UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Regressão simbólica em redes complexas

Beatriz Regina Brum

Tese de Doutorado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP
SÃO CARLOS

Beatriz Regina Brum

Regressão simbólica em redes complexas

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Doutora em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO FINAL*

Área de Concentração: Estatística

Orientador: Prof. Dr. Francisco Aparecido Rodrigues

USP – São Carlos
Março de 2026

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Brum, Beatriz
B892r Regressão simbólica em redes complexas / Beatriz Brum; orientador Francisco Rodrigues. -- São Carlos, 2026.
105 p.

Tese (Doutorado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2026.

1. Regressão Simbólica. 2. Sistemas dinâmicos. 3. Redes complexas. I. Rodrigues, Francisco , orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Beatriz Regina Brum

Symbolic regression in complex networks

Doctoral dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP and to the Department of Statistics – DEs-UFSCar, in partial fulfillment of the requirements for the degree of the Doctorate Interagency Program Graduate in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Francisco Aparecido Rodrigues

USP – São Carlos
March 2026

UFSCar

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas

Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Defesa de Tese de Doutorado do(a) candidato(a) Beatriz Regina Brum, realizada em 20/02/2026.

Comissão Julgadora:

Prof(a). Dr(a). Francisco Aparecido Rodrigues (ICMC-USP)

Prof(a). Dr(a). Rafael Izbicki (DEs-UFSCar)

Prof(a). Dr(a). Eniuce Menezes de Souza (UEM)

Prof(a). Dr(a). Matheus Henrique Dal Molin Ribeiro (UTFPR)

Prof(a). Dr(a). Willian Luís de Oliveira (UEM)

O Relatório de Defesa produzido pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação.

À memória dos mais de 700 mil brasileiros que perderam a vida para a COVID-19. Que esta pesquisa honre suas histórias e contribua para políticas públicas mais eficazes em saúde.

AGRADECIMENTOS

Agradeço à coordenação e aos professores do Programa de Pós-Graduação em Estatística (PIPGEs), pela dedicação e pelo compromisso com a formação acadêmica e ao CNPq pelo um ano de bolsa concedida. Manifesto um agradecimento especial ao meu orientador Francisco Rodrigues pela generosidade na transmissão de seus conhecimentos e pela paciência. À Isolde Previdelli, professora e amiga, pelo apoio e incentivo contínuo.

Registro meu sincero agradecimento aos membros da banca examinadora, pela disponibilidade, pela análise cuidadosa e pelas contribuições intelectuais que aprimoraram a qualidade científica desta tese.

Por fim, agradeço à minha querida irmã, Bruna Brum, pelo apoio incondicional, e às minhas amigas Edilenia Queiros, Luiza Lober e Márcia Lorena Alves dos Santos, pelo incentivo e companheirismo ao longo desta jornada.

“Não há nada escondido que não venha a ser descoberto, ou oculto que não venha a ser conhecido.”

(Salmo “Lucas 12:2”)

RESUMO

BRUM, B. R. **Regressão simbólica em redes complexas** . 2026. 105 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2026.

Descobrir as equações que regem um sistema a partir de observações é fundamental em diversas áreas da ciência, permitindo tanto compreender suas propriedades quanto prever comportamentos futuros. Recentemente, técnicas de aprendizado de máquina baseadas em regressão simbólica emergiram como alternativa automatizada para essa tarefa, com a vantagem de não exigir conhecimento prévio do domínio para descrever sistemas dinâmicos de forma eficaz, embora possam se beneficiar desse conhecimento para refinar as equações obtidas. Este estudo busca desenvolver uma nova metodologia para a análise e inferência de fenômenos complexos, analisando e comparando a eficácia de múltiplos algoritmos de regressão simbólica na identificação de sistemas dinâmicos epidêmicos em diferentes meios de propagação. Tradicionalmente, as análises de dinâmica temporal baseiam-se em dados sequenciais organizados linearmente, abordagem que se mostra limitada para o estudo de sistemas nos quais as interconexões entre elementos e as interações recorrentes influenciam decisivamente sua evolução. Para superar essa limitação, a pesquisa organizou-se em duas vertentes: a primeira utilizando dados de meio homogêneo e a segunda, de meios heterogêneos. Em ambas, aplicaram-se algoritmos de regressão simbólica, utilizando como referência o ajuste obtido por modelos de Floresta Aleatória. A eficácia dos métodos foi avaliada por meio de análise descritiva e da aplicação do teste de Wilcoxon, a fim de verificar diferenças significativas entre as dinâmicas reais e as estimadas. Na segunda vertente, incluiu-se análise estatística inferencial para investigar a influência da topologia da rede na precisão dos regressores. Os resultados demonstram que alguns algoritmos, como SINDy, SR e MultKAN, apresentaram alta capacidade de reconstrução das dinâmicas subjacentes em ambas as vertentes, sendo o SR o que apresentou melhor desempenho entre eles. Contudo, não foram detectados efeitos significativos da topologia da rede no desempenho dos algoritmos, exceto em relação ao R^2, o que já era esperado. Esses achados indicam que as equações aprendidas revelaram-se generalizáveis para os cenários aqui descritos, independentemente das características estruturais do meio de propagação, o que viabiliza sua potencial implementação em contextos reais de saúde pública.

Palavras-chave: Regressão Simbólica, Sistemas dinâmicos complexos e redes complexas.

ABSTRACT

BRUM, B. R. **Symbolic regression in complex networks** . 2026. 105 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2026.

Discovering the equations that govern a system from observations is fundamental in several areas of science, as it enables both understanding its properties and predicting future behavior. Recently, machine learning techniques based on symbolic regression have emerged as an automated alternative for this task, with the advantage of not requiring prior domain knowledge to effectively describe dynamical systems, although they may benefit from such knowledge to refine the obtained equations. This study aims to develop a new methodology for the analysis and inference of complex phenomena by analyzing and comparing the performance of multiple symbolic regression algorithms in identifying epidemic dynamical systems across different propagation environments. Traditionally, temporal dynamics analyses rely on sequentially organized linear data, an approach that is limited when studying systems in which interconnections among elements and recurrent interactions decisively influence their evolution. To overcome this limitation, the research was structured into two approaches: the first using data from homogeneous environments and the second from heterogeneous environments. In both cases, symbolic regression algorithms were applied, using the fit obtained from Random Forest models as a reference. The effectiveness of the methods was evaluated through descriptive analysis and the application of the Wilcoxon test to assess significant differences between real and estimated dynamics. In the second approach, inferential statistical analysis was included to investigate the influence of network topology on the accuracy of the regressors. The results show that some algorithms, such as SINDy, SR, and MultKAN, demonstrated a high capacity to reconstruct the underlying dynamics in both approaches, with SR achieving the best performance among them. However, no significant effects of network topology were detected on algorithm performance, except with respect to R^2, which was already expected. These findings indicate that the learned equations proved to be generalizable to the scenarios described here, regardless of the structural characteristics of the propagation environment, thereby supporting their potential implementation in real-world public health settings.

Keywords: Symbolic regression, Complex dynamic systems and complex networks.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

1.1 Contextualização

A descoberta de equações a partir de dados observacionais constitui um dos pilares fundamentais do método científico tradicional. Desde os trabalhos de Johannes Kepler, que inferiu as leis do movimento planetário a partir de observações astronômicas meticulosas (CAMPS-VALLS *et al.*, 2023), até as formulações teóricas de Isaac Newton que consolidaram a mecânica clássica, o processo de identificação de relações matemáticas subjacentes a fenômenos naturais caracterizou-se historicamente por sua natureza manual, fundamentando-se essencialmente em procedimentos sistemáticos de tentativa e erro.

Nas últimas décadas, contudo, o advento do Big Data, caracterizado por volumes massivos de informações, alta dimensionalidade e interações predominantemente não lineares, redefiniu os limites da modelagem científica. A complexidade estrutural desses dados exige métodos capazes de extrair, de maneira automatizada, representações simbólicas que capturem a dinâmica integral dos sistemas observados. Nesse contexto, a descoberta de leis passa a depender não apenas da formulação teórica, mas também da capacidade computacional de explorar vastos espaços funcionais.

Impulsionado por avanços em Inteligência Artificial, esse cenário favoreceu o desenvolvimento da Regressão Simbólica (RS), uma abordagem que automatiza a busca simultânea pela estrutura funcional e pelos parâmetros de um modelo. Diferentemente da regressão paramétrica tradicional, que ajusta coeficientes em formas previamente especificadas e requer extensos diagnósticos estatísticos, a RS realiza uma exploração combinatória de operadores matemáticos, variáveis e constantes, permitindo identificar diretamente a arquitetura matemática mais compatível com os dados. Essa flexibilidade a torna particularmente adequada para sistemas complexos.

Nesta tese, a RS foi aplicada a dados simulados provenientes de sistemas dinâmicos

propagando-se em contextos homogêneos e heterogêneos, com o objetivo de recuperar, de maneira semi-supervisionada, a forma estrutural das equações governantes. Essa estratégia foi adotada para evitar a subestimação do potencial inferencial de determinados algoritmos, reduzindo vieses associados a restrições excessivas de busca. A análise evidenciou o desempenho consistente de três bibliotecas, PySINDy, PySR e PyKAN, com destaque para a PySR, que apresentou superioridade na recuperação estrutural das dinâmicas avaliadas. Assim, esta pesquisa não apenas contribui para o avanço metodológico da descoberta automatizada de equações, mas também reforça o papel da RS como ferramenta promissora na modelagem de sistemas dinâmicos complexos.

1.2 Motivação

Historicamente, a busca por leis físicas capazes de explicar os dados obtidos por estudos observacionais tem sido um grande desafio para a ciência. A abundância de dados gerados por diversas áreas do conhecimento ampliou consideravelmente o campo de estudo da ciência. Dados provenientes de processos físicos naturais têm incentivado cientistas há décadas na busca por padrões subjacentes aos dados. Uma iniciativa visando a automatização da identificação de expressões no âmbito da Inteligência Artificial foi apresentada por Langley (1977), dando origem ao sistema BACON, batizado em homenagem ao renomado cientista Francis Bacon. Esse método foi empregado na descoberta de equações por meio de técnicas de detecção de regularidades e criação de novos atributos.

No entanto, a base para o desenvolvimento do primeiro modelo de regressão simbólica foi possível graças a algoritmos genéticos (GA), uma metodologia desenvolvida por John Holland e seus colegas da Universidade de Michigan em 1975 (GOLBERG, 1989). Algoritmos genéticos são técnicas de otimização não tradicionais inspiradas no mecanismo de seleção natural e genética de Darwin (1910), destinadas a ajustar hiperparâmetros de uma forma que imite o processo de evolução (GOLBERG, 1989). Este método culminou no que pode ser considerado a primeira abordagem de regressão simbólica, apresentada por Koza (1992). Desenvolvido por meio de programação genética (GP), um ramo dos algoritmos genéticos, o método adota um esquema de codificação com base em árvores de decisão, que representam as expressões (DIVEEV; SHMALKO, 2021).

Nos últimos anos, algoritmos baseados nessas técnicas ganharam destaque à medida que se concentram na otimização e exploração de soluções aproximadas para problemas de busca (SIVANANDAM; DEEPA, 2008). Na busca de regressão por meio de um domínio simbólico e com base em uma análise de dissimilaridade e acasalamento, Gustafson, Burke e Krasnogor (2005) aprimoraram a GP, encontrando melhorias estatisticamente significativas na qualidade da solução.

Refinando a abordagem acima mencionada, Schmidt e Lipson (2009a) desenvolveram

um método também baseado em GP, mas aumentado pela otimização de Pareto, ou programação genética de Pareto. Dados experimentais foram coletados a partir de movimentos rastreados computacionalmente para encontrar expressões para Lagrangianos e Hamiltonianos, essenciais na descrição de um determinado sistema físico por meio de equações governantes. O processo de busca no espaço das expressões simbólicas acelerou-se então, onde modelos mais simples foram usados como base para representações mais complexas. Como exemplo, o software *Eureqa* visava não apenas buscar uma única expressão simbólica, mas sim o conjunto de melhores expressões de acordo com sua complexidade. Uma avaliação deste programa em medições aleatórias pode ser vista em (DUBČÁKOVÁ, 2011), onde o desempenho e a implementação do algoritmo são descritos.

Nesse contexto de algoritmos evolutivos, Stephens (2016) desenvolveu o GPlearn, uma biblioteca de código aberto escrita na linguagem de programação Python desenvolvida com base na biblioteca *scikit-learn*, permitindo a integração do GPlearn, baseado em um algoritmo genético, com outras bibliotecas de aprendizado de máquina.

Brunton, Proctor e Kutz (2016a) combinaram técnicas de promoção de esparsidade e aprendizado de máquina com sistemas dinâmicos para descobrir equações governantes a partir de dados ruidosos. Esse aprimoramento permitiu que o algoritmo proposto pelos autores fosse utilizado tanto em sistemas simples quanto em sistemas de alta dimensão. Abordagem útil a várias aplicações em dinâmica oscilatória, tanto lineares quanto não lineares, como sistemas caóticos de Lorenz e vórtice formado pelo derramamento de fluido. Impulsionados pelo desejo de descobrir sistemas dinâmicos em redes metabólicas e regulatórias, que exibem muitas vezes dinâmicas não lineares com não-linearidades de funções racionais em sua formulação, Mangan *et al.* (2016) desenvolveram um algoritmo para equações diferenciais ordinárias implícitas (SINDy implícito). Este algoritmo pode ser aplicado a redes biológicas como: a cinética enzimática de Michaelis-Menten, a rede reguladora de competência em bactérias e a rede metabólica para glicólise de leveduras. Ainda utilizando técnicas de promoção de esparsidade, Rudy *et al.* (2017) desenvolveram um método capaz de descobrir equações diferenciais parciais com base em medidas de séries temporais. A técnica pode identificar modelos dinâmicos mais complexos e não lineares.

No trabalho de Sahoo, Lampert e Martius (2018), foi apresentado um método implementado em um sistema carrinho-pêndulo. A metodologia emprega uma arquitetura de rede neural que permitiu a compreensão das relações funcionais e sua generalização a partir de dados observados para regiões anteriormente inexploradas do espaço de parâmetros. Combinando redes neurais com um conjunto de técnicas inspiradas na física, Udrescu e Tegmark (2020) desenvolveram um algoritmo SR recursivo multidimensional (AI-Feynman). O algoritmo foi aplicado a 100 equações das palestras de (Feynman, 2024), descobrindo todas elas e sendo considerado um dos melhores algoritmos disponíveis para a investigação de sistemas físicos.

Explorando o método proposto por Zheng *et al.* (2018), regressão regularizada esparsa

relaxada (SR3) SINDy-SR3, Champion *et al.* (2020) apresentaram uma estrutura de otimização esparsa capaz de aprender modelos parcimoniosos de sistemas dinâmicos a partir de dados. Esta formulação de regressão esparsa visa descobrir as equações de um sistema dinâmico usando dados e selecionando termos relevantes de uma biblioteca de funções possíveis. O desenvolvimento de uma estrutura abrangente de modelos de SR, impulsionada por dados de séries temporais, resultou na implementação do pacote de regressão esparsa em Python, PySINDy. A biblioteca abrange várias implementações e é conhecida como método de Identificação Esparsa de Sistemas Dinâmicos Não Lineares (SINDy) (SILVA *et al.*, 2020; KAPTANOGLU *et al.*, 2021; BRUNTON; PROCTOR; KUTZ, 2016a; BRUNTON; PROCTOR; KUTZ, 2016b).

Baseado em GP e GA, o PySR, desenvolvido por Cranmer (2023), é uma ferramenta poderosa para gerar modelos interpretáveis. Projetado para ter programação de alto desempenho em Julia, o algoritmo é altamente otimizado e normalmente usa Python como interface. O método PyKAN (LIU *et al.*, 2024b) também apresenta uma alternativa interessante aos algoritmos mencionados. Sua metodologia é baseada no teorema da representação de Kolmogorov-Arnold. Mas para este algoritmo, ainda havia espaço para melhorias, o que levou ao desenvolvimento do MultKAN (LIU *et al.*, 2024a), um dos modelos de SR mais recentes até o momento.

A aplicação destes modelos em múltiplos domínios tem contribuído de forma significativa para o avanço da ciência. Várias investigações têm sido realizadas adotando estas abordagens, por exemplo, na modelação de sistemas climáticos (STANISLAWSKA; KRAWIEC; KUNDZEWICZ, 2012), bem como na procura de algoritmos híbridos em ciência dos materiais (WANG; WAGNER; RONDINELLI, 2019). Chen, Angulo e Liu (2019) empregaram métodos de SR para compreender a dinâmica de ecossistemas complexos. Além disso, a técnica tem sido usada na previsão da velocidade do vento (ABDELLAOUI; MEHRKANOON, 2021), e também na busca da superfície de volatilidade implícita no mercado financeiro (LUO; YU, 2023).

Kiyani *et al.* (2023) utilizaram o GPlearn para descobrir componentes desconhecidas de PDEs não lineares complexas utilizando a abordagem de decomposição de domínio, visando encontrar a forma fechada da componente não linear desconhecida. Gudetti *et al.* (2023) demonstraram que o SINDy pode ser utilizado em aplicações NVH, para controlar e reduzir a vibração em produtos, auxiliando na análise de conjuntos de dados, o que é útil na engenharia automóvel e em vários outros campos. Miyazaki *et al.* (2023) utilizaram a IA-Feynman para descobrir o modelo de desconto hiperbólico como uma função de desconto que os humanos não conseguiam encontrar. Wong e Cranmer (2022) relataram que o uso de regressão simbólica com PySR pode ser adotado como uma estratégia eficaz para identificar modelos interpretáveis de população de ondas gravitacionais. Almeida (2024) empregou KANs para investigar a capacidade de condução autônoma de veículos com o objetivo de melhorar a previsão de tráfego. Para além da essência destas aplicações, ainda há muito a ser explorado. Dependendo do contexto em que os dados foram produzidos, certos modelos de RS podem ser mais apropriados, e o mapeamento torna-se crucial para a precisão da inferência.

No que tange a comparação e desempenho desses algoritmos, Udrescu e Tegmark (2020), na busca de tecnologia de ponta, também desenvolveu uma comparação completa entre seu algoritmo, AI-Feynman, e o Eureqa (SCHMIDT; LIPSON, 2014), empregando 120 conjuntos de dados sintéticos para esta análise comparativa. Cava *et al.* (2021) introduziram uma plataforma de *benchmarking* para avaliar o desempenho de 14 métodos contemporâneos de RS em 252 conjuntos de dados, comparando-os com 7 métodos de aprendizagem automática. Os autores criaram o SRBench para ser um projeto de *bench* reprodutível e de código aberto. O artigo publicado avaliou o desempenho do modelo de regressão e a sua capacidade para aprender equações e sistemas simples.

Landajuela *et al.* (2022) apresentaram uma estrutura de RS que é modular e unificada, juntando várias estratégias. Eles combinaram cinco bibliotecas diferentes, incluindo GPLearn, AI-Feynman e PySINDy, que demonstraram desempenho de alto nível quando avaliados pelo SRBench (CAVA *et al.*, 2021). Além disso, mostraram que esses modelos de regressão trabalham super bem juntos. Já Cranmer (2023) trouxe o *"EmpiricalBench"*, uma ferramenta feita para comparar e avaliar a qualidade de diferentes softwares de regressão simbólica em aplicações científicas. Por fim, Thing e Koksbang (2024) elaboraram um *benchmark* de 12 algoritmos de regressão simbólica aplicados a 28 conjuntos de dados representando seis diferentes configurações astrofísicas.

No contexto de redes complexas, Hu, Cui e Yang (2025) avançaram na aplicação da regressão simbólica ao automatizar a descoberta das equações que governam a dinâmica de sistemas em redes a partir de dados observados. Para isso, os autores combinam aprendizado profundo com regressão simbólica, utilizando o primeiro para capturar padrões dinâmicos complexos e o segundo para extrair, posteriormente, formas funcionais interpretáveis. A abordagem foi validada em múltiplas dinâmicas de propagação em redes, contemplando distintos mecanismos de interação e evolução temporal, o que reforça o caráter geral e a robustez do método proposto. O estudo privilegiou a identificação direta da estrutura matemática intrínseca aos dados simulados, incorporando explicitamente conhecimento físico prévio acerca da dinâmica em redes, particularmente o princípio de que a mudança de estado de um nó pode depender tanto de seu próprio estado quanto dos estados de seus vizinhos.

Essa dependência entre dinâmica e topologia encontra respaldo em resultados consolidados da teoria espectral de grafos. É bem estabelecido que o maior autovalor da matriz de adjacência está diretamente relacionado à estrutura da rede e aos momentos da distribuição de grau (CHUNG; LU; VU, 2004). Ademais, já se demonstrou que o número básico de reprodução de uma doença mantém relação direta com a matriz de adjacência da rede de contatos (GÓMEZ *et al.*, 2010). Tal resultado é particularmente relevante para modelos epidemiológicos de propagação, como SIR e SIS, pois evidencia que a estrutura da rede não apenas influencia, mas impõe restrições formais à dinâmica de transmissão. Nesse sentido, a incorporação explícita dessas propriedades estruturais no processo de descoberta de equações não constitui apenas um refina-

mento metodológico, mas um alinhamento necessário entre dados observados e fundamentos matemáticos da dinâmica em redes.

Diante disso, emerge uma questão ainda pouco investigada: se a topologia da rede exerce influência direta sobre a dinâmica de propagação, poderia ela também impactar o próprio processo de descoberta das equações que governam o sistema subjacente aos dados observados? Embora redes complexas e técnicas de regressão simbólica sejam amplamente utilizadas em diversas áreas do conhecimento, ainda persiste uma lacuna na literatura quanto à avaliação sistemática do desempenho desses métodos na recuperação da forma funcional de sistemas dinâmicos epidêmicos estruturados em redes complexas. Investigar essa relação não apenas contribui para o avanço metodológico na descoberta de equações a partir de dados, mas também pode subsidiar estratégias mais fundamentadas para a gestão em saúde pública, com potenciais impactos diretos sobre a sociedade.

1.3 O objetivo geral

O presente estudo visa desenvolver uma nova metodologia, fundamentada na regressão simbólica, para a inferência de equações de sistemas dinâmicos em diferentes meios de propagação, abrangendo desde ambientes homogêneos até contextos heterogêneos, como as redes complexas.

Os objetivos específicos são:

1. Simular conjuntos de dados provenientes de diversos sistemas dinâmicos que se propagam em campo homogêneo quanto heterogêneo (redes complexas).

2. Aplicar a regressão simbólica ao conjunto de dados sintéticos e avaliar a sua eficiência na geração da representação correta da forma estrutural geradora dos dados, bem como no cálculo do erro médio absoluto (MAE) e do coeficiente de determinação (R^2).

3. Comparar o desempenho dos modelos de regressão simbólica com o modelo de Floresta Aleatória, considerado este como linha de base, a fim de verificar o potencial dessa metodologia.

4. Explorar a capacidade dos regressores quando aplicados a dados provenientes de sistemas dinâmicos epidêmicos mais complexos que se propagam em redes regulares.

5. Analisar o potencial da regressão simbólica como uma nova metodologia para inferir a dinâmica de epidemias que se propagam em redes complexas. Avaliamos como diferentes topologias de rede, impactam a eficácia e a precisão dos modelos gerados por essa técnica.

1.4 Contribuições no campo de redes complexas

- Uma nova metodologia para a ciência de redes: O desenvolvimento de uma nova metodologia para inferir equações dinâmicas em sistemas não homogêneos, com base na regressão simbólica semi-supervisionada, superando limitações de abordagens neste campo que não descrevem a forma simbólica estrutural da dinâmica.

- A geração de um novo conjunto de dados *"benchmark"*: A criação e disponibilização pública de um conjunto de dados sintéticos ou um *benchmark* para validar metodologias de inferência de equações em redes complexas (ver no Apêndice A), preenchendo uma lacuna de avaliação no campo.

- A aplicação pioneira em um novo contexto: A demonstração, pela primeira vez, da aplicabilidade da regressão simbólica semisupervisionada bem-sucedida para inferir expressões simbólicas interpretáveis de sistemas dinâmicos epidêmicos complexos em diferentes meios de propagação.

