UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
PRÓ-REITORIA DE GRADUAÇÃO
DEPARTAMENTO DE ENGENHARIAS E TECNOLOGIA
CURSO BACHARELADO INTERDISCIPLINAR EM TECNOLOGIA DA INFORMAÇÃO

MARIA ISABELLY LIMA DE SOUSA

DETECÇÃO DE EXOPLANETAS POR TRÂNSITO PLANETÁRIO: FUNDAMENTOS FÍSICOS E DESAFIOS COMPUTACIONAIS NA ERA DO BIG DATA

PAU DOS FERROS
2026

MARIA ISABELLY LIMA DE SOUSA

DETECÇÃO DE EXOPLANETAS POR TRÂNSITO PLANETÁRIO: FUNDAMENTOS FÍSICOS E DESAFIOS COMPUTACIONAIS NA ERA DO BIG DATA

Monografia apresentada à Universidade Federal Rural do Semi-Árido como requisito para obtenção do título Bacharela em Interdisciplinar em Tecnologia da Informação .

Orientador: Francisco Ernandes Matos Costa, Prof. Dr.

PAU DOS FERROS

2026

© Todos os direitos estão reservados a Universidade Federal Rural do Semi-Árido. O conteúdo desta obra é de inteira responsabilidade do (a) autor (a), sendo o mesmo, passível de sanções administrativas ou penais, caso sejam infringidas as leis que regulamentam a Propriedade Intelectual, respectivamente, Patentes: Lei n° 9.279/1996 e Direitos Autorais: Lei n° 9.610/1998. O conteúdo desta obra tomar-se-á de domínio público após a data de defesa e homologação da sua respectiva ata. A mesma poderá servir de base literária para novas pesquisas, desde que a obra e seu (a) respectivo (a) autor (a) sejam devidamente citados e mencionados os seus créditos bibliográficos.

S725d Sousa, Maria Isabelly.
Detecção de exoplanetas por trânsito planetário: fundamentos físicos e fesafios computacionais na era do big data / Maria Isabelly Sousa. - 2026.
60 f. : il.

Orientador: Francisco Ernandes Matos Costa.
Monografia (graduação) - Universidade Federal Rural do Semi-árido, Curso de Tecnologia da Informação, 2026.

1. Exoplanetas. 2. Método de trânsito . 3. Transit Least Squares. 4. Big Data. 5. Detecção automatizada. I. Matos Costa, Francisco Ernandes, orient. II. Título.

Ficha catalográfica elaborada por sistema gerador automático em conformidade com AACR2 e os dados fornecidos pelo) autor(a).
Biblioteca Campus Pau dos Ferros
Bibliotecária: Erlanda Maria Lopes da Silva
CRB: 15/966

O serviço de Geração Automática de Ficha Catalográfica para Trabalhos de Conclusão de Curso (TCC´s) foi desenvolvido pelo Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (USP) e gentilmente cedido para o Sistema de Bibliotecas da Universidade Federal Rural do Semi-Árido (SISBI-UFERSA), sendo customizado pela Superintendência de Tecnologia da Informação e Comunicação (SUTIC) sob orientação dos bibliotecários da instituição para ser adaptado às necessidades dos alunos dos Cursos de Graduação e Programas de Pós-Graduação da Universidade.

MARIA ISABELLY LIMA DE SOUSA

DETECÇÃO DE EXOPLANETAS POR TRÂNSITO PLANETÁRIO: FUNDAMENTOS FÍSICOS E DESAFIOS COMPUTACIONAIS NA ERA DO BIG DATA

Monografia apresentada à Universidade Federal Rural do Semi-Árido como requisito para obtenção do título Bacharela em Interdisciplinar em Tecnologia da Informação .

Defendida em: 08 / 06 / 2026

BANCA EXAMINADORA

Francisco Ernandes Matos Costa, Prof. Dr. (UFERSA)
Presidente

Rosana Cibely Batista Rego, Prof. Dr. (UFERSA)
Membro Examinador

Kennedy Rurison Lopes, Prof. Dr. (UFERSA)
Membro Examinador

DEDICATÓRIA

À minha versão de 14 anos, que sonhava em estudar astronomia *(presente)*.

AGRADECIMENTOS

Agradeço, primeiramente, a Deus, pela oportunidade de estudar na UFERSA e por me conceder força, sabedoria e perseverança para concluir esta jornada.

Ao meu orientador, Ernandes, agradeço pela oportunidade de ser orientada, pela confiança e pelas contribuições essenciais para a realização deste trabalho.

Aos meus familiares, minha eterna gratidão: à minha mainha Nitinha, à minha mãe Aparecida, ao meu pai e avô Eneas, que, mesmo sem ter tido a oportunidade de concluir seus estudos, criou seus sete filhos com muito esforço, dedicação e amor. Foi através do seu suor, junto à minha mãe Nitinha, que tive a oportunidade de chegar até aqui e conquistar o sonho de estudar em uma universidade.

Às minhas tias Nalva, Gracinha e Maria, ao meu irmão Pablo e a todos os demais familiares que nunca mediram esforços para me apoiar. Obrigada por acreditarem em mim, por estarem ao meu lado e por tornarem possível a realização deste sonho. Se hoje chego até aqui, é graças a cada um de vocês.

Aos amigos que a UFERSA me presenteou, minha gratidão por terem caminhado comigo durante essa trajetória. Stheffany, Andrey, Aparecida, minha parceira nessa jornada, e tantos outros fizeram parte de momentos que levarei para sempre comigo. Vocês foram essenciais, especialmente nos períodos de maior dificuldade, transformando dias de estresse em boas risadas, aprendizados e memórias inesquecíveis.

Em especial, agradeço a Kayque, por sua gentileza, dedicação e apoio ao longo dessa caminhada. Sua presença e incentivo foram muito importantes durante essa etapa.

A todos que fizeram parte desta jornada e deixaram sua marca na minha história, o meu mais sincero obrigada.

EPÍGRAFE

“Às estrelas que ouvem e aos sonhos que são atendidos.”

(Sarah J. Maas)

RESUMO

A detecção de exoplanetas pelo método de trânsito planetário consolidou-se como a principal técnica de descoberta na exoplanetologia contemporânea, impulsionada por missões espaciais como *Kepler* e *TESS*, que geram volumes massivos de dados astronômicos. Este trabalho tem como objetivo analisar o método de trânsito planetário e discutir como técnicas de ciência de dados contribuem para sua eficácia na era dos grandes volumes de dados. A pesquisa combina revisão bibliográfica sobre fundamentos físicos do método de trânsito e desafios computacionais com procedimentos empíricos utilizando o algoritmo Transit Least Squares a 6807 curvas de luz do catálogo de Objetos de Interesse da Kepler, incluindo 2694 planetas confirmados e 4113 falsos positivos. Os resultados demonstraram que falsos positivos alcançaram eficiência média de detecção de sinal (46,47) superior ao de planetas confirmados (41,85), e valores máximos de até 289,69 contra 250,99, evidenciando que a métrica de eficiência de detecção de sinal, isoladamente, não discrimina sinais reais de artefatos periódicos. O processamento das 6807 curvas de luz demandou aproximadamente 4 meses em hardware convencional, ilustrando os desafios de escala na análise de grandes volumes de dados astronômicos. Os resultados empíricos confirmam as três dimensões dos grandes volumes de dados propostas por Zhang & Zhao: veracidade (60,4% de falsos positivos na amostra), variedade (integração de metadados e séries temporais) e volume (processamento massivo de dados). Embora o Transit Least Squares seja uma ferramenta robusta para detecção de periodicidade, sua integração com classificadores baseados em aprendizado profundo, como redes neurais convolucionais, é essencial para reduzir falsos positivos em fluxos de processamento automatizados de detecção de exoplanetas.

Palavras-chave: exoplanetas; método de trânsito; transit least squares; big data; detecção automatizada.

ABSTRACT

The detection of exoplanets using the planetary transit method has become the main discovery technique in contemporary exoplanetology, driven by space missions such as *Kepler* and *TESS*, which generate massive volumes of astronomical data. This study aims to analyze the planetary transit method and discuss how data science techniques contribute to its effectiveness in the big data era. The research combines a literature review on the physical foundations of the transit method and computational challenges with empirical procedures using the Transit Least Squares algorithm applied to 6807 light curves from the Kepler Object of Interest catalog, including 2694 confirmed planets and 4113 false positives. The results demonstrated that false positives achieved a higher mean signal detection efficiency (46.47) than confirmed planets (41.85), with maximum values reaching 289.69 compared to 250.99, showing that the signal detection efficiency metric alone does not discriminate real signals from periodic artifacts. Processing the 6807 light curves required approximately 4 months on conventional hardware, illustrating the scalability challenges in analyzing large volumes of astronomical data. The empirical results confirm the three dimensions of big data proposed by Zhang & Zhao: veracity (60.4% of false positives in the sample), variety (integration of metadata and time series), and volume (massive data processing). Although Transit Least Squares is a robust tool for periodicity detection, its integration with deep learning-based classifiers, such as convolutional neural networks, is essential to reduce false positives in automated exoplanet detection pipelines.

Keywords: exoplanets; transit method; transit least squares; big data; automated detection.

LISTA DE FIGURAS

Figura 1 – Número de exoplanetas descobertos por ano, discriminados por método de detecção. Observa-se o crescimento expressivo das detecções pelo método de trânsito a partir da missão *Kepler*. 27

Figura 2 – Ilustração do método de trânsito, mostrando a relação entre a passagem do planeta em frente à estrela e a correspondente queda no brilho observada na curva de luz. 29

Figura 3 – Local view do planeta confirmado Kepler-2 b (K00002.01), evidenciando o formato em "U" característico do trânsito planetário. 51

Figura 4 – Global view do planeta confirmado Kepler-2 b (K00002.01), mostrando os trânsitos periódicos a cada 2,20 dias. 51

Figura 5 – Local view do falso positivo K07010.01, mostrando o formato em "V" típico de sistemas binários eclipsantes. 52

Figura 6 – Global view do falso positivo K07010.01, mostrando o período extremamente curto de 0,76 dias. 52

LISTA DE TABELAS

Tabela 1 – Estimativa do volume de dados de missões e levantamentos astronômicos . . 31
Tabela 2 – Parâmetros do algoritmo TLS 44
Tabela 3 – Distribuição da amostra por classificação 46
Tabela 4 – Estatísticas do SDE por classe 47
Tabela 5 – Maiores valores de SDE para planetas confirmados 48
Tabela 6 – Maiores valores de SDE para falsos positivos 48
Tabela 7 – Estatísticas do período orbital (dias) 48
Tabela 8 – Estatísticas da duração do trânsito (dias) 49

LISTA DE ABREVIATURAS E SIGLAS

BLS	Box Least Squares
CNN	Convolutional Neural Network
IA	Inteligência Artificial
MAST	Mikulski Archive for Space Telescopes
NASA	National Aeronautics and Space Administration
PLATO	Planetary Transits and Oscillations of stars
SDE	Signal Detection Efficiency
TESS	Transiting Exoplanet Survey Satellite
TLS	Transit Least Squares

SUMÁRIO

1 INTRODUÇÃO 14

1.1 Problemática 15

1.2 Justificativa 16

1.3 Objetivos 17

1.3.1 Objetivo Geral 18

1.3.2 Objetivos Específicos 18

1.4 Estrutura do Trabalho 18

2 FUNDAMENTAÇÃO TEÓRICA 20

2.1 Mecânica Celeste 20

2.1.1 As Leis de Newton e o Problema de Dois Corpos 21

2.1.1.1 O Problema de Dois Corpos 21

2.1.2 Lei da Gravitação Universal e Leis de Kepler 22

2.2 Leis da Radiação Térmica Aplicada a Corpos Celestes 23

2.3 Efeito Doppler da Luz 25

2.4 Detecção de Exoplanetas 26

2.5 Método de Trânsito para Detecção de Exoplanetas 28

2.6 Grandes Volumes de Dados na Astronomia 30

2.7 Ciência de Dados Aplicada em Conjunto com o Método de Trânsito 32

2.7.1 Box Least Squares 33

2.8 Transit Least Squares 34

2.8.1 Otimizações Computacionais do TLS 36

2.8.2 Limitações e Compromissos Técnicos do TLS 37

3 METODOLOGIA 39

3.1 Delineamento e classificação da pesquisa 39

3.2 Procedimentos Computacionais com o TLS 40

3.2.1 Objetivo 40
3.2.2 Fonte e Seleção dos Dados 41
3.2.3 Procedimentos de Processamento 41
3.2.4 Descrição dos Parâmetros do TLS 43
3.2.5 Limitações e Escopo 43
4 RESULTADOS 46
4.1 Caracterização do conjunto de dados 46
4.1.1 Desafios da big data observados na amostra 46
4.2 Análise estatística dos parâmetros 47
4.2.1 SDE 47
4.2.2 Período orbital 48
4.2.3 Duração do trânsito 49
4.3 Desempenho do algoritmo Transit Least Squares 49
4.3.1 Análise morfológica: casos ilustrativos 50
4.3.2 Tempo de processamento 52
4.4 Discussão dos resultados 53
5 CONCLUSÕES E TRABALHOS FUTUROS 55
5.1 Conclusões 55
5.2 Limitações do Estudo 56
5.3 Trabalhos Futuros 56
REFERÊNCIAS 58
APÊNDICE A FLUXOGRAMA DO ALGORITMO TRANSIT LEAST SQUARES (TLS) 60
APÊNDICE B CÓDIGO DO PIPELINE DE PROCESSAMENTO 61

1 INTRODUÇÃO

A exploração espacial representa um dos marcos mais significativos do avanço científico e tecnológico da humanidade. Desde o lançamento do primeiro satélite artificial, o *Sputnik 1*, em 1957, até as missões que exploraram Marte, Vênus e os confins do Sistema Solar, a humanidade ampliou continuamente seus horizontes. A chegada do homem à Lua, em 1969, foi um marco emblemático, mas outras conquistas, como as missões *Voyager*, os telescópios espaciais e as sondas que pousaram em cometas e asteroides, também transformaram nossa compreensão do universo. A partir desse momento, a astronomia passou por uma transformação profunda, impulsionada pelo desenvolvimento de instrumentos de alta precisão e pelo uso intensivo da computação.

Historicamente, os princípios que hoje fundamentam a detecção de exoplanetas remontam aos séculos XVIII e XIX. A metodologia proposta por Edmond Halley para determinar a Unidade Astronômica (UA) baseava-se na observação do trânsito de Vênus. Esse mesmo princípio geométrico (o bloqueio parcial da luz de uma estrela por um corpo que cruza a linha de visada) foi posteriormente adaptado pela astronomia moderna para a detecção de exoplanetas (Santos; Amorim, 2017).

O método de trânsito consiste na observação da diminuição periódica do brilho estelar, gerando assinaturas características conhecidas como curvas de luz. Embora a primeira detecção de trânsito em um sistema extrassolar tenha ocorrido no final do século XX com o planeta HD 209458b, foi no século XXI que a área atingiu sua maturidade com o lançamento de missões espaciais dedicadas a descoberta de exoplanetas. (Santos; Amorim, 2017).

