data. **Physical Review D**, v. 88, n. 6, p. 062003, 2013. 4

BLONDEL, V. D.; GUILLAUME, J.-L.; LAMBIOTTE, R.; LEFEBVRE, E. Fast unfolding of communities in large networks. **Journal of Statistical Mechanics: Theory and Experiment**, v. 2008, n. 10, p. P10008, 2008. 54

BROEKGAARDEN, F. S.; BERGER, E. Formation of the first two black hole–neutron star mergers (gw200115 and gw200105) from isolated binary evolution. **The Astrophysical Journal Letters**, v. 920, n. 1, p. L13, 2021. 16

BUONANNO, A. Gravitational waves. **arXiv preprint arXiv:0709.4682**, 2007. 9

CABERO, M. et al. Blip glitches in advanced ligo data. **Classical and Quantum Gravity**, v. 36, n. 15, p. 155010, 2019. 39

CAVAGLIA, M.; STAATS, K.; GILL, T. Finding the origin of noise transients in ligo data with machine learning. **arXiv preprint arXiv:1812.05225**, 2018. 4

CHATTERJI, S.; BLACKBURN, L.; MARTIN, G.; KATSAVOUNIDIS, E. Multiresolution techniques for the detection of gravitational-wave bursts. **Classical and Quantum Gravity**, v. 21, n. 20, p. S1809, 2004. 32

CHRISTENSEN, N. Stochastic gravitational wave backgrounds. **Reports on Progress in Physics**, v. 82, n. 1, p. 016903, 2018. 11

CLASSICAL AND QUANTUM GRAVITY. **How do we know LIGO detected gravitational waves?** 2016. Disponível em: <https://cqgplus.com/2016/06/06/how-do-we-know-ligo-detected-gravitational-waves>. Acesso em: 23 out. 2021. 33

COLLABORATION, L. S. et al. Observation of gravitational waves from two neutron star–black hole coalescences. **Astrophysical Journal Letters**, v. 915, n. 1, 2021. 16

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, n. 3, p. 273–297, 1995. 75

COUGHLIN, S. et al. Classifying the unknown: discovering novel gravitational-wave detector glitches using similarity learning. **Physical Review D**, v. 99, n. 8, p. 082002, 2019. 4

CREIGHTON, T. Tumbleweeds and airborne gravitational noise sources for ligo. **Classical and Quantum Gravity**, v. 25, n. 12, p. 125011, 2008. 24

CUOCO, E. et al. Enhancing gravitational-wave science with machine learning. **Machine Learning: Science and Technology**, v. 2, n. 1, p. 011002, 2020. 4

DAVIS, D. et al. Ligo detector characterization in the second and third observing runs. **Classical and Quantum Gravity**, v. 38, n. 13, p. 135014, 2021. 25

DETCHAR. **Detchar Wiki**. 2021. Disponível em: <https://wiki.ligo.org/DetChar/WebHome>. Acesso em: 23 out. 2021. 3

D'INVERNO, R. **Introducing Einstein's relativity**. USA: Oxford University Press, 1992. 8

EINSTEIN, A. Näherungsweise integration der feldgleichungen der gravitation. **Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften**, p. 688–696, 1916. 1

_____. Über gravitationswellen. **Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften**, p. 154–167, 1918. 1

FACELI, K. et al. **Inteligência Artificial: Uma abordagem de aprendizado de máquina**. [S.l.: s.n.], 2011. 72

FERREIRA, T. A. **Análise multidimensional de transientes nos detectores LIGO**. Dissertação (Mestrado em Astrofísica) — Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 2018. 35

FERREIRA, T. A.; COSTA, C. A. Comparison between t-sne and cosine similarity for ligo glitches analysis. **Classical and Quantum Gravity**, v. 39, n. 16, p. 165013, 2022. 83

FLANAGAN, E. E.; HUGHES, S. A. The basics of gravitational wave theory. **New Journal of Physics**, v. 7, n. 1, p. 204, 2005. 8

FORD, K. W.; WHEELER, J. A. **Geons, black holes, and quantum foam: A life in physics**. [S.l.: s.n.], 1999. 7