- Validação da robustez: confirmação da robustez dos modelos de regressão simbólica nos diferentes cenários simulados sob distintas topologias de rede, não sendo observada influência significativa da estrutura da rede no desempenho dos algoritmos.

- Artigo publicado: BRUM, B. R. LOBER. L, PREVIDELLI, I, Rodrigues, F. A. **Discovering equations from data: symbolic regression in dynamical systems**. O artigo referente ao Capítulo 5 foi publicado no Journal of Physics: Complexity. Disponível em: [<https://doi.org/10.48550/arXiv.2508.20257>].

- Artigo sob revisão: LOBER. L. BRUM, B. R, PREVIDELLI, I, Rodrigues, F. A. **Discovering equations from data: symbolic regression in dynamic systems on complex networks.** O artigo referente ao Capítulo 6 está passando pelo processo de revisão interna pelos autores.

1.5 Organização da Tese

Captulo 1

1. INTRODUÇAO

1.1. Contextualização	1.3. Objetivos	1.5. Organização da Tese
1.2. Motivação	1.4. Contribuições	

Capítulo 2

2. FUNDAMENTAÇÃO TEÓRICA

   2.1. Regressão Simbólica
   2.2. Redes Complexas
   2.3. Processos em redes
   2.4. Processos dinâmicos

Capítulo 3

3. MATERIAIS E MÉTODOS

   3.1. Método para análise da RS em dinâmicas em meio homogêneo.
   3.2. Método para análise da RS em dinâmicas em meios não homogêneos.
   3.3. Caráter semi-supervisionado.

Capítulo 4

4. REGRESSÃO SIMBÓLICA EM PROCESSOS DINÂMICOS EM REDES REGULARES

   4.1. Métricas e definição de uma linha de base.
   4.2. Resultados e discussões.
   4.3. Conclusões e perspectivas.

Capítulo 5

5. REGRESSÃO SIMBÓLICA EM PROCESSOS DINÂMICOS EM REDES NÃO REGULARES

   5.1. Panorama Geral.
   5.2. Do Problema Epidemiológico à Descoberta de Equações.
   5.3. Regressão simbólica para extração de equações em dados complexos.
   5.4. Resultados e discussões.

Capítulo 6

6. CONCLUSÕES, LIMITAÇÕES E PROPOSTA DE PESQUISA

   6.1. Conclusões e perspectivas.
   6.2. Limitação.
   6.3. Proposta de pesquisa.

CAPÍTULO

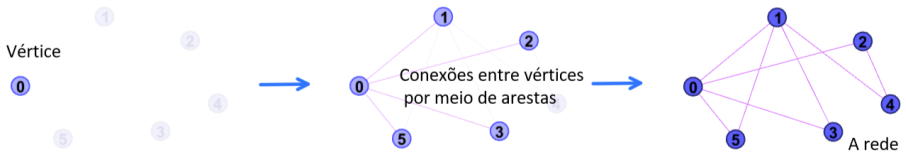
2

FUNDAMENTAÇÃO TEÓRICA

2.1 Redes Complexas

Newman (2003) define uma rede como um conjunto de nós em que cada par pode, ou não, estar conectado por uma aresta (ver Figura 1). A estrutura de uma rede é representada por um grafo $G(V,E)$, em que V e E representam os conjuntos de nós (ou vértices) e arestas respectivamente, conceito amplamente abordado em matemática discreta. Nesse contexto, os nós representam os elementos do sistema, enquanto as arestas descrevem as relações ou interações existentes entre esses elementos (VESPIGNANI, 2012).

Figura 1 – Exemplo de rede e seus componentes.



Fonte: Elaborada pela autora.

Redes estão presentes em diversos contextos do cotidiano. Exemplos de redes reais incluem redes de distribuição, redes de citações científicas, redes metabólicas e redes sociais que representam interações entre indivíduos (NEWMAN, 2003). Tais estruturas são classificadas como redes complexas devido à abundância de nós e conexões que as compõem. Diferentemente dos grafos matemáticos tradicionais, cujas propriedades estruturais podem ser facilmente analisadas, essas redes apresentam uma complexidade significativa, tornando sua compreensão e caracterização mais desafiadoras.

Redes complexas não podem ser descritas de maneira trivial devido à sua alta dimensionalidade e intricada organização. Geralmente modelam comportamentos do mundo real e, à primeira vista, assemelham-se a um emaranhado denso de nós interligados. Diante dessa

complexidade estrutural, torna-se essencial aplicar técnicas de redução de dimensionalidade, utilizando medidas que capturem e resumam as principais características topológicas da rede.

2.1.1 Topologia da Rede

A topologia de uma rede é frequentemente caracterizada por um conjunto conciso de métricas que fornecem informações essenciais sobre sua estrutura e comportamento. Entre essas medidas, destacam-se: o grau, a distribuição do grau, transitividade, distância média entre os nós, correlações estruturais, centralidades e detecção de comunidades, entre outras.

O grau: Representa a medida de conectividade e é uma quantidade fundamental, é possível descrever duas abordagens para o grau, dependendo se a rede é não direcionada ou direcionada.

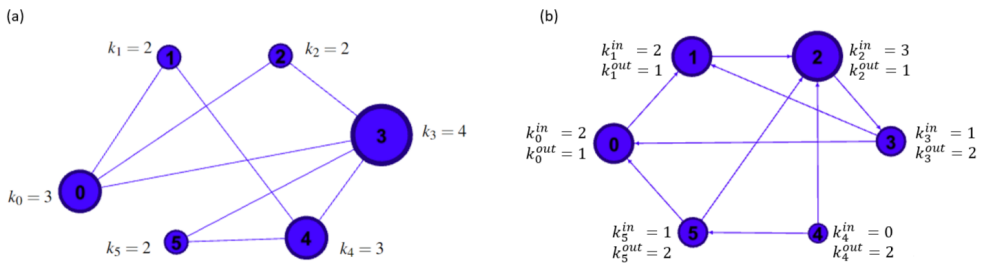
Em uma rede não direcionada, o grau k_i de um vértice i é definido como o número de arestas conectadas a esse vértice (Figura 2 (a)). Isso é equivalente ao número de vizinhos aos quais o vértice está conectado. Pode ser calculado somando-se os elementos da linha correspondente na matriz de adjacência A, composta de zeros e uns. Se o vértice i estiver conectado ao vértice j, atribui-se o valor 1 a A_{ij}; caso contrário, atribui-se o valor 0. Assim, o grau do vértice i pode ser descrito por:

$$k_i = \sum_{j=1}^{N} A_{ij}, \tag{2.1}$$

em que N é o número total de vértices na rede.

No caso de uma rede dirigida, existem duas análises relacionadas ao grau: o grau de entrada (k^{in}) de um vértice e o grau de saída (k^{out}) do mesmo. O grau de entrada de um vértice é definido pelo número de arestas que apontam para esse vértice, enquanto o grau de saída é o número de arestas que partem desse vértice (Figura 2 (b)). O grau total de um vértice em uma rede dirigida é dado pela soma do grau de entrada e do grau de saída, ou seja, $k^i = k^{in} + k^{out}$ (BARABÁSI, 2013).

Figura 2 – (a) Representação de uma rede não dirigida com grau de vértices k. (b) Modelo de uma rede dirigida com grau de vértices k^{in} e k^{out}.



Fonte: Elaborada pela autora.

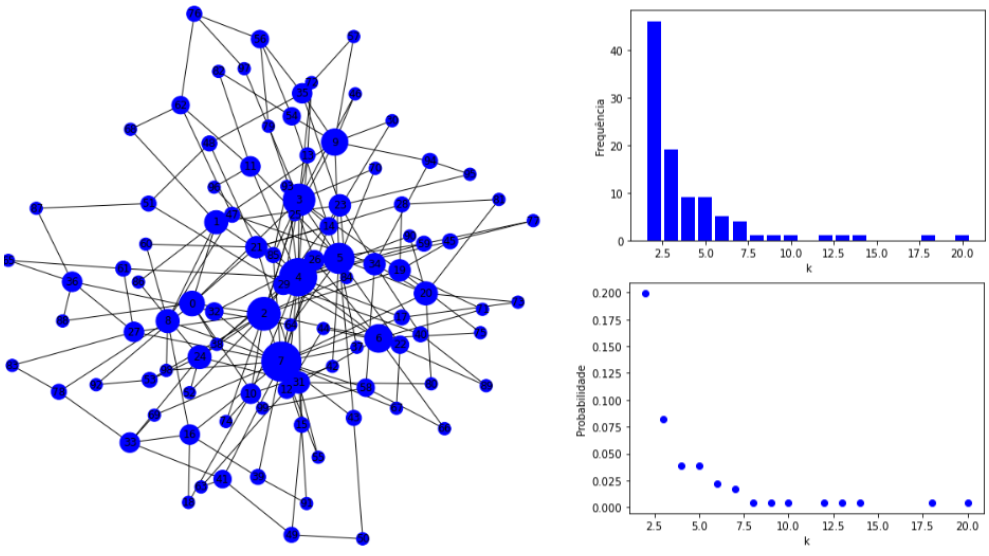
Com base na determinação dos graus, é possível extrair medidas estatísticas da rede, como o grau médio representado por $\langle k \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i$, os momentos dos graus $\langle k^n \rangle$, e avaliar a complexidade da rede. O grau médio da rede pode ser descrito pela seguinte formulação:

$$\langle k \rangle = \sum_{k=0}^{k_{max}} k p_k, \tag{2.2}$$

onde k_{max} representa o grau máximo presente na rede. Essa fórmula permite calcular o valor médio do grau considerando a distribuição do grau em toda a rede, visto na Figura 3.

A distribuição do grau: Baseada no grau, esta é uma métrica fundamental para identificar propriedades relevantes em redes complexas, como mostra a Figura 3. Ela revela que nós com poucas conexões são muito prováveis que aqueles altamente conectados. Esse padrão, observado em redes reais como a web (ALBERT; JEONG; BARABÁSI, 1999), roteadores (FALOUTSOS; FALOUTSOS; FALOUTSOS, 1999) e propagação de epidemias (BARTHÉLEMY *et al.*, 2004), segue uma **lei de potência**, ou seja, $p_k = k^{-\lambda}$ (NEWMAN, 2005). Essa lei, típica de redes **livres de escala**, indica a presença de **hubs**, poucos nós com muitas conexões, e implica invariância de escala e comportamento linear em escala log-log (ADAMIC, 2000), características estruturais cruciais para entender a robustez e a dinâmica de redes sociais.

Figura 3 – Representação de uma rede não direcionada com a distribuição de frequência do grau representada em um gráfico de barras e a função de massa de probabilidade p_k exibida em um diagrama de dispersão.



Fonte: Elaborada pela autora.

Assortatividade: O índice é um fator relevante na previsão de links. Quando nós com características semelhantes tendem a se conectar, a acurácia na previsão de conexões aumenta. Isso pode ser visto na influência de atributos como a preferência política (CONOVER *et al.*, 2011). A forma mais comum de assortatividade está relacionada ao grau dos nós, permitindo prever se aqueles com muitas conexões estão ligados entre si. Essa correlação pode ser quantificada

por meio do coeficiente de correlação de Pearson (FOSTER *et al.*, 2011). Entretanto, devido à complexidade das redes, outras abordagens analíticas têm sido amplamente empregadas. Um exemplo é a assortatividade grau a grau, que se baseia no grau médio dos vizinhos (k_{nn}), permitindo descrever k_{nn} em função do grau k (BARABÁSI, 2013), como representado na equação (2.3).

$$k_{nn}(k_i) = \sum_{k'} k' P(k'|k), \tag{2.3}$$

em que a probabilidade condicional $P(k'|k)$ representa a chance de que um nó com grau k esteja conectado a outro nó com grau k'.

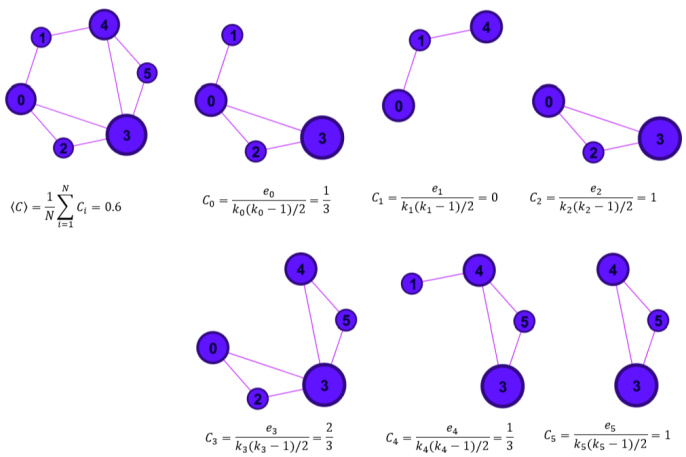
Transitividade: A transitividade, também conhecida como coeficiente de aglomeração, é uma medida comumente utilizada em redes para avaliar a tendência dos vizinhos de um determinado nó estarem conectados entre si. Ela indica o grau de conexão ou força das relações entre os vizinhos de um nó específico. Alguns dos coeficientes utilizados para quantificar matematicamente a transitividade de um nó são os coeficientes de cluster local e global (BARRAT; BARTHELEMY; VESPIGNANI, 2008).

Coeficiente de cluster local: Dada uma rede $G(V,E)$, o coeficiente de agrupamento local de um nó $v_i \in V$, denotado por C_i, é definido como:

$$C_i = \frac{2e_i}{k_i(k_i-1)}, \tag{2.4}$$

onde e_i é o número de conexões entre os vizinhos de v_i, e k_i é o número de vizinhos desse nó. Esse coeficiente mostra a tendência dos vizinhos de v_i se conectarem. O valor de C_i varia entre 0 e 1. Se $C_i = 1$, todos os vizinhos estão conectados; se $C_i = 0$, nenhum vizinho está conectado. Na ilustração da Figura 4, por exemplo, o nó 3 tem $C_3 = 2/3$, maior que o nó 4, que tem $C_4 = 1/3$, indicando que os vizinhos do nó 3 estão mais interligados. A média dos coeficientes locais para todos os nós é calculada por $\langle C \rangle = \frac{1}{N}\sum_{i=1}^{N} C_i$, onde N é o número total de nós.

Figura 4 – Coeficiente de agrupamento C_i para $i = 0, \cdots, 5$ e média local.



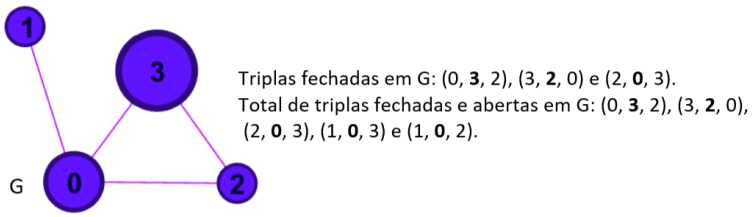
Fonte: Elaborada pela autora.

O coeficiente global C é calculado como:

$$C = \frac{3 \times \text{Número de triângulos}}{\text{Número de triplas conectadas de vértices}}, \tag{2.5}$$

onde triplas são grupos de três vértices que podem estar conectados ou não. O coeficiente varia de 0 a 1 (NEWMAN, 2003), um valor alto de C indica maior densidade da rede, isto é, com vizinhos bem conectados. Na Figura 5, o grafo G tem cinco triplas, sendo três fechadas.

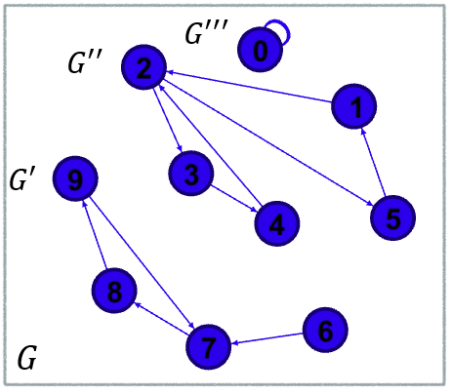
Figura 5 – Exemplo de triplas em uma rede.



Fonte: Elaborada pela autora.

Componentes conectados: Em redes sociais, é comum o interesse na acessibilidade de um vértice. Em uma rede conectada, cada nó pode ser alcançado a partir de outro por meio das arestas (BARRAT; BARTHELEMY; VESPIGNANI, 2008). Essa acessibilidade pode ser compreendida melhor definindo o caminho $\mathscr{P} i_0, i_n$ no grafo $G(V,E)$. Para um caminho, são definidos o conjunto de vértices $V_{\mathscr{P}} = i_0, i_1, \cdots, i_n$ e, o conjunto de arestas $E_{\mathscr{P}} = (i_0, i_1), \cdots, (i_{n-1}, i_n)$. Esse caminho conecta o vértice i_0 a i_n, e para cada par de nós, é descrito um caminho de comprimento n. Portanto, um grafo é dito conectado se houver pelo menos um caminho mais curto entre cada par de nós. Um componente de um grafo é um subgrafo conectado em uma rede (BARRAT; BARTHELEMY; VESPIGNANI, 2008). Por outro lado, pode não haver conexão. Por exemplo, o grafo G da Figura 6, os componentes são desconexos. Em $G'(V',E')$ e $G''(V'',E'')$, não há um caminho mais curto de distância finita que os conecte.

Figura 6 – Grafo G e seus componentes: G', G'' e G'''.



Fonte: Elaborada pela autora.

Nos componentes de um grafo, ao considerar a conexão, é importante levar em conta a direcionalidade das arestas. A direção das arestas permite verificar se um componente do grafo é fortemente conectado. Um componente fortemente conectado de um grafo G é definido como o maior subconjunto de vértices contido no grafo, no qual para cada par de vértices existe um caminho do vértice i_0 para i_n e um caminho de i_n para i_0 (MILLER; RANUM, 2013).

No grafo G da Figura 6, considerando-se o componente G'', é possível verificar que para cada par de vértices existe um caminho entre eles. Por exemplo, o caminho de 3 para 1 é descrito por $3,4,2,5,1$, e o caminho de 1 para 3 é $1,2,3$. Portanto, G'' é fortemente conectado, o que significa que é possível mover-se de um nó para outro por meio das arestas. Por outro lado, no componente G', existe um caminho de 6 para 9, mas não há um caminho de 9 para 6, devido à direção da aresta.

Centralidade: O trabalho de Rodrigues (2019) introduz uma compreensão fundamental das principais medidas utilizadas para caracterizar a centralidade de vértices em redes complexas. A centralidade, um conceito de extrema relevância, delineia o grau de importância de um vértice específico na disseminação de informações (RODRIGUES, 2019). Em cenários de propagação de epidemias, por exemplo, os vértices mais centrais são aqueles que promoverão a disseminação da doença com maior rapidez. Assim, no âmbito deste estudo, serão abordadas algumas dessas medidas para compreender mais profundamente como os vértices se manifestam globalmente na estrutura da rede.

A *Centralidade do grau* é, em essência, a representação do grau de um nó, o número de seus vizinhos mais próximos. Quando normalizada, quanto maior essa medida, maior será a centralidade do grau, indicando uma maior probabilidade de o nó ter o grau máximo (MEYBORG, 2014). Embora essa métrica seja capaz de identificar os nós mais interconectados na rede, ela possui suas próprias limitações. Primeiramente, trata-se de uma medida local, o que a torna inadequada para avaliar a centralidade global, especialmente quando esses nós altamente conectados (*hubs*) estão situados nas margens da rede (RODRIGUES, 2019).

Em contrapartida, o conceito de $k-core$ se refere a um subgrafo no qual todos os vértices possuem um grau mínimo de k. Essa métrica é amplamente empregada quando o objetivo é compreender a estrutura periférica e central de uma rede. No contexto da região periférica, é amplamente reconhecido que a rede é composta por nós com graus baixos, alguns dos quais têm grau zero, enquanto outros possuem grau um. A medida que se adentra mais profundamente na rede, o valor de k aumenta. A ideia subjacente a esse método é lapidar a rede de modo que o subgrafo inicial represente o componente central da mesma. Dessa forma, a abordagem consiste em remover, do núcleo de grau zero da rede, todos os nós com esse mesmo grau. Em seguida, procedemos à remoção, do núcleo de grau um, de todos os nós que possuem esse mesmo grau. Ao continuar avançando em direção ao centro da rede, se alcança seu componente principal, que é o subgrafo central.

A *Centralidade de proximidade* avalia a proximidade de um nó central em relação aos

demais nós no grafo. Essa métrica é definida em termos da distância do caminho mais curto, em outras palavras, o vértice mais central é aquele cuja distância para os demais é a menor. A centralidade de proximidade é determinada pela fração entre o número total de nós na rede e a soma das distâncias mais curtas, e é representada como:

$$C_i = \frac{N}{\sum_{j=1, j\neq i}^{N} d_{ij}} \tag{2.6}$$

Uma medida maior indica maior centralidade do nó. Entretanto, essa métrica tem suas limitações. Devido à variação estreita dos diâmetros das redes, muitos nós podem apresentar a mesma métrica, dificultando distinguir um nó central de outro não central (RODRIGUES, 2019).

Na *Centralidade do autovetor*, mensuramos a centralidade de um nó com base na centralidade dos demais. Portanto, a centralidade do autovetor de um nó é a soma das centralidades do autovetor de cada um dos seus vizinhos. Denotamos a centralidade do nó i como x_i, expressa por:

$$x_i = \frac{1}{\lambda} \sum_{k=1}^{N} A_{k,i} x_k, \tag{2.7}$$

em que $A_{k,i}$ é a matriz de adjacências, x_k é a centralidade do autovetor do nó k, e λ é uma constante de normalização. Em sua representação matricial, dada por $AX = \lambda X$, em que X é o vetor coluna cuja i -ésima entrada é x_i, ou seja, cada elemento de X indica a centralidade dos autovetores de diferentes nós. Dessa forma, o principal autovetor é o autovetor associado ao maior autovalor λ. Essa métrica sugere que, se um nó estiver conectado a outro de grande importância, sua própria importância tende a aumentar.

2.1.2 Modelos de Crescimento de Rede

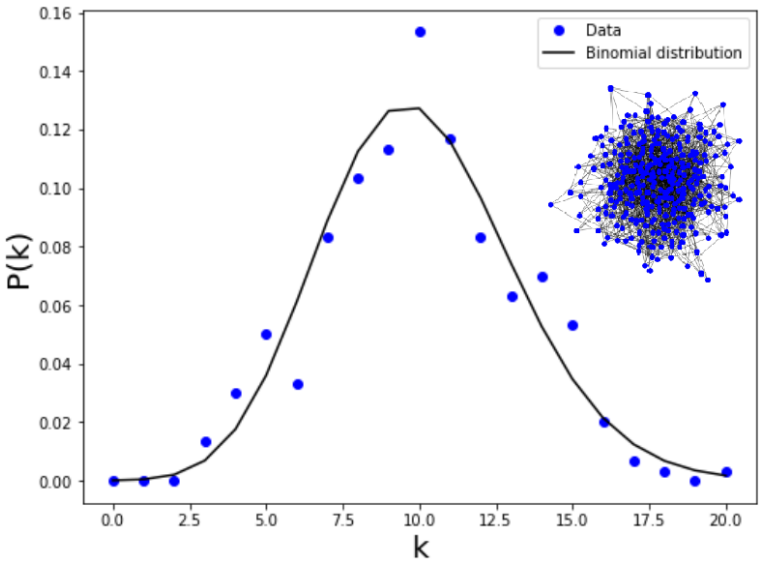
Neste tópico será discutido brevemente alguns dos mais citados modelos de rede. Esses modelos têm por finalidade gerar redes artificiais que apresentam propriedades já vistas e analisadas em redes reais. Para que redes artificiais sejam geradas, é preciso compreender os procedimentos pelos quais uma rede real evolui ao longo do tempo. Assim, modelos com fundamentação matemática foram criados visando imitar os procedimentos pelos quais estas se desenvolvem e, por meio computacional, criar redes artificiais para predizer como esta evoluirá. Dessa forma, métricas topológicas já revisadas anteriormente, ou um subconjunto delas, como distribuição do grau, coeficiente de agrupamento, entre outras, também devem estar presentes na rede simulada. As principais e mais famosas propriedades verificadas em uma rede do mudo real são: **alta média de agrupamento local**; propriedade de **mundo pequeno** e a distribuição do grau ser **livre de escala**.