A missão Kepler, da NASA, foi especificamente projetada para aplicar o método de trânsito em larga escala. Com um fotômetro de alta precisão e um campo de visão de aproximadamente 115 graus quadrados, o telescópio monitorou simultaneamente o brilho de mais de 150.000 estrelas na região das constelações de Cygnus e Lyra (Koch *et al.*, 2010). Esse esforço resultou na descoberta de milhares de candidatos a exoplanetas, demonstrando que sistemas planetários são comuns em nossa galáxia.

Dando continuidade a esse legado, a missão TESS expandiu a busca para quase todo o céu, focando em estrelas cerca de 30 a 100 vezes mais brilhantes do que aquelas observadas pelo Kepler. O objetivo do TESS é catalogar planetas em torno das estrelas mais próximas e brilhantes, facilitando estudos de acompanhamento voltados à caracterização de atmosferas planetárias. (Ricker *et al.*, 2015).

O avanço dessas missões introduziu a astronomia definitivamente na chamada Era do Big Data. Segundo Zhang e Zhao (2015), a astronomia moderna apresenta as características clássicas dos chamados "V's" do Big Data: Volume, Velocidade, Variedade e Veracidade. O volume massivo de dados gerado por telescópios como o Kepler e o TESS exige o uso de algoritmos de mineração de dados e técnicas de computação de alto desempenho, uma vez que as informações sobre trânsitos planetários frequentemente estão imersas em ruídos instrumentais e variações naturais das estrelas. (Zhang; Zhao, 2015).

Assim, a transição da observação visual que começou com grandes figuras da astronomia, para a análise automatizada de grandes volumes de dados simboliza a convergência entre a mecânica celeste clássica e a ciência da computação moderna, permitindo a exploração de novos mundos em uma escala sem precedentes.

1.1 Problemática

O advento das missões espaciais modernas, como *Kepler* e *TESS*, resultou em um volume de dados que supera drasticamente a capacidade de análise manual. Repositórios como o *NASA Exoplanet Archive* (Aeronautics; Archive, 2026) concentram anos de observações fotométricas de milhares de estrelas, onde potenciais sinais de trânsitos planetários estão imersos em ruídos instrumentais, variabilidade estelar e falsos positivos. Como apontado por Zhang; Zhao (2015), a astronomia entrou na era do *Big Data*, onde o volume e a velocidade de geração de dados exigem novas infraestruturas de processamento e algoritmos de mineração de dados.

A inspeção visual dessas informações tornou-se inviável. A missão *Kepler*, por exemplo, monitorou mais de 150.000 estrelas simultaneamente, gerando curvas de luz com cadências de minutos ao longo de anos (Koch *et al.*, 2010). A sutileza dos sinais (pequenas quedas de luminosidade que podem representar menos de 0,01% do brilho total da estrela) frequentemente se confunde com flutuações naturais da própria fotosfera estelar ou com artefatos gerados pelo sistema de detectores CCD (*Charge-Coupled Device*) (Koch *et al.*, 2010).

Além do desafio quantitativo, destaca-se também a complexidade qualitativa dos dados. Cada curva de luz pode apresentar múltiplos eventos sobrepostos, além de interferências externas que dificultam a identificação de padrões claros. O ruído estatístico, aliado à presença de falsos positivos (como sistemas binários eclipsantes que podem imitar ou mascarar o sinal de um planeta), torna o processo de validação particularmente delicado (Koch *et al.*, 2010). Nesse contexto, torna-se indispensável o uso de técnicas de pré-processamento, incluindo normalização,

detrending e filtragem de ruído, a fim de garantir que o sinal de trânsito seja corretamente identificado e interpretado (Zhang; Zhao, 2015).

Nesse contexto, o problema central consiste em compreender, com base na literatura científica, como técnicas de Ciência de Dados e Inteligência Artificial têm sido empregadas na filtragem, no processamento e na classificação de grandes volumes de dados astronômicos. A aplicação de algoritmos de *Machine Learning* mostra-se fundamental para automatizar a identificação e distinção entre sinais planetários reais e ruídos de fundo (Shallue; Vanderburg, 2018).

Portanto, o desafio fundamental é transformar dados brutos em informação científica válida, reduzindo a incidência de falsos positivos e aumentando a precisão na detecção. Sem o desenvolvimento de abordagens automatizadas e robustas, o enorme potencial científico acumulado pelas missões espaciais e pelos grandes levantamentos astronômicos permanecerá subutilizado, limitando o avanço da ciência exoplanetária e a compreensão da diversidade de sistemas planetários na galáxia.

1.2 Justificativa

O estudo da detecção de exoplanetas apresenta relevância científica, pois contribui para a compreensão da formação e evolução de sistemas planetários, além de ampliar o debate acerca da possibilidade de existência de vida fora da Terra. A astronomia contemporânea, impulsionada por missões espaciais como *Kepler* e *TESS*, consolidou o método de trânsito planetário como uma das principais técnicas para a identificação de mundos extrassolares (Ricker *et al.*, 2015; Koch *et al.*, 2010). No entanto, o crescimento exponencial do volume de dados gerados por essas missões trouxe um desafio: distinguir sinais reais de trânsito de ruídos instrumentais e de fenômenos astrofísicos capazes de produzir falsos positivos (Zhang; Zhao, 2015).

Nesse cenário, o conceito de *Big Data* tornou-se central. O volume massivo de curvas de luz coletadas exige abordagens que superam a capacidade da análise manual ou de algoritmos estatísticos simples. A literatura recente demonstra que o uso de redes neurais convolucionais (*Convolutional Neural Networks* – CNNs) contribui para a classificação dessas informações. (Shallue; Vanderburg, 2018), por exemplo, demonstraram com o modelo *Astronet* que o aprendizado profundo (*Deep Learning*) é capaz de identificar planetas em sinais com baixa razão sinal-ruído que haviam sido negligenciados por métodos tradicionais, resultando na descoberta de novos planetas em sistemas já conhecidos, como o sistema Kepler-90.

A integração entre fundamentos físicos e abordagens computacionais revela-se, portanto, uma necessidade imediata. O método de trânsito não pode mais ser analisado apenas sob a ótica da fotometria clássica; ele deve ser compreendido dentro de um *pipeline* de Ciência de Dados. (Pearson; Palafox; Griffith, 2017) ressaltam que as redes neurais são altamente generalizáveis e capazes de aprender características físicas dos planetas diretamente dos dados, superando métodos tradicionais baseados em ajuste por mínimos quadrados (*least squares*), especialmente em conjuntos de dados ruidosos. Etapas como normalização e *detrending* tornam-se, assim, processos essenciais de pré-processamento para que algoritmos de Inteligência Artificial transformem dados brutos em descobertas científicas confiáveis.

Justifica-se, portanto, a realização desta pesquisa de natureza mista (teórico-experimental) por sua capacidade de integrar conceitos de física e computação em uma perspectiva interdisciplinar, especialmente no contexto da detecção de exoplanetas pelo método de trânsito. Diante do crescente volume de dados provenientes de missões astronômicas e das limitações inerentes aos métodos tradicionais de análise, torna-se necessário compreender como técnicas de Ciência de Dados e Inteligência Artificial podem contribuir para a melhoria da eficiência, sensibilidade e confiabilidade desses processos.

Ao discutir contribuições consolidadas e inovações recentes, este trabalho busca não apenas apresentar fundamentos teóricos, mas também analisar desafios práticos associados à detecção de trânsitos planetários, como a presença de ruído, ocorrência de falsos positivos e custos computacionais. Nesse contexto, o uso de algoritmos como o TLS permite investigar, em dados reais, as implicações dessas limitações e o papel da automação na análise de grandes volumes de dados.

O fortalecimento dessa abordagem interdisciplinar fornece subsídios teóricos e metodológicos relevantes para futuras investigações, bem como para a otimização de sistemas automatizados de detecção em missões astronômicas contemporâneas e futuras.

1.3 Objetivos

Nesta seção, são apresentados o objetivo geral e os objetivos específicos que orientam o desenvolvimento da pesquisa. O objetivo geral define a direção principal do estudo, enquanto os objetivos específicos detalham as etapas necessárias para alcançar esse propósito.

1.3.1 Objetivo Geral

Analisar o método de trânsito planetário e discutir como técnicas de ciência de dados contribuem para sua eficácia na astronomia contemporânea.

1.3.2 Objetivos Específicos

O presente trabalho desdobra-se nos seguintes objetivos específicos:

1. Revisar os fundamentos da mecânica celeste, as Leis de Kepler e a física envolvida no trânsito planetário.
2. Analisar o algoritmo TLS para detecção de trânsitos periódicos, com ênfase em suas limitações físicas, estatísticas e computacionais, especialmente no que se refere à sensibilidade ao ruído.
3. Discutir o papel da big data e da automação no processamento e análise de grandes volumes de dados astronômicos, destacando sua importância na detecção de trânsitos planetários.
4. Utilizar procedimentos computacionais com o algoritmo TLS como ferramenta de investigação empírica, permitindo observar em dados reais os desafios teóricos discutidos, tais como tempo de processamento, ocorrência de falsos positivos e necessidade de automação diante de grandes volumes de dados astronômicos.

1.4 Estrutura do Trabalho

Este trabalho está dividido em 5 capítulos, além das referências. O primeiro capítulo é a Introdução, que apresenta a problemática, a justificativa da pesquisa e os objetivos gerais do estudo.

O segundo capítulo traz a Fundamentação Teórica, onde são discutidos os fundamentos físicos do método de trânsito planetário, os desafios da Big Data na astronomia e os principais algoritmos de detecção de trânsitos periódicos, como o BLS e o TLS.

O terceiro capítulo descreve a Metodologia adotada na pesquisa, incluindo o delineamento do estudo, a seleção dos dados e os procedimentos computacionais realizados com o algoritmo TLS.

O quarto capítulo apresenta os Resultados obtidos a partir da aplicação do pipeline, incluindo a caracterização da amostra, a análise estatística descritiva dos parâmetros orbitais e de detecção, o desempenho do algoritmo e a discussão dos achados.

Por fim, o quinto capítulo traz as Conclusões do trabalho, bem como as limitações do estudo e perspectivas para trabalhos futuros na área.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os conceitos fundamentais que embasam o desenvolvimento da pesquisa. São discutidos os fundamentos da mecânica celeste os princípios da radiação térmica, que permitem compreender a dinâmica orbital dos planetas e a emissão de energia das estrelas. Em seguida, aborda-se o método de trânsito planetário, destacando sua relevância para a detecção de exoplanetas na era da big data e os avanços computacionais associados.

2.1 Mecânica Celeste

A mecânica celeste é o ramo da astronomia que utiliza os princípios da mecânica (clássica ou relativística) para estudar o movimento dos corpos e realizar previsões orbitais (Araujo; Ramos, 2023). Seu objetivo central é descrever o comportamento de planetas, satélites e estrelas a partir das forças que atuam sobre eles, sendo a interação gravitacional a força predominante e, em muitos modelos, a única variável significativa (Araujo; Ramos, 2023). Essa abordagem permite modelar sistemas astronômicos com elevado grau de precisão, tratando os corpos celestes como pontos de massa sujeitos a campos de força central.

Nesse contexto, a interação gravitacional entre dois corpos constitui o ponto de partida para a compreensão da dinâmica orbital (Araujo; Ramos, 2023). A partir das leis de Newton e da lei da gravitação universal, estabelecem-se equações de movimento que descrevem como massas distintas se influenciam mutuamente no espaço.

Além disso, a mecânica celeste fornece a base para a análise de sistemas com diferentes complexidades. Enquanto o problema de dois corpos admite soluções analíticas exatas (nas quais os corpos orbitam um centro de massa comum (baricentro)), sistemas com três ou mais corpos (problema de n-corpos) exigem métodos numéricos e computacionais para sua resolução (Araujo; Ramos, 2023). No entanto, para a detecção de exoplanetas, a aproximação de dois corpos é amplamente satisfatória, permitindo extrair parâmetros fundamentais a partir de observações indiretas.

Dessa forma, a mecânica celeste não é apenas um estudo teórico, mas o arcabouço de engenharia que sustenta a interpretação de dados observacionais. Ela permite relacionar variações mínimas no brilho estelar a propriedades físicas robustas, como o raio e a distância orbital dos exoplanetas detectados.

2.1.1 As Leis de Newton e o Problema de Dois Corpos

Para a fundamentação teórica do método de trânsito planetário, é imprescindível a compreensão das Leis do Movimento formuladas por Isaac Newton, com sua versão antecedente produzida em 1666 mas apenas publicadas em 1687 na obra *Philosophiæ Naturalis Principia Mathematica*. Essas leis constituem o núcleo da mecânica clássica e fornecem a base necessária para descrever a dinâmica de massas sob a influência de forças gravitacionais (Filho; Saraiva, 2014).

A **primeira lei de Newton** (lei da inércia) foi formulada com base nas conclusões de Galileu Galilei, obtidas, por meio de seus experimentos, Newton concluiu que um corpo mantém seu estado de movimento (repouso ou movimento retilíneo uniforme) se a resultante das forças que atuam sobre ele for nula. Como consequência dessa lei, o movimento orbital não ocorre espontaneamente, mas resulta da ação contínua da gravidade (Filho; Saraiva, 2014). No contexto astronômico, isso evidencia que a órbita de um planeta não é um estado "natural" de movimento, mas o resultado de uma força central (gravidade) que desvia continuamente o planeta de sua trajetória retilínea.

A **segunda lei de Newton** (lei fundamental) Afirma que a resultante das forças que atuam sobre um corpo é proporcional a aceleração adquirida por esse corpo. Dessa forma, quanto maior for a força aplicada, maior será a aceleração produzida; por outro lado, quanto maior for a massa do objeto, menor será a aceleração para uma mesma força. Além disso, a aceleração possui a mesma direção e sentido da força resultante que atua sobre o corpo (Filho; Saraiva, 2014).

$$\vec{F} = m\vec{a}. \tag{2.1}$$

A **terceira lei de Newton** (ação e reação) A toda ação corresponde uma reação de mesma intensidade, mesma direção e sentido contrário. Essa lei evidencia que a interação gravitacional é mútua, com planeta e estrela orbitando em torno de um centro de massa comum (Filho; Saraiva, 2014).