FRUCHTERMAN, T. M.; REINGOLD, E. M. Graph drawing by force-directed placement. **Software: Practice and experience**, v. 21, n. 11, p. 1129–1164, 1991. 52

GABOR, D. Theory of communication. part 1: the analysis of information. **Journal of the Institution of Electrical Engineers-part III: Radio and Communication Engineering**, v. 93, n. 26, p. 429–441, 1946. 26

GALAUDAGE, S. **Cosmic lighthouses and continuous gravitational waves**. 2022. Disponível em: <https://www.zooniverse.org/>. Acesso em: ago. de 2022. 11

GÉRON, A. **Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems**. [S.l.]: O'Reilly Media, 2019. 64, 75

GLANZER, J. et al. Data quality up to the third observing run of advanced ligo: Gravity spy glitch classifications. **arXiv preprint arXiv:2208.12849**, 2022. 40

GONZÁLEZ, G. Suspensions thermal noise in the ligo gravitational wave detector. **Classical and Quantum Gravity**, v. 17, n. 21, p. 4409, 2000. 22

GRAVITATIONAL Wave Detectors and Sources. 2022. Disponível em: <http://gwplotter.com/>. Acesso em: ago. 2022. 18

GRAVITYSPY. **Help scientists at LIGO search for gravitational waves, the elusive ripples of spacetime**. 2022. Disponível em: <https://www.zooniverse.org/projects/zooniverse/gravity-spy/>. Acesso em: 2022. 4, 34, 41, 43

GWOPENSCIENCE. **L1 lines cleaning file for O1**. 2017. Disponível em: <https://www.gw-openscience.org/static/speclines/o1/O1LinesToBeCleaned_L1_v3.txt>. Acesso em: set. 2022. 25

_____. **O3 instrumental lines**. 2020. Disponível em: <https://www.gw-openscience.org/O3/o3aspeclines/>. Acesso em: set. 2022. 25

GWOSC. **GW open data workshops**. 2022. Disponível em: <https://www.gw-openscience.org/workshops/>. Acesso em: 2022. 33

GWPY. **What is GWpy?** 2022. Disponível em: <https://gwpy.github.io/docs/stable/overview/>. Acesso em: 2022. 33, 84

HARRY, G. M. et al. Advanced ligo: the next generation of gravitational wave detectors. **Classical and Quantum Gravity**, v. 27, n. 8, p. 084006, 2010. 22

HINTON, G. E.; ROWEIS, S. Stochastic neighbor embedding. **Advances in Neural Information Processing Systems**, v. 15, 2002. 67

HOREWICZ, M. C.; NASCIMENTO JUNIOR, C. L.; PERRELLA, W. J. Reconhecimento automático de modulação digital de sinais de comunicações. In: SIMPÓSIO DE APLICAÇÕES OPERACIONAIS EM ÁREAS DE DEFESA. **Anais...** [S.l.]: ITA, 2007. 76

HUGHES, S. A.; THORNE, K. S. Seismic gravity-gradient noise in interferometric gravitational-wave detectors. **Physical Review D**, v. 58, n. 12, p. 122002, 1998. 24

HULSE, R. A.; TAYLOR, J. H. Discovery of a pulsar in a binary system. **The Astrophysical Journal**, v. 195, p. L51–L53, 1975. 1

JADHAV, S.; MUKUND, N.; GADRE, B.; MITRA, S.; ABRAHAM, S. Improving significance of binary black hole mergers in advanced ligo data using deep learning: confirmation of gw151216. **Physical Review D**, v. 104, n. 6, p. 064051, 2021. 4

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning**. [S.l.: s.n.], 2013. 76

JARANOWSKI, P.; KRÓLAK, A. **Analysis of gravitational-wave data**. [S.l.]: Cambridge University Press, 2009. 7

KAGRA: 2.5 generation interferometric gravitational wave detector. **Nature Astronomy**, v. 3, n. 1, p. 35–40, 2019. 17

KENYON, I. R. **General relativity**. USA: Oxford University Press, 1990. 1

KOBOUROV, S. G. Spring embedders and force directed graph drawing algorithms. **arXiv preprint arXiv:1201.3011**, 2012. 52