Em dados históricos, a primeira propriedade foi inicialmente observada em uma rede que descrevia as relações entre os membros de um clube (ZACHARY, 1977), posteriormente verificada em redes sociais (MISLOVE *et al.*, 2007) e, mais tarde, em uma rede do Facebook (UGANDER *et al.*, 2011). A segunda, foi investigada mediante experimentos que mediram a

distância média entre indivíduos em uma rede social de norte-americanos (MILGRAM, 1967). Em seguida, Korte e Milgram (1970) aplicaram a mesma metodologia em um estudo com remetentes brancos de Los Angeles e destinatários brancos e negros em Nova York, o que popularizou o conceito como o fenômeno do **mundo pequeno**. Por fim, a terceira propriedade, denominada **livre de escala**, como já mencionado, propõe que a distribuição de graus segue uma lei de potência, ou seja, uma função sem parâmetro de escala. Dessa forma, uma rede artificial deve exibir uma distribuição de graus dada por $p(k) \sim k^{-\lambda}$, onde $2 \leq \lambda \leq 3$ (PASTOR-SATORRAS *et al.*, 2015).

Modelo de Crescimento de Rede Aleatória de Erdös-Rényi: Este modelo de rede pressupõe que, começando com N nós, é possível gerar uma rede conectando aleatoriamente pares de nós até que M bordas sejam formadas. No entanto, esse modelo apresentou algumas limitações, principalmente em relação à falta de clareza sobre quando criar bordas entre os nós (DOROGOVTSEV; MENDES, 2022). Aprimorando esse modelo, (GILBERT, 1959) propôs substituir o número fixo de bordas M por uma probabilidade p. Isso permite que cada par de nós selecionados seja conectado com base na probabilidade p, dando origem a um processo probabilístico Bernoulli. No entanto, o modelo não satisfaz o alto coeficiente de agrupamento local médio observado em redes reais (DOROGOVTSEV; MENDES, 2022). Na Figura 7, é mostrada a distribuição de grau de um ensaio da rede aleatória proposta por Erdös-Rényi com $\langle k \rangle = 8$, $p \sim 0.03$ e $N = 300$. Nela, é possível verificar que o comportamento da distribuição não segue uma lei de potência.

Figura 7 – Exemplo das distribuições de grau de um ensaio da rede aleatória proposta por Erdös-Rényi com $\langle k \rangle = 8$, $p \sim 0.03$ e $N = 300$.

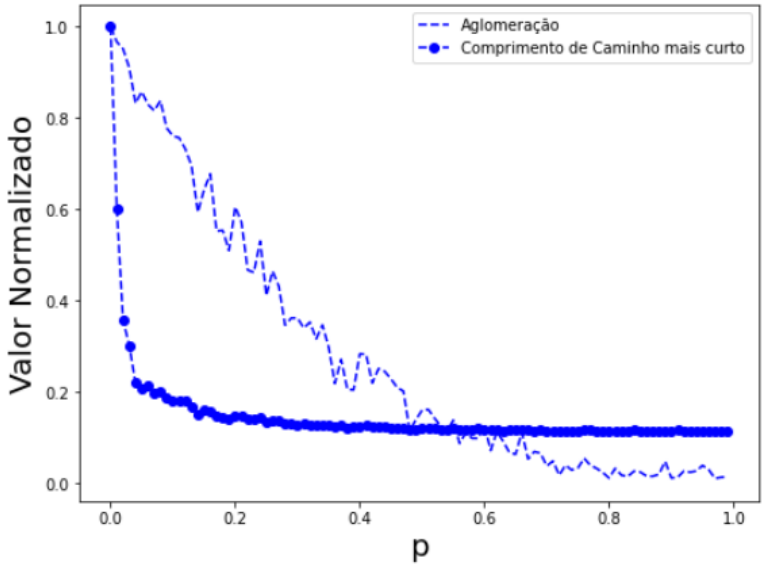


Fonte: Elaborada pela autora.

Modelo de Crescimento de Watts-Strogats: Formado inicialmente por uma estrutura de rede regular (WATTS; STROGATZ, 1998), em que os nós têm o mesmo grau e estão

conectados apenas aos seus vizinhos. A partir dessa lista de nós, o algoritmo começa escolhendo nós aleatoriamente e conectando uma de suas arestas a outro vértice que ainda não tenha sido conectado ao nó inicial com probabilidade p, evitando autoconexões e duplicação de arestas. Esse processo de reconexão de bordas continua até que a estrutura regular da rede seja quebrada e a rede se torne mais aleatória. Embora o modelo satisfaça a propriedade de mundo pequeno, o alto coeficiente de agrupamento e o baixo comprimento médio do caminho (veja Figura 8), ele não preserva a propriedade de distribuição de grau sem escala.

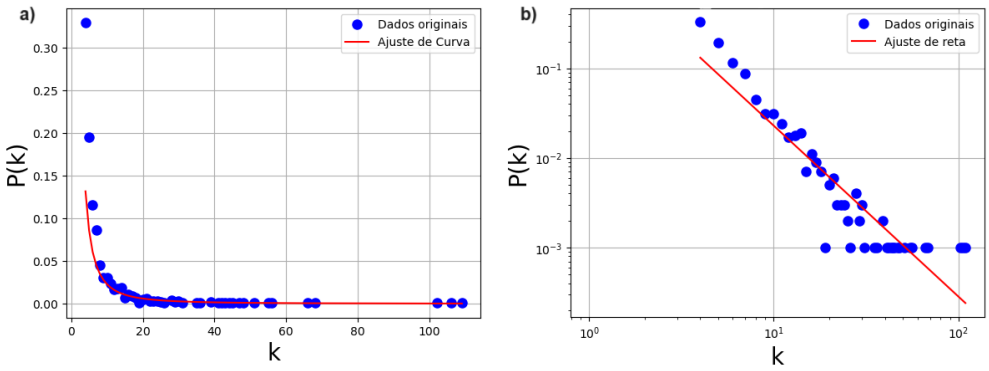
Figura 8 – No exemplo de simulação de redes de Watts e Strogatz (1998), quando variando p, é possível verificar que o comportamento do coeficiente de aglomeração é alto, ao passo que o comprimento médio de caminho mais curto é sempre inferior.



Fonte: Elaborada pela autora.

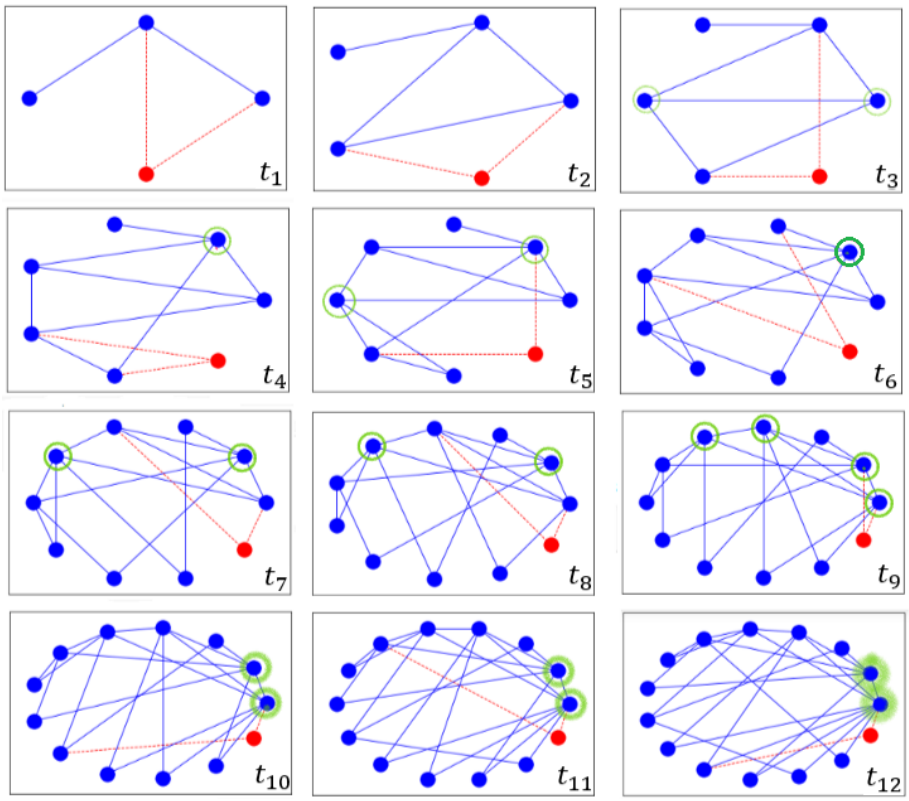
Modelo de Crescimento de Rede de Barabasi-Albert: Quanto ao terceiro modelo, um nó se aproxima da rede com um grau fixo de m e se junta a ela, começando com m_0 nós iniciais, um de cada vez. Quando um nó com m de arestas entra na rede, ele seleciona os nós aos quais se conectará. Esse modelo também é conhecido como modelo de conexão preferencial. A seleção é baseada no grau atual dos nós já conectados na rede. Assim, a probabilidade de esses nós adquirirem uma nova borda depende de seu grau, ou seja, é proporcional ao seu grau atual. Essa preferência por nós de nível superior leva ao surgimento de redes sem escala, em que a distribuição do grau segue uma lei de potência (BARABÁSI, 2013), veja Figura 9. A regra adotada pelo modelo durante o crescimento da rede permite o surgimento de nós altamente conectados, conhecidos como *hubs*. Em redes reais, é comum observar esses *hubs* (Figura 10), que são nós que recebem muitas conexões e, à medida que a rede cresce, esses nós se tornam cada vez mais conectados.

Figura 9 – Distribuição de grau da rede de Barabási–Albert. a) Os dados são visualizados em escala normal. b) Os dados são visualizados em escala log-log.



Fonte: Elaborada pela autora.

Figura 10 – Simulação da rede BA, iniciada com $m_0 = 3$ nós e $m = 2$ arestas. Mediante a evolução temporal e na medida em que nós exteriores, em vermelho, se unem a rede emergem os *hubs*, destacados em cor verde.



Fonte: Elaborada pela autora.

2.2 Processos Dinâmicos

Dentre a vasta gama de processos dinâmicos, são apresentados quatro categorias de sistemas considerados clássicos: um sistema caótico, um oscilatório e um populacional predador-presa. A quarta categoria compreende seis modelos epidemiológicos, os quais descrevem a

propagação de doenças por meio de diferentes estruturas compartimentais.

2.2.1 Sistema caótico: Atrator de Lorenz

Um exemplo clássico de dinâmica caótica que se origina de um modelo de convecção atmosférica (LORENZ, 1963), e independentemente da dinâmica de laser de modo único (HAKEN; SAUERMANN, 1963), é dado pelas equações de Lorenz-Haken, que são descritas por

$$\begin{aligned} \dot{x} &= \sigma(y-x) \\ \dot{y} &= -x(\rho-z)-y, \\ \dot{z} &= xy-\beta z \end{aligned} \tag{2.8}$$

onde, considerando a dedução de Lorenz para a atmosfera, x é proporcional à taxa de convecção de um fluido, y à variação de temperatura horizontal e z à variação de temperatura vertical. σ, ρ e β são constantes que se relacionam com propriedades desse sistema físico.

2.2.2 Sistema oscilatório: O pêndulo não linear

O pêndulo é um dos modelos fundamentais da física, servindo de base para o entendimento do Movimento Harmônico Simples (MHS). O sistema de equações diferenciais que rege a dinâmica do pêndulo simples é dado pela seguinte equação:

$$\ddot{\theta} = -\frac{g}{l}\sin\theta \tag{2.9}$$

onde θ é o deslocamento angular (BELÉNDEZ *et al.*, 2007).

2.2.3 Sistema populacional: Dinâmica predador-presa de Lotka-Volterra

Um modelo fundamental na ecologia, as equações de Lotka-Volterra são frequentemente utilizadas (DIZ-PITA; OTERO-ESPINAR, 2021) para simular e compreender a interação entre populações de predadores e presas em uma cadeia alimentar. O sistema de equações que descreve este modelo é dado por:

$$\begin{aligned} \dot{u} &= \alpha u - \beta uv \\ \dot{v} &= -\gamma v + \delta uv, \end{aligned} \tag{2.10}$$

em que u representa a densidade populacional das presas e v a dos predadores. $\alpha, \beta, \gamma, \delta \in \mathbb{R}^+$, com α sendo a taxa máxima de crescimento per capita das presas, β o efeito dos predadores na taxa de mortalidade das presas e, complementando os dois anteriores, γ é a taxa de mortalidade per capita dos predadores e δ o efeito das presas na taxa de crescimento dos predadores.

Ao considerar o equilíbrio entre as duas populações, isto é, $u = \gamma\delta$ e $y = \alpha\beta$, respectivamente, para as presas e predadores (WANGERSKY, 1978), ambos dependem dos parâmetros

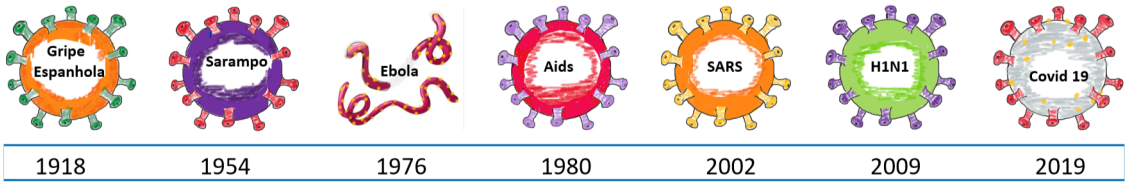
um do outro, o que, como consequência, significa que aumentar a taxa de crescimento das presas beneficia os predadores.

2.2.4 Sistemas Epidêmicos

Ao longo dos anos, a humanidade tem sido confrontada com uma série de epidemias significativas. A resistência observada em diversos agentes, incluindo vírus, bactérias, parasitas e outros microrganismos patogênicos, que possuem a habilidade de se propagar rapidamente entre indivíduos, frequentemente resultando em surtos epidêmicos, encontra paralelos em ícones históricos marcantes. Entre esses exemplos, destaca-se a peste ou praga de Atenas, considerada um dos eventos mais antigos registrados. Essa epidemia ocorreu por volta de 430 a.C. e dizimou cerca de 25% da população da cidade (LITTMAN, 2009). Outra epidemia de destaque é a peste-negra, uma das mais famosas da história, que ocorreu na Europa durante o século XIV, período da Idade Média. Estima-se que essa epidemia tenha causado a morte de aproximadamente um quarto a um terço da população, com a maioria dos indivíduos infectados sobrevivendo apenas de dois a três dias (MURRAY, 2001). Esses são apenas dois exemplos de epidemias antigas, mas ao longo da história, diversas outras ocorrências foram registradas.

A cronologia de surgência de algumas infecções virais nos últimos dois séculos pode ser vista na Figura 11. Nessa figura, o Sarampo e Ebola foram considerados epidemias, enquanto outras doenças alcançaram proporções pandêmicas. Para obter mais detalhes sobre a cronologia de vírus, epidemias e pandemias, é possível consultar as obras de: Montenegro, Batista e Stroppa (2021), Costa e Merchan-Hamann (2016).

Figura 11 – Cronologia de vírus populares nos séculos XX e XXI.



Fonte: Elaborada pela autora.

Essas memórias históricas fornecem fortes evidências da necessidade de aprimorar os estudos voltados ao enfrentamento desses tipos de eventos. Nesse contexto, naturalmente, surgem questionamentos sobre como ocorre a disseminação de doenças e de que forma elas se espalham. Essas questões têm estimulado o interesse de diversas áreas do conhecimento. Embora biólogos, médicos e epidemiologistas tenham se empenhado em descrever os diferentes patógenos, suas mutações e formas de contágio, a necessidade de modelagem matemática para compreender como doenças contagiosas se espalham e a taxa desse processo são pontos-chave para conter pandemias.

Modelos matemáticos como o proposto por Kermack e McKendrick (1927), que descrevem um sistema de equações diferenciais ordinárias, têm se tornado não apenas uma ferramenta útil, mas também imprescindível na compreensão da propagação de doenças.

A modelagem de sistemas epidemiológicos vem sendo amplamente difundida por físicos e cientistas de redes que tem muito a contribuir, principalmente quando se trata de responder questões que tangem a disseminação de doenças em meios heterogêneos. Todas as doenças contagiosas basicamente se espalham por meio de uma rede. Então, a modelagem de uma doença que está se espalhando por ela deve considerar fatores fundamentais que são óbvios: o patógeno que se espalha e a própria rede. O patógeno permite descrever quão transmissível ele é, pois, algumas doenças são mais epidêmicas que outras. A rede, no que lhe concerne, é de fundamental importância, pois redes mais densas têm uma maior chance de propagar doenças mais rapidamente quando comparadas com redes de menor densidade.

2.2.4.1 Modelos Epidemiológicos Clássicos

Modelos teóricos matemáticos voltados à epidemiologia foram a princípio propostos visando compreender a forma pela qual uma epidemia pode espalhar uma doença em uma população específica. Tais modelos, como já mencionados, foram introduzidos inicialmente por Kermack e McKendrick (1927), mediante um sistema de equações diferenciais ordinárias em que duas probabilidades a princípio eram consideradas. A primeira corresponde a probabilidade que um indivíduo (nó) possa infectar outro (taxa de infecção β) e a segunda é de que o indivíduo infectado seja curado (taxa de recuperação γ). Assim, por meio desse sistema de equações diferenciais ordinárias foi possível determinar o comportamento da curva de infectados ao passo que a epidemia se espalha ao longo do tempo. Muitos modelos têm sido propostos até hoje, mas os modelos considerados clássicos e mais simples são: SIR e SIS resumidos abaixo.

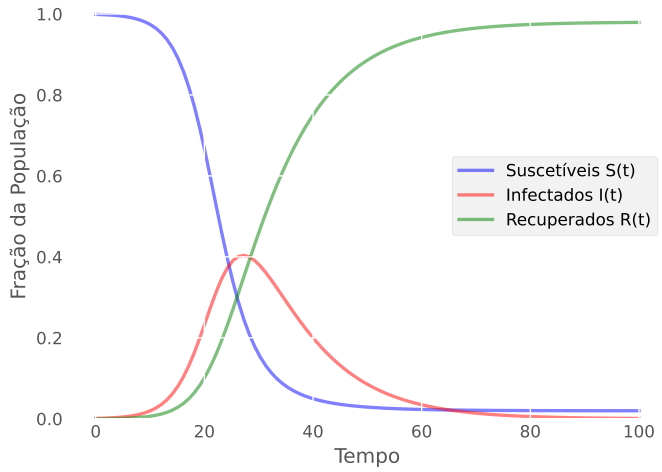
SIR: O modelo SIR assume três estados no ciclo de vida de uma infecção, suscetível (S), infectado (I) e Recuperado (R). Este pode ser descrito pelo sistema de equações 2.11.

$$\begin{cases} \frac{dS}{dt} = -\beta SI \\ \frac{dI}{dt} = \beta SI - \gamma I, \\ \frac{dR}{dt} = \gamma I \end{cases} \tag{2.11}$$

em que $S+I+R$, corresponde ao número total N da população. Note que o sistema é simples, no estado S, o indivíduo está sujeito a ser infectado com uma taxa β. Assim, cada nó infectado I, infecta um nó suscetível S com probabilidade β, de modo que há um número S de nós suscetíveis a taxa β, resultando na taxa de variação da população suscetível no tempo, $-\beta IS$. Para a taxa de infecção, o mesmo número de nós que se tornarão infectados, mas alguns poderiam ser recuperados com probabilidade γ, isto é, a taxa de população recuperada é γI e taxa de mudança da população infectada é $\beta IS - \gamma I$. O gráfico simulado na Figura 12, mostra o padrão de comportamento das populações S, I e R ao longo do tempo (KEELING; ROHANI, 2011).

Nesse, o desempenho da fração das populações S, I e R ao longo do tempo mostra que enquanto a população S é infectada, sua proporção diminui ao passo que ocorre um aumento da população infectada até chegar ao pico da curva $I(t)$, assim como o número de recuperados tende a aumentar com o passar do tempo. Esse modelo admite que no estado recuperado os indivíduos adquirem imunidade e não são mais infectados.

Figura 12 – Gráfico do modelo SIR para propagação de uma epidemia, nele é descrita a relação entre a fração da população e o tempo. A solução numérica das equações para a população de $N = 1000$ indivíduos, com o início da epidemia dado por um indivíduo infectado no tempo, $I(0) = 1$, número de recuperados em t_0, $R(0) = 0$ e total de indivíduos suscetíveis em t_0 de $S(0) = N - I(0) - R(0)$, com probabilidades, $\beta = 0.4$ e $\gamma = 0.1$.



Fonte: Elaborada pela autora.

SIS: Deriva diretamente do modelo SIR, nele não há mais o estado recuperado. O modelo atual pode ser descrito pela Equação 2.12, após se tornar infectado, o indivíduo pode novamente se tornar suscetível, a dinâmica ode ser visto na Figura 13.

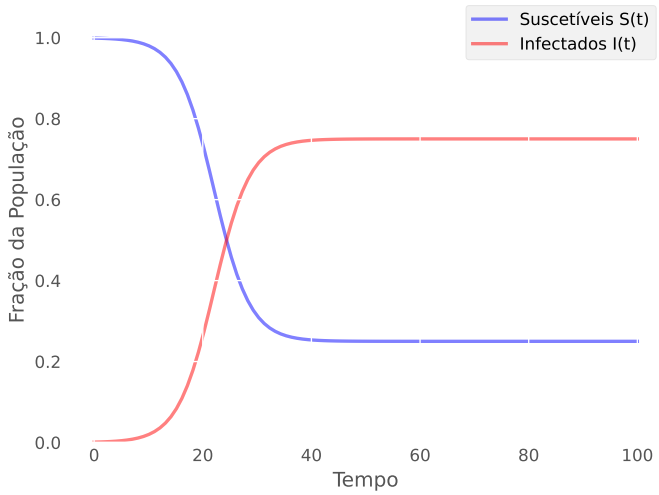
$$\begin{cases} \frac{dS}{dt} = -\beta IS + \gamma I \\ \frac{dI}{dt} = \beta IS - \gamma I. \end{cases} \tag{2.12}$$

Note que a variação dos suscetíveis ao longo do tempo mostra que enquanto a população S é infectada, sua proporção diminui, ao passo que ocorre um aumento da população infectada no decorrer do tempo.

Expandindo ainda mais os sistemas SIR e SIS mencionados anteriormente, vários outros modelos compartimentais podem ser criados usando outros estados possíveis para os indivíduos em uma rede. Neste estudo, são usados os seguintes

- SEIR (ANNAS *et al.*, 2020), adicionar uma categoria exposta (E) à SIR padrão;

Figura 13 – Gráfico do modelo SIS para propagação de uma epidemia, onde é descrita a relação entre a fração da população e o tempo. A solução numérica das equações para a população de $N = 1000$ indivíduos, com o início da epidemia dada por um indivíduo infectado no tempo t_0, $I(0) = 1$, e total de indivíduos suscetíveis em t_0 de $S(0) = N - I(0)$, com probabilidades, $\beta = 0.4$ e $\gamma = 1/10$.



Fonte: Elaborada pela autora.

- SEIRD (KOROLEV, 2021), expandir o modelo SEIR adicionando uma contagem de falecidos (D);
- SIRV (OKE *et al.*, 2019), em que os indivíduos vacinados (V) também são considerados na dinâmica;
- SIRS (HU *et al.*, 2019), permitindo a reinfecção no modelo SIR.

As equações desses sistemas podem ser vistas na Tabela 1.

Tabela 1 – Sistemas de equações diferenciais ordinárias de modelos epidêmicos.