2.1.1.1 O Problema de Dois Corpos

O problema dos dois corpos constitui o único cenário, dentro da mecânica celeste, que admite solução analítica exata. Newton, em seus Principia, demonstrou que duas massas sujeitas

exclusivamente à atração gravitacional mútua descrevem trajetórias cônicas (elipses, parábolas ou hipérboles) dependendo da energia total do sistema. Esse problema reduz-se a uma formulação de força central, conhecida como Problema de Kepler, no qual a força varia com o inverso do quadrado da distância. (Araujo; Ramos, 2023)

A completeza da solução decorre da existência de integrais primeiras suficientes para tornar o sistema completamente integrável. Além das dez integrais associadas ao centro de massa, ao momento angular e à energia, o chamado vetor de Laplace (ou vetor de Lenz) fornece duas integrais adicionais, o que permite descrever plenamente o movimento. (Araujo; Ramos, 2023)

No contexto do sistema solar, a solução do problema de dois corpos serve como aproximação de primeira ordem: como a massa do Sol domina sobre a dos planetas, cada planeta pode ser tratado, em uma abordagem inicial, como se estivesse isolado, orbitando em elipse ao redor do astro central. Embora as interações planetárias introduzam perturbações que tornam o problema real muito mais complexo, a base elíptica oferece um ponto de partida essencial para estudos mais avançados. (Araujo; Ramos, 2023)

2.1.2 Lei da Gravitação Universal e Leis de Kepler

A dinâmica de sistemas gravitacionais, como um planeta orbitando uma estrela ou um satélite orbitando a terra, é descrita pela interação gravitacional entre a trajetória orbital e a força que a governa. De acordo com a lei da gravitação Universal, dois corpos de massas M e m se atraem mutuamente com uma força diretamente proporcional ao produto de suas massas e inversamente proporcional ao quadrado da distância r que os separa, isto é:

$$\vec{F} = -\frac{GMm}{r^2}\hat{r}, \tag{2.2}$$

onde G é a constante gravitacional e $\hat{r}$ é um versor radial. Em sistemas gravitacionalmente ligados, a força gravitacional atua como uma força centrípeta responsável por manter os planetas em órbita, impedindo que sigam em movimento retilíneo uniforme. O conceito de campo gravitacional $\vec{g}$ permite descrever como a massa de uma estrela central domina a região ao seu redor, determinando a trajetória de corpos menores, como exoplanetas, em sua vizinhança.

Mesmo antes de Newton explicar as causas do movimento orbital, Johannes Kepler, já tinha estabelecido um conjunto de três leis empíricas, a partir da análise das observações feitas

por Tycho Brahe, para descrever o movimento dos planetas.

Primeira lei (lei das órbitas): estabelece que cada planeta descreve uma trajetória elíptica, com o sol ocupando um dos focos da elipse. A forma dessa elipse é definida pela excentricidade e, parâmetro essencial em sistemas de exoplanetas, pois determina a proximidade mínima e máxima entre o planeta e sua estrela ao longo da órbita.

Segunda lei (lei das áreas): afirma que o raio vetor que conecta o planeta à estrela varre áreas iguais em intervalos de tempo iguais. Como consequência, a velocidade orbital não é uniforme: o planeta se desloca mais rapidamente nas proximidades do periastro (ponto de menor distância) e mais lentamente no apoastro (ponto de maior distância). Essa lei é uma consequência da conservação do movimento angular.

Terceira lei (lei harmônica): estabelece que o quadrado do período orbital τ é proporcional ao cubo do semieixo maior a da órbita (Filho; Saraiva, 2014):

$$\tau^2 \propto a^3. \tag{2.3}$$

Essas leis são ferramentas essenciais para o estudo de exoplanetas. Segundo (Santos; Amorim, 2017), a terceira lei de Kepler é essencial para caracterizar sistemas planetários. Quando observamos o tempo entre dois trânsitos (o período τ) e estimamos a massa da estrela (M) por meio de modelos estelares, podemos calcular a distância do planeta (a) através da relação:

$$\tau^2 = \frac{4\pi^2}{GM} a^3. \tag{2.4}$$

Além disso, a variação de velocidade descrita pela Segunda Lei de Kepler tem impacto direto na duração do trânsito. Um trânsito que ocorre próximo ao periastro de uma órbita excêntrica será mais rápido do que um trânsito em uma órbita circular de mesmo período (Santos; Amorim, 2017). Dessa forma, a análise da curva de luz (o gráfico de brilho da estrela ao longo do tempo) permite aos astrônomos interpretar o sistema e extrair sua arquitetura fís xica completa.

2.2 Leis da Radiação Térmica Aplicada a Corpos Celestes

Em 1859, o físico alemão Gustav Kirchhoff definiu o corpo negro como um objeto ideal que absorve toda a radiação incidente, sem refletir ou transmitir qualquer parcela dela (Filho; Saraiva, 2014). Por ser um absorvedor perfeito, esse corpo também deve emitir energia na

mesma taxa em que a absorve, mantendo o equilíbrio termodinâmico. Assim, o corpo negro é considerado um emissor perfeito de radiação eletromagnética. Embora um corpo negro perfeito não exista na natureza, as estrelas constituem boas aproximações, pois são objetos opacos em equilíbrio termodinâmico local, permitindo a aplicação dessas leis ao estudo de sua radiação (Filho; Saraiva, 2014).

A base teórica moderna para a radiação emitida por um corpo negro foi estabelecida por Max Planck, em 1900, ao propor que a energia eletromagnética é emitida em quantidades discretas, denominadas fótons (Filho; Saraiva, 2014). A lei de Planck descreve como a densidade de radiação de um corpo negro se distribui em função do comprimento de onda (λ) e da temperatura (T) e pode ser escrita como:

$$\rho(\lambda, T) = \frac{8\pi hc}{\lambda^5} \frac{1}{e^{\frac{hc}{\lambda kT}} - 1}, \tag{2.5}$$

onde h é a constante de Planck, c é a velocidade da luz e k é a constante de Boltzmann. Essa relação é fundamental para a interpretação do espectro de radiação das estrelas.

Derivando a expressão acima em relação ao comprimento de onda e igualando o resultado a zero, obtém-se uma relação entre o comprimento de onda de máxima emissão e a temperatura do corpo negro, isto é:

$$\lambda_{\max} T = 2898 \times 10^{-3}\,\mathrm{m \cdot K}. \tag{2.6}$$

Esse resultado é conhecido como lei de Wien.

A lei de Wien permite determinar a temperatura superficial de uma estrela a partir do comprimento de onda em que a emissão é máxima. Note que estrelas mais quentes (mais azuis) emitem mais radiação em comprimentos de onda menores, enquanto que estrelas mais frias (mais avermelhadas), emitem mais radiação em comprimento de ondas maiores.

Agora integrando a lei de Planck sobre todo o espectro de frequências(ou comprimento de onda) pode-se mostrar que a energia emitida por um corpo negro é proporcional a quarta potência da temperatura, isto é:

$$F = \sigma T^4, \tag{2.7}$$

onde F é o fluxo de energia emitido e σ a constante de Stefan-Boltzmann.

A luminosidade (L) e o fluxo de energia de uma estrela estão relacionadas por

$$F = \frac{L}{4\pi r^2}. \tag{2.8}$$

onde r é a distância da estrela ao oberservador. Combinando as equações 2.7 e 2.8, obtém-se que:

$$L = 4\pi\sigma r^2 T^4. \tag{2.9}$$

Essa relação permite determinar a temperatura superficial de uma estrela a partir de sua luminosidade.

Para estrelas, essa relação permite determinar sua luminosidade total a partir de sua temperatura de seu raio (Filho; Saraiva, 2014).

No contexto do método do trânsito, as leis da radiação térmica desempenham um papel fundamental na caracterização da estrela hospedeira. Como a profundidade do trânsito está diretamente relacionada à razão entre as áreas do planeta e da estrela, a determinação do raio estelar, que depende da luminosidade e da temperatura, é essencial para a estimativa correta do tamanho do exoplaneta (Santos; Amorim, 2017).

Além disso, a energia emitida pela estrela, descrita pelas leis de radiação de corpo negro, juntamente com a distância orbital do planeta, permite estimar sua temperatura de equilíbrio. Dessa forma, essas leis fornecem a base teórica necessária para interpretar as variações de brilho observadas durante o trânsito e inferir propriedades físicas dos exoplanetas (Santos; Amorim, 2017).

2.3 Efeito Doppler da Luz

O efeito Doppler consiste na alteração da frequência observada de uma onda em decorrência do movimento relativo entre a fonte emissora e o observador. Na astronomia, esse fenômeno constitui a base para a medição da velocidade radial dos astros, sendo uma ferramenta diagnóstica fundamental para o estudo de exoplanetas e da dinâmica estelar (Filho; Saraiva, 2014).

Se uma fonte luminosa emite luz com frequência (f_0), a frenquência observada (f), é dada por:

$$f \approx f_0 \left(1 \pm \frac{v}{c}\right), \tag{2.10}$$

onde v é a velocidade relativa entre a fonte e o observador e foi assumido que $v \ll c$.

Na expressão anterior usa-se o sinal positivo quando a fonte e observador estão se aproximando e o sinal negativo quando estão se afastando.

Note que a partir da expressão acima pode-se concluir que, ocorre:

- **Desvio para o vermelho (redshift):** ocorre quando o objeto se afasta do observador, resultando em um aumento do comprimento de onda, o que desloca a radiação observada para regiões de maior comprimento de onda, ou seja, para o vermelho (Filho; Saraiva, 2014).
- **Desvio para o azul (blueshift):** ocorre quando o objeto se aproxima do observador, resultando na diminuição do comprimento de onda, deslocando a radiação observada para regiões de menor comprimento de onda, ou seja, para o azul (Filho; Saraiva, 2014).

O efeito Doppler também permite a determinação da rotação estelar, pois em uma estrela em rotação, uma das bordas se aproxima do observador enquanto a outra borda se afasta. Essa diferença nas velocidades radiais ocasiona um aumento característico nas linhas de absorção do espectro estelar (Filho; Saraiva, 2014).

Esse conceito é fundamental para distinguir variações no brilho estelar causadas por fenômenos intrínsecos, como manchas na superfície da estrela, daquelas provocadas pela passagem de planetas em trânsito, uma vez que a confirmação de um exoplaneta requer a observação de quedas periódicas e regulares no fluxo luminoso, evitando interpretações equivocadas associadas a outros processos estelares (Santos; Amorim, 2017).

2.4 Detecção de Exoplanetas

A confirmação da existência de planetas fora do Sistema Solar ocorreu a partir da década de 1990, quando as primeiras detecções foram realizadas. Em 1992, Aleksander Wolszczan e Dale Frail identificaram dois planetas orbitando o pulsar PSR B1257+12 por meio do método de variações no tempo de pulsar (Rice, 2014). Três anos depois, Michel Mayor e Didier Queloz descobriram 51 Pegasi b, o primeiro exoplaneta orbitando uma estrela semelhante ao Sol, utilizando o método da velocidade radial (Rice, 2014).

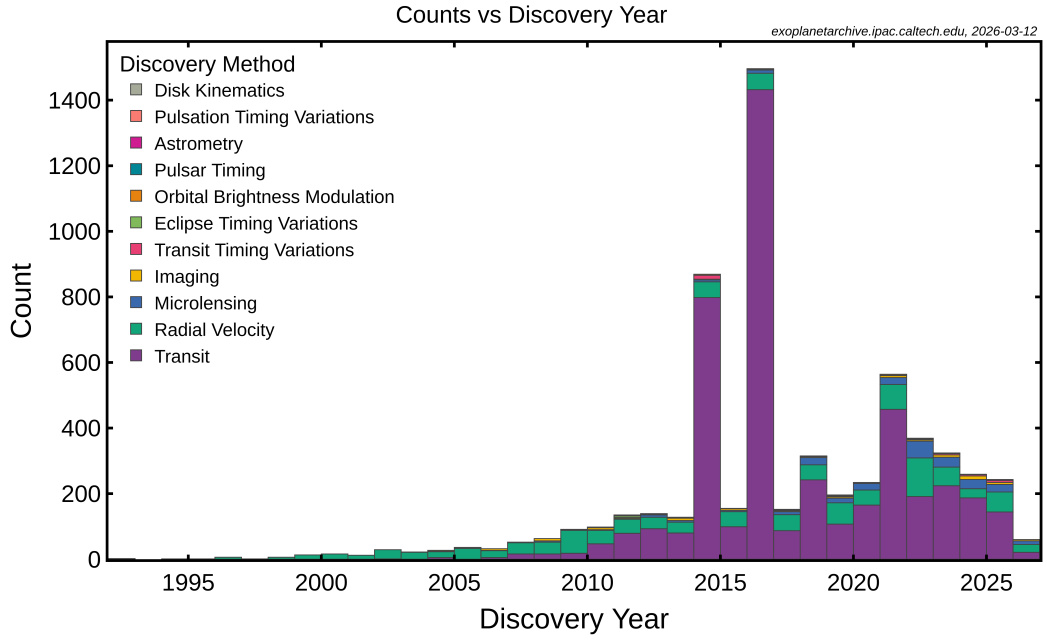
Desde então, diferentes técnicas indiretas foram desenvolvidas para a detecção de exoplanetas. Os principais métodos são:

- **Velocidade radial:** mede as oscilações da estrela causadas pela atração gravitacional do planeta, permitindo estimar a massa mínima do planeta e seu período orbital (Rice, 2014).

- **Trânsito:** detecta pequenas quedas periódicas no brilho da estrela quando um planeta passa à sua frente, possibilitando a determinação do raio planetário (Rice, 2014).
- **Microlente gravitacional:** observa o aumento temporário do brilho de uma estrela distante quando outra estrela (com possível planeta) atua como lente gravitacional (Rice, 2014).
- **Imageamento direto:** obtém imagens do próprio planeta, embora seja limitado a planetas massivos e distantes de sua estrela (Rice, 2014).

A figura1 mostra a contagem de exoplanetas em função do ano de descoberta, com a distribuição de métodos de detecção, elaborada a partir dos dados do NASA Exoplanet Archive. O método de trânsito começou a se consolidar como a principal técnica de detecção a partir do lançamento da missão *Kepler*, em 2009, enquanto a técnica de velocidade radial predominou nas décadas anteriores. Métodos como microlente gravitacional e imageamento direto contribuem com um número menor de detecções, mas permitem explorar regiões complementares do espaço de parâmetros (Rice, 2014).

Figura 1 – Número de exoplanetas descobertos por ano, discriminados por método de detecção. Observa-se o crescimento expressivo das detecções pelo método de trânsito a partir da missão *Kepler*.



Fonte: NASA Exoplanet Archive (2026).

Observa-se um pico significativo nas detecções por volta de 2014–2016, associado à validação em larga escala de candidatos identificados pela missão *Kepler*, o que explica a concentração de descobertas nesse período. Dados mais recentes confirmam essa tendência e

evidenciam a predominância do método de trânsito ao longo dos anos.

A diversidade de técnicas ilustrada na figura 1 permite investigar diferentes configurações planetárias. Até a publicação do trabalho de Rice (2014), mais de 1800 exoplanetas haviam sido identificados, revelando uma ampla variedade de massas, períodos orbitais e composições, muitas vezes distintas das observadas no Sistema Solar (Rice, 2014). Desde então, esse número cresceu significativamente, ultrapassando a marca de 6.000 exoplanetas confirmados na atualidade, de acordo com dados da NASA (Aeronautics; Administration, 2026b).

2.5 Método de Trânsito para Detecção de Exoplanetas

O método de trânsito consiste em observar a diminuição periódica do brilho de uma estrela quando um planeta passa diante de seu disco, em relação à linha de visada do observador (Santos; Amorim, 2017). Esse fenômeno tornou-se uma ferramenta poderosa para a descoberta de mundos além do Sistema Solar.