LATEST Update on Start of Next Observing Run (O4). 2022. Disponível em: <https://www.ligo.caltech.edu/news/ligo20220617>. Acesso em: ago. 2022. 16, 17

LIGO. **LIGO DV**. 2021. Disponível em: <https://ldvw.ligo.caltech.edu/ldvw/view>. Acesso em: set. 2022. 34, 36

_____. **Introduction to LIGO gravitational waves**. 2022. Disponível em: <https://www.ligo.org/science/GW-Inspiral.php>. Acesso em: mar. 2022. 12

LIGO SCIENTIFIC COLLABORATION. **LIGO**. 2022. Disponível em: <https://www.ligo.org/>. Acesso em: mar. 2022. 3

LIGODCC. **Observing scenario timeline graphic**. 2022. Disponível em: <https://dcc.ligo.org/LIGO-G2002127/public>. Acesso em: ago. 2022. 17

MAATEN, L. v. d.; HINTON, G. Visualizing data using t-sne. **Journal of Machine Learning Research**, v. 9, n. Nov, p. 2579–2605, 2008. 67

MACLEOD, D. M.; FAIRHURST, S.; HUGHEY, B.; LUNDGREN, A. P.; PEKOWSKY, L.; ROLLINS, J.; SMITH, J. R. Reducing the effect of seismic noise in ligo searches by targeted veto generation. **Classical and Quantum Gravity**, v. 29, n. 5, p. 055006, 2012. 23

MAGGIORE, M. **Gravitational waves: volume 1: theory and experiments**. [S.l.]: OUP Oxford, 2007. 7

_____. **Gravitational waves: Volume 1: Theory and experiments**. [S.l.]: Oxford university press, 2008. 7, 21, 23

MAUCELI, E.; GENG, Z.; HAMILTON, W.; JOHNSON, W.; MERKOWITZ, S.; MORSE, A.; PRICE, B.; SOLOMONSON, N. The allegro gravitational wave detector: Data acquisition and analysis. **Physical Review D**, v. 54, n. 2, p. 1264, 1996. 13

MEDIUM. 2019. Disponível em: <https://medium.com/deep-math-machine-learning-ai/different-types-of-machine-learning-and-their-types-34760b9128a2>. Acesso em: nov. 2019. 64

METCALF, L.; CASEY, W. Metrics, similarity, and sets. In: _____ **(Ed.). Cybersecurity and applied mathematics**. [S.I.]: Elsevier, 2016. p. 3–22. Disponível em: <https://doi.org/10.1016/b978-0-12-804452-0.00002-6>. 51

MUKUND, N.; ABRAHAM, S.; KANDHASAMY, S.; MITRA, S.; PHILIP, N. S. Transient classification in ligo data using difference boosting neural network. **Physical Review D**, v. 95, n. 10, p. 104059, 2017. 4

NANOGRAV. **Gravitational wave detectors and sources**. 2022. Disponível em: <http://nanograv.org/>. Acesso em: ago. 2022. 18

NETWORKX. **Network analysis in Python**. 2014–2022. Disponível em: <https://networkx.org/>. Acesso em: set. 2022. 52

NEWMAN, M. **Networks**. [S.l.]: Oxford University Press, 2018. 49

NEWTON, I. **The Principia: mathematical principles of natural philosophy**. [S.l.]: University of California Press, 1999. 1

NOBEL. 2017. Disponível em: <https://www.nobelprize.org/prizes/physics/2017/summary/>. Acesso em: mar. 2022. 2, 15

NUTTALL, L. Characterizing transient noise in the ligo detectors. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, v. 376, n. 2120, p. 20170286, 2018. 29

NUTTALL, L. et al. Improving the data quality of advanced ligo based on early engineering run results. **Classical and Quantum Gravity**, v. 32, n. 24, p. 245005, 2015. 45

NVIDIA. **SuperVize Me: What's the difference between supervised, unsupervised, semi-supervised and reinforcement learning?** 2018. Disponível em: <https://blogs.nvidia.com/blog/2018/08/02/supervised-unsupervised-learning/>. Acesso em: abr. 2019. 65, 72