SIS	SIR	SEIR	SEIRD	SIRV	SIRS
$\dot{S} = -\frac{\beta}{N}IS + \gamma I$ $\dot{I} = \frac{\beta}{N}IS - \gamma I$	$\dot{S} = -\frac{\beta IS}{N}$ $\dot{I} = \frac{\beta IS}{N} - \gamma I$ $\dot{R} = \gamma I$	$\dot{S} = -\frac{\beta IS}{N}$ $\dot{E} = \frac{\beta IS}{N} - \sigma E$ $\dot{I} = \sigma E - \gamma I$ $\dot{R} = \gamma I$	$\dot{S} = -\frac{\beta SI}{N}$ $\dot{E} = \frac{\beta SI}{N} - \sigma E$ $\dot{I} = \sigma E - \gamma I$ $\dot{R} = \gamma(1-\mu)I$ $\dot{D} = \gamma\mu I$	$\dot{S} = -\frac{\beta SI}{N} - \varepsilon S$ $\dot{I} = \frac{\beta SI}{N} - \gamma I$ $\dot{R} = \gamma I$ $\dot{V} = \varepsilon S$	$\dot{S} = -\frac{\beta SI}{N} + \delta R$ $\dot{I} = \frac{\beta SI}{N} - \gamma I$ $\dot{R} = \gamma I - \delta R$

2.2.5 Processos dinâmicos epidemiológicos em redes complexas

Ainda no campo dos estudos epidêmicos, um dos primeiros dilemas perante uma nova patologia é se ela de fato irá se propagar e que valores assumem os parâmetros do sistema que descreve a dinâmica. Toda epidemia expõe o número reprodutivo da infecção que pode ser medido e definido por

$$R_0 = \frac{\beta}{\gamma}, \tag{2.13}$$

sendo interpretado como a intensidade com que uma infecção ou vírus se espalha. Essa métrica é útil na identificação da ocorrência ou não de uma epidemia. Sempre que $R_0 < 1$ a doença tende desaparecer e quando $R_0 > 1$, a doença persiste e se propaga (BARRAT; BARTHELEMY; VESPIGNANI, 2008).

Em Gómez *et al.* (2010), descobriu-se que existe uma relação entre o R_0 e a matriz de adjacência da rede de contatos. Quando uma epidemia se espalha, é importante considerar, não somente aspectos referentes à patogenia da doença, mas também a estrutura da rede, a qual essa doença está se propagando. Em função disso, esses modelos epidêmicos que até então eram considerados simples, perdem essa característica. Quando um processo dinâmico, como, por exemplo, a propagação de epidemias, é moldados sob uma rede, onde cada um de seus nós, indivíduo, pertencem a um estado, os modelos epidemiológicos tornam-se relativamente complicados, sendo tratados como modelos epidemiológicos em redes complexas.

As pesquisas dedicadas à disseminação de epidemias em redes complexas se deu devido as descobertas de que a estrutura de conexão entre indivíduos em tais redes está relacionado ao número reprodutivo da infecção, isto é, a força com que o vírus se espalha. Como base na matriz de adjacência é possível identificar indivíduos (nós), que sejam potenciais propagadores. Essa identificação possibilita alterar a estrutura da rede, a qual pode ser modificada mediante a políticas de vacinação ou imposição de *lockdown*. A mudança nas características da rede permitem o controle de β, reduzindo-o e alcançando limites para os quais o sistema de saúde possa suprir a demanda.

Dessa forma, a propagação epidêmica em redes complexas vem sendo modelada por meio de abordagens como Cadeia de Markov de tempo discreto (GÓMEZ *et al.*, 2010) e Teoria de Campo médio Heterogêneo (KISS; MILLER; SIMON, 2017a). Entretanto, como nenhuma dessas abordagens gera uma representação simbólica de fácil interpretação, abre-se uma lacuna entre a dinâmica observada e a agilidade de inferir sobre ela.

2.3 Regressão Simbólica

Segundo Cava *et al.* (2021), a Regressão Simbólica é tradicionalmente classificada como uma técnica de aprendizado de máquina. No entanto, também é pertinente enquadrá-la como uma abordagem do campo da Inteligência Artificial, uma vez que faz uso de algoritmos evolutivos, os quais são característicos dessa área. De acordo com Schmidt e Lipson (2009b), trata-se de uma técnica de aprendizado supervisionado que busca emular o raciocínio científico humano na descoberta de sistemas dinâmicos a partir de dados, definição corroborada por Cranmer (2023). Ainda assim, é importante destacar que a Regressão Simbólica pode transitar entre os paradigmas supervisionado e não supervisionado, especialmente quando incorpora conhecimentos prévios sobre a dinâmica geradora dos dados. Atualmente, tais algoritmos têm demonstrado a capacidade de reduzir significativamente o tempo necessário para identificar leis físicas ocultas nos dados,

desde que estes atendam às premissas do modelo adotado. O aprimoramento contínuo dessa técnica tem sido impulsionado tanto pelos avanços recentes em algoritmos de aprendizado de máquina quanto pelo progresso do arcabouço estatístico nas últimas décadas (LANGLEY, 1977).

Do ponto de vista matemático, o principal objetivo dos modelos de Regressão Simbólica (SR) é encontrar uma função $\hat{y}(\mathbf{x}) = \hat{\phi}(\mathbf{x}, \hat{\theta})$, com $\hat{\phi} : \mathbb{R}^d \longrightarrow \mathbb{R}$, a partir de um conjunto de dados $\mathscr{D}$, definido como

$$\mathscr{D} = (\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_i, y_i), \cdots, (\mathbf{x}_N, y_N), \tag{2.14}$$

em que $\mathbf{x}_i \in \mathbb{R}^d$ representa os vetores de entrada, e y_i o valor-alvo ou resultado que se deseja prever por meio do processo de aprendizagem (CAVA *et al.*, 2021).

As subseções a seguir são dedicadas à descrição de cinco modelos contemporâneos de regressão simbólica analisados nesta tese.

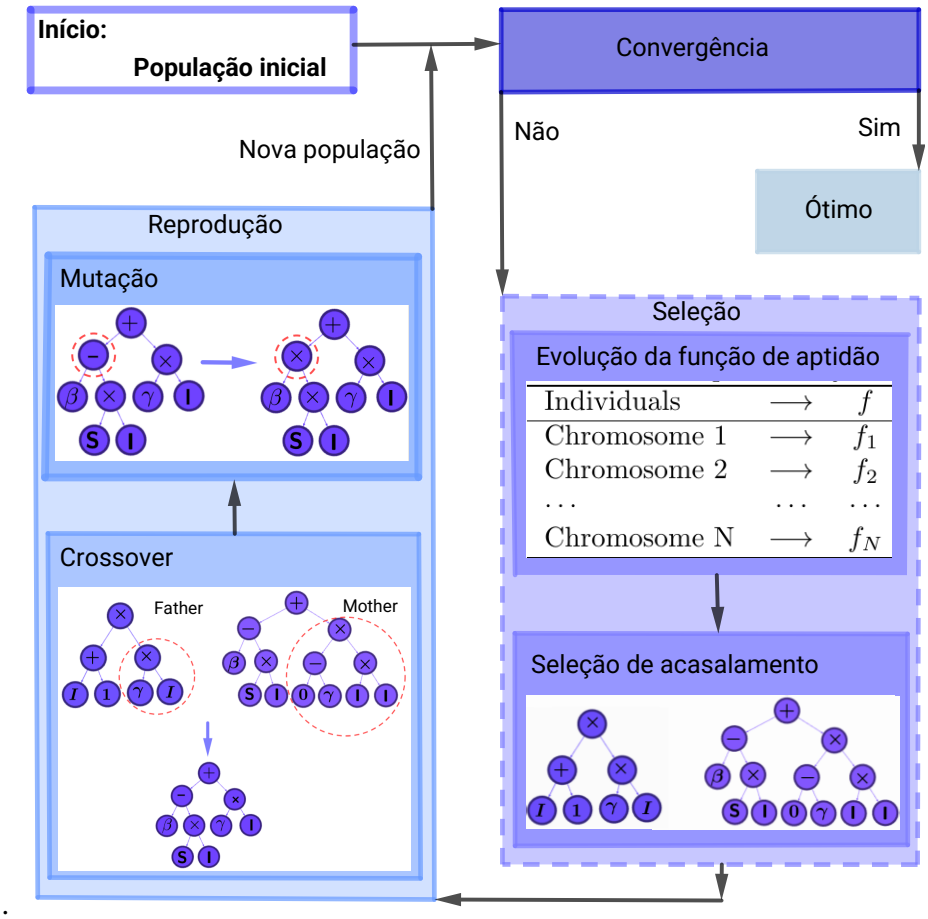
2.3.1 GPLearn

Baseado em Algoritmos Genéticos, o mecanismo inicia seu processo de busca por meio de uma população de expressões simbólicas geradas aleatoriamente. Essas expressões correspondem a combinações dos elementos do espaço de busca, que é composto por operadores e funções definidos pelo usuário ($\{+, -, *, \cdots, \sin, \cdots\}$). As diversas combinações geram expressões simbólicas variadas, cada uma com um código genético único representado por um cromossomo codificado, descrevendo a composição genética de cada indivíduo.

A avaliação da adequação de cada indivíduo, $f(\boldsymbol{x})$, é realizada por meio de transformações, ou seja, os mapeamentos de genótipo-fenótipo (f_g) e fenótipo-adequação (f_p) (ROTHLAUF, 2006), em que $f = f_p(f_g(\boldsymbol{x}^g))$ é determinado. Com base na qualidade do ajuste, são utilizados os conceitos propostos por Darwin, em que os indivíduos mais aptos são selecionados e os menos aptos a sobreviver são eliminados (SIVANANDAM; DEEPA, 2008). Agora, com uma população nova e um pouco menor, esses indivíduos passam por um processo de seleção em pares. Em seguida, vem a reprodução, com base nos princípios fundamentais das teorias de herança genética desenvolvidas por Mendel em 1865. Nessa etapa, o algoritmo realiza cruzamentos e mutações, gerando uma nova população que será submetida novamente a todo o processo até que ocorra a convergência e a expressão desejada seja encontrada. O algoritmo além de eficiente, permite a implementação de funções externas ao código padrão.

Uma visão geral simplificada dessa configuração pode ser vista em Figura 14. Para utilizar esse algoritmo, é necessário adaptar a série temporal a um problema de regressão, em que cada linha dos dados está relacionada a um passo de tempo. A taxa de variação do sistema dinâmico é ajustada uma de cada vez com as variáveis pertinentes ao problema.

Figura 14 – Uma ilustração simples da pesquisa conduzida pelo GPLearn



Fonte: Elaborada pela autora.

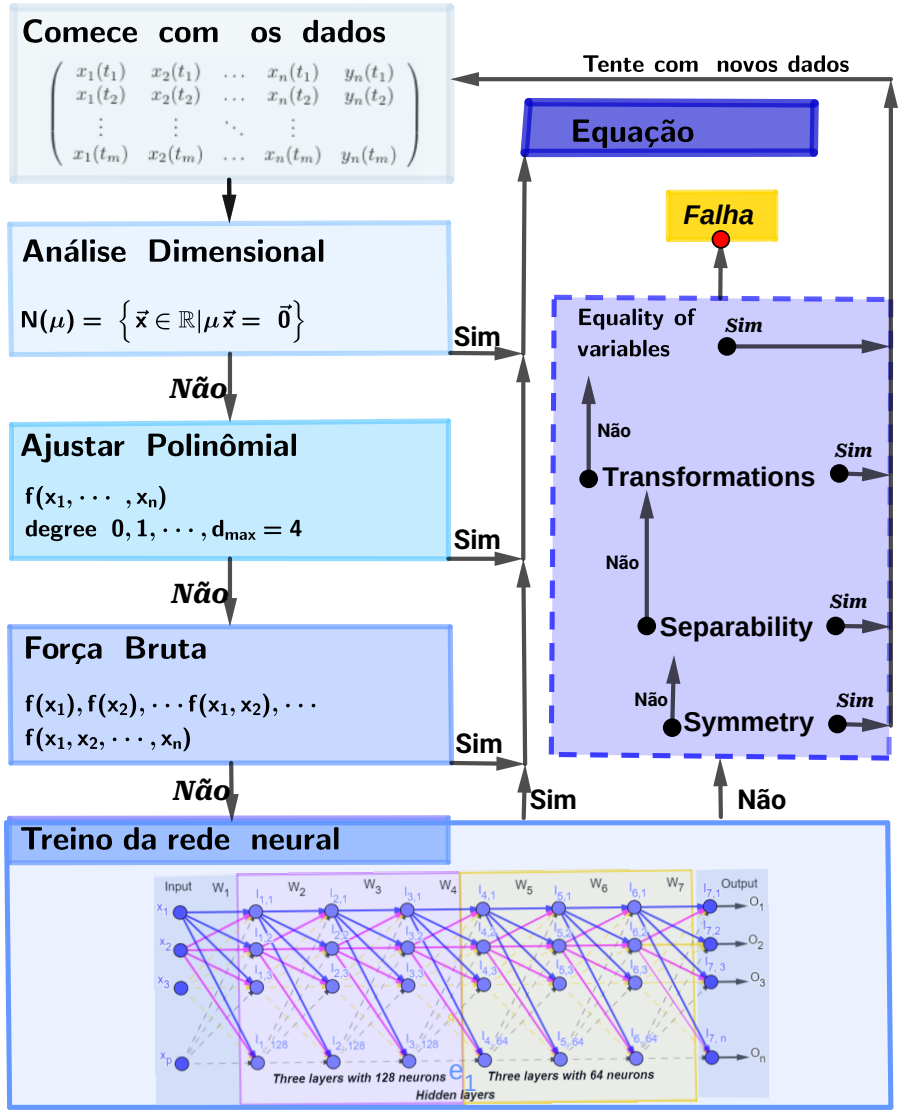
2.3.2 AI-Feynman

O AI-Feynman, desenvolvido no Laboratório de Inteligência Artificial do MIT, é um dos métodos RS mais conhecidos, devido à sua alta precisão no domínio da física. A principal característica desse algoritmo é sua semelhança com a modelagem física. Sua eficácia se atribui ao fato de que as expressões comumente estudadas em física são, em sua maioria, composicionais, ou seja, uma função f pode ser vista como uma combinação de pequenos conjuntos de funções elementares, e o AI-Feynman consegue operar semelhantemente. O conjunto de ferramentas utilizado por esse método varia desde processos simples na busca por representações simbólicas até a aplicação de técnicas de aprendizado de máquina de última geração. A expressão é gerada por meio de uma codificação em árvore utilizando notação polonesa reversa (UDRESCU; TEGMARK, 2020). O algoritmo foi aprimorado por meio de estratégias que exploram propriedades simplificadoras comuns em processos físicos naturais, como a análise dimensional, uma abordagem direta que resulta em uma redução considerável do espaço inicial de variáveis.

Além disso, o método inclui ajuste polinomial, de graus zero a quarto, por mínimos quadrados não lineares. Também utiliza um algoritmo de busca exaustiva, gerando um espaço

que abrange todas as expressões possíveis, no espaço de tempo especificado para a busca, das mais simples às mais complexas, com diferentes combinações de parâmetros (força bruta). O ajuste também pode ser realizado utilizando estruturas baseadas em redes neurais, o que facilita a investigação de propriedades de simetria e separabilidade. Por fim, o algoritmo também contempla verificação de igualdade e transformações dos dados. Um esquema simplificado da funcionalidade do algoritmo pode ser visto na Figura 15.

Figura 15 – Esquema do procedimento utilizado pelo método AI-Feynman, inspirado em princípios da física.



Fonte: Elaborada pela autora.

Conforme os desenvolvedores, o algoritmo se mostrou muito superior ao Eureka, o melhor algoritmo até aquele momento.

2.3.3 SINDy

O PySINDy é uma biblioteca em Python desenvolvida para a identificação esparsa de sistemas dinâmicos não lineares, seguindo o método conhecido como SINDy (Sparse Identification of Nonlinear Dynamics). Essa abordagem se baseia em dados experimentais e em um conjunto de ferramentas que permitem inferir as equações que regem sistemas físicos, ou seja, extrair as leis dinâmicas diretamente a partir de dados observados.

A estrutura fundamental do método pode ser representada pela forma geral de um sistema dinâmico que descreve a evolução temporal de um vetor de estado:

$$\frac{d}{dt}x(t) = f(\mathbf{x}(t)). \tag{2.15}$$

Nesse contexto, assume-se que a dinâmica f é esparsa em relação às variáveis de estado $\mathbf{x}(t) \in \mathbb{R}^n$ (QUADE *et al.*, 2018). Isso significa que a derivada de cada variável tende a depender de poucas funções em um espaço maior de possíveis termos candidatos. Por exemplo, na seguinte combinação linear:

$$f\_i(x) = \xi\_1\theta\_1(x) + \xi\_2\theta\_2(x) + \cdots + \xi\_k\theta\_k(x), \tag{2.16}$$

se as funções básicas θ_j forem bem escolhidas, a maioria dos coeficientes ξ_j se anula, evidenciando a estrutura esparsa. O SINDy utiliza essa característica para aplicar uma regressão esparsa e identificar os coeficientes relevantes.

O algoritmo inicia-se, então, com a construção das seguintes matrizes:

$$\mathbf{X} = \begin{bmatrix} x_1(t_1) & x_2(t_1) & \cdots & x_n(t_1) \\ x_1(t_2) & x_2(t_2) & \cdots & x_n(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(t_m) & x_2(t_m) & \cdots & x_n(t_m) \end{bmatrix}, \quad \dot{\mathbf{X}} = \begin{bmatrix} \dot{x_1}(t_1) & \dot{x_2}(t_1) & \cdots & \dot{x_n}(t_1) \\ \dot{x_1}(t_2) & \dot{x_2}(t_2) & \cdots & \dot{x_n}(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ \dot{x_1}(t_m) & \dot{x_2}(t_m) & \cdots & \dot{x_n}(t_m) \end{bmatrix},$$

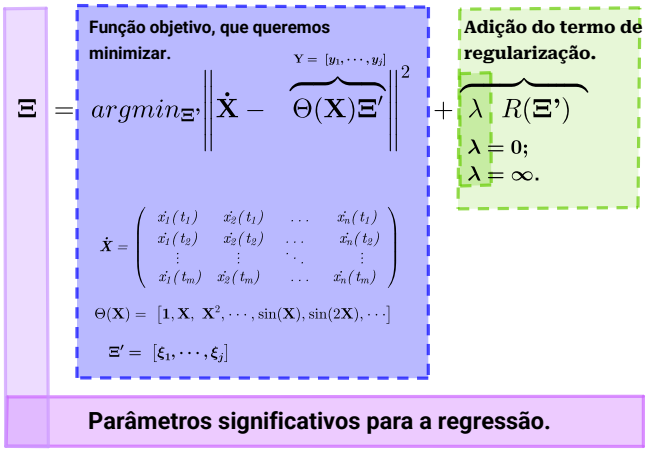
em que a matriz $\mathbf{X}$ é composta pelas variáveis medidas pelo sistema, com n variáveis medidas em m tempos, ou seja, a matriz com os dados de uma série temporal. Ao diferenciar $\mathbf{X}$, obtém-se $\dot{\mathbf{X}}$, a matriz alvo de derivadas no tempo e objeto do SINDy.

Na sequência, é necessário compor a matriz $\Theta(X) = [\theta_1(\mathbf{X}), \theta_2(\mathbf{X}), \cdots, \theta_l(\mathbf{X})]$, onde está uma lista de bibliotecas candidatas na formulação da expressão simbólica procurada. Cada $\theta(\mathbf{X})$ dessa biblioteca denota uma matriz com vetores coluna de dados para todas as séries temporais possíveis correspondentes a essa função. A aproximação da regressão esparsa é dada por $\dot{X} \approx \Theta(X)\Xi$, em que Ξ é um conjunto de coeficientes que determina os termos ativos na f, isto é, um vetor de coeficientes esparsos que satisfaz a aproximação (SILVA *et al.*, 2020; KAPTANOGLU *et al.*, 2021). Um esquema ilustrativo desse processo foi apresentado por Kaptanoglu *et al.*

(2021) e está representado na Figura 16. A partir de uma matriz de dados do sistema, é possível calcular numericamente suas derivadas para compor a matriz de derivadas. Com isso, pode-se então ajustar os dados observados às funções de uma biblioteca pré-definida, que pode incluir polinômios de até quinto grau, funções trigonométricas, exponenciais, entre outras. A escolha da biblioteca depende do conhecimento prévio do usuário sobre o comportamento do sistema: por exemplo, se os dados indicam uma dinâmica suave e contínua, uma base polinomial pode ser suficiente para capturar a estrutura simbólica aproximada do sistema.

Dessa forma, o algoritmo constrói uma série de modelos candidatos que descrevem como os dados evoluem ao longo do tempo. A regressão esparsa aplicada nesses modelos resulta em uma matriz de coeficientes na qual apenas alguns termos apresentam valores significativamente diferentes de zero. Esses coeficientes revelam os elementos relevantes da biblioteca funcional, permitindo identificar a forma simbólica que melhor representa a dinâmica observada. Uma das alternativas para se obter uma regressão esparsa é por meio da aplicação de um regularizador esparso, como a norma L_0 ou L_1, resolvendo assim o problema de otimização apresentado na ilustração.

Figura 16 – Ilustração esquemática da pesquisa realizada pelo SINDy: com base na matriz de derivadas $\dot{\mathbf{X}}$, o objetivo é minimizar a diferença entre $\dot{\mathbf{X}}$ e, em que $\Theta(\mathbf{X})$ é a biblioteca que contém uma lista de funções candidatas para capturar o modelo, enquanto Ξ' representa os coeficientes a serem ajustados. Por fim, as restrições podem ser aplicadas aos coeficientes dos parâmetros do modelo por meio de $R(\Xi')$.



Fonte: Elaborada pela autora.

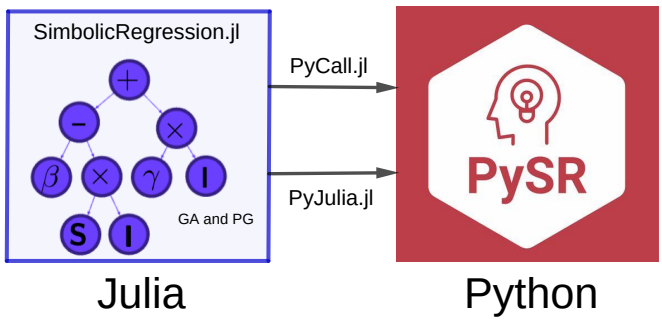
O parâmetro λ controla o grau de esparsidade desejado na regressão. Para calibrá-lo adequadamente, o software implementa o algoritmo da curva de Pareto, que avalia o desempenho do modelo para diferentes valores de λ. Quando λ é muito pequeno, há o risco de incluir termos irrelevantes no modelo; por outro lado, valores muito altos podem eliminar termos importantes, comprometendo a qualidade da representação. Assim, a análise da curva de Pareto permite encontrar um valor intermediário ideal para λ, promovendo um modelo mais parcimonioso e evitando o sobre ajuste.

2.3.4 SR

Desenvolvido com base em algoritmos genéticos e implementado na linguagem Julia, a biblioteca PySR destaca-se em relação à alternativa GPLearn por empregar um *framework* com desempenho computacional significativamente superior, característica atribuída à eficiência da linguagem Julia. Além disso, o algoritmo sobressai-se por sua robustez e alta eficiência em tarefas de regressão simbólica. Sua arquitetura interna é similar ao fluxograma apresentado na Figura 14 e pode ser resumida no diagrama da Figura 17.

O algoritmo pode ser executado tanto em ambiente Python quanto Julia, adaptando-se às preferências e necessidades do usuário. No ambiente Julia, por meio da biblioteca *SymbolicRegression.jl*, é possível incorporar funções personalizadas ao código utilizando *kernels* SIMD. Esses *kernels* são aplicados em tempo de execução e permitem alto desempenho computacional, ao possibilitar a realização simultânea de múltiplas operações sobre os dados, a diferenciação automática das expressões e a manipulação eficiente de populações de equações, tudo isso com suporte à computação paralela (CRANMER, 2023). No estudo mencionado, os autores também introduzem o conceito de *EmpiricalBench*, uma ferramenta projetada para comparar e avaliar o desempenho de diferentes algoritmos, além de mensurar a capacidade da regressão simbólica em aplicações científicas.