Historicamente, os trânsitos de Vênus e Mercúrio diante do Sol, observados nos séculos XVIII e XIX, foram fundamentais para o cálculo da unidade astronômica (UA) (Santos; Amorim, 2017). Observações realizadas em diferentes pontos da Terra permitiram determinar a distância ao Sol com base em princípios trigonométricos. Embora o trânsito de Vênus ocorra apenas quatro vezes em 243 anos, esse fenômeno serviu como base para as técnicas que hoje aplicamos a estrelas distantes (Santos; Amorim, 2017).

Na astronomia moderna, o primeiro exoplaneta estudado por esse método foi o HD 209458b. Em 2002, vieram as primeiras descobertas diretas: OGLE-TR-56b e OGLE-TR-10b, com massas de 0,6 e 1,3 vezes a massa de Júpiter, respectivamente (Santos; Amorim, 2017). Esses casos inauguraram a categoria dos chamados “Júpiteres quentes”, caracterizados por seus grandes raios. Como consequência, produzem trânsitos mais profundos e, portanto, são mais facilmente detectáveis pelo método de trânsito.

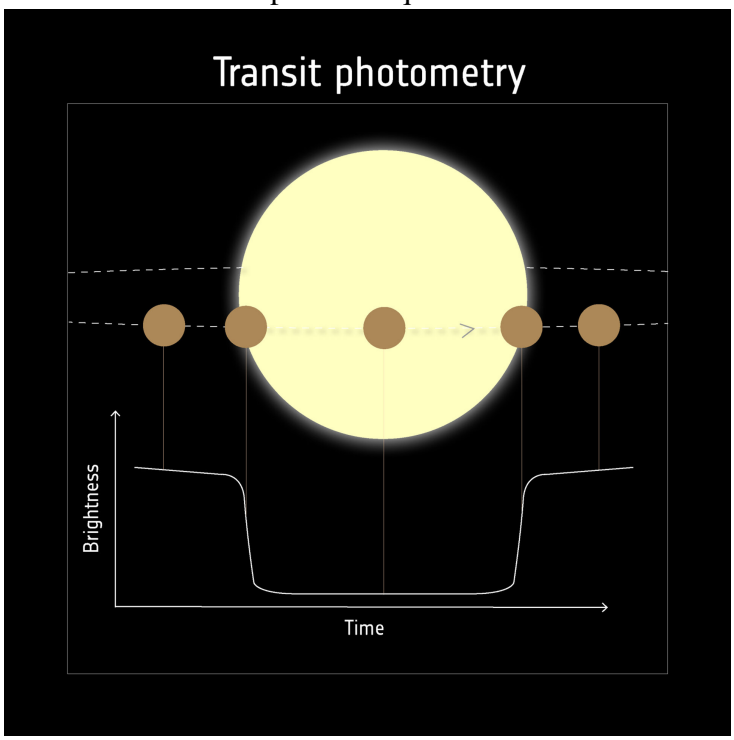
O método consolidou-se com missões espaciais como COROT e Kepler, lançada em 2009, esta última responsável pela confirmação de milhares de exoplanetas e pela revelação de que sistemas planetários múltiplos são comuns (Santos; Amorim, 2017). Atualmente, missões como Transiting Exoplanet Survey Satellite (TESS) e a futura PLATO buscam identificar planetas em estrelas mais próximas do Sistema Solar.

Estatisticamente, o método de trânsito é o mais produtivo na atualidade. Segundo o National Aeronautics and Space Administration (NASA) Exoplanet Archive, cerca de 73%

dos exoplanetas confirmados foram descobertos por essa técnica (Aeronautics; Administration, 2026a).

Além da detecção, esse método permite estimar o raio do planeta a partir da profundidade da curva de luz, que está diretamente relacionada à razão entre as áreas do planeta e da estrela $(R_p/R_*)^2$ (Santos; Amorim, 2017).

Figura 2 – Ilustração do método de trânsito, mostrando a relação entre a passagem do planeta em frente à estrela e a correspondente queda no brilho observada na curva de luz.



Fonte: ESA (2026).

A figura 2 apresenta um exemplo típico de curva de luz, evidenciando a queda de brilho durante o trânsito planetário e os principais parâmetros observáveis.

Embora o método favoreça a detecção de planetas de grande porte e exija que o sistema esteja alinhado a aproximadamente 90° em relação à linha de visada, sua complementaridade com outras técnicas, como a espectroscopia Doppler, é essencial para a determinação da densidade e da composição planetária (Santos; Amorim, 2017). Além disso, o método de trânsito é inerentemente mais sensível a planetas com períodos orbitais curtos, uma vez que estes produzem uma maior frequência de eventos de trânsito em um dado intervalo de tempo, aumentando a significância estatística da detecção e a confiabilidade da validação..

2.6 Grandes Volumes de Dados na Astronomia

A astronomia contemporânea atravessou uma mudança de paradigma, na qual projetos de observação modernos produzem volumes de dados que atingem a escala de petabytes, impulsionando o surgimento de disciplinas híbridas como a Astroestatística e a Astroinformática (Zhang; Zhao, 2015).

Os autores destacam que a astroestatística, e de forma mais ampla a astroinformática, emergiu em um contexto particularmente favorável, impulsionado pelo crescimento acelerado do volume e da complexidade dos dados astronômicos (Zhang; Zhao, 2015). Com a evolução dos levantamentos do céu, os repositórios passaram de escalas de gigabytes para terabytes e até petabytes, exigindo novas abordagens para armazenamento, processamento e análise.

Nesse cenário, a mineração de dados tornou-se uma ferramenta essencial, permitindo a extração eficiente de padrões, relações e conhecimentos relevantes a partir de grandes bases de dados astronômicos. Entre as principais tarefas destacam-se a classificação, regressão, agrupamento, análise de séries temporais e detecção de anomalias, fundamentais para a interpretação de fenômenos astrofísicos (Zhang; Zhao, 2015).

Além disso, o aumento da dimensionalidade dos dados, decorrente de levantamentos mais profundos, multiespectrais e no domínio do tempo, introduz desafios como a chamada "maldição da dimensionalidade", que pode comprometer o desempenho de diversos algoritmos. Para mitigar esse problema, técnicas de seleção e extração de atributos são amplamente empregadas, sendo a seleção frequentemente preferida por preservar o significado físico das variáveis originais. Dessa forma, a mineração de dados consolida-se como um elemento central na análise de dados astronômicos em larga escala (Zhang; Zhao, 2015).

Ao analisar a taxonomia proposta, observa-se que as dimensões do Big Data na astronomia impõem desafios que transcendem a infraestrutura de armazenamento, incidindo diretamente na precisão dos modelos analíticos (Zhang; Zhao, 2015). A variedade, por exemplo, manifesta-se na necessidade de integrar fluxos heterogêneos, como metadados estelares e séries temporais de brilho, para a validação de um sinal. Já a veracidade constitui um dos principais obstáculos técnicos na detecção de exoplanetas: o ruído intrínseco das observações e as flutuações naturais da estrela, como manchas estelares, podem mimetizar trânsitos reais, exigindo que as técnicas de mineração de dados possuam robustez estatística para mitigar falsos positivos (Zhang; Zhao, 2015).

O advento de projetos como o *Large Synoptic Survey Telescope* (LSST) introduz uma

escala na qual o processamento em tempo real torna-se obrigatório. Conforme descrito na literatura, o LSST deverá detectar cerca de mil novas supernovas por noite ao longo de dez anos, implicando a geração de dezenas a centenas de milhares de alertas (Zhang; Zhao, 2015). A mineração eficiente, a classificação correta dos candidatos e a realização de observações de acompanhamento representam um desafio significativo para a comunidade astronômica. Essa infraestrutura redefine a mineração de dados como a principal via para a extração de padrões em populações estelares em larga escala.

Tabela 1 – Estimativa do volume de dados de missões e levantamentos astronômicos

Missão / Projeto	**Volume de dados**
Hubble Space Telescope	$\sim$ 434 TB (Aeronautics; Administration, 2024)
Kepler Space Telescope	$\sim$ 20–50 TB (Aeronautics; Administration, 2018)
Sloan Digital Sky Survey (SDSS)	$\sim$ 750 TB (Collaboration, 2022)
Gaia	$\sim$ 500 TB (Agency, 2025a)
LSST (Vera Rubin Observatory)	$\sim$ 15 PB (Desc; Issc, 2026)

Fonte: NASA/ESA (2026).

Como mostra a tabela 1, os volumes de dados correspondem a estimativas acumuladas dos acervos das missões, considerando todo o período de coleta até as versões mais recentes dos arquivos públicos. Ressalta-se que tais valores são dinâmicos e tendem a crescer continuamente com novas observações e reprocessamentos. Ainda assim, observa-se uma variação que vai de centenas de terabytes a dezenas de petabytes.

Conforme discutido por (Zhang; Zhao, 2015), a astronomia na era do *big data* é caracterizada pelos chamados “4 Vs”: volume, variedade, velocidade e valor dos dados. Esses aspectos implicam desafios relacionados ao armazenamento, processamento, integração e análise de grandes conjuntos de dados heterogêneos, provenientes de diferentes instrumentos e levantamentos. Além disso, a alta dimensionalidade e a complexidade dos dados tornam necessária a adoção de métodos de mineração de dados e aprendizado de máquina para a extração eficiente de conhecimento científico.

A missão *Kepler* representou um marco na aplicação de Big Data à exoplanetologia, destacando-se pela elevada estabilidade fotométrica de seu sistema observacional (Koch *et al.*, 2010). Entretanto, o principal desafio concentrou-se no processamento contínuo de grandes volumes de dados, exigindo pipelines capazes de tratar séries temporais extensas de milhares de estrelas simultaneamente.

O TESS expandiu esse cenário ao adotar uma estratégia de levantamento quase completo

do céu, com foco em estrelas brilhantes e próximas. A espaçonave foi colocada em uma órbita altamente elíptica ao redor da Terra, com período de aproximadamente 13,7 dias. Durante sua missão principal, o satélite realiza um mapeamento sistemático do céu em setores, nos quais cada estrela é observada por intervalos que variam de cerca de 27 dias até quase um ano, dependendo de sua posição em relação aos polos eclípticos (Ricker *et al.*, 2015). Esse modelo observacional resulta na geração contínua de grandes volumes de dados fotométricos, incluindo imagens completas e séries temporais, o que demanda infraestruturas robustas para armazenamento, processamento e análise em larga escala.

Para evitar que esses dados se tornem subutilizados, a comunidade científica tem desenvolvido diversas ferramentas e plataformas que possibilitam a aplicação de técnicas de aprendizado de máquina na identificação de padrões e anomalias em grandes catálogos astronômicos (Zhang; Zhao, 2015).

A integração desses dados é viabilizada por iniciativas como o *Observatório Virtual*. Conforme descrito por Zhang e Zhao, esse ambiente fornece acesso distribuído e interoperável a diferentes bases de dados astronômicos, permitindo a combinação eficiente de informações provenientes de múltiplas missões (Zhang; Zhao, 2015).

2.7 Ciência de Dados Aplicada em Conjunto com o Método de Trânsito

O método de trânsito planetário consolidou-se como o principal motor de descobertas de exoplanetas, mas o volume massivo de dados gerado por missões espaciais exigiu uma transição para a análise automatizada. Nesse cenário, a Inteligência Artificial (IA) desempenha um papel fundamental. Em 2017, o engenheiro de software Christopher Shallue (Google Brain) e o astrônomo Andrew Vanderburg (Harvard/UT Austin) desenvolveram uma Rede Neural Convolucional, denominada Astronet, projetada especificamente para automatizar a distinção entre trânsitos planetários reais e falsos positivos (Shallue; Vanderburg, 2018).

Antes de detalhar o algoritmo, é preciso compreender como as Redes Neurais Artificiais (RNAs) auxiliam na identificação desses mundos. O telescópio espacial Kepler identifica milhares de sinais periódicos conhecidos como "Eventos de Cruzamento de Limiar"(Threshold Crossing Events- TCEs) (Koch *et al.*, 2010). Para classificar esses sinais, o modelo Astronet processa as curvas de luz utilizando duas representações distintas: a global view e a local view (Shallue; Vanderburg, 2018).

A global view permite que a rede observe a curva completa para identificar eclipses

secundários ou variações estelares de longo período. Já a local view foca exclusivamente no momento do trânsito, permitindo a análise de detalhes sutis no formato do sinal. Essa combinação é crucial para distinguir o formato em "U", característico de um planeta, do formato em "V", típico de sistemas de estrelas binárias de eclipse. Além disso, através da visão global, o modelo aprende a detectar eclipses secundários, evidência forte de que o sinal provém de uma estrela secundária passando atrás da primária, e não de um exoplaneta (Shallue; Vanderburg, 2018).

Embora a rede neural Astronet seja altamente eficaz na classificação de sinais, ela depende de uma etapa prévia de varredura sistemática para identificar candidatos potenciais em curvas de luz brutas. O método padrão para essa detecção primária, especialmente no contexto de sistemas multiplanetários e sinais de baixa amplitude, é o algoritmo Box Least-Squares (BLS) (Shallue; Vanderburg, 2018).

Para garantir a robustez, Shallue e Vanderburg (Shallue; Vanderburg, 2018) utilizaram a técnica de Ensemble, treinando dez cópias independentes do modelo para reduzir a variância na classificação de sinais difíceis.

A eficácia dessa abordagem foi demonstrada pela capacidade da Astronet em classificar candidatos com uma precisão de 96% no conjunto de testes do Kepler (Shallue; Vanderburg, 2018). O diferencial do modelo reside na sua alta performance em sinais de baixa Razão Sinal-Ruído (SNR), onde métodos anteriores falhavam. Ao aplicar a rede a sistemas já conhecidos, o modelo identificou novos planetas, como o Kepler-90i (o oitavo planeta do sistema Kepler-90) e o Kepler-80g. Em última análise, o aprendizado profundo provou não apenas acelerar a validação, mas ampliar as fronteiras da descoberta ao resgatar sinais extremamente tênues imersos no ruído estelar (Shallue; Vanderburg, 2018).

2.7.1 Box Least Squares

Antes do desenvolvimento do BLS, a detecção de periodicidades em séries temporais astronômicas baseava-se em métodos como a Transformada de Fourier e a Minimização de Dispersão de Fase (PDM), que são mais adequados a sinais aproximadamente senoidais e, portanto, pouco eficientes para a detecção de trânsitos, os quais apresentam quedas abruptas e de curta duração (Kovacs; Zucker; Mazeh, 2002).

O algoritmo BLS foi proposto por (Kovacs; Zucker; Mazeh, 2002) com o objetivo de identificar trânsitos periódicos em grandes volumes de dados fotométricos. Sua premissa central é modelar o trânsito como uma função em degrau (“caixa”), simplificando o problema de ajuste

e tornando o método computacionalmente eficiente.