OLIVEIRA, N. F.; AGUIAR, O. D. The Mario Schenberg gravitational wave antenna. **Brazilian Journal of Physics**, v. 46, n. 5, p. 596–603, 2016. 13

OSA. **LIGO: Finally poised to catch elusive gravitational waves?** 2019. Disponível em: <https://www.osa-opn.org/home/articles/volume_26/march_2015/features/ligo_finally_poised_to_catch_elusive_gravitational/>. Acesso em: maio 2019. 19

PEDREGOSA, F. et al. Scikit-learn: machine learning in python. **the Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. 68

PLATT, E. L. **Network science with python and networkX quick start guide: explore and visualize network data effectively**. [S.l.: s.n.], 2019. 50, 52

POWELL, J.; TORRES-FORNÉ, A.; LYNCH, R.; TRIFIRÒ, D.; CUOCO, E.; CAVAGLIÀ, M.; HENG, I. S.; FONT, J. A. Classification methods for noise transients in advanced gravitational-wave detectors ii: performance tests on advanced ligo data. **Classical and Quantum Gravity**, v. 34, n. 3, p. 034002, 2017. 4

RAMIREZ, J. **Optical cavity inference techniques for low noise interferometry**. 2019. Disponível em: <https://dcc.ligo.org/public/0161/P1900188/002/JorgeRamirez_Interferometer_Inference_Poster.pdf>. Acesso em: 2022. 19

REITZE, D.; SAULSON, P. R.; GROTE, H. **Advanced interferometric gravitational-wave detectors**. [S.l.]: World Scientific, 2019. 3

RILES, K. Gravitational waves: sources, detectors and searches. **Progress in Particle and Nuclear Physics**, v. 68, p. 1–54, 2013. 19

ROBINET, F.; ARNAUD, N.; LEROY, N.; LUNDGREN, A.; MACLEOD, D.; MCIVER, J. Omicron: a tool to characterize transient noise in gravitational-wave detectors. **arXiv preprint arXiv:2007.11374**, 2020. 26

RUSSELL, S. J.; NORVIG, P. **Artificial intelligence: a modern approach**. [S.l.]: Malaysia; Pearson Education,, 2016. 73, 74

SAMUEL, A. L. Machine learning. **The Technology Review**, v. 62, n. 1, p. 42–45, 1959. 63

SATHYAPRAKASH, B. S.; SCHUTZ, B. F. Physics, astrophysics and cosmology with gravitational waves. **Living Reviews in Relativity**, v. 12, n. 1, p. 1–141, 2009. 10

SAULSON, P. R. **Fundamentals of interferometric gravitational wave detectors**. USA: Syracuse University Press, 2017. 21, 22, 23

SCHUTZ, B. F.; RICCI, F. Gravitational waves, sources, and detectors. **arXiv preprint arXiv:1005.4735**, 2010. 11

SENGUPTA, A. **The sensitivity of the advanced LIGO detectors at the beginning of gravitational wave Astronomy**. [S.l.]: Cornell University Library, 2016. 44, 45

SHANNON, C. E. A mathematical theory of communication. **The Bell System Technical Journal**, v. 27, n. 3, p. 379–423, 1948. 69

SHAO, Y. On the neutron star/black hole mass gap and black hole searches. **Research in Astronomy and Astrophysics**, 2022. 16

SMITH, J. R.; ABBOTT, T.; HIROSE, E.; LEROY, N.; MACLEOD, D.; MCIVER, J.; SAULSON, P.; SHAWHAN, P. A hierarchical method for vetoing noise transients in gravitational-wave detectors. **Classical and Quantum Gravity**, v. 28, n. 23, p. 235005, 2011. 29

SONI, S. **Identification and reduction of scattered light noise in LIGO**. Tese (Doutorado) — Louisiana State University and Agricultural & Mechanical College, 2021. 27

SONI, S. **Fast scattering corner station coupling at LLO**. 2022. Disponível em: <https://dcc.ligo.org/DocDB/0179/G2102369/003/Fast%20Scattering%20Corner%20Station%20Coupling%20at%20LLO%20V2.pdf>. Acesso em: set. 2022. 98