Figura 17 – Método SR, cuja execução eficiente do código é viabilizada pelas implementações em Julia, enquanto a interface Python facilita a interação com o usuário.



Fonte: Elaborada pela autora.

Outro diferencial relevante do SR é o controle sobre operadores aninhados, o que impede a formação de composições indesejadas entre funções. Em essência, o SR configura-se como uma ferramenta extremamente poderosa, que combina alta desempenho computacional do Julia com a acessibilidade da interface Python. Essa estrutura permite, ainda, a implementação de operadores unários baseados em conhecimentos teóricos prévios sobre os dados, além da geração e visualização das expressões matemáticas resultantes.

2.3.5 KAN/MultKAN

O cerne da abordagem desenvolvida por LIU *et al.* consiste na incorporação das Redes de Kolmogorov-Arnold em substituição às redes neurais do tipo perceptron multicamadas (MLPs). Enquanto os MLPs se baseiam no teorema da aproximação universal, o qual garante que, dado um conjunto de dados e uma arquitetura de rede específica, é possível aproximar funções contínuas com precisão arbitrária, tais redes dependem de uma topologia adequada para aprender os dados com exatidão.

Em contraste, o Teorema da Representação de Kolmogorov-Arnold assegura que toda função contínua pode ser expressa como uma combinação de funções mais simples. Ou seja, dada uma função contínua $f : [0,1]^n \to \mathbb{R}$ derivada de algum processo físico definido, ela pode ser representada como:

$$f(x_1, x_2, \ldots, x_n) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^{n} \phi_{q,p}(x_p) \right), \tag{2.17}$$

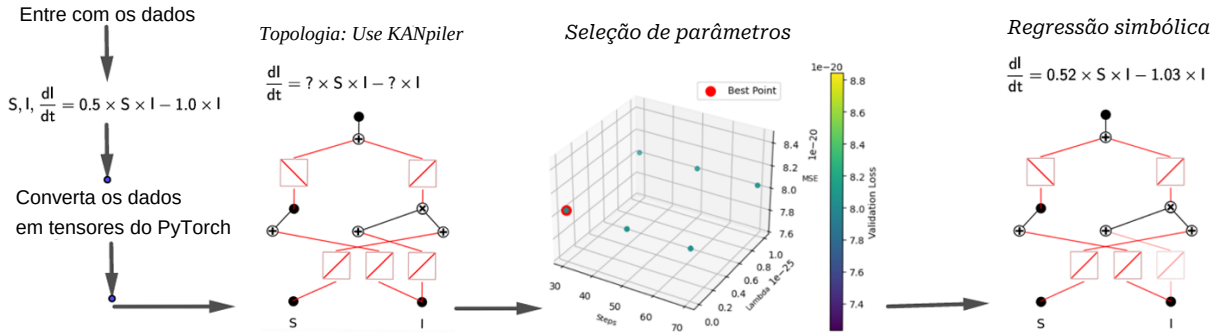
em que $\Phi_q : [0,1] \to \mathbb{R}$ e $\phi_{q,p} : \mathbb{R} \to \mathbb{R}$ são funções contínuas. Este teorema serve como fundamentação para as redes Kolmogorov–Arnold (LIU *et al.*, 2024b). Os autores generalizam a representação original, cuja profundidade era de duas camadas e a largura igual a $2n+1$, sendo n o número de variáveis de entrada. A possibilidade de adicionar camadas adicionais a essas redes e treiná-las por meio de retropropagação permite descobrir relações mais complexas, frequentemente encontradas em problemas do mundo real (LIU *et al.*, 2024b).

As KANs não utilizam funções de ativação fixas. Em vez de pesos convencionais, elas adotam funções unidimensionais ao longo das arestas, parametrizadas por **splines** durante o treinamento, permitindo que a rede aprenda suas próprias funções de ativação e oferecendo maior flexibilidade na modelagem.

A precisão das KANs é notável devido à sua capacidade de evitar o esquecimento catastrófico. Isso é possível graças aos pontos de controle das B-*splines*, que permitem uma adaptação local sem comprometer o que já foi aprendido, algo frequentemente observado em MLPs.

A predominância da operação de multiplicação em sistemas físicos reais motivou os autores a aprimorar o algoritmo, resultando na criação do MultKAN. Como destacado por Liu *et al.* (2024a), essa nova abordagem tem o potencial de revelar estruturas multiplicativas presentes nos dados. Além disso, os autores expandiram a biblioteca com ferramentas que permitem a utilização de conhecimento prévio ou empírico sobre o sistema estudado, como o KANCompiler, o qual representa uma vantagem significativa na busca por expressões simbólicas. É um algoritmo que favorece uma alta interação entre humano e máquina, conforme ilustrado na Figura 18.

Figura 18 – O PyKAN utiliza um compilador de fórmulas que permite incorporar conhecimento prévio na busca por expressões. As bordas indicam funções com maior correlação com os dados, enquanto os nós representam operações como adição e multiplicação. Com a escolha adequada de parâmetros e funções de ativação, o modelo resultante torna-se interpretável.



Fonte: Elaborada pela autora.

2.3.6 ODEFormer

Para inferir sistemas dinâmicos usando modelo de linguagem neural simbólica, o ODEFormer, proposto por D'Ascoli *et al.* (2023), é o método transformador codificador-decodificador mais atualizado, sendo suportado por uma base de código aberta. Em comparação com os primeiros algoritmos baseados em transformadores, o ODEFormer tem a vantagem de conseguir inferir as equações que regem sistemas multidimensionais, provando ser um método particularmente adequado para sistemas tridimensionais e outros com várias variáveis independentes. Ao contrário dos algoritmos SR mencionados acima, o ODEFormer opera sob um paradigma diferente: sua capacidade preditiva é baseada na dinâmica aprendida durante o treinamento em grandes conjuntos de dados, permitindo que o algoritmo identifique de forma probabilística o comportamento de vários sistemas, posteriormente transferido para o conjunto de dados de interesse.

O ODEFormer inclui 16 cabeças de atenção (VASWANI *et al.*, 2017) e 512 dimensões de incorporação, totalizando aproximadamente 86 milhões de parâmetros. Esses parâmetros foram treinados usando um extenso conjunto de dados sintéticos de vários sistemas dinâmicos. Esses dados passam por uma estratégia de incorporação na qual as expressões são tokenizadas e representadas como vetores em um espaço, denotado por $\mathbb{R}^{((D+1)\times 3)\times d_{emb}}$, onde D é a dimensão do sistema original que gerou os dados [1], e d_{emb} é a dimensão de incorporação associada à equação original. A capacidade de tratar os dados como uma sequência de tokens, um princípio fundamental da tradução automática moderna, é o que torna os algoritmos baseados em transformadores únicos em comparação com outras abordagens de regressão simbólica. Para decodificar e inferir equações, o modelo usa um método de amostragem de feixe (GAEL *et al.*, 2008).

Os autores também propõem um extenso conjunto de dados para comparar seu algoritmo, chamado ODEBench (D'ASCOLI *et al.*, 2023). Nele, 63 equações com diferentes números de di-

[1] $D_{max} = 6$ para o modelo pré-treinado usado neste estudo

mensões, algumas apresentando comportamento caótico, são usadas para verificar o desempenho do modelo treinado e divulgadas publicamente para futuras comparações.

2.3.7 Visão Geral dos Algoritmos Utilizados implementados pelas bibliotecas em Python

Neste tópico, apresenta-se uma síntese dos algoritmos de regressão simbólica utilizados ao longo desta tese.

Tabela 2 – Matriz comparativa de algoritmos de regressão simbólica para descoberta de EDOs.

Características	**GPLearn**	**AI-Feynman**	**SINDy**	**SR**	**PyKAN**	**ODEFormer**
Avaliação interna com divisão treino-teste	Não	Sim	Não	Não	Sim	Parcial
Controle de hiperparâmetros e biblioteca de funções	Alto	Baixo/Médio	Alto	Alto	Alto	Baixo
Convergência para representação simbólica	Robusta	Falível	Robusta	Robusta	Robusta	Falível
Específico para EDOs	Adaptável	Adaptável	Sim	Adaptável	Adaptável	Sim
Ranking de desempenho computacional	Lento	Intermediário	Veloz	Veloz	Veloz	Intermediário
Integração de conhecimento prévio humano	Parcial	Não	Total	Parcial	Total	Não
Critério de seleção	$\min(MSE(f), C(f))$	min(MSE) min(MDL)	$\min\ (MSE) + \lambda_{reg}$	$min(MSE(f) + C(f))$ [multiobjetivo]	$\min\ (MSE) + \lambda_{reg}$	min(Cross-Entropy) MSE
Seleção	Única	Lista	Única	Lista	Única	Única/Lista

CAPÍTULO

3

MATERIAIS E MÉTODOS

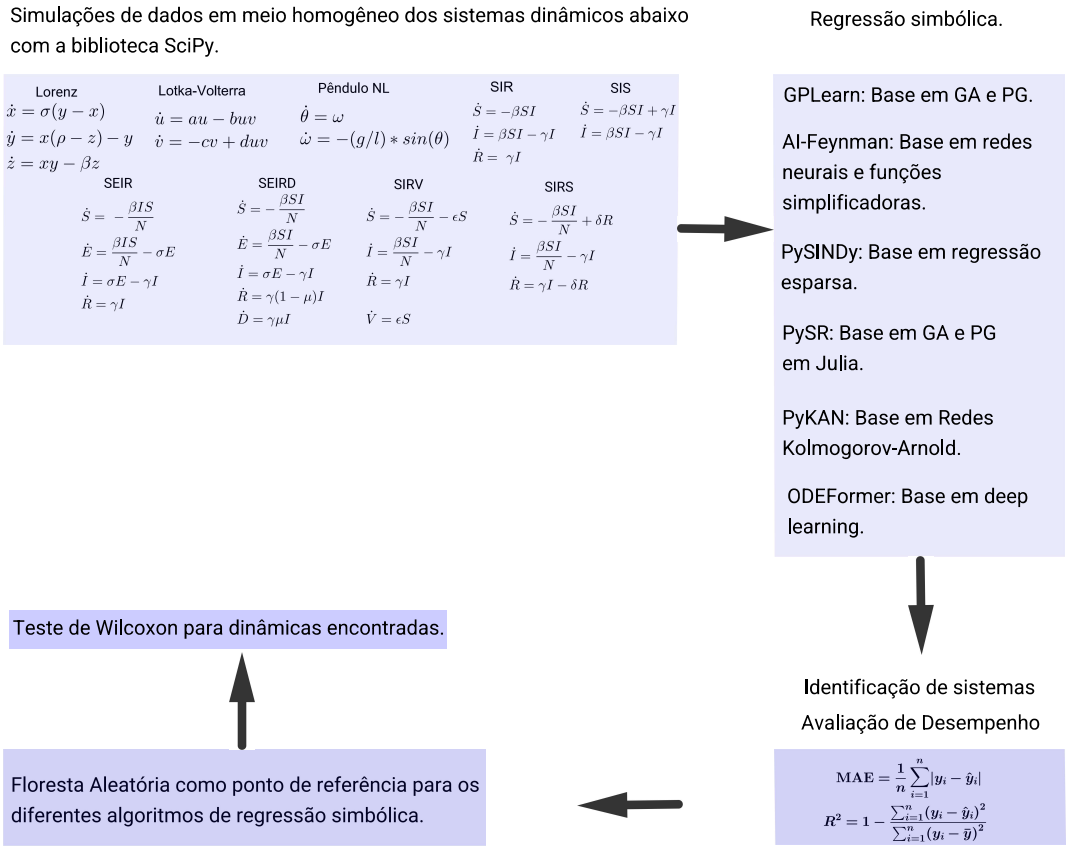
3.1 Método para análise da RS em dinâmicas em meio homogêneo

Os seis métodos de RS descritos na Seção 2.3 foram aplicados na identificação das dinâmicas descritas na Seção 2.2, conforme o esquema metodológico da Figura 19 na página 60. Os resultados obtidos da análise com base no método apresentado podem ser verificados no Capítulo 4.

Para aplicar os modelos de Regressão Simbólica selecionados, foram criados conjuntos de dados que representam diferentes processos dinâmicos. Esses conjuntos foram simulados a partir de modelos populares de sistemas dinâmicos simples e complexos, tais como o modelo de propagação de epidemia SIR e SIS. Adicionalmente, foram incluídos na análise os sistemas dinâmicos caóticos de Lorenz e Pêndulo não-linear, o sistema que descreve a dinâmica entre espécies Lotka-Volterra, além das variantes SEIR, SEIRD, SIRV e SIRS. O critério de avaliação visual foi utilizado para verificar se a equação estimada reproduz a forma funcional original responsável pela geração dos dados. A seguir, é descrito em detalhes o processo de geração dos dados:

Nas Tabelas 11 e 12, do Apêndice A.2, são fornecidos parâmetros detalhados, condições iniciais e intervalos necessários para as simulações de dados. A função SOLVE_IVP da biblioteca SciPy foi utilizada para realizar a integração e resolver as equações diferenciais de cada sistema. Ainda neste Apêndice estão as Tabelas 13 e 14 com os parâmetros utilizados em cada modelo de SR. Para a maioria dos algoritmos a escolha dos parâmetros se deu pelo método de tentativa e erro, isto devido ao alto-custo computacional destes. Para o algoritmo SINDy realizamos um mapeamento entre os melhores otimizadores assim como o parâmetro de regularização mais adequado. Quanto ao KAN foi realizada uma busca exaustiva dos parâmetros que foram empregados. Os resultados da análise podem ser verificados no Capítulo 4.

Figura 19 – Representação esquemática da metodologia. Dados de sistemas dinâmicos foram gerados e usados como entradas para diferentes algoritmos de regressão simbólica, o modelo de Floresta Aleatória foi utilizado como linha de base. As formas estruturais identificadas com sucesso foram submetidas ao teste de Wilcoxon.



3.1.1 Uso de conhecimento dentro do domínio

Conforme discutido anteriormente, alguns algoritmos de SR têm maneiras de integrar características conhecidas do sistema em investigação em seu procedimento de regressão. Quando o sistema é conhecido em sua totalidade, incluindo suas variáveis e suas respectivas derivadas, o objetivo de inferir a equação que originou os dados completos pode ser simplificado usando o conhecimento do domínio das relações entre essas variáveis.

Neste estudo, foram consideradas cinco formas de incorporar informações prévias durante o treinamento, cada uma implementada de maneira diferente nos modelos, quando disponível. Tabela 3 abaixo exibe os métodos disponíveis para fazer isso e se eles foram usados neste estudo. Eles foram: (i) se as variáveis no conjunto de dados foram explicitamente selecionadas ou divididas para refletir os dados que seriam descritivos da dinâmica conhecida; (ii) ajuste de equação única em vez de investigar todo o sistema de uma só vez; (iii) seleção do conjunto de operadores, que inclui operandos e funções matemáticas; e (iv) escolha da melhor equação da lista de resultados do processo de ajuste, consoante o que se sabe sobre o sistema, se elas não fossem selecionadas de acordo com a pontuação de precisão interna do modelo. Para este último caso, apenas o KAN exigiu escolha manual em alguns casos. O teste de Wilcoxon foi empregado

para avaliar a ausência de evidências estatísticas de diferença entre a dinâmica original e aquela recuperada pelos algoritmos de regressão simbólica. Adicionalmente, foi especificado o nível de complexidade de cada sistema analisado.

Tabela 3 – Uso do conhecimento interno dos sistemas investigados durante o treinamento de SR. Um ✓indica que o método está disponível para o algoritmo, e ✓✓denota o uso para todos os sistemas, salvo indicação em contrário.

Method \ Algorithm	Variable selection	Single equation fitting	Operator selection	Equation fixing
GPLearn	✓✓	✓✓*	✓✓	
AI Feynman	✓✓	✓✓	✓✓	
PySINDy	✓		✓✓	✓
PySR	✓	✓✓	✓✓	✓
PyKAN	✓✓	✓✓	✓✓	✓✓
ODEFormer	✓	✓		

*Não utilizado para os sistemas Lorenz, Lotka-Volterra, Pêndulo e SIS.

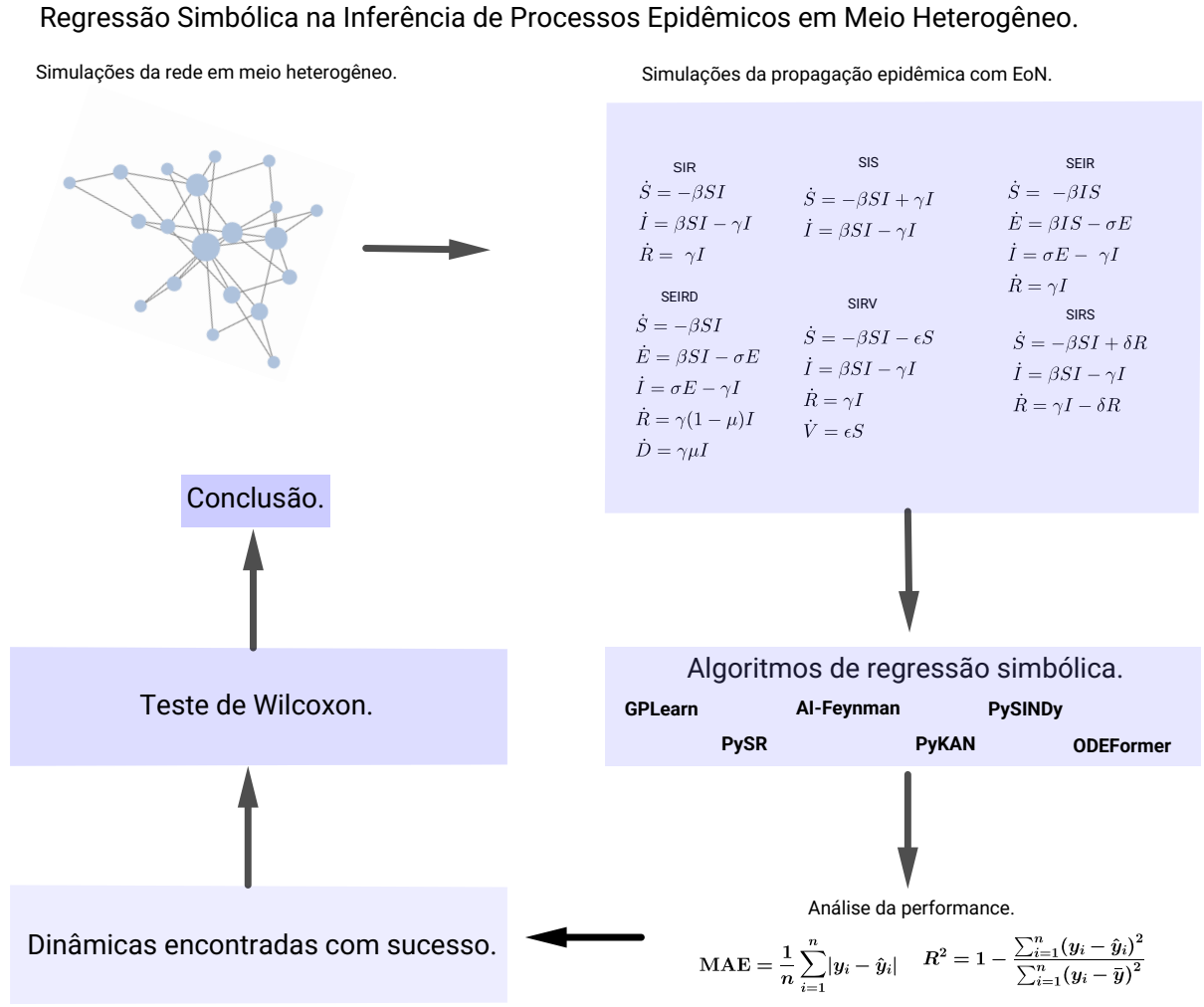
É importante mencionar que, na maioria dos cenários comuns do mundo real, as informações sobre o sistema investigado provavelmente estarão ausentes. Além da incapacidade de ajustar os algoritmos usando propriedades conhecidas, isso também introduz um desafio significativo: durante o treinamento, o cálculo de derivadas por métodos de diferença finita pode introduzir (ou amplificar) ruído nos dados, o que, por sua vez, complica a identificação de um sistema. Schaeffer e McCalla (2017) abordou essa questão para algoritmos baseados em esparsidade por meio de formulações integrais que contornam a necessidade de diferenciação numérica, embora sejam necessários mais estudos para complementar outros métodos.

3.2 Método para análise da RS em dinâmicas para meios não homogêneos

O método adotado para a análise apresentada no Capítulo 5, está representado no fluxograma da Figura 20. A partir das redes geradas, foram simulados seis processos epidêmicos, os quais foram submetidos a diferentes modelos de regressão simbólica. Em seguida, avaliou-se a capacidade de cada modelo em recuperar a estrutura funcional do sistema e sua adequação aos dados quando comparada ao modelo de Floresta Aleatória. Por fim, realizou-se uma análise comparativa para identificar os modelos mais apropriados para a inferência de sistemas dinâmicos em redes complexas, bem como testes estatísticos.

A geração de dados sintéticos de processos dinâmicos em redes e a realização de análises para cada modelo de regressão simbólica foram conduzidos em um ambiente Python, usando as bibliotecas **networkx** (HAGBERG; CONWAY, 2020) e **EoN** (MILLER; TING, 2020) com o método Gillespie (KISS; MILLER; SIMON, 2017b). O primeiro foi usado na geração de topologias de rede distintas, como a Erdos-Renyi, rede aleatória com padrões de conexões mais

Figura 20 – O fluxo metodológico desta abordagem para meios heterogêneos é ilustrado. Inicialmente, um conjunto de redes distintas é gerado para servir de meio à propagação de processos epidêmicos. Os dados resultantes dessas simulações são, então, submetidos a algoritmos de regressão simbólica. Por fim, para cada forma estrutural subjacente à dinâmica encontrada com sucesso, aplicou-se o teste de Wilcoxon.



homogêneas e baixa estruturação comunitária, e a Barabási-Albert, rede com *hubs* centrais conectando várias partes menos conectadas, ideal para modelar fenômenos como redes sociais, ambas com 1.000 nós.

Essas redes servem como meio para o processo de propagação de epidemias, no qual o pacote EoN é usado para modelagem e simulação da propagação de doenças infecciosas em redes complexas. Foram gerados dados sintéticos para os seguintes processos epidêmicos: SIR, SIS, SEIR, SEIRD, SIRV e SIRS. Nesta pesquisa, o cálculo das derivadas baseou-se no método analítico, possível devido ao conhecimento do modelo teórico. Em contraste, estudos observacionais, onde os parâmetros são desconhecidos, recorrem frequentemente a aproximações numéricas como diferenças finitas, suavização por splines ou filtros de séries temporais. Detalhes sobre a rede, o modelo epidemiológico e suas respectivas taxas de transição estão descritos na Tabela 4.

Tabela 4 – Parâmetros usados na simulação do processo epidêmico. Em todos os casos, o número inicial de indivíduos infectados foi de 10 vértices, e a função de diferenciação usada foi do tipo Savitzky-Golay de primeira ordem.

Net	Modelo	**Parâmetros de transição**
ER	SIR	$t_{I\to R} = 0.1, t_{I\to R} = 0.2$
	SIS	$t_{I\to S} = 0.2$
	SEIR	$t_{E\to I} = 0.3, t_{I\to R} = 0.2, t_{S\to E} = 0.1$
	SEIRD	$t_{E\to I} = 0.3, t_{I\to R} = 0.2, t_{I\to D} = 0.1, t_{S\to E} = 0.3$
	SIRV	$t_{S\to V} = 0.5, t_{I\to R} = 1.0, t_{S\to I} = 0.5$
	SIRS	$t_{I\to R} = 1.0, t_{S\to R} = 0.2, t_{S\to I} = 0.2$
BA	SIR	$\gamma = 0.1, \tau = 0.2$
	SIS	$\tau = 0.2$
	SEIR	$t_{E\to I} = 0.3, t_{I\to R} = 0.2, t_{S\to E} = 0.1$
	SEIRD	$t_{E\to I} = 0.3, t_{I\to R} = 0.2, t_{I\to D} = 0.1, t_{S\to E} = 0.3$
	SIRV	$t_{S\to V} = 0.5, t_{I\to R} = 1.0, t_{S\to I} = 0.5$
	SIRS	$t_{I\to R} = 1.0, t_{S\to R} = 0.2, t_{S\to I} = 0.2$

Adicionalmente, para viabilizar a aplicação dos modelos de regressão simbólica que, em sua maioria, não são projetados para lidar diretamente com dados de séries temporais, as equações que descrevem a evolução temporal do sistema foram reformuladas em um formato compatível com regressão supervisionada. Cada equação diferencial foi ajustada separadamente a partir dos dados derivados. A única exceção foi o modelo PySINDy, que permite operar diretamente sobre séries temporais dinâmicas, dispensando essa transformação.