O método segue algumas etapas principais: (i) pré-processamento e normalização dos dados; (ii) varredura em períodos e dobramento em fase (*phase folding*); (iii) binagem dos dados (*binning*), reduzindo o ruído; (iv) ajuste de um modelo em caixa às curvas de luz; e (v) construção de um espectro de potência, no qual o melhor período é identificado pelo maior pico. A significância da detecção é avaliada por meio da estatística Signal Detection Efficiency (SDE), sendo valores com relação sinal-ruído efetiva $\alpha > 6$ considerados robustos (Kovacs; Zucker; Mazeh, 2002).

Apesar de sua eficiência, o BLS apresenta limitações importantes. A principal delas é a simplificação do trânsito como uma função em caixa, o que ignora efeitos físicos relevantes, como o escurecimento de limbo (*limb darkening*) e as fases de ingresso e egresso. Além disso, o método pode perder sensibilidade para sinais de baixa amplitude e é suscetível a efeitos de discretização decorrentes da binagem, bem como a instabilidades na fase (Kovacs; Zucker; Mazeh, 2002).

Missões como a HATNet realizam o monitoramento fotométrico contínuo de campos estelares, produzindo curvas de luz para dezenas de milhares de estrelas simultaneamente. Essas curvas são então analisadas utilizando o algoritmo BLS, que permite a identificação de sinais periódicos característicos de trânsitos planetários, ou seja, pequenas quedas no brilho estelar causadas pela passagem de um planeta em frente à sua estrela hospedeira (Pál *et al.*, 2008).

Assim, embora seja robusto na identificação de periodicidades, o BLS possui limitações na descrição detalhada do sinal de trânsito, o que motiva o uso de métodos mais sofisticados, como o Transit Least Squares (TLS).

Embora o algoritmo Box Least Squares (BLS) não tenha sido aplicado diretamente nos procedimentos computacionais deste trabalho, sua apresentação é fundamental para compreender a evolução dos métodos de detecção de trânsitos e as limitações que motivaram o desenvolvimento do TLS, utilizado na etapa empírica desta pesquisa.

2.8 Transit Least Squares

Avançando no cronograma de desenvolvimento de ferramentas astronômicas, surge em 2019 o algoritmo TLS. Desenvolvido pelos pesquisadores Michael Hippke e René Heller, o TLS foi projetado com o objetivo explícito de superar as limitações intrínsecas do *BLS* na detecção de planetas de pequeno porte (Hippke; Heller, 2019). Embora o BLS tenha sido amplamente adotado

na literatura desde (Kovacs; Zucker; Mazeh, 2002), e sua influência se estenda a classificadores modernos, como a rede neural *AstroNet* (Shallue; Vanderburg, 2018), sua premissa de modelar o trânsito como uma "função de caixa"(*boxcar*) ignora a complexidade física do evento.

Como discutido anteriormente, a aproximação geométrica do BLS negligencia o escurecimento de limbo estelar e as fases graduais de ingresso e egresso do planeta. Essa simplificação introduz um ruído sistemático que pode mascarar sinais débeis, especialmente de exoplanetas com dimensões semelhantes à Terra. Para mitigar esse problema, Hippke e Heller buscaram criar um método baseado em um modelo de trânsito fisicamente realista (Hippke; Heller, 2019).

A filosofia de desenvolvimento do TLS priorizou a sensibilidade estatística em detrimento da velocidade computacional extrema, visando maximizar o retorno científico de missões como *Kepler*, TESS e a futura missão Planetary Transits and Oscillations of stars (PLATO). Essa mudança de paradigma foi viabilizada pelo expressivo crescimento do poder de processamento das CPUs desde o início do século XXI. O que em 2002 exigia simplificações severas para viabilizar o processamento, hoje permite que cálculos complexos (como a análise de dados sem agrupamento (*unbinned*)) sejam realizados em tempos razoáveis em computadores convencionais (Hippke; Heller, 2019). O processo segue estas etapas:

1. **Template físico e escurecimento de limbo:** diferentemente do BLS, que utiliza uma aproximação simplificada em forma de caixa, o TLS emprega um modelo físico mais realista para descrever o trânsito. Esse modelo é baseado nas leis de Mandel & Agol (Hippke; Heller, 2019) e incorpora efeitos astrofísicos relevantes, como o escurecimento de limbo (*limb darkening*) e as fases suaves de ingresso e egresso. Além disso, o algoritmo utiliza um *template* construído a partir da morfologia média de trânsitos de planetas pequenos confirmados, permitindo uma representação mais fiel do sinal esperado.

2. **Definição da grade de parâmetros e amostragem:** o TLS realiza a busca por sinais periódicos explorando um espaço de parâmetros que inclui o período orbital, a época do trânsito e sua duração. Essa busca é restringida por critérios físicos, utilizando estimativas de massa e raio estelar provenientes de catálogos observacionais, o que limita a análise a durações de trânsito compatíveis com sistemas planetários reais. Esse procedimento reduz o espaço de busca e aumenta a eficiência do algoritmo.

3. **Dobramento e otimização de memória:** para cada período testado, a curva de luz é convertida para o domínio de fase por meio do *phase folding*, permitindo a sobreposição

de múltiplos eventos periódicos. A organização dos dados é realizada com o algoritmo *Mergesort*, escolhido por sua eficiência computacional, o que contribui para a redução do tempo total de processamento.

4. **Análise não agrupada (*unbinned*):** ao contrário do BLS, o TLS não realiza a discretização dos dados em *bins*. Em vez disso, a análise é feita diretamente sobre todos os pontos da curva de luz, preservando a estrutura fina do sinal. Essa característica é especialmente importante para a detecção de trânsitos rasos, cuja morfologia pode ser degradada por processos de agrupamento.

5. **Minimização por χ^2 e escalonamento:** o ajuste entre o modelo teórico e os dados observacionais é realizado ponto a ponto, utilizando uma métrica baseada no qui-quadrado. Para otimizar o desempenho computacional, o TLS utiliza modelos pré-calculados e ajusta analiticamente apenas a profundidade do trânsito, reduzindo significativamente o número de operações necessárias durante a busca.

6. **Signal residue (SR) e SDE:** a qualidade do ajuste é quantificada por meio do *Signal Residue*, que mede o grau de correspondência entre o modelo e os dados. A partir dessa métrica, calcula-se o (SDE), definido de forma análoga ao utilizado no BLS, ou seja, é utilizado para avaliar a significância estatística dos sinais detectados e identificar candidatos robustos a trânsitos planetários.

O fluxograma do algoritmo TLS, ilustrando as etapas descritas acima, encontra-se no Apêndice A.

2.8.1 Otimizações Computacionais do TLS

O TLS incorpora estratégias de otimização voltadas à redução do custo computacional da busca por sinais periódicos. Uma das principais abordagens é o uso de compilação *Just-in-Time* (JIT), implementada por meio da biblioteca *Numba* (Hippke; Heller, 2019).

Diferentemente da execução interpretada padrão em Python, a compilação JIT converte trechos críticos do código para linguagem de máquina em tempo de execução, permitindo desempenho próximo ao de linguagens compiladas como C. Conforme descrito por (Hippke; Heller, 2019), a implementação de referência do TLS é escrita em Python puro, mas as partes críticas em termos de tempo de processamento foram implementadas com o pacote *Numba*, que traduz o código Python para código de máquina por meio de compilação *just-in-time*, resultando

em um ganho de aproximadamente duas ordens de grandeza no tempo de computação (Hippke; Heller, 2019).

Além disso, o TLS adota otimizações adicionais no processo de *phase folding* e ordenação dos dados. Para cada período orbital testado, o algoritmo dobra a curva de luz em fase e ordena os dados, uma operação computacionalmente intensiva. Os autores identificaram que o algoritmo *Mergesort* é consistentemente cerca de 20% mais rápido que o *Quicksort* para essa finalidade específica, sendo portanto adotado na implementação do TLS (Hippke; Heller, 2019).

Outra estratégia importante diz respeito ao armazenamento em *cache* de modelos de trânsito pré-calculados. Em vez de recalcular o modelo de trânsito para cada combinação de parâmetros, o TLS pré-computa e armazena em memória os modelos para todas as durações de trânsito testadas, escalando-os em profundidade dinamicamente durante a busca. Essa abordagem reduz operações repetitivas, utilizando uma relação de compromisso entre tempo de processamento e memória RAM (Hippke; Heller, 2019).

2.8.2 Limitações e Compromissos Técnicos do TLS

Embora o *Transit Least Squares* represente um avanço na detecção de sinais débeis, a sua aplicação prática impõe desafios que devem ser ponderados, especialmente em grandes volumes de dados ou cenários de ruído complexo:

- **Carga de processamento e escalabilidade:** o TLS exige um esforço computacional que escala de forma quadrática (S^2) em relação à duração da série temporal. Ao abdicar do agrupamento de dados (*binning*) para preservar a resolução morfológica, o algoritmo demanda um processamento intensivo da CPU, o que pode tornar a análise morosa em missões de longa duração sem o suporte de aceleração por *hardware* ou compiladores JIT (Hippke; Heller, 2019).
- **Rigidez do modelo e viés geométrico:** a eficácia do algoritmo está intrinsecamente ligada à semelhança entre o sinal real e o **modelo circular padrão**. Trânsitos que fogem à norma, como órbitas altamente excêntricas ou configurações de impacto que resultam em curvas em forma de "V" (trânsitos rasantes), apresentam um desajuste que reduz a métrica SDE e pode levar à não detecção (Hippke; Heller, 2019).
- **Dependência cromática:** a vantagem competitiva do TLS sobre modelos simplificados reduz-se conforme o comprimento de onda aumenta. No regime do infravermelho próximo, o fenômeno do escurecimento de limbo é menos proeminente, tornando o perfil do trânsito

mais abrupto e similar a uma caixa. Nesses casos, a complexidade matemática do TLS pode não justificar o custo computacional, já que o BLS oferece resultados comparáveis (Hippke; Heller, 2019).

- **Vulnerabilidade a ruídos não-brancos:** o cálculo da significância do sinal (SDE) parte do pressuposto de um ruído puramente aleatório. No entanto, curvas de luz reais frequentemente apresentam ruído causado por manchas estelares ou oscilações instrumentais. Nesses cenários, o TLS pode superestimar a confiança de uma detecção, exigindo uma análise de resíduos mais criteriosa (Hippke; Heller, 2019).
- **Efeitos de amostragem e integração:** o algoritmo não corrige automaticamente a distorção temporal que ocorre porque o instrumento coleta luz durante um intervalo fixo para cada ponto de dado (cadência). Esse tempo de exposição contínuo acaba "misturando"diferentes níveis de brilho se o sinal variar rapidamente durante a integração. Em missões com cadência lenta, a integração da luz durante a exposição suaviza as bordas do trânsito, criando uma discrepância entre o modelo físico idealizado pelo TLS e a realidade observada pelo telescópio. (Hippke; Heller, 2019)

Esses compromissos técnicos não inviabilizam o uso do TLS, mas indicam que sua aplicação requer escolhas conscientes por parte do pesquisador. Em contextos de alta precisão fotométrica e busca por planetas de pequeno porte, o ganho em sensibilidade justifica o custo computacional. No entanto, em cenários de ruído correlacionado ou comprimentos de onda mais longos, o BLS ainda pode ser uma alternativa mais eficiente. A escolha entre os métodos deve, portanto, considerar as características específicas dos dados e os objetivos científicos da análise.

Em síntese, o BLS e o TLS representam abordagens complementares: enquanto o primeiro prioriza eficiência computacional e robustez na detecção inicial, o segundo maximiza a sensibilidade a sinais de baixa amplitude por meio de uma modelagem física mais realista.

3 METODOLOGIA

3.1 Delineamento e classificação da pesquisa

A presente pesquisa possui natureza mista (qualitativa e quantitativa), com ênfase predominante na análise conceitual e interpretativa do método de trânsito planetário e de sua aplicação no contexto da astronomia contemporânea, especialmente diante do advento da Big Data. O componente quantitativo restringe-se aos procedimentos computacionais com o algoritmo TLS, que geram métricas numéricas (SDE, períodos, durações) utilizadas como evidências ilustrativas para apoiar a discussão teórica, sem pretensão de inferência estatística ou generalização. O estudo busca compreender e discutir a importância das tecnologias computacionais e da ciência de dados no avanço da astronomia e da astrofísica, bem como descrever os fundamentos físicos que sustentam o método de trânsito.

Para fundamentar experimentalmente as discussões teóricas sobre os desafios da análise de grandes volumes de dados astronômicos, foram conduzidos procedimentos computacionais com o algoritmo TLS, aplicado ao arquivo *KOI template* disponibilizado pela NASA. O objetivo desses procedimentos não é a realização de uma busca sistemática por novos exoplanetas. O propósito restringe-se a analisar, em um contexto prático, os desafios discutidos ao longo do trabalho, tais como o tempo de processamento em grandes volumes de dados, a ocorrência de falsos positivos mesmo em sinais com valores elevados de SDE e a necessidade de filtros automatizados para tornar a análise viável em larga escala. Dessa forma, o uso do TLS atua como um complemento empírico à análise teórica, permitindo observar, em dados reais, os fenômenos e limitações abordados conceitualmente. Trata-se, portanto, de uma etapa metodológica que integra a análise qualitativa, cujo papel é apoiar e sustentar as discussões teóricas por meio de evidências concretas, sem se configurar como uma busca exaustiva ou conclusiva.

Adicionalmente, ressalta-se que os resultados obtidos não possuem caráter de inferência estatística (não há testes de hipótese, intervalos de confiança ou generalização para a população de exoplanetas), devendo ser interpretados como estatísticas descritivas que evidenciam empiricamente os desafios práticos associados à análise de dados astronômicos em larga escala.

Quanto aos seus objetivos, a pesquisa caracteriza-se como exploratória e descritiva. Exploratória, por investigar e sistematizar conhecimentos sobre o método de trânsito planetário e suas abordagens modernas de análise de dados; e descritiva, por apresentar e organizar os principais conceitos físicos, matemáticos e computacionais envolvidos nesse método, conforme

abordados na literatura científica. A componente quantitativa, embora presente, não busca estabelecer relações causais ou validar hipóteses estatísticas, servindo exclusivamente como suporte empírico à análise qualitativa.

Do ponto de vista dos procedimentos técnicos, a pesquisa combina abordagens bibliográfica e experimental. A fundamentação teórica foi desenvolvida exclusivamente a partir da análise de fontes secundárias, incluindo livros clássicos de mecânica celeste e astrofísica, artigos científicos publicados em periódicos especializados, bem como documentação técnica relacionada a algoritmos empregados na detecção de exoplanetas. Complementarmente, os procedimentos com o TLS constituíram uma análise prática controlada, conduzida com o propósito de examinar, em dados reais, os aspectos discutidos na revisão bibliográfica, não tendo pretensão de generalização ou validação estatística.

Dessa forma, o estudo adota uma abordagem predominantemente teórica, com um componente prático de apoio. O foco principal está na revisão, na discussão crítica e na integração de conhecimentos provenientes da física, da astronomia e da ciência de dados. Essa metodologia permite compreender de que maneira o método de trânsito planetário, aliado ao processamento automatizado de grandes volumes de dados, desempenha um papel fundamental na astronomia moderna e na descoberta de exoplanetas.