SONI, S. et al. Reducing scattered light in ligo's third observing run. **Classical and Quantum Gravity**, v. 38, n. 2, p. 025016, 2020. 93

_____. Discovering features in gravitational-wave data through detector characterization, citizen science and machine learning. **Classical and Quantum Gravity**, v. 38, n. 19, p. 195016, 2021. 38

STATQUEST. **t-SNE, Clearly Explained**. 2017. Disponível em: <https://www.youtube.com/watch?v=NEaUSP4YerM>. Acesso em: dez. 2022. 66

TAYLOR, J. H.; WEISBERG, J. M. A new test of general relativity-gravitational radiation and the binary pulsar psr 1913+ 16. **The Astrophysical Journal**, v. 253, p. 908–920, 1982. 1

THORNE, K. S.; WINSTEIN, C. J. Human gravity-gradient noise in interferometric gravitational-wave detectors. **Physical Review D**, v. 60, n. 8, p. 082001, 1999. 24

USHIZIMA, D. M.; LORENA, A. C.; CARVALHO, A. D. Support vector machines applied to white blood cell recognition. In: INTERNATIONAL CONFERENCE ON HYBRID INTELLIGENT SYSTEMS, 5. **Proceeding...** [S.l.], 2005. 76

VIRGO. **Virgo**. 2019. Disponível em: <http://www.virgo-gw.eu/>. Acesso em: mar. 2022. 2

WEBER, J. Detection and generation of gravitational waves. **Physical Review**, v. 117, n. 1, p. 306, 1960. 1, 12

WEISBERG, J. M.; NICE, D. J.; TAYLOR, J. H. Timing measurements of the relativistic binary pulsar psr b1913+ 16. **The Astrophysical Journal**, v. 722, n. 2, p. 1030, 2010. 1, 2

YAKUNIN, K. N. et al. Gravitational waves from core collapse supernovae. **Classical and Quantum Gravity**, v. 27, n. 19, p. 194005, 2010. 11

ZEVIN, M. et al. Gravity spy: integrating advanced ligo detector characterization, machine learning, and citizen science. **Classical and Quantum Gravity**, v. 34, n. 6, p. 064003, 2017. 33

ZOONIVERSE. **People-powered research**. 2022. Disponível em: <https://www.zooniverse.org/>. Acesso em: 2022. 4, 32

PUBLICAÇÕES TÉCNICO-CIENTÍFICAS EDITADAS PELO INPE

Teses e Dissertações (TDI)

Teses e Dissertações apresentadas nos Cursos de Pós-Graduação do INPE.

Manuais Técnicos (MAN)

São publicações de caráter técnico que incluem normas, procedimentos, instruções e orientações.

Notas Técnico-Científicas (NTC)

Incluem resultados preliminares de pesquisa, descrição de equipamentos, descrição e ou documentação de programas de computador, descrição de sistemas e experimentos, apresentação de testes, dados, atlas, e documentação de projetos de engenharia.

Relatórios de Pesquisa (RPQ)

Reportam resultados ou progressos de pesquisas tanto de natureza técnica quanto científica, cujo nível seja compatível com o de uma publicação em periódico nacional ou internacional.

Propostas e Relatórios de Projetos (PRP)

São propostas de projetos técnico-científicos e relatórios de acompanhamento de projetos, atividades e convênios.

Publicações Didáticas (PUD)

Incluem apostilas, notas de aula e manuais didáticos.

Publicações Seriadas

São os seriados técnico-científicos: boletins, periódicos, anuários e anais de eventos (simpósios e congressos). Constam destas publicações o Internacional Standard Serial Number (ISSN), que é um código único e definitivo para identificação de títulos de seriados.

Programas de Computador (PDC)

São a seqüência de instruções ou códigos, expressos em uma linguagem de programação compilada ou interpretada, a ser executada por um computador para alcançar um determinado objetivo. Aceitam-se tanto programas fonte quanto os executáveis.

Pré-publicações (PRE)

Todos os artigos publicados em periódicos, anais e como capítulos de livros.