No Apêndice A.2 as Tabelas 15 e 16 especificam os parâmetros utilizados por cada algoritmo de regressão simbólica na busca pela forma estrutural do sistema dinâmico que se propaga tanto em redes ER quanto BA. Os resultados da análise podem ser verificados no Capítulo 5.

3.3 Caráter semi-supervisionado

A abordagem adotada neste estudo pode ser caracterizada como semi-supervisionada, pois as variáveis de entrada foram previamente definidas com base no problema investigado, enquanto o algoritmo de regressão simbólica atuou exclusivamente na busca por relações matemáticas entre esses termos, sem realizar seleção automática de variáveis comum na maioria dos estudos. Essa estratégia difere de métodos não supervisionados, como técnicas de agrupamento ou redução de dimensionalidade, em que o modelo identifica padrões sem qualquer orientação prévia. Nesse contexto, a supervisão assegura que as expressões geradas permaneçam diretamente vinculadas às variáveis de interesse, aspecto fundamental para a identificação de fórmulas interpretáveis e fisicamente consistentes. Tal característica é particularmente relevante, uma vez que o pesquisador da área se apoia em conhecimento teórico previamente consolidado e amplamente difundido na literatura. Em modelos epidemiológicos, por exemplo, identificar parâmetros que descrevam com precisão a força de propagação de um vírus ou a taxa efetiva de infecção é fundamental para entender e prever o comportamento da epidemia.

CAPÍTULO

4

REGRESSÃO SIMBÓLICA EM PROCESSOS DINÂMICOS EM REDES REGULARES

Neste capítulo, são apresentados os resultados do método detalhado na Seção 3.1 do Capítulo 3, esquematicamente ilustrado pela Figura 19.

4.1 Métricas e definição de uma linha de base

A avaliação dos resultados de um método de regressão simbólica consiste em determinar se ele é capaz de recuperar a expressão matemática exata subjacente à dinâmica observada, ou seja, verificar se as expressões identificadas são matematicamente equivalentes à equação original, juntamente com os parâmetros estimados que se aproximam numericamente dos valores reais. Para isso, este estudo comparou algoritmos usando três critérios: (i) comparar as formas estruturais obtidas pelos algoritmos de regressão simbólica com as originais; (ii) quantificar a diferença dos dados resultantes nas equações diferenciais em relação ao sistema original; e (iii) verificar a complexidade das equações obtidas pelos algoritmos.

Para quantificar a recuperação da forma estrutural da equação, foi utilizado um teste de postos sinalizados de Wilcoxon (CONOVER, 1999) para avaliar se as expressões geradas pelo SR produziam dados estatisticamente indistinguíveis dos do sistema real. A recuperação foi considerada bem-sucedida quando a hipótese nula de ausência de diferença não foi rejeitada ao nível de significância de $\alpha = 0,05$ (isto é, p-valor$> 0,05$), indicando ausência de evidência estatística suficiente para detectar diferenças entre os conjuntos de dados.

Para ancorar o desempenho do SR a outros métodos de aprendizado de máquina, um algoritmo de floresta aleatória (BREIMAN, 2001) foi incluído como uma linha de base não interpretável. Tanto para as equações originais quanto para aquelas inferidas pelos algoritmos SR, as soluções foram integradas na mesma grade temporal usada para a geração de dados, e o

desempenho preditivo foi quantificado usando o coeficiente de determinação (R2), seguindo a prática padrão na literatura SR (CAVA *et al.*, 2021).

Por fim, a complexidade de cada equação inferida foi calculada seguindo a discussão em (CAVA *et al.*, 2021), onde é definida como a quantidade de variáveis, operadores e constantes presentes na equação dada. Para equações multidimensionais, essa métrica mede a soma da complexidade de cada dimensão.

4.2 Resultados e discussões

Seguindo as especificações da Seção 3.1 no Capítulo 3, foram obtidos os resultados apresentados nas Tabelas 5 e 6. Em todas as tabelas apresentadas abaixo, os sistemas dinâmicos cujas formas estruturais foram corretamente identificadas, quando comparadas com as equações originais, estão marcados com um "sinal de verificação" (✓), enquanto as aproximações ou descrições incorretas desses sistemas não apresentam essa marcação.

Os resultados da Tabela 7 mostram que todos os modelos de regressão simbólica empregados, exceto ODEFormer e AI Feynman, demonstraram capacidade suficiente para identificar as representações simbólicas da maioria dos sistemas dinâmicos investigados, especialmente quando se considera sua eficácia na descrição de formas estruturais da dinâmica epidemiológica.

O PySR foi o algoritmo com melhor desempenho em todos os aspectos, identificando com sucesso a forma estrutural correta de todos os sistemas, e todos, exceto um, com diferenças significativas nos parâmetros resultantes, tornando este algoritmo o mais versátil e preciso de todos os testados.

PySINDy e PyKAN também identificaram todos os sistemas, mas não com tanta precisão quando se consideram os resultados do teste de Wilcoxon, com o PyKAN obtendo o sistema de equações do SIR sem diferenças significativas. Embora esses métodos produzam resultados ligeiramente inferiores de acordo com a Figura 21, eles se mostram suficientemente precisos na descrição das equações de movimento do sistema. No entanto, vale a pena notar que eles foram mais sensíveis às mudanças no parâmetro usado durante a pesquisa: enquanto no KAN uma regularização mais forte favorece a recuperação da forma estrutural aproximada, em alguns casos ela parece impedir a otimização dos parâmetros. No PySINDy, por outro lado, otimizar os parâmetros de regularização pode ser a solução para esse problema.

O GPLearn não conseguiu recuperar apenas um conjunto de equações (SEIRD) e, embora a maioria dos seus resultados tenha se mostrado significativamente diferente da dinâmica original segundo o teste de Wilcoxon, ele ainda é capaz de fornecer a forma estrutural correta de outros sistemas. No entanto, sem a seleção adequada de parâmetros, suas equações de saída tendem a ser mais complexas do que as formas estruturais originais, apresentando multiplicação e combinação excessivas de termos, uma característica comum nos algoritmos tradicionais de programação

Tabela 5 – Forma estrutural original dos sistemas dinâmicos, juntamente com os parâmetros relevantes usados pelos sistemas dinâmicos para gerar seus dados sintéticos (primeira coluna) e as equações encontradas por cada algoritmo de regressão simbólica (todas as outras colunas). Os sistemas dinâmicos cujas formas estruturais foram identificadas corretamente estão marcados com "marca de verificação"(✓).

Governing equation	GPLearn	AI-Feynman	PySINDy	PySR	PyKAN	ODEFormer
Non linear pendulum $\dot{\theta} = \omega$ $\dot{\omega} = -9.8\sin(\theta)$	✓ $\dot{\theta} = \omega$ $\dot{\omega} = -9.17sin(1.08\theta)$	✓ $\dot{\theta} = \omega$ $\dot{\omega} = -9.8\sin(\theta)$	✓ $\dot{\theta} = \omega$ $\dot{\omega} = -9.8\sin(\theta)$	✓ $\dot{\theta} = \omega$ $\dot{\omega} = -9.8\sin(\theta)$	✓ $\dot{\theta} = \omega$ $\dot{\omega} = -9.8\sin(\theta)$	$\dot{\theta} = 1.06\omega - 0.04\omega(0.05\omega - 13.14\theta)$ $\dot{\omega} = -9.41\theta - \frac{1.2027}{(10.73-15.52\theta)}$
Lotka-Volterra $\dot{u} = 2u - 0.5uv$ $\dot{v} = -v + 0.375uv$	✓ $\dot{u} = 2u - 0.5uv$ $\dot{v} = -v + 0.18uv$	✓ $\dot{u} = 2u - 0.5uv$ $\dot{v} = -0.99v + 0.33uv$	✓ $\dot{u} = 1.94u - 0.49uv$ $\dot{v} = -0.95v + 0.37uv$	✓ $\dot{u} = 2.0u - 0.5uv$ $\dot{v} = 0.37uv - 1.0v$	✓ $\dot{u} = 2u - 0.5uv$ $\dot{v} = -v + 0.19uv$	✓ $\dot{u} = 1.5u - 0.4uv$ $\dot{v} = -1.4v + 0.2uv$
Lorenz $\dot{x} = 2(y - x)$ $\dot{y} = x(1 - z) - y$ $\dot{z} = xy - 2.6z$	$\dot{x} = 2(y - x)$ $\dot{y} = x(1 - z) - y$ $\dot{z} = 0.17y - 0.92z$	$\dot{x} = 2(y - x)$ $\dot{y} = x(1 - z)) - y$ $\dot{z} = 0.02 + z\sqrt{z}$	✓ $\dot{x} = 2(y - x)$ $\dot{y} = x(1 - z) - y$ $\dot{z} = xy - 2.6z$	✓ $\dot{x} = 2.0(y - x)$ $\dot{y} = x(1 - z) - y$ $\dot{z} = xy - 0.6z$	✓ $\dot{x} = 2(y - x)$ $\dot{y} = x(1 - z) - y$ $\dot{z} = -xy - 2.6z$	$\dot{x} = 1.9(y - x)$ $\dot{y} = -1.2yz^2$ $\dot{z} = 0.05y - 0.63z$

Tabela 6 – Descrição das representações simbólicas da dinâmica de propagação da epidemia aproximada. Os sistemas dinâmicos cujas formas estruturais foram identificadas corretamente estão marcados com "marca de verificação"(✓).

	Governing equation	GPLearn	AI-Feynman	PySINDy	PySR	PyKAN	ODEFormer
SIS	$\dot{S} = -0.3SI + 0.1I$ $\dot{I} = 0.3SI - 0.1I$	✓ $\dot{S} = -0.29SI + 0.1I$ $\dot{I} = 0.29SI - 0.1I$	$\dot{S} = -0.3SI + 0.1I^2$ $\dot{I} = 0.2SI - 0.1I^2$	✓ $\dot{S} = -0.3SI - 0.1S$ $\dot{I} = 0.3SI + 0.1S$	✓ $\dot{S} = -0.3SI + 0.1I$ $\dot{I} = 0.3SI - 0.1I$	✓ $\dot{S} = -0.3SI + 0.1I$ $\dot{I} = 0.3SI - 0.1I$	$\dot{S} = 9.01SI(0.03(-1+0.14S)^2 - 0.11S)$ $\dot{I} = 0.09I(2.1 - 3.44I)$
SIR	$\dot{S} = -0.5SI$ $\dot{I} = 0.5SI - 0.1I$ $\dot{R} = 0.1I$	✓ $\dot{S} = -0.5SI$ $\dot{I} = 0.44SI - 0.1I$ $\dot{R} = 0.1I$	$\dot{S} = -0.5SI$ $\dot{I} = SI - I^2 - 0.0007$ $\dot{R} = 0.1I$	✓ $\dot{S} = -0.5SI$ $\dot{I} = 0.5SI - 0.1I$ $\dot{R} = 0.1I$	✓ $\dot{S} = -0.5SI$ $\dot{I} = 0.5SI - 0.1I$ $\dot{R} = 0.1I$	✓ $\dot{S} = -0.5SI$ $\dot{I} = 0.5SI - 0.1I$ $\dot{R} = 0.1I$	$\dot{S} = -0.1S$ $\dot{I} = 0.5SI - 0.1R$ $\dot{R} = 0.1I$
SIRV	$\dot{S} = -0.5SI - 0.2S$ $\dot{I} = 0.5SI - 0.1I$ $\dot{R} = 0.1I$ $\dot{V} = 0.2S$	✓ $\dot{S} = -SI - 0.146S$ $\dot{I} = 0.32SI - 0.09I$ $\dot{R} = 0.1I$ $\dot{V} = 0.2S$	$\dot{S} = -0.4SI - 0.2S\sqrt{I+1}$ $\dot{I} = -0.09I + 0.09S\sqrt{\sqrt{I}+1}$ $\dot{R} = 0.1I$ $\dot{V} = 0.2S$	$\dot{S} = -0.5SI - 0.2S$ $\dot{I} = 0.4SI - 0.2IV$ $\dot{R} = 0$ $\dot{V} = 0.2S$	✓ $\dot{S} = -0.5SI - 0.2S$ $\dot{I} = 0.5SI - 0.1I$ $\dot{R} = 0.1I$ $\dot{V} = 0.2S$	✓ $\dot{S} = -0.5SI - 0.2S$ $\dot{I} = 0.5SI - 0.1I$ $\dot{R} = 0.1I$ $\dot{V} = 0.2S$	$\dot{S} = -0.3S$ $\dot{I} = 0.1S - 0.1I$ $\dot{R} = 1.0V - 2.7R$ $\dot{V} = 0.2S$
SIRS	$\dot{S} = -0.5SI + 0.2R$ $\dot{I} = 0.5SI - 0.1I$ $\dot{R} = 0.1I - 0.2R$	✓ $\dot{S} = -0.47SI + 0.19R$ $\dot{I} = 0.49SI - 0.1I$ $\dot{R} = 0.10I - 0.21R$	$\dot{S} = -0.02SI(S + I^2 - R) - 0.02S$ $\dot{I} = 1.08I(0.27S - 0.9)$ $\dot{R} = I - 0.19R$	$\dot{S} = -0.5SI + 0.2R$ $\dot{I} = 0.1SI$ $\dot{R} = I + 0.9R$	✓ $\dot{S} = -0.5SI + 0.2R$ $\dot{I} = 0.5SI - I$ $\dot{R} = 0.1I - 0.2R$	✓ $\dot{S} = -0.317SI + 0.2R - 0.2$ $\dot{I} = 0.3SI - 1.0I$ $\dot{R} = 1.0I - 0.2R$	$\dot{S} = -0.2S^2$ $\dot{I} = 0.3I$ $\dot{R} = -1.7R$
SEIR	$\dot{S} = -0.5SI$ $\dot{E} = 0.5SI - 0.5E$ $\dot{I} = 0.5E - 0.1I$ $\dot{R} = 0.1I$	✓ $\dot{S} = -0.5SI$ $\dot{E} = 0.48SI - 0.48E$ $\dot{I} = 0.45E - 0.096I$ $\dot{R} = 0.096I$	$\dot{S} = -0.5SI$ $\dot{E} = 0.5SI - 0.5E$ $\dot{I} = 0.27\left(E + \frac{E^2}{I+E}\right)$ $\dot{R} = \arcsin(0.09I)$	✓ $\dot{S} = -0.5SI$ $\dot{E} = 0.5SI - 0.5E$ $\dot{I} = 0.5E - 0.1I$ $\dot{R} = 0.1I$	✓ $\dot{S} = -0.5SI$ $\dot{E} = 0.5SI - 0.5E$ $\dot{I} = 0.5E - 0.1I$ $\dot{R} = 0.1I$	✓ $\dot{S} = -0.5IS$ $\dot{E} = -0.5E + 0.5IS$ $\dot{I} = 0.5E - 0.1I$ $\dot{R} = 0.1I$	$\dot{S} = -1.3SE$ $\dot{E} = 0.1E - 0.7ER$ $\dot{I} = 0.4E - 0.1I$ $\dot{R} = 1.5I - 28.1SIR$
SEIRD	$\dot{S} = -0.5SI$ $\dot{E} = 0.5SI - 0.2E$ $\dot{I} = 0.2E - 0.2I$ $\dot{R} = 0.1I$ $\dot{D} = 0.1I$	$\dot{S} = -0.5SI$ $\dot{E} = 0.1SE - 0.06I$ $\dot{I} = (E + 0.06) * (E - I)$ $\dot{R} = 1.0I$ $\dot{D} = 0.1I$	$\dot{S} = -0.3SI$ $\dot{E} = -0.01(E * ((E/I) + 1))$ $\dot{I} = 0.2 * E - 0.2I$ $\dot{R} = 0.1I$ $\dot{D} = 0$	✓ $\dot{S} = -0.5SI$ $\dot{E} = 0.5SI - 0.2E$ $\dot{I} = 0.2E - 0.2I$ $\dot{R} = 0.1I$ $\dot{D} = 0.1I$	✓ $\dot{S} = -0.5SI$ $\dot{E} = 0.5SI - 0.2E$ $\dot{I} = 0.2E - 0.2I$ $\dot{R} = 0.1I$ $\dot{D} = 0.1I$	✓ $\dot{S} = -0.5IS$ $\dot{E} = -0.2E + 0.5IS$ $\dot{I} = 0.2E - 0.2I$ $\dot{R} = 0.1I$ $\dot{D} = 0.1I$	$\dot{S} = -0.4SI$ $\dot{E} = 1.0I - 0.9E$ $\dot{I} = 0.1I - 0.8I(-0.2 - 1.0D)^2$ $\dot{R} = 0.1E$ $\dot{D} = 0.1I$

genética (VANNESCHI; CASTELLI; SILVA, 2010).

O AI Feynman recupera a maioria das equações dos sistemas Lorenz, Lotka-Volterra e pêndulo não linear, faltando apenas a forma estrutural correta no eixo z do primeiro. No entanto, o desempenho para os sistemas epidemiológicos compartimentais foi abaixo da média, com nenhum dos resultados obtidos refletindo a dinâmica original. Ao inspecioná-los mais detalhadamente, pode-se observar que, para SIS, SIR e SIRV, o AI Feynman gera equações de maior complexidade do que as expressões originais, com o padrão invertido para SIRS, SEIR e SEIRD. Essa instabilidade nas soluções indica que o modelo não é tão eficaz para esse tipo específico de dados.

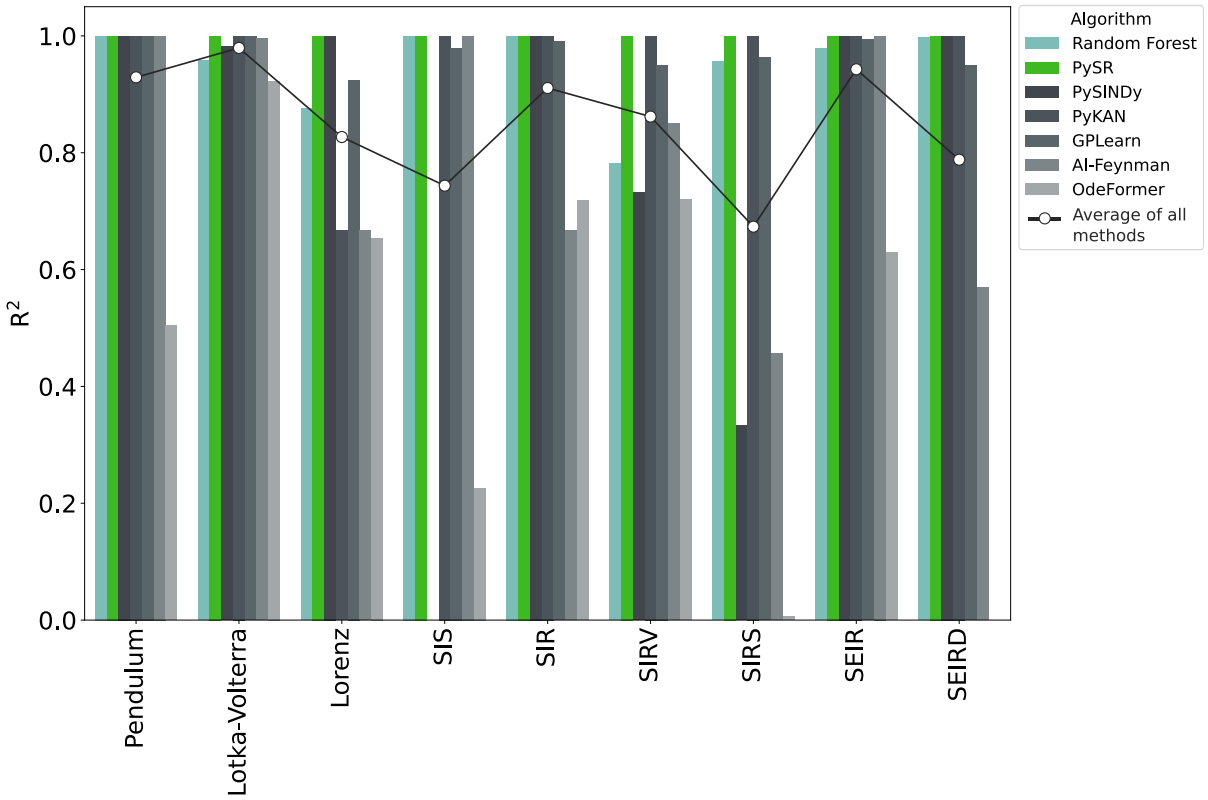
Por fim, o ODEFormer identificou apenas um dos sistemas incluídos neste estudo, além de apresentar um R^2 abaixo da média na maioria dos casos. Esse algoritmo também inferiu equações com graus de complexidade que, em alguns casos, variaram significativamente quando comparadas às equações originais, com uma redução notável no desempenho para sistemas com dimensões mais altas.

Tabela 7 – Resumo dos principais resultados de Tabela 5 e Tabela 6. As marcas de seleção indicam que a forma estrutural do sistema foi identificada com sucesso. As células sombreadas em cinza destacam os resultados que não apresentaram diferenças estatisticamente significativas em comparação com a dinâmica original, segundo o teste de Wilcoxon. As células com números adicionados mostram um aumento (ou diminuição) na complexidade quando comparadas ao sistema original.

System		**Symbolic regression method**					
Name	Complexity	GPLearn	AI-Feynman	PySINDy	PySR	PyKAN	ODEFormer
Non-linear pendulum	8	✓	✓	✓	✓	✓	+13
Lotka-Volterra	19	✓	✓	✓	✓	✓	✓
Lorenz	22	✓	-2	✓	✓	✓	-3
SIS	20	✓	+4	✓	✓	✓	+3
SIR	21	✓	+2	✓	✓	✓	+2
SIRV	27	✓	+8	✓	✓	✓	+5
SIRS	28	✓	-2	✓	✓	✓	+5
SEIR	35	✓	-4	✓	✓	✓	-15
SEIRD	39	- 10	-11	✓	✓	✓	-6

No geral, embora algumas das equações resultantes dos algoritmos de regressão simbólica fossem apenas aproximações das reais, a comparação com os sistemas originais revelou que essas diferenças eram geralmente pequenas, o que pode ser verificado ao integrar as equações resultantes no mesmo intervalo de tempo que os dados originais e calcular R^2 para inferir a eficiência da regressão. Os resultados mostrados em Figura 21 indicam que a maioria dos algoritmos pode, pelo menos, capturar efetivamente a dinâmica geral e fornecer uma visão geral das equações que regem os diversos fenômenos.

Figura 21 – Desempenho dos modelos de regressão simbólica em termos de R^2, com uma linha preta exibindo uma média de todos os métodos para um determinado sistema.