3.2 Procedimentos Computacionais com o TLS

3.2.1 Objetivo

Embora a presente pesquisa tenha caráter predominantemente bibliográfico e qualitativo, foram realizados procedimentos computacionais com o algoritmo TLS aplicados a dados da missão *Kepler*. O objetivo desses procedimentos não consiste na busca sistemática por novos exoplanetas, nem na construção de modelos de aprendizado de máquina, mas sim na investigação, em contexto prático, do comportamento do algoritmo diante de grandes volumes de dados reais.

Busca-se observar como desafios amplamente discutidos na literatura (como custo computacional, presença de ruído, ocorrência de falsos positivos e necessidade de automação) se manifestam durante o processamento de dados astronômicos em larga escala.

Os experimentos resultam na geração de curvas de luz processadas e métricas associadas, utilizadas nesta pesquisa como material empírico para análise qualitativa dos desafios computacionais e observacionais envolvidos na detecção de exoplanetas.

3.2.2 Fonte e Seleção dos Dados

Os dados utilizados foram obtidos a partir do catálogo KOI, disponibilizado pelo (Aeronautics; Archive, 2026), no formato *koi_template.csv*. Para os procedimentos computacionais, foram considerados apenas objetos classificados como *CONFIRMED* e *FALSE POSITIVE*,

As curvas de luz foram obtidas a partir do Mikulski Archive for Space Telescopes (MAST) utilizando a biblioteca `lightkurve`. Apenas objetos com dados disponíveis foram considerados.

Foram aplicados critérios mínimos de qualidade para garantir que apenas curvas de luz confiáveis fossem processadas:

- **Mínimo de 1.500 pontos válidos:** curvas de luz com menos pontos teriam pouca informação para uma detecção confiável, pois o número insuficiente de medições dificulta a identificação de um padrão periódico de trânsito.
- **Cobertura temporal superior a 5%:** Este critério mede a fração do tempo total de observação que realmente contém medições válidas. Uma cobertura muito baixa indica muitos dados perdidos (por falhas instrumentais, gaps na observação, etc.), o que compromete a detecção. O valor foi calculado dividindo-se o número de pontos válidos pelo número máximo teórico de amostras possível no intervalo de tempo observado.
- **Curva com variação significativa de fluxo:** Foram descartadas curvas de luz que apresentavam desvio padrão nulo, ou seja, sem nenhuma variação de brilho ao longo do tempo. Esses dados geralmente indicam arquivos corrompidos, estrelas constantes sem atividade ou problemas no download, sendo inviáveis para qualquer detecção de trânsito.

3.2.3 Procedimentos de Processamento

O algoritmo TLS foi empregado em ambiente Python 3.11, com uso das bibliotecas `lightkurve`, `numpy`, `pandas` e `astropy`. O processamento seguiu as etapas descritas a seguir.

Inicialmente, as curvas de luz foram submetidas a um processo de agregação temporal com o objetivo de reduzir ruído e custo computacional. Este processo consiste em agrupar pontos consecutivos da série temporal e substituí-los por sua média, reduzindo o número total de pontos da curva.

O algoritmo TLS foi então aplicado às séries temporais processadas. O intervalo de busca de períodos foi definido dinamicamente:

- Na ausência de informação prévia, o intervalo varia de 0,5 dias até metade do intervalo

temporal da curva de luz;

- Quando disponível, o período orbital conhecido (informação prévia) é utilizado para restringir a busca a uma janela estreita (detalhada adiante);

O valor mínimo de período também foi ajustado para garantir consistência com a resolução temporal, respeitando a condição $P_{\text{min}} \geq 2\times$ agregação temporal.

Ressalta-se que nenhum dos parâmetros apresentados foi definido de forma definitiva no início do desenvolvimento. Ao longo de meses de experimentação, os valores foram sendo reiteradamente ajustados com base em testes realizados em subamostras representativas de curvas de luz, buscando um equilíbrio entre sensibilidade, especificidade e custo computacional. A configuração final aqui documentada reflete, portanto, o resultado desse processo empírico e iterativo.

A métrica principal utilizada para avaliação dos sinais foi a **Eficiência de Detecção de Sinal** (SDE), que quantifica a significância estatística do sinal detectado em relação ao ruído de fundo. Foram adotados dois critérios distintos:

- Busca cega: $SDE \geq 5{,}0$;
- Busca guiada por *informação prévia*: $SDE \geq 3{,}0$.

Sinais que não atendem ao respectivo limiar (SDE < 5,0 para busca cega; SDE < 3,0 para busca guiada) são descartados e não prosseguem para as etapas de geração das curvas de luz dobradas e empilhadas.

Busca guiada por informação prévia: Quando o período orbital do objeto é conhecido a partir do catálogo KOI (por exemplo, por confirmação prévia via velocidade radial), este valor é utilizado como período conhecido para restringir o intervalo de busca do TLS. A janela de busca é centrada no valor do período conhecido com largura proporcional de até 30% do período, respeitando os limites globais de $P_{\text{min}} = 0,1$ dias e $P_{\text{max}} = 750$ dias. Essa estratégia reduz o espaço de parâmetros a ser explorado pelo algoritmo, diminuindo o custo computacional e aumentando a sensibilidade na vizinhança do período esperado. O critério de aceitação para objetos com hint é menos restritivo (SDE $\geq$ 3,0) por tratar-se de sinais já conhecidos na literatura.

Após a detecção do período, as curvas de luz foram dobradas para geração da global view, posteriormente normalizada para média zero e desvio padrão unitário.

A local view foi construída por meio da extração e empilhamento dos trânsitos individuais, centrados no evento detectado, utilizando interpolação para padronização do número de amostras (3.000 pontos) e uma janela de extração proporcional à duração do trânsito (fator de meia janela

= 4,0) onde o fator 4,0 significa que a janela de extração estende-se por um intervalo equivalente a quatro vezes a duração do trânsito para cada lado do centro do evento, garantindo a inclusão de todo o trânsito e de partes suficientes da linha de base adjacente para estimativa do fluxo fora do trânsito.

3.2.4 Descrição dos Parâmetros do TLS

Os principais parâmetros do algoritmo TLS utilizados neste trabalho são descritos na Tabela 2, com ênfase em seu papel na detecção de sinais de trânsito e no desempenho computacional do método.

3.2.5 Limitações e Escopo

Os procedimentos descritos não constituem uma análise estatística exaustiva, mas sim uma investigação qualitativa apoiada em dados reais.

Limitações observadas incluem:

- Restrições de memória no processamento de curvas de luz extensas;
- Impossibilidade de processamento ininterrupto, devido ao uso do computador para outras finalidades;
- Tempo elevado de execução em uma grade de busca maior, exigindo múltiplos ajustes nos parâmetros ao longo do processamento;
- Redução da grade de busca de parâmetros, o que pode ter resultado em menor qualidade nas curvas de luz;
- Além disso, algumas curvas de luz apresentaram lacunas significativas nos dados, resultando em global views e local views de baixa qualidade, com formas irregulares que dificultam a identificação de trânsitos planetários.

Adicionalmente, optou-se pelo uso direto do TLS, sem a etapa prévia com o algoritmo Box-fitting Least Squares (BLS). Esta decisão baseia-se no foco qualitativo da análise, que prioriza a ilustração de desafios computacionais em detrimento da exaustividade na detecção. O TLS já incorpora internamente uma versão otimizada do BLS, de modo que a ausência de uma etapa separada não compromete a capacidade de detecção para os propósitos deste estudo.

O uso de agregação temporal com resolução de 0,01 dias foi adotado apenas na fase final do processamento, como parte do ajuste iterativo dos parâmetros para aumentar a resolução da grade de busca. Durante as etapas anteriores, valores de binning maiores foram utilizados

Tabela 2 – Parâmetros do algoritmo TLS

Parâmetro	**Valor**	**Descrição**
SDE	—	Métrica central que quantifica a significância estatística do sinal detectado em relação ao ruído de fundo. Valores mais elevados indicam maior probabilidade de um evento físico real.
Fator de superamostragem	3	Controla a resolução da grade de períodos testados. Valores mais elevados aumentam a densidade de amostragem, favorecendo sinais sutis, mas com maior custo computacional.
Agregação temporal	0,01 dias	Agrupa pontos consecutivos da série temporal, reduzindo ruído e volume de dados. Pode suavizar características finas do trânsito (ingressos e egressos). Conforme discutido na Seção 3.2.7, mostrou-se inefetivo na prática.
Remoção de tendências (*flatten*)	Janela de 501 pontos	Remove variações de longo período da curva de luz, preservando os sinais transitórios de interesse (trânsitos planetários).
Profundidade mínima de trânsito	10^{-5}	Menor decréscimo relativo no fluxo considerado pelo modelo, permitindo a detecção de sinais rasos (exoplanetas de pequeno porte).
Número mínimo de trânsitos	2	Exigência de pelo menos dois eventos periódicos para validação, reduzindo falsos positivos por flutuações aleatórias.
Passo da grade de duração	1,1	Controla a resolução da busca pela duração do trânsito, formando uma grade geométrica com precisão relativa constante para trânsitos curtos e longos.
Período conhecido (*hint*)	—	Permite restringir o intervalo de busca em torno de valores conhecidos, reduzindo custo computacional e aumentando a sensibilidade.

Fonte: Autoria própria (2026).

e mostraram-se efetivos na redução de ruído e custo computacional. No entanto, conforme observado por (Hippke; Heller, 2019), o uso de binning maior (ou seja, com resolução mais grossa) pode suavizar as fases de ingresso e egresso do trânsito, comprometendo a qualidade da curva de luz e reduzindo a sensibilidade do algoritmo para planetas pequenos. Por esta razão, optou-se por reduzir o valor do binning ao longo do processamento, chegando ao valor final de 0,01 dias. Este valor final, no entanto, mostrou-se inefetivo por ser inferior à cadência original da missão Kepler (aproximadamente 29,4 minutos ou 0,0204 dias), conforme a Tabela 1 de (Koch *et al.*, 2010). Consequentemente, os dados foram processados em sua resolução temporal original, e nenhuma redução significativa adicional de ruído ou de custo computacional foi alcançada com este parâmetro final. A limitação referente ao valor de 0,01 dias é assumida, e os resultados devem ser interpretados à luz da manutenção da cadência original da missão.

4 RESULTADOS

4.1 Caracterização do conjunto de dados

O conjunto de dados analisado é composto por 6807 registros provenientes do catálogo de Objetos de Interesse Kepler (KOIs), após remoção de objetos classificados como candidatos. A amostra final inclui 2694 planetas confirmados e 4113 falsos positivos.

Conforme mostrado na tabela 3, há um desbalanceamento significativo entre as classes, com predominância de falsos positivos (60,4%). Esse cenário reflete um dos principais desafios da exoplanetologia moderna: o tratamento de grandes volumes de dados nos quais apenas uma fração corresponde a detecções reais.

Tabela 3 – Distribuição da amostra por classificação

Classificação	Quantidade	Proporção (%)
CONFIRMADOS	2694	39,6
FALSOS POSITIVOS	4113	60,4

Fonte: MAST/NASA (2026).

[1] Nesse contexto, conforme discutido por Zhang & Zhao, a astronomia insere-se no paradigma do Big Data, caracterizado não apenas pelo grande volume de dados, mas também por aspectos como veracidade e validade. A presença majoritária de falsos positivos na amostra evidencia a necessidade de métodos automatizados capazes de filtrar e classificar sinais com alta confiabilidade.

4.1.1 Desafios da big data observados na amostra

Os dados analisados neste trabalho exemplificam diretamente os desafios do Big Data na astronomia discutidos na seção 2.6. Três dimensões propostas por Zhang & Zhao manifestam-se de forma particularmente evidente:

Veracidade: a tabela 3 revela que 60,4% dos objetos no catálogo KOI são falsos positivos. Conforme apontado na fundamentação teórica, o ruído intrínseco das observações e as flutuações estelares naturais (como manchas na superfície da estrela) podem mimetizar trânsitos reais. A análise da seção 4.3.2 demonstra que falsos positivos alcançam SDE tão elevados quanto planetas

[1] As métricas estatísticas apresentadas neste capítulo (médias, mediana) têm caráter puramente descritivo, restrito a esta amostra, sem pretensão de inferência ou generalização.

confirmados (até 289,69), evidenciando que a veracidade é um dos principais obstáculos para a automação da detecção.

Variedade: a necessidade de integrar diferentes fontes de informação fica evidente na metodologia adotada: o pipeline combina metadados do catálogo KOI (períodos orbitais, classificações) com séries temporais de brilho obtidas do MAST, além de parâmetros estelares utilizados para restringir a grade de períodos do TLS. Essa heterogeneidade de dados é característica da astronomia moderna.

Volume: o processamento das 6807 curvas de luz (cada uma com centenas a milhares de pontos temporais demandou aproximadamente 4 meses de tempo real (seção 4.3.4)), ilustrando o desafio de escala mesmo para uma amostra modesta em comparação com missões como LSST (que produzirá 200 PB de dados, conforme a tabela 1).

Esses resultados empíricos confirmam a afirmação de Zhang & Zhao de que a astronomia entrou na era da big data, onde o tratamento do ruído e a automação não são conveniências, mas requisitos estruturais para a produção de ciência.

4.2 Análise estatística dos parâmetros

4.2.1 SDE

A análise do SDE revelou diferenças importantes entre as classes, embora com sobreposição significativa. A tabela 4 resume as estatísticas descritivas.

Tabela 4 – Estatísticas do SDE por classe

Classe	Média	Mínimo	Máximo
CONFIRMADOS	41,85	3,43	250,99
FALSOS POSITIVOS	46,47	3,93	289,69

Fonte: MAST/NASA (2026).

Observa-se que a média do SDE para falsos positivos (46,47) é ligeiramente superior à dos confirmados (41,85), contradizendo a expectativa intuitiva de que sinais reais apresentariam maior significância estatística. Esse resultado evidencia que o SDE, isoladamente, não discrimina de forma confiável planetas reais de falsos positivos.

As tabelas 5 e 6 apresentam os dez maiores valores de SDE em cada classe.

Tabela 5 – Maiores valores de SDE para planetas confirmados

Objeto	Período (dias)	SDE
K00002.01	2,20	250,99
K01916.03	3,34	227,20
KIC9941662	1,76	208,59
KIC8191672	3,55	202,90
K00776.01	3,73	202,74
K00010.01	3,52	198,81
K00760.01	4,96	181,23
KIC7778437	5,01	180,04
KIC9885417	0,60	173,17
K00767.01	2,82	167,93

Fonte: MAST/NASA (2026).

Tabela 6 – Maiores valores de SDE para falsos positivos

Objeto	Período (dias)	SDE
KIC8573168	0,90	289,69
KIC12403968	0,86	248,22
KIC7031430	0,82	244,15
K00743.01	19,40	243,38
K04339.01	0,99	224,92
KIC6778008	1,42	214,59
KIC9101208	1,81	213,38
K01232.02	119,40	205,68
KIC3228986	0,73	203,14
KIC8560285	14,22	200,27

Fonte: MAST/NASA (2026).