4.2.1 Os efeitos do ruído na regressão simbólica

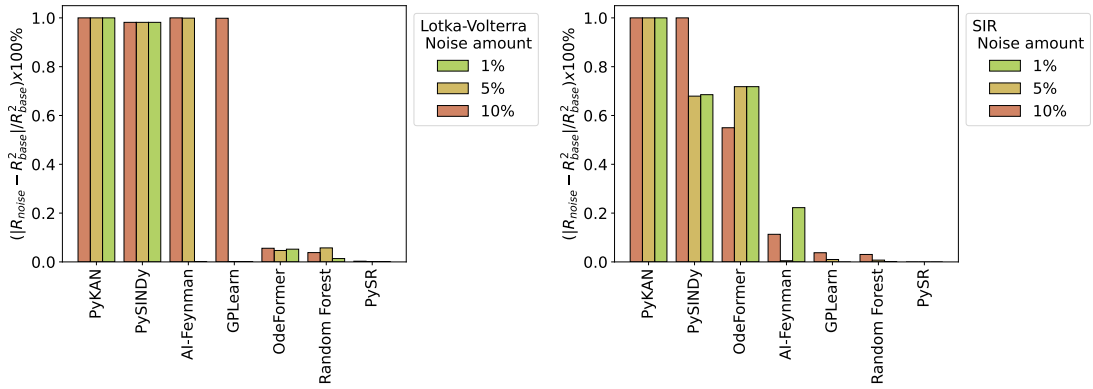
Comparar os algoritmos escolhidos com um conjunto de sistemas dinâmicos também requer compreender os efeitos dos dados corrompidos no desempenho desses métodos. Para isso, dois desses sistemas, o modelo Lotka-Volterra para a dinâmica predador-presa e o modelo compartimental SIR, foram escolhidos para comparar os efeitos da adição de ruído aos dados e a eficácia resultante da regressão realizada por todos os algoritmos. A escolha desses dois sistemas foi baseada nos resultados de Tabela 7, onde quase todos os algoritmos obtiveram efetivamente as equações corretas para ambos os sistemas. Além de ajustar os conjuntos de dados para adicionar ruído distribuído gaussiano, nenhum hiperparâmetro (ver Tabelas 11 e 12) foi alterado.

Notavelmente, o PyKAN foi o mais afetado pelos dados ruidosos, apresentando uma perda completa em R^2 para ambos os sistemas usando qualquer quantidade de ruído, apontando para uma generalização bastante pobre dos resultados por esse método. Em contraste, o PySR foi o menos afetado pelo ruído nos dados, com diferenças quase insignificantes em R^2 para todos os casos.

AI-Feynman, PySINDy e GPLearn também apresentaram picos de perda de desempenho, embora não tão graves quanto no caso do PyKAN e principalmente contidos no sistema Lotka-Volterra.

O ODEFormer apresentou uma perda de R^2 abaixo da média, com exceção do SIR com

Figura 22 – Diferenças resultantes em R^2 ao adicionar ruído gaussiano aos dados sintéticos de um sistema Lotka-Volterra (em cima) e SIR (em baixo).



1% de ruído adicionado: nele, o modelo apresentou $R^2 = 0$ para dois componentes desse sistema, que apresentou um aumento significativo na complexidade (37 contra 21) em comparação com o sistema SIR de linha de base.

É importante mencionar que os resultados da Figura 22 não são equivalentes a uma avaliação sistemática dessas técnicas: o objetivo deste estudo específico era apenas aumentar as irregularidades nos dados originais e verificar as diferenças de desempenho. Para se obter uma avaliação completa, seria essencial considerar as extensões e configurações que podem ser usadas para atenuar o efeito do ruído durante sua execução, que estão presentes na maioria dos métodos de SR investigados. Abaixo, são discutidos algoritmos que incluem tais estratégias para lidar com dados ruidosos:

- Os autores originais do ODEFormer (D'ASCOLI *et al.*, 2023) fazem comparações completas adicionando diferentes quantidades de ruído a vários sistemas, mostrando que o algoritmo é robusto ao ruído até $\sigma = 0,05$ (ou 5% de ruído adicionado) em vários sistemas;
- O PySINDy fornece estratégias de mitigação de ruído em seus métodos de diferenciação (KAPTANOGLU *et al.*, 2022), que é o estágio mais sensível ao ruído deste algoritmo;
- Raghav *et al.* (2024) estendeu recentemente o método GPLearn original para ter um desempenho mais preciso em tarefas sensíveis ao ruído, ao mesmo tempo, em que aprimorou a estratégia original para acomodar a interatividade do usuário;
- O PySR possui um módulo de redução de ruído integrado que pode ser aplicado definindo o sinalizador `denoise=True` ao inicializar o modelo. Mais detalhes sobre a implementação desse método são fornecidos em (CRANMER, 2023).

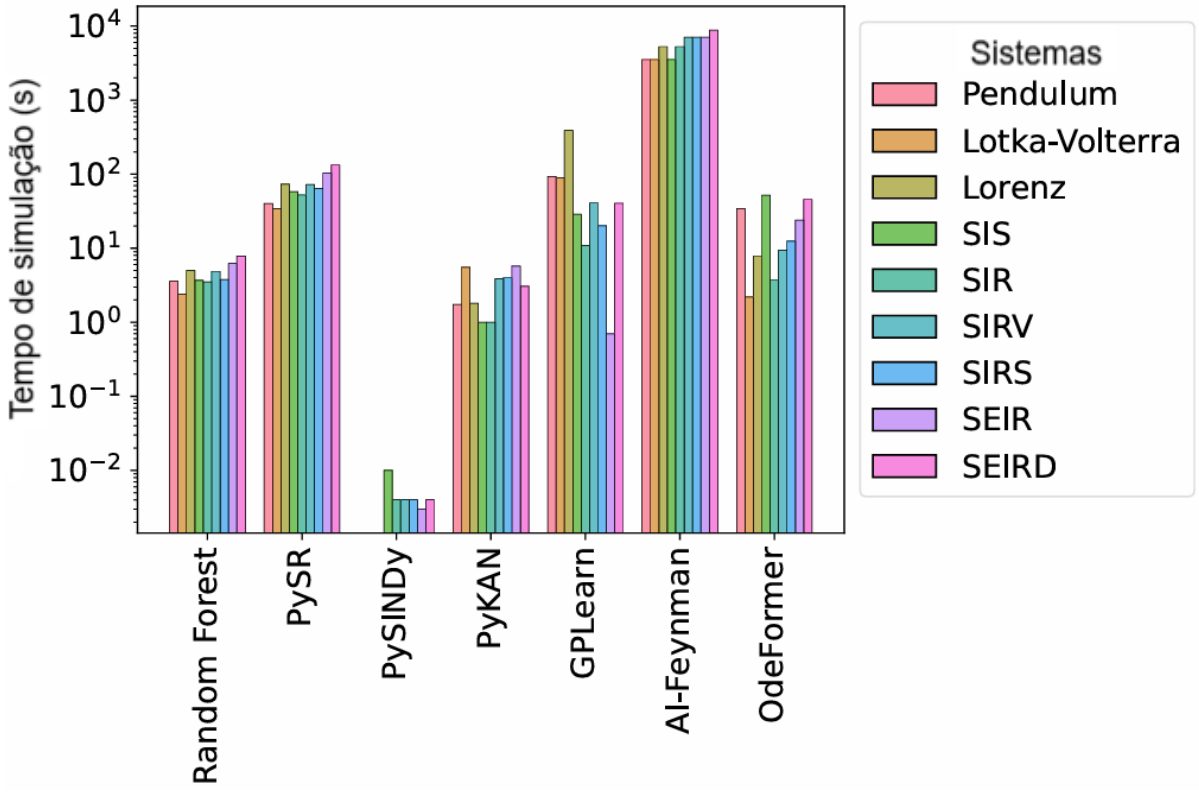
4.2.2 Computational complexity

Todos os algoritmos são implementados de maneiras diferentes e, como tal, a quantidade de recursos, incluindo o tempo de operação, pode variar. Para determinar quantitativamente sua

complexidade computacional de acordo com essas necessidades, cada simulação individual foi cronometrada, com os resultados apresentados em Figura 23.

O AI-Feynman é, de longe, o mais lento dos métodos analisados, com cada sistema levando até duas horas e meia para ser executado. O segundo método mais lento é o GPLearn, que levou em média 75 segundos para cada simulação. Em contraste com esses números, o PySINDy foi executado quase instantaneamente para alguns sistemas e manteve uma média de tempo muito baixa de 5×10^{-3} segundos para outros. Os demais algoritmos foram executados em alguns segundos até dois minutos. Esses números são proporcionalmente semelhantes aos listados no benchmark por Cava *et al.* (2021), embora os tamanhos dos conjuntos de dados e as especificações dos computadores [1] dos dois estudos são diferentes.

Figura 23 – Tempo de processamento para cada combinação de algoritmo e sistema dinâmico, em segundos. Para o PySINDy, $t = 0$s ao simular sistemas, Lorenz, Lotka-Volterra e pêndulo simples.



4.3 Conclusões e perspectivas

A metodologia empregada, juntamente com os resultados obtidos por meio dela, permitiu a identificação dos melhores algoritmos de regressão simbólica em geral, ao mesmo tempo, em que apresentou uma aplicação eficiente de tais técnicas em um novo domínio.

[1] Especificações do computador: Fedora Linux 42 (64 bits), 24 × AMD Ryzen 9 9900X 12-Core Processor, 64 GB de RAM @3600 MHz, NVIDIA GeForce RTX 2060 SUPER, placa-mãe B850M D3HP.

O PySR provou ser o algoritmo mais robusto, recuperando a forma estrutural correta em todos os sistemas testados e também apresentando o melhor desempenho em todas as métricas: sua capacidade de reconstruir consistentemente a estrutura subjacente a partir dos dados, combinada com sua eficiência computacional, consolidou seu desempenho superior. Apenas em um caso as expressões resultantes mostram diferenças estatisticamente significativas em relação à dinâmica original, embora essas discrepâncias fossem numericamente insignificantes.

PySINDy e PyKAN, algoritmos que permitem maior interatividade em seu processo de treinamento, também tiveram um bom desempenho, identificando corretamente a maioria dos sistemas dinâmicos. No entanto, em um número ligeiramente maior de casos, os resultados gerados por esses métodos apresentaram diferenças significativas em comparação com as equações originais. Também note-se que ambos, e especialmente o PyKAN, tiveram uma queda significativa no desempenho quando corrupção na forma de ruído foi adicionada aos dados originais, apontando para uma menor capacidade de generalização quando comparado ao método de melhor desempenho.

Outros métodos tiveram graus variados de sucesso, com casos de uso que podem superar as abordagens de melhor desempenho ou complementá-las. No entanto, ODEFormer e AI Feynman foram dois outliers notáveis quando comparados a todos os outros. O primeiro, sendo a primeira metodologia multivariada baseada em transformador disponível, permitindo assim flexibilidade de inferência que pode ser ainda mais aprimorada, tem sua função atual limitada à de um gerador de hipóteses, o que corrobora (D'ASCOLI *et al.*, 2023). Além disso, embora esse método de aprendizagem por transferência evite problemas de amplificação de ruído dos métodos de diferenciação (ver Subseção 3.1.1), ele também pode representar uma desvantagem comparativa em relação aos algoritmos que incorporam uma fatia do sistema diretamente em seus estágios de treinamento (D'ASCOLI *et al.*, 2023).

Quanto ao AI Feynman, deve-se observar que ele é o que mais consome recursos, considerando o tempo necessário para realizar simulações. Eles também são os algoritmos mais antigos implementados nesta seleção e, no caso do AI-Feynman, é importante mencionar que o código-fonte, disponibilizado pelos próprios autores da metodologia em (**??**), não é atualizado há alguns anos, com a última atualização datando de 2020. Os futuros usuários do código devem estar cientes de que podem surgir dificuldades técnicas devido ao algoritmo depender de dependências desatualizadas.

Os resultados deste estudo destacam o potencial da regressão simbólica, sugerindo seu potencial como uma nova metodologia para a modelagem de dados relacionados à dinâmica epidêmica e, provavelmente, sendo flexível o suficiente para outros domínios do conhecimento, conforme indica a literatura atual. Além disso, atualizações desses algoritmos ou o desenvolvimento de novos algoritmos poderiam trazer resultados ainda melhores quando comparados aos métodos já de alto desempenho, superando as limitações atuais e permitindo uma recuperação dinâmica mais precisa e melhores resultados de previsão no futuro. A continuação deste estudo

é essencial para avaliar a capacidade desses métodos em descrever sistemas dinâmicos e, consequentemente, sua contribuição para a descrição da propagação de epidemias em populações reais.

Quanto às limitações atuais dessa abordagem, é importante mencionar que, mesmo com os avanços atuais nesse campo do aprendizado de máquina, a recuperação de equações a partir de dados por meio da regressão simbólica não pode ser feita em tempo polinomial, caracterizando-a como um problema NP-difícil (VIRGOLIN; PISSIS, 2022). Isso deve ser levado em consideração ao adicionar grandes conjuntos de funções básicas e suas várias permutações a esses métodos.

Além disso, este estudo não comparou exaustivamente todos os métodos incluídos e, dada a natureza orientada por dados da abordagem, seria necessário expandir as comparações para uma gama maior de parâmetros e condições iniciais dos sistemas escolhidos, além de adicionar outros, para discutir mais a fundo se um determinado método é superior a outro e em quais casos. Isso é especialmente verdadeiro quando se lida com sistemas com comportamento caótico conhecido, o que inclui o atrator de Lorenz investigado. Trabalhos futuros podem expandir esse tópico, possivelmente como uma plataforma de benchmark interativa análoga ou como um complemento ao SRBench (CAVA *et al.*, 2021).

Por fim, embora os resultados atuais apontem para um uso amplamente bem-sucedido desses métodos para vários sistemas, o desempenho para conjuntos de dados do mundo real pode ser prejudicado tanto pelo conhecimento limitado do comportamento real do sistema quanto pela possível imprecisão dos métodos de diferenciação numérica. Em modelos de regressão simbólica, a multicolinearidade pode resultar na inclusão de termos estruturalmente redundantes ou quase equivalentes. Assim, recomenda-se a avaliação da colinearidade por meio de métricas como o Fator de Inflação da Variância (VIF), que permite quantificar a inflação da variância associada à correlação entre preditores, além de outras abordagens estatísticas complementares (CHAN *et al.*, 2022; CASTILLO; VILLA, 2005), previamente à aplicação desses algoritmos em conjuntos de dados desconhecidos. Em conclusão, pesquisas futuras que empreguem esse método para a descoberta de equações devem considerar as técnicas mais robustas e atualizadas para lidar com essas questões, juntamente com uma investigação completa da identificabilidade do sistema em questão, antes de fornecer evidências conclusivas sobre a utilidade da regressão simbólica para dados do mundo real.

CAPÍTULO

5

REGRESSÃO SIMBÓLICA EM PROCESSOS DINÂMICOS EM REDES NÃO REGULARES

5.1 Panorama Geral

A propagação de doenças infecciosas é um tema central em diversas áreas da ciência, fato especialmente evidente nas últimas décadas, com surtos e pandemias que desafiaram a saúde pública global. Para compreender os mecanismos subjacentes a esses fenômenos, modelos epidemiológicos baseados em redes complexas tornaram-se ferramentas valiosas. Eles permitem representar a estrutura de interações entre indivíduos e o processo de disseminação, analisando como a topologia da rede influencia a dinâmica da propagação.

Esses sistemas, compostos por elementos interconectados que geram comportamentos globais não triviais, foram inicialmente caracterizadas por (NEWMAN, 2003). Em contextos epidemiológicos, vértices representam indivíduos e arestas indicam contatos potencialmente infecciosos. Essa abordagem tende a superar a limitação dos modelos tradicionais de população homogênea ao incorporar a heterogeneidade das interações sociais. Entre os modelos de rede mais estudados, destacam-se o de Erdős-Rényi (ER) (ERDŐS; RÉNYI *et al.*, 1960; GILBERT, 1959), que gera uma distribuição de grau aproximadamente Poisson, e o de Barabási-Albert (BA) (BARABÁSI; ALBERT, 2011), que introduz o mecanismo de ligação preferencial, resultando em redes com hubs e distribuição de grau em lei de potência. Ambos modelos discutidos anteriormente no Capítulo 2, ao qual o leitor pode retornar para retomar esses conceitos.

As características estruturais das redes, medidas por métricas como grau médio, variância do grau e coeficiente de aglomeração, influenciam diretamente os dados observados em processos epidêmicos. Modelos como a Cadeia de Markov (GÓMEZ *et al.*, 2010) e campo médio heterogêneo (HMF) (PASTOR-SATORRAS; VESPIGNANI, 2001) incorporam essas características para representar com maior fidelidade a dinâmica de propagação. No entanto, a

tradução de dados observados em equações analíticas permanece um desafio. Para enfrentá-lo, destaca-se no campo da inteligência artificial: a regressão simbólica.

Diferentemente dos métodos tradicionais de regressão, a regressão simbólica busca inferir a própria forma da equação a partir dos dados, combinando operadores básicos para encontrar expressões matemáticas interpretáveis. Essa característica a torna ideal para explorar sistemas complexos com relações não triviais, como epidemias em redes heterogêneas.

Diante desse contexto, o objetivo central deste capítulo é contextualizar, desenvolver e aplicar uma metodologia baseada em algoritmos de regressão simbólica para inferir dinâmicas epidemiológicas em redes complexas. Neste ponto, fazem-se uso dos materiais e métodos descritos no Capítulo 3, Seção 3.2, visando identificar os modelos com melhor desempenho e examinar eventuais limitações decorrentes das distintas topologias de rede.

5.2 Do Problema Epidemiológico à Descoberta de Equações: Uma Abordagem com Regressão Simbólica

A história da humanidade sempre esteve entrelaçada com crises de saúde decorrentes de padrões de contágio. Os efeitos observados da contaminação por vários agentes, incluindo vírus, bactérias, parasitas e outros microrganismos patogênicos, devem-se à sua capacidade de se espalhar rapidamente entre os hospedeiros, resultando frequentemente em surtos epidêmicos que estão correlacionados com eventos históricos.

Como discutido no Capítulo 2, entre esses episódios históricos, a peste de Atenas (LITTMAN, 2009) e a peste-negra figuram entre os mais impactantes das eras anteriores. Estima-se que esta última tenha causado a morte de aproximadamente um terço da população europeia durante o período medieval, sendo que a maioria dos indivíduos infectados sobrevivia apenas dois a três dias (MURRAY, 2001).

Além dessas grandes epidemias, doenças como o sarampo e o ebola permaneceram restritas às regiões afetadas, enquanto outras, como a COVID-19, alcançaram escala pandêmica em virtude de sua elevada transmissibilidade, intensificada pela mobilidade global contemporânea. Uma discussão mais aprofundada sobre a cronologia de vírus, epidemias e pandemias é apresentada no Capítulo 2, Seção 2.2.4, assim como nas obras de MONTENEGRO; BATISTA; STROPPA e COSTA; MERCHAN-HAMANN COSTA; MERCHAN-HAMANN.

Tanto os registros históricos quanto o contexto recente da pandemia da COVID-19 fornecem fortes evidências da necessidade de desenvolver estudos voltados ao manejo desse tipo de evento, a fim de auxiliar na otimização de respostas imediatas e auxiliar na tomada de decisões das agências de saúde durante possíveis surtos futuros. Nesse contexto, surgem naturalmente questões sobre como as doenças se disseminam e de que maneiras elas se propagam. Essas questões têm estimulado o interesse de vários campos do conhecimento: embora biólogos,

médicos e epidemiologistas tenham se empenhado em descrever diferentes patógenos e suas mutações e modos de transmissão, os avanços na modelagem matemática e computacional têm se mostrado ferramentas úteis para melhor compreender e prevenir o surgimento de contágios semelhantes.

A base da epidemiologia quantitativa tem origem na modelagem estatística, com modelos compartimentais surgindo como resultado da análise da propagação de doenças em escala global em um determinado sistema. Nesse contexto, abordagens como SIS e SIR (KERMACK; MCKENDRICK, 1927), tornaram-se indispensáveis como base para a compreensão da propagação de doenças e foram posteriormente expandidas com mais estados de transição para refletir melhor os processos de contágio no mundo real.

Assim, por meio desse sistema de equações diferenciais ordinárias foi possível determinar o comportamento da curva de infectados ao passo que a epidemia se espalha ao longo do tempo. Muitos modelos têm sido propostos até hoje, mas os modelos considerados clássicos e mais simples são: SIR e SIS. Ambos permitiram a análise profunda de quando uma nova infecção surge é se ela de fato irá se propagar. Isso por que toda epidemia expõe o número reprodutivo da infecção que pode ser medido e definido por R_0, razão descrita pelas taxas de infecção e recuperação. Essa métrica fundamental tem por finalidade descreve a intensidade com que uma infecção ou vírus se espalha e é útil na identificação da ocorrência ou não de uma epidemia. Sempre que $R_0 < 1$, a doença tende a desaparecer e quando $R_0 > 1$, a doença persiste e se propaga (BARRAT; BARTHELEMY; VESPIGNANI, 2008).

5.2.1 Modelos epidêmicos em Redes Complexas

Quando uma epidemia se espalha, torna-se essencial considerar a estrutura da rede sobre a qual a doença se propaga. Isso porque, modelos epidemiológicos que antes eram considerados simples, devido à suposição de meio homogêneo de propagação, revelam-se inadequados para capturar a heterogeneidade dos contatos reais. A incorporação de topologias complexas de redes introduz propriedades como nós altamente conectados (hubs), que aceleram as transições entre estados nos modelos epidemiológicos, resultando em uma dinâmica de propagação mais complexa.

Pesquisas, tais como as realizadas por Chung, Lu e Vu (2004) através de uma análise espectral de grafos aleatórios, revelaram que o maior autovalor da matriz adjacente está ligado à estrutura da rede e aos momentos do grau. Gómez *et al.* (2010), simularam a propagação epidêmica em redes complexas por meio da abordagem de Cadeia de Markov de tempo discreto, onde foi possível verificar a existência da relação entre o R_0 e a matriz de adjacência da rede de contatos. Como base na matriz de adjacência foi possível identificar indivíduos (nós), potencialmente propagadores. Essa identificação possibilitou alterar a estrutura da rede, a qual pode ser modificada mediante a políticas de vacinação ou imposição de *lockdown*. A mudança nas características da rede permite o controle de β, reduzindo-o e alcançando limites em que o

sistema de saúde possa suprir a demanda.

Atualmente, há duas abordagens estabelecidas para modelar a propagação de epidemias em redes complexas: uma baseia-se em cadeias de Markov e a outra em teorias de Campo Médio Heterogêneo. Estes métodos constituem a base da análise processos epidêmicos complexos, sendo frequentemente utilizados para compreender a relação entre a estrutura da rede e a dinâmica do sistema.

Nesse cenário, a inteligência artificial surge como uma ferramenta promissora para compreender a dinâmica desses sistemas. Diferente das abordagens analíticas tradicionais, métodos de regressão simbólica não apenas realizam previsões sobre estados futuros de transmissão, mas também geram equações interpretáveis. Tais equações podem ser diretamente comparadas a moda, modelos compartimentais clássicos, o que potencialmente viabiliza respostas mais ágeis a novos surtos epidêmicos.

5.2.2 Regressão Simbólica

Como já descrito no Capítulo 2, a regressão simbólica é basicamente um modelo de regressão, cujo objetivo é identificar a expressão matemática subjacente a um conjunto de dados. Suas aplicações têm se mostrado amplas e diversificadas, indo da área de sistemas dinâmicos (QUADE *et al.*, 2016), em contextos energéticos e meteorológicos (PAIVA *et al.*, 2018; ABDELLAOUI; MEHRKANOON, 2021) até o campo da psicologia (MIYAZAKI *et al.*, 2023).

A metodologia tem raízes no método empírico de Kepler, que no século XVII identificou as órbitas elípticas dos planetas com base em dados observacionais (CAMPS-VALLS *et al.*, 2023). No campo da Inteligência Artificial, avanços começaram com o sistema BACON (LANGLEY, 1977), seguido pelo uso dos algoritmos genéticos desenvolvidos por Holland (GOLBERG, 1989). A consolidação da técnica ocorreu com a introdução da programação genética por Koza (1992), que utilizou árvores de expressão para representar equações matemáticas, estabelecendo o primeiro modelo funcional de regressão simbólica (como discutido na Seção 2.3).