A sobreposição entre as classes é evidente: falsos positivos alcançam SDE tão elevados quanto (e até superiores aos) planetas confirmados. O maior SDE entre falsos positivos (289,69) supera o maior SDE entre confirmados (250,99). Esse resultado sugere que o TLS, embora sensível, gera detecções de alta significância estatística tanto para sinais reais quanto para artefatos instrumentais ou binárias eclipsantes.

4.2.2 Período orbital

Tabela 7 – Estatísticas do período orbital (dias)

Classe	Média	Mediana
CONFIRMADOS	31,09	12,30
FALSOS POSITIVOS	53,33	18,70

Fonte: MAST/NASA (2026).

Como mostra a tabela 7 falsos positivos apresentam período médio (53,33 dias) superior ao dos confirmados (31,09 dias). Esse comportamento pode estar associado à maior dificuldade de validação de sinais de longo período, que exigem mais tempo de observação para confirmação, ou à presença de configurações orbitais específicas em sistemas binários eclipsantes.

4.2.3 Duração do trânsito

Tabela 8 – Estatísticas da duração do trânsito (dias)

Classe	Média	Mediana
CONFIRMADOS	0,0787	0,0428
FALSOS POSITIVOS	0,0917	0,0231

Fonte: MAST/NASA (2026).

A tabela 8 apresenta as estatísticas da duração do trânsito (duration_days) por classe. Observa-se que os falsos positivos apresentam média de duração (0,0917 dias) superior à dos confirmados (0,0787 dias), enquanto a mediana dos falsos positivos (0,0231 dias) é menor que a dos confirmados (0,0428 dias), indicando uma distribuição assimétrica.

4.3 Desempenho do algoritmo Transit Least Squares

O algoritmo Transit Least Squares (TLS) foi aplicado a todas as 6807 curvas de luz da amostra. Conforme discutido na fundamentação teórica (seção 2.8), o TLS foi desenvolvido por Hippke & Heller com o objetivo explícito de superar as limitações do algoritmo Box Least-Squares (BLS), particularmente no que se refere à modelagem do escurecimento de limbo e das fases graduais de ingresso e egresso do trânsito.

Embora o BLS tenha sido o algoritmo padrão por muito tempo, sendo utilizado, por exemplo, na etapa de triagem inicial usada em classificadores como o AstroNet, a literatura recente recomenda o TLS para buscas que priorizam sensibilidade a planetas de pequeno porte, ainda que com maior custo computacional (Hippke; Heller, 2019). Neste trabalho, optou-se pelo TLS exatamente por sua maior fidelidade física, ainda que sem uma comparação quantitativa direta com o BLS, uma vez que o objetivo principal é analisar os desafios de detecção em grandes volumes de dados, e não estabelecer uma competição entre algoritmos.

O processamento das curvas de luz foi realizado por meio de um pipeline automatizado, conforme descrito na seção 3.2. Um trecho representativo da execução, evidenciando a etapa de

filtragem dos dados e a organização da amostra utilizada neste trabalho, encontra-se no Apêndice B.

4.3.1 Análise morfológica: casos ilustrativos

As figuras 3 e 4 apresentam as curvas de luz do planeta confirmado Kepler-2 b (K00002.01), enquanto as Figuras 5 e 6 apresentam as curvas de luz do falso positivo K07010.01 (KIC 8288719). Ambos os objetos apresentam SDE elevado.

O planeta confirmado selecionado foi o **Kepler-2 b** (K00002.01), também conhecido como HAT-P-7 b. Descoberto em 2008 pelo projeto HATNet, este objeto é um Júpiter quente (*Hot Jupiter*) com período orbital de 2,20 dias, massa de aproximadamente 1,8 vezes a massa de Júpiter e raio de cerca de 1,6 vezes o raio de Júpiter (Pál *et al.*, 2008). Sua curva de luz (figuras 3 e 4) exibe o formato em "U" característico de um trânsito planetário, com ingressos e egressos suaves resultantes do efeito de escurecimento de limbo estelar.

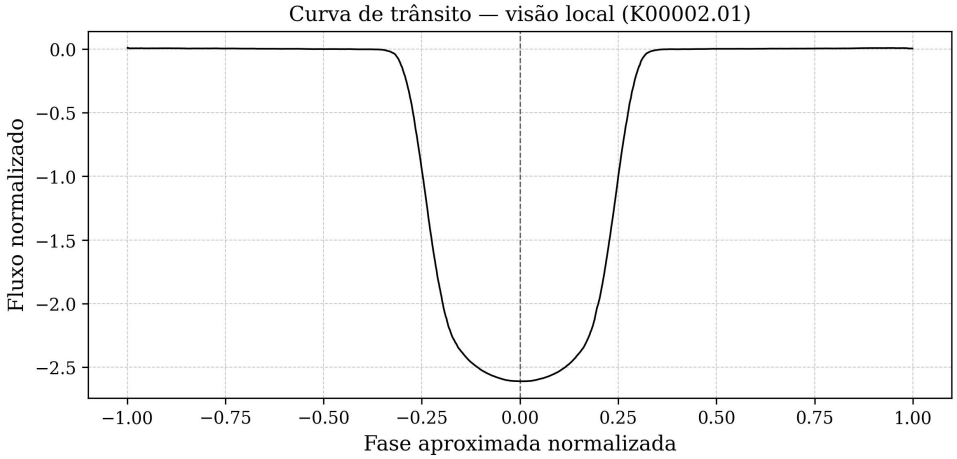
O falso positivo selecionado foi o **K07010.01** (KIC 8288719). A aplicação do algoritmo *Transit Least Squares* (TLS) a esse KOI indicou um período orbital extremamente curto, de aproximadamente 0,76 dias (cerca de 18 horas), além de um valor elevado de SDE igual a 118,79. No entanto, é importante considerar que a estimativa do período depende diretamente dos parâmetros adotados na execução do algoritmo, o que pode levar a discrepâncias em relação aos valores reportados em bases oficiais, como o NASA Exoplanet Archive.

A curva de luz correspondente (figuras 5 e 6) apresenta um formato em "V" mais abrupto, característico de sistemas binários eclipsantes, nos quais o objeto secundário provoca uma obstrução mais rápida e profunda da luz estelar em comparação com trânsitos planetários. Períodos tão curtos, quando associados a quedas de brilho mais pronunciadas, são geralmente indicativos da presença de companheiros estelares (como anãs M ou anãs brancas) e não de planetas.

Observa-se que, embora ambos os objetos tenham SDE elevado (250,99 para Kepler-2 b e 118,79 para K07010.01), a morfologia do trânsito é distintamente diferente. A curva de luz do exoplaneta apresenta o formato em "U" esperado, enquanto a curva do falso positivo exibe o formato em "V" característico de binárias eclipsantes. Essa diferença morfológica, embora perceptível ao olho humano, é difícil de ser capturada automaticamente por algoritmos baseados apenas em ajuste de modelo, como o TLS. Conforme discutido por Shallue & Vanderburg , redes neurais convolucionais (como a AstroNet) são capazes de aprender esses padrões morfológicos

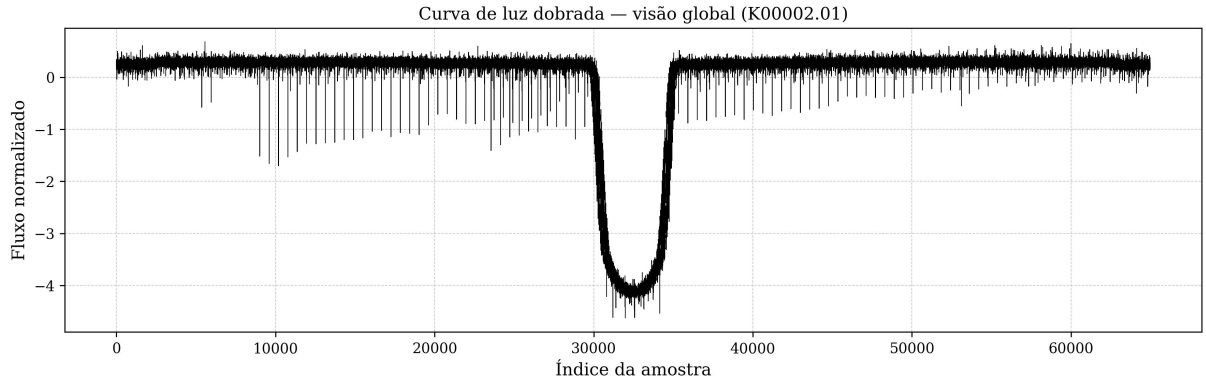
e atuar como uma segunda etapa de validação, aumentando a especificidade do pipeline de detecção.

Figura 3 – Local view do planeta confirmado Kepler-2 b (K00002.01), evidenciando o formato em “U” característico do trânsito planetário.



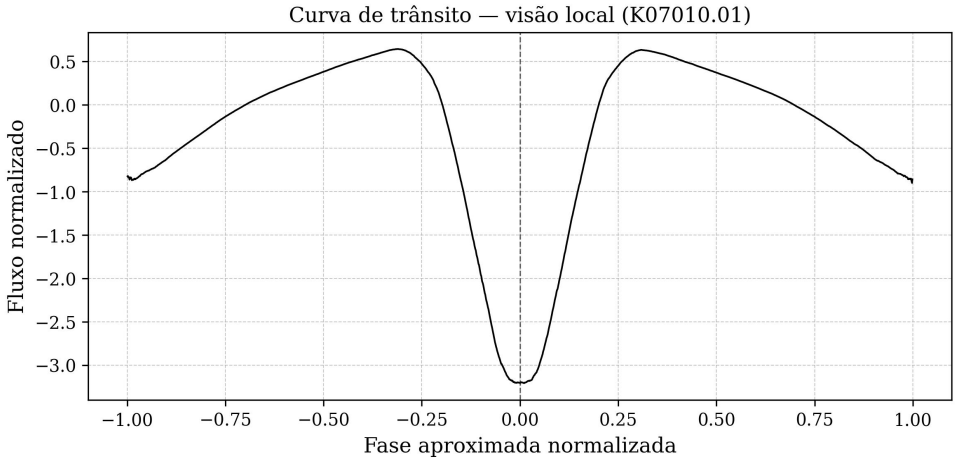
Fonte: Kepler (2026).

Figura 4 – Global view do planeta confirmado Kepler-2 b (K00002.01), mostrando os trânsitos periódicos a cada 2,20 dias.



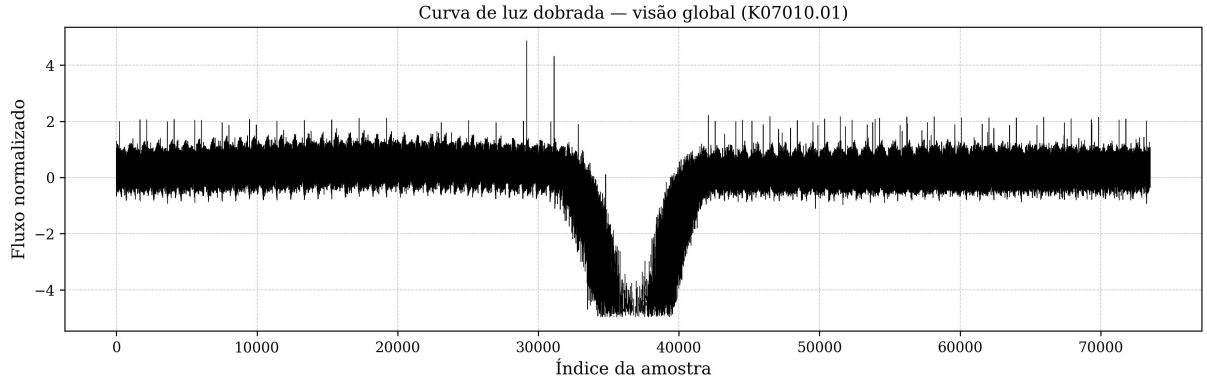
Fonte: Kepler (2026).

Figura 5 – Local view do falso positivo K07010.01, mostrando o formato em “V” típico de sistemas binários eclipsantes.



Fonte: Kepler (2026).

Figura 6 – Global view do falso positivo K07010.01, mostrando o período extremamente curto de 0,76 dias.



Fonte: Kepler (2026).

A análise morfológica aqui apresentada é qualitativa e ilustrativa, baseada na inspeção visual de dois casos extremos. Não foram extraídas métricas quantitativas de morfologia (como assimetria ou curtose), uma vez que o objetivo é meramente exemplificar os padrões observados, não classificar automaticamente os sinais.

4.3.2 Tempo de processamento

O tempo total de processamento das 6807 curvas de luz com o TLS foi de aproximadamente 4 meses em hardware convencional (Ryzen 5, 8 GB RAM, 256 GB SSD), considerando o tempo real de execução. Esse período incluiu interrupções, reinicializações do pipeline, limita-

ções de memória que eventualmente exigiam o reprocessamento de alguns objetos, e a execução em intervalos ao longo de vários meses conforme a disponibilidade de recursos computacionais. Não foi possível mensurar separadamente o tempo gasto em cada etapa do processamento (pré-processamento, detrending, execução do TLS, etc.), uma vez que o pipeline foi executado de forma integrada e sujeito a condições variáveis de hardware e interrupções. O objetivo do trabalho não é a avaliação de desempenho bruto do algoritmo, mas sim a identificação de desafios práticos em cenário real de processamento.

Conforme reportado por Hippke & Heller , o TLS é computacionalmente mais custoso que o BLS, com diferença estimada de aproximadamente uma ordem de grandeza. No entanto, para missões como PLATO (Agency, 2025b), que monitorarão centenas de milhares de estrelas, esse custo torna-se um desafio significativo, especialmente quando se consideram as dificuldades práticas de processamento observadas neste trabalho, que incluem falhas de download, curvas de luz corrompidas e a necessidade de reprocessamento. Tais desafios evidenciam que, em larga escala, a automação robusta e a paralelização não são apenas conveniências, mas requisitos estruturais para viabilizar a análise, conforme discutido por Zhang & Zhao no contexto do Big Data astronômico.

4.4 Discussão dos resultados

Os resultados obtidos evidenciam que a detecção de exoplanetas por meio do método de trânsito envolve desafios que vão além da simples identificação de sinais periódicos. A análise dos dados do TLS demonstrou que:

1. **O SDE não é um discriminador confiável**: os falsos positivos apresentam SDE médio superior ao de exoplanetas confirmados (46,47 vs. 41,85), e os maiores valores de SDE ocorrem em ambas as classes.
2. **Falsos positivos de curta duração são especialmente problemáticos**: muitos objetos espúrios com SDE extremamente alto têm durações de trânsito muito curtas (0,02–0,03 dias), compatíveis com binárias eclipsantes.
3. **O custo computacional é significativo**: o processamento de 6807 objetos exigiu cerca de 4 meses em tempo real, evidenciando a necessidade de otimizações ou paralelização para aplicações em larga escala.

Essas observações estão alinhadas com as limitações teóricas do TLS discutidas por Hippke & Heller na seção 2.8.2, particularmente no que se refere à vulnerabilidade a ruídos

não-brancos e à dificuldade de distinguir trânsitos planetários de outros fenômenos periódicos.

Conforme apresentado por (Shallue; Vanderburg, 2018), abordagens complementares baseadas em aprendizado profundo (como a rede neural AstroNet) podem atuar na etapa de revalidação, aprendendo padrões morfológicos que o TLS não captura. A integração entre algoritmos determinísticos de busca (TLS) e classificadores baseados em inteligência artificial representa, portanto, o caminho mais promissor para o tratamento de grandes volumes de dados astronômicos.

Os desafios observados neste trabalho, alta taxa de falsos positivos, necessidade de integração de dados heterogêneos (metadados + séries temporais) e custo computacional elevado, ilustram diretamente as três dimensões do Big Data discutidas na seção 2.6: veracidade, variedade e volume. Conforme argumentado por Zhang & Zhao , a solução para esses desafios não reside apenas no aumento da capacidade computacional bruta, mas no desenvolvimento de métodos estatisticamente robustos e algoritmos escaláveis. Nesse sentido, a transição de abordagens puramente determinísticas (como o TLS) para sistemas híbridos que incorporam aprendizado de máquina representa a direção natural da exoplanetologia contemporânea, alinhando-se à tendência observada em outras áreas da astronomia intensiva em dados.

5 CONCLUSÕES E TRABALHOS FUTUROS

5.1 Conclusões

Este trabalho teve como objetivo analisar o método de trânsito planetário e discutir como técnicas de ciência de dados contribuem para sua eficácia na astronomia contemporânea, especialmente no contexto da era do Big Data. A pesquisa combinou uma revisão bibliográfica aprofundada dos fundamentos físicos do método de trânsito com procedimentos computacionais empíricos utilizando o algoritmo TLS aplicado a 6807 curvas de luz do catálogo KOI.

A análise dos resultados permitiu estabelecer as seguintes conclusões principais:

1. O SDE não é um discriminador confiável: a análise estatística demonstrou que falsos positivos apresentam SDE médio (46,47) superior ao de planetas confirmados (41,85), e ambos os grupos alcançam valores extremamente elevados (até 289,69 para falsos positivos e 250,99 para confirmados). Esse resultado evidencia que a métrica SDE, isoladamente, não permite distinguir sinais planetários reais de artefatos periódicos como binárias eclipsantes.

2. A amostra analisada reflete os desafios do Big Data: a distribuição dos dados (60,4% de falsos positivos) e o tempo de processamento de aproximadamente 4 meses para 6807 curvas de luz ilustram diretamente as três dimensões do Big Data propostas por Zhang & Zhao (2015): veracidade (alta taxa de falsos positivos), variedade (integração de metadados e séries temporais) e volume (processamento massivo de dados).

3. A análise morfológica qualitativa revela padrões distintivos: o comparação entre o planeta confirmado Kepler-2 b (formato em “U”) e o falso positivo K07010.01 (formato em “V”) demonstrou que a morfologia do trânsito é um discriminador potencialmente eficaz, mas que não é capturado por algoritmos puramente determinísticos como o TLS.

4. O custo computacional é um desafio significativo: o processamento em hardware convencional (Ryzen 5, 8 GB RAM, 256 GB SSD) demandou aproximadamente 4 meses de tempo real, evidenciando que, para missões futuras como PLATO, que monitorarão centenas de milhares de estrelas, a automação robusta e a paralelização não são conveniências, mas requisitos estruturais.

Em síntese, esta pesquisa contribuiu para a compreensão das limitações e potencialidades do algoritmo TLS no contexto da detecção automatizada de exoplanetas, demonstrando que, embora o método de trânsito seja extremamente produtivo, a transição para a era do Big Data exige a integração de abordagens determinísticas com técnicas de aprendizado de máquina e

inteligência artificial.

5.2 Limitações do Estudo

As principais limitações deste trabalho incluem:

- **Alta ocorrência de falsos positivos**: a amostra analisada é composta por 4113 falsos positivos (60,4%) contra 2694 planetas confirmados (39,6%), evidenciando que a maioria dos sinais detectados não corresponde a planetas reais. Esse desbalanceamento reforça a necessidade de métodos adicionais de validação para filtrar os falsos positivos.
- **Agregação temporal inefetiva:** o uso de agregação temporal mostrou-se inefetivo tanto com resolução de 0,01 dias (inferior à cadência original da Kepler, resultando em nenhuma redução significativa de ruído) quanto com valores maiores (que diminuíam a resolução da grade de busca para sinais mais sutis).
- **Tempo de processamento não otimizado**: o pipeline não operou ininterruptamente, e o tempo de processamento reflete condições reais de execução, não um ambiente controlado.
- **Limitações do hardware**: os experimentos foram realizados em um computador com processador Ryzen 5, 8 GB de memória RAM e 256 GB SSD, o que impôs restrições práticas ao processamento, como a necessidade de reprocessar curvas de luz extensas (mais de 60.000 pontos) que exigiam tratamento especial devido a limitações de memória.

5.3 Trabalhos Futuros

Com base nos resultados e limitações identificadas, sugerem-se as seguintes direções para pesquisas futuras:

1. **Integração do TLS com redes neurais convolucionais**: conforme mostrado por (Shallue; Vanderburg, 2018), redes neurais como a AstroNet são capazes de aprender padrões morfológicos (formatos “U” vs “V”) que o TLS não captura. Trabalhos futuros podem desenvolver um pipeline híbrido no qual o TLS atua como detector primário de periodicidade e uma Convolutional Neural Network (CNN) atua como classificador secundário para filtrar falsos positivos.
2. **Otimização computacional para larga escala**: o tempo de processamento de 4 meses para 6807 objetos evidencia a necessidade de paralelização e otimização para aplicações em missões como PLATO. Sugere-se a implementação do TLS em arquiteturas de computação

de alto desempenho (HPC) ou o uso de GPUs para aceleração.

3. **Análise quantitativa da morfologia de trânsitos**: a extração de métricas como assimetria, curtose e razão entre duração do ingresso e duração total pode permitir a classificação automatizada de trânsitos sem a necessidade de redes neurais complexas.
4. **Aplicação do pipeline a outras missões**: o pipeline desenvolvido pode ser adaptado para processar dados do TESS e, futuramente, do PLATO, permitindo a comparação do desempenho do TLS em diferentes regimes observacionais.
5. **Análise de sensibilidade por subgrupos**: estudos futuros podem estratificar a amostra por tamanho planetário, período orbital e profundidade de trânsito para avaliar a sensibilidade do TLS em diferentes regimes, especialmente para planetas terrestres e órbitas longas.
6. **Comparação sistemática entre TLS e BLS**: embora este trabalho tenha optado pelo uso direto do TLS, uma comparação quantitativa controlada entre os dois algoritmos na mesma amostra permitiria avaliar empiricamente as vantagens e desvantagens de cada método em diferentes cenários de razão sinal-ruído.

Por fim, a continuidade desta linha de pesquisa contribuirá para o avanço da exoplanetologia na era do Big Data, demonstrando que a união entre os fundamentos físicos do método de trânsito e as técnicas de ciência de dados é essencial para transformar o crescente volume de dados astronômicos em descobertas científicas confiáveis e escaláveis.

REFERÊNCIAS

AERONAUTICS, N.; ADMINISTRATION, S. **Kepler Mission Archive**. 2018. MAST. Disponível em: https://archive.stsci.edu/kepler/.

AERONAUTICS, N.; ADMINISTRATION, S. **Hubble by the Numbers**. 2024. NASA Science. Dados do arquivo da missão Hubble. Acesso em: 19 maio 2026. Disponível em: https://science.nasa.gov/mission/hubble/overview/hubble-by-the-numbers/.

AERONAUTICS, N.; ADMINISTRATION, S. **Exoplanet Discoveries Dashboard**. 2026. Acesso em: 28 mar. 2026. Disponível em: https://science.nasa.gov/exoplanets/discoveries-dashboard/.

AERONAUTICS, N.; ADMINISTRATION, S. **Exoplanets**. 2026. Acesso em: 04 abr. 2026. Disponível em: https://science.nasa.gov/exoplanets/.

AERONAUTICS, N.; ARCHIVE, S. A. E. **NASA Exoplanet Archive**. 2026. Disponível em: https://exoplanetarchive.ipac.caltech.edu/. Acesso em: 30 mar. 2026.

AGENCY, E. S. **Gaia Mission Numbers**. 2025. ESA Cosmos. Acesso em: 22 de mai. 2026. Disponível em: https://www.cosmos.esa.int/web/gaia/mission-numbers.

AGENCY, E. S. **Plato: Terrestrial planet hunter**. 2025. ESA. Disponível em: https://www.esa.int/Science_Exploration/Space_Science/Plato. Acesso em: 13 abr. 2026.

ARAUJO, G. C.; RAMOS, F. V. Um breve ensaio sobre a história da mecânica celeste. **Studies in Engineering and Exact Sciences**, v. 4, n. 1, p. 154–170, 2023.

COLLABORATION, S. **Sloan Digital Sky Survey Data Release 19**. 2022. https://www.sdss.org/dr19/. Acesso em 2026.

DESC, L.; ISSC. **LSST Data Science Overview**. 2026. Acesso em: 22 mai. 2026. Disponível em: https://issc.science.lsst.org/pages/DataScienceOverview.html.

FILHO, K. d. S. O.; SARAIVA, M. d. F. O. **Astronomia e Astrofísica**. Porto Alegre: Departamento de Astronomia do Instituto de Física da UFRGS, 2014. Disponível em: http://astro.if.ufrgs.br/. Acesso em: 05 mar. 2026. Disponível em: http://astro.if.ufrgs.br/.

HIPPKE, M.; HELLER, R. Optimized transit detection algorithm to search for periodic transits of small planets. **Astronomy & Astrophysics**, v. 623, n. A39, p. 1–13, 2019.

KOCH, D. G. *et al.* Kepler mission design, realized photometric performance, and early science. **The Astrophysical Journal Letters**, v. 713, n. 2, p. L79–L86, 2010.

KOVACS, G.; ZUCKER, S.; MAZEH, T. A box-fitting algorithm in the search for periodic transits. **Astronomy & Astrophysics**, v. 391, n. 1, p. 369–377, 2002.

PÁL, A. *et al.* Hat-p-7b: An extremely hot massive planet transiting a bright star in the kepler field. **The Astrophysical Journal**, v. 680, n. 2, p. 1450–1456, 2008.

PEARSON, K. A.; PALAFOX, L.; GRIFFITH, C. A. Searching for exoplanets using artificial intelligence. **Monthly Notices of the Royal Astronomical Society**, v. 474, n. 1, p. 478–491, 2017.

RICE, K. The detection and characterization of extrasolar planets. **Challenges**, v. 5, n. 2, p. 296–323, 2014.

RICKER, G. R. *et al.* Transiting exoplanet survey satellite. **Journal of Astronomical Telescopes, Instruments, and Systems**, v. 1, n. 1, p. 014003, 2015.

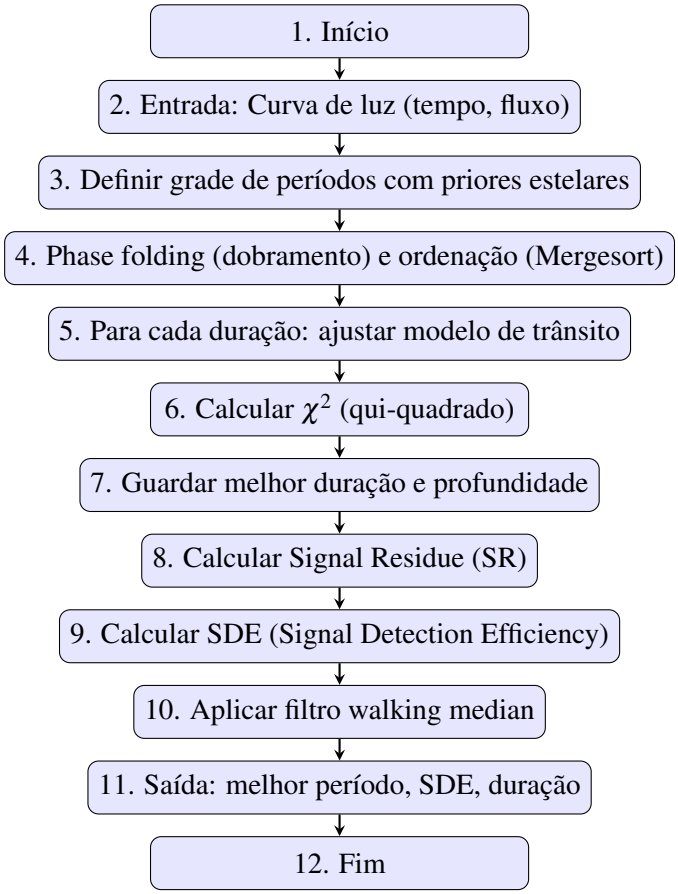
SANTOS, W. C.; AMORIM, R. G. G. Descobertas de exoplanetas pelo método do trânsito. **Revista Brasileira de Ensino de Física**, v. 39, n. 2, p. e2308, 2017.

SHALLUE, C. J.; VANDERBURG, A. Identifying exoplanets with deep learning: A five-planet resonant chain around kepler-80 and an eighth planet around kepler-90. **The Astronomical Journal**, v. 155, n. 2, p. 94, 2018.

ZHANG, Y.; ZHAO, Y. Astronomy in the big data era. **Data Science Journal**, v. 14, n. 11, p. 1–9, 2015.

APÊNDICE A – FLUXOGRAMA DO ALGORITMO TRANSIT LEAST SQUARES (TLS)

Este apêndice apresenta o fluxograma do algoritmo *Transit Least Squares* (TLS), conforme descrito por (Hippke; Heller, 2019). O fluxograma ilustra as principais etapas do algoritmo, desde a entrada da curva de luz até o cálculo do SDE (Signal Detection Efficiency).



Fonte: Hippke e Heller (2019).

APÊNDICE B – CÓDIGO DO PIPELINE DE PROCESSAMENTO

Este apêndice apresenta um trecho do código do pipeline desenvolvido para processamento das curvas de luz, destacando a etapa de filtragem e organização da amostra analisada.

```
try:
    print("Starting tls_dataset.py")

    df = load_csv(TEMPLATE_CSV)
    df = df[df["koi_disposition"].isin(["CONFIRMED", "FALSE POSITIVE"])]
    df = df.reset_index(drop=True)
    print("Size:", len(df))

    kic_column = next((c for c in df.columns if "kepid" in c.lower() or "kic" in c.lower()), None)
    if not kic_column:
        kic_column = 'kepid' if 'kepid' in df.columns else df.columns[0]

    processed_kois = set()
    processed_old_kics = set()
    if os.path.exists(META_PATH):
        try:
            meta = pd.read_csv(META_PATH)
            if "koi_name" in meta.columns:
                processed_kois = set(meta["koi_name"].dropna().values)
            if "kic_id" in meta.columns:
                processed_old_kics = set(meta["kic_id"].dropna().values)

        except:
            pass
```

Fonte: Autoria própria (2026).

O código completo está disponível em repositório público: https://github.com/isabellylimals/katniss-data-processing.