Hoje, os modelos de regressão simbólica abrangem uma ampla variedade de abordagens, desde algoritmos genéticos até métodos baseados em aprendizado de máquina, como redes neurais e técnicas de regularização. Com o objetivo de investigar o desempenho dos modelos simbólicos de código aberto desenvolvidos em Python que conciliem desempenho competitivo e interpretabilidade, quando aplicados a dados provenientes de redes complexas, os algoritmos **GPLearn** (STEPHENS, 2016), **AI-Feynman** (UDRESCU; TEGMARK, 2020), **PySINDy** (SILVA *et al.*, 2020), **PySR** (CRANMER, 2023), **PyKAN** (KAPTANOGLU *et al.*, 2021; LIU *et al.*, 2024b) e **ODEFormer** (D'ASCOLI *et al.*, 2023) foram aplicados. A avaliação desses métodos permitiu comparar suas capacidades de recuperação de equações, robustez frente a ruído, eficiência computacional e potencial de generalização, oferecendo uma análise sistemática

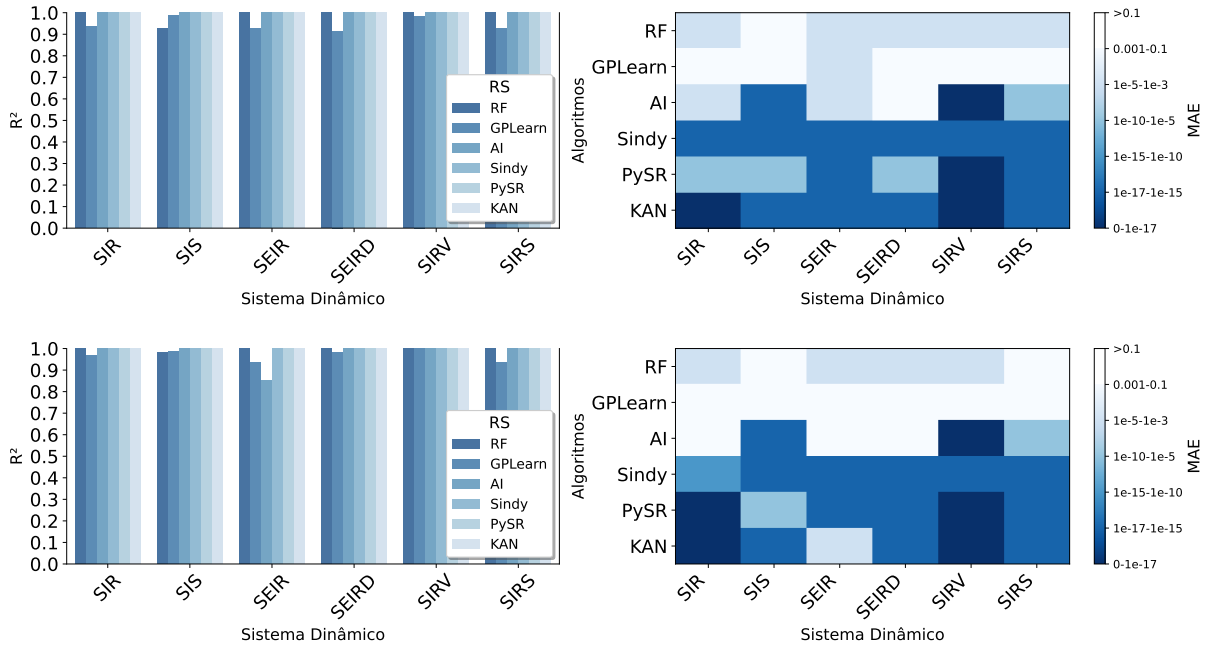
e inédita no contexto de sistemas dinâmicos epidêmicos em redes complexas, um domínio no qual ainda carecem estudos comparativos estruturados dessa natureza.

Embora redes complexas e técnicas de regressão simbólica sejam amplamente empregadas em diversas áreas do conhecimento, persistia até então uma lacuna na literatura referente à avaliação do desempenho de modelos de regressão simbólica na recuperação da forma funcional de sistemas dinâmicos epidêmicos em redes complexas. Nas seções seguintes, estão detalhados os resultados obtidos com base na implementação do método descrito na Seção 3.2.

5.3 Regressão Simbólica para Extração de Equações em Dados Complexos

A análise do desempenho dos modelos de regressão simbólica, em relação ao R^2 e MAE, estratificada por tipo de rede, detectou padrões de desempenho notavelmente consistentes. Esta uniformidade mediante diferentes cenários de transmissão ressalta uma possível robustez e generalizabilidade dos algoritmos investigados. Essa similaridade de desempenho pode ser verificado na Figura 24.

Figura 24 – Comparações pós-hoc foram realizadas por meio de efeitos marginais estimados a partir do modelo fractional logit, permitindo avaliar diferenças médias de desempenho entre os métodos de regressão simbólica em relação ao Random Forest, separadamente por tipo de rede. A significância estatística foi avaliada com base em intervalos de confiança dos contrastes marginais.



Para facilitar a visualização, os valores de MAE foram categorizados em intervalos, permitindo uma melhor identificação de padrões de desempenho dos modelos. O modelo de Floresta Aleatória foi utilizado apenas como linha de base, por isso se assumiu uma avaliação

otimista em relação a ela. A ideia não é comparar seu desempenho com os modelos de regressão, mas sim, verificar se eles conseguem se aproximam dela em seu melhor cenário possível.

Com base na análise gráfica, verifica-se o bom desempenho dos algoritmos de regressão simbólica, os quais apresentam proximidade com o cenário mais otimista do modelo de Floresta Aleatória. Esse desempenho é corroborado pela capacidade de capturar a variabilidade dos dados (alto R^2) e pelos baixos valores de MAE.

As expressões simbólicas identificadas pelo algoritmo de regressão para cada processo epidêmico nas diferentes redes estão apresentadas nas Tabelas 9 e 10. A Tabela 8 resume a análise, na qual as equações recuperadas corretamente são indicadas por uma marca de seleção (✓). Para dinâmicas reconstruídas com sucesso, foi aplicado o teste de Wilcoxon, e aquelas sem diferenças estatisticamente significativas em relação aos dados originais estão marcadas em cinza.

Tabela 8 – Resumo dos principais resultados de Tabela 9 e Tabela 10. As marcas de seleção indicam que a forma estrutural do sistema foi identificada com sucesso. As células sombreadas em cinza destacam a dinâmica que não apresentou diferenças estatisticamente significativas em comparação com a dinâmica original, de acordo com o teste de Wilcoxon ($\alpha = 0.05$).

Net	SR model	**Modelos epidemiológicos**					
		SIR	SIS	SEIR	SEIRD	SIRV	SIRS
ER	GPLearn	✓				✓	
	AI-Feynman	✓				✓	✓
	PySINDy	✓	✓	✓	✓	✓	✓
	PySR	✓	✓	✓	✓	✓	✓
	PyKAN	✓	✓	✓	✓	✓	✓
	ODEFormer*						
BA	GPLearn	✓				✓	
	AI-Feynman	✓				✓	✓
	PySINDy	✓	✓	✓	✓	✓	✓
	PySR	✓	✓	✓	✓	✓	✓
	PyKAN	✓	✓	✓	✓	✓	✓
	ODEFormer*						

*O algoritmo ODEFormer apresentou falha durante a etapa de busca, não gerando resultados válidos.

5.3.1 O efeito da topologia da rede no desempenho da regressão simbólica

Para avaliar o impacto da rede de geração dos dados e do algoritmo de regressão simbólica no desempenho preditivo, medido pelo coeficiente de determinação (R^2), uma variável fracionária limitada ao intervalo [0,1], o modelo Beta não foi viável devido à forte concentração próxima ao

Tabela 9 – A tabela descreve as representações simbólicas dos sistemas epidêmicos em redes complexas ER aproximadas, geradas pelos modelos de regressão. Sistemas dinâmicos cujas formas estruturais foram identificadas corretamente são marcados com "check mark"(✓). Nota: O algoritmo ODEFormer falhou durante o processo de busca, por esse motivo ele foi omitidos na tabela.

A expressão procurada	GPLearn	AI-Feynman	PySINDy	PySR	PyKAN
$\dot{S} = -0.2SI$ $\dot{I} = 0.2SI - 1.0I$ $\dot{R} = 1.0I$	✓ $\dot{S} = -3.905SI$ $\dot{I} = 4SI - 1.01I$ $\dot{R} = I$	✓ $\dot{S} = -0.2SI$ $\dot{I} = -0.001 + (I/(-((I*S)+1)))$ $\dot{R} = 1.0I$	✓ $\dot{S} = -0.200SI$ $\dot{I} = -1.0I + 0.2SI$ $\dot{R} = 1.0I$	✓ $\dot{S} = -0.2SI$ $\dot{I} = 0.2SI - 1.0I$ $\dot{R} = 1.0I$	✓ $\dot{S} = -0.2SI$ $\dot{I} = 0.2SI - 1.0I$ $\dot{R} = 1.0I$
$\dot{S} = -0.2SI + 1.0I$ $\dot{I} = 0.2SI - 1.0I$	$\dot{S} = S(-0.5I - 0.64)$ $\dot{I} = 0.7S + 0.09$	✓ $\dot{S} = -0.2SI + 0.6I + 0.4I^2$ $\dot{I} = 0.2SI - 0.6I - 0.4I^2$	✓ $\dot{S} = 1.0I - 0.2SI$ $\dot{I} = -1.0I + 0.2SI$	✓ $\dot{S} = -0.2SI + 1.0I$ $\dot{I} = 0.2SI - 1.0I$	✓ $\dot{S} = -0.2IS + 0.2I - 0.8S + 0.8$ $\dot{I} = 0.2IS - 0.2I + 0.8S - 0.8$
$\dot{S} = -0.1SI$ $\dot{E} = 0.1SI - 0.3E$ $\dot{I} = 0.3E - 0.2I$ $\dot{R} = 0.2I$	$\dot{S} = (S^2 + 2S)(I^2 - 0.9I)$ $\dot{E} = 0.24E - 0.24I$ $\dot{I} = 0.299E - 0.199134I$ $\dot{R} = 0.2II$	$\dot{S} = (-SI + \sqrt{(}2) + 2)^{(}1/4) - 1.36$ $\dot{E} = -E/(S + I + 3) + 0.0007$ $\dot{I} = (E - (E+3) * (I + 0.007507113095)/5)/(E+3)$ $\dot{R} = 0.2I$	✓ $\dot{S} = -0.1SI$ $\dot{E} = -0.3E + 0.1SI$ $\dot{I} = 0.3E - 0.2I$ $\dot{R} = 0.2I$	✓ $\dot{S} = -0.1SI$ $\dot{E} = 0.1SI - 0.3E$. $\dot{I} = 0.3E - 0.2I$ $\dot{R} = 0.2I$	✓ $\dot{S} = -0.1SI$ $\dot{E} = 0.1SI - 0.3E$ $\dot{I} = 0.3E - 0.2I$ $\dot{R} = 0.2I$
$\dot{S} = -0.3SI$ $\dot{E} = 0.3SI - 0.2E$ $\dot{I} = 0.2E - (0.2 + 0.05)I$ $\dot{R} = 0.2I)$ $\dot{D} = 0.05I$	$\dot{S} = -5.7SI$ $\dot{E} = -0.28E(-S + 3I)(E + 0.18)/(E + 0.6I + 0.18)$ $\dot{I} = 0.2 * E - 0.25I$ $\dot{R} = 0.05 - 0.07I$ $\dot{D} = 0.05I$	$\dot{S} = 2(-SI + 2)^{(}1/4) - 2.3$ $\dot{E} = (S*I)^{(}3/2) - 0.2E + 0.002$ $\dot{I} = 0.2 * E - 0.25 * I - 9.08e - 5$ $\dot{R} = 0.2I$ $\dot{D} = 1.14 - (3 - I)^{(}1/8)$	✓ $\dot{S} = -0.3S * I$ $\dot{E} = -0.2E + 0.3S * I$ $\dot{I} = 0.2E - 0.250I$ $\dot{R} = 0.2I$ $\dot{D} = 0.05I$	✓ $\dot{S} = -0.3SI$ $\dot{E} = 0.3SI - 0.2E$ $\dot{I} = 0.2E - 0.25I$ $\dot{R} = 0.2I$ $\dot{D} = 0.05I$	✓ $\dot{S} = -0.3SI$ $\dot{E} = 0.3SI - 0.2E$ $\dot{I} = 0.2E - 0.25I$ $\dot{R} = 0.2I$ $\dot{D} = 0.05I$
$\dot{S} = -0.3SI - 0.5S$ $\dot{I} = 0.3SI - 1.0I$ $\dot{R} = 1.0I$ $\dot{V} = 0.5S$	✓ $\dot{S} = -8.8SI - 0.88S$ $\dot{I} = 10SI - 0.6I$ $\dot{R} = 1.0I$ $\dot{V} = 0.49S$	✓ $\dot{S} = -0.5SI - 0.5S$ $\dot{I} = 0.5SI - 1.0I$ $\dot{R} = 1.0I$ $\dot{V} = 0.5S$	✓ $\dot{S} = -0.5S + -0.5SI$ $\dot{I} = -1.0I + 0.5SI$ $\dot{R} = 1.0I$ $\dot{V} = 0.5S$	✓ $\dot{S} = -0.5SI - 0.5S$ $\dot{I} = 0.5SI - 1.0I$ $\dot{R} = 1.0I$ $\dot{V} = 0.5S$	✓ $\dot{S} = -0.5SI - 0.5S$ $\dot{I} = 0.5SI - 1.0I$ $\dot{R} = 1.0I$ $\dot{V} = 0.5S$
$\dot{S} = -0.2SI + 0.2R$ $\dot{I} = 0.2SI - 1.0I$ $\dot{R} = 1.0I - 0.2R$	$\dot{S} = -R - 0.617$ $\dot{I} = 4SI - 1.012I$ $\dot{R} = I - 0.207R$	✓ $\dot{S} = -0.2SI + 0.2R$ $\dot{I} = 0.2SI - 1.0I$ $\dot{R} = I - 0.2R + 5.2e - 6$	✓ $\dot{S} = 0.2R + -0.2SI$ $\dot{I} = -1.0I + 0.2SI$ $\dot{R} = 1.0I + -0.2R$	✓ $\dot{S} = -0.2SI - 0.2R$ $\dot{I} = 0.2SI - 1.0I$ $\dot{R} = 1.0I - 0.2R$	✓ $\dot{S} = -1.0SI + 1.5R$ $\dot{I} = 3.972SI - 1.0I$ $\dot{R} = 1.0I - 0.2R$

Tabela 10 – A tabela descreve as representações simbólicas dos sistemas epidêmicos em redes complexas BA aproximadas, geradas pelos modelos de regressão. Sistemas dinâmicos cujas formas estruturais foram identificadas corretamente são marcados com "check mark"(✓).

A expressão procurada	GPLearn	AI-Feynman	PySINDy	PySR	PyKAN
$\dot{S} = -0.2SI$ $\dot{I} = 0.2SI - 1.0I$ $\dot{R} = 1.0I$	✓ $\dot{S} = -4.56SI$ $\dot{I} = 4SI - 0.971I$ $\dot{R} = I$	✓ $\dot{S} = -0.2SI$ $\dot{I} = I(I(S-I)-1)+0.008$ $\dot{R} = 0.1I$	✓ $\dot{S} = -0.2SI$ $\dot{I} = -1.0I + 0.2SI$ $\dot{R} = 1.0I$	✓ $\dot{S} = -0.2SI$ $\dot{I} = 0.2SI - 1.0I$ $\dot{R} = 1.0I$	✓ $\dot{S} = -0.2SI$ $\dot{I} = 0.2SI - 1.0I$ $\dot{R} = 1.0I$
$\dot{S} = -0.2SI + 1.0I$ $\dot{I} = 0.2SI - 1.0I$	$\dot{S} = -I - 0.981 + 0.44/S$ $\dot{I} = 0.768 - I^2$	$\dot{S} = 0.2I(I+4)$ $\dot{I} = -0.2I(I+4)$	✓ $\dot{S} = 1.0I - 0.2SI$ $\dot{I} = -1.0I + 0.2SI$	✓ $\dot{S} = -0.2SI + 1.0I$ $\dot{I} = 0.1SI - 1.0I$	✓ $\dot{S} = -0.2IS + 1.0I$ $\dot{I} = 0.2IS - 1.0I$
$\dot{S} = -0.1SI$ $\dot{E} = 0.1SI - 0.3E$ $\dot{I} = 0.3E - 0.2I$ $\dot{R} = 0.2I$	$\dot{S} = -1.998SI$ $\dot{E} = (E-I)/(0.76S+2.68)$ $\dot{I} = -0.59E(I-0.564)$ $\dot{R} = 0.2I$	$\dot{S} = 0.9 - \sqrt{(}5SI+25)/5$ $\dot{E} = (-0.0002 * S^(1/4) - I - 0.0007)/(S^(1/4)+3)$ $\dot{I} = ((E(E+1)+1)/(I+1))^(1/4) - 0.99$ $\dot{R} = 0.2I$	✓ $\dot{S} = -0.1SI$ $\dot{E} = -0.3E + 0.1SI$ $\dot{I} = 0.3E - 0.2I$ $\dot{R} = 0.2I$	✓ $\dot{S} = -0.1SI$ $\dot{E} = 0.1SI - 0.3E$ $\dot{I} = 0.3E - 0.2I$ $\dot{R} = 0.2I$	✓ $\dot{S} = -0.099SI$ $\dot{E} = -0.3E + 0.099IS$ $\dot{I} = 0.3E - 0.2I$ $\dot{R} = 0.2I$
$\dot{S} = -0.3SI$ $\dot{E} = 0.3SI - 0.2E$ $\dot{I} = 0.2E - (0.2+0.05)I$ $\dot{R} = 0.2I$ $\dot{D} = 0.05I$	$\dot{S} = -0.258S(I+0.586)$ $\dot{E} = 0.32E(-I-0.88)(-16.13SI+1.07I)$ $\dot{I} = 0.2E - 0.25I$ $\dot{R} = 0.05778 - 0.107I$ $\dot{D} = 0.05I$	$\dot{S} = 1.6 - \sqrt{(}SI+1+\sqrt{(}3))$ $\dot{E} = -0.2E + 1.0\sqrt{(}1+2/\sqrt{(}-S*I+1)) - 1.7$ $\dot{I} = (E+I-0.5(I+2)^2+2)/(I+2)^2$ $\dot{R} = 0.2I$ $\dot{D} = 0.05I$	✓ $\dot{S} = -0.3SI$ $\dot{E} = 0.3SI - 0.2E$ $\dot{I} = 0.2E - 0.25I$ $\dot{R} = 0.2I$ $\dot{D} = 0.05I$	✓ $\dot{S} = -0.3SI$ $\dot{E} = 0.3SI - 0.2E$ $\dot{I} = 0.2E - 0.25I$ $\dot{R} = 0.2I$ $\dot{D} = 0.05I$	✓ $\dot{S} = -0.3SI$ $\dot{E} = 0.3SI - 0.2E$ $\dot{I} = 0.2E - 0.25I$ $\dot{R} = 0.2I$ $\dot{D} = 0.05I$
$\dot{S} = -0.3SI - 0.5S$ $\dot{I} = 0.3SI - 1.0I$ $\dot{R} = 1.0I$ $\dot{V} = 0.5S$	✓ $\dot{S} = -8.8SI - 0.8S$ $\dot{I} = 9.8SI - 0.98I$ $\dot{R} = 1.0I$ $\dot{V} = 0.49S$	✓ $\dot{S} = -0.5SI - 0.5S$ $\dot{I} = 0.5SI - 1.0I$ $\dot{R} = 1.0I$ $\dot{V} = 0.5S$	✓ $\dot{S} = -0.5SI - 0.5S$ $\dot{I} = 0.5SI - 1.0I$ $\dot{R} = 1.0I$ $\dot{V} = 0.5S$	✓ $\dot{S} = -0.5SI - 0.5S$ $\dot{I} = 0.5SI - 1.0I$ $\dot{R} = 1.0I$ $\dot{V} = 0.5S$	✓ $\dot{S} = -0.5SI - 0.5S$ $\dot{I} = 0.5SI - 1.0I$ $\dot{R} = 1.0I$ $\dot{V} = 0.5S$
$\dot{S} = -0.2SI + 0.2R$ $\dot{I} = 0.2SI - 1.0I$ $\dot{R} = 1.0I - 0.2R$	$\dot{S} = I(IR+0.28) + R - 0.73$ $\dot{I} = -0.732I + (S-0.064)(SI+4.07742I)$ $\dot{R} = I - 0.192R$	$\dot{S} = -0.2SI + 0.2R$ $\dot{I} = 0.2SI - 1.0I$ $\dot{R} = I - 0.2R - 2.5e-5$	✓ $\dot{S} = -0.2SI + 0.2R$ $\dot{I} = 0.2SI - 1.0I$ $\dot{R} = 1.0I - 0.2R$	✓ $\dot{S} = -0.2SI + 0.2R$ $\dot{I} = 0.2SI - 1.0I$ $\dot{R} = 1.0I - 0.2R$	✓ $\dot{S} = -1.0IS + 1.0R$ $\dot{I} = 0.2IS - 1.0I$ $\dot{R} = 1.0I - 0.2R$

valor máximo. Assim, optou-se pelo modelo Linear Generalizado *Fractional Logit*, proposto por Papke e Wooldridge (1996).

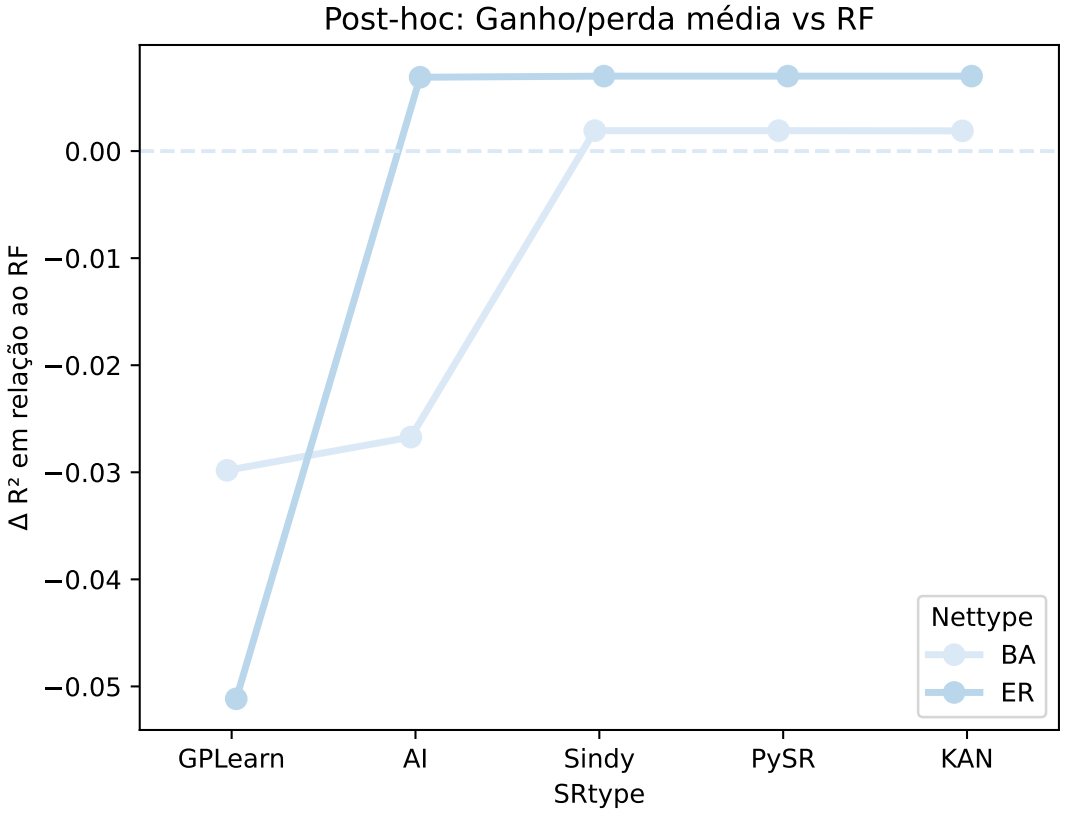
$$\text{logit}\left(\mathbb{E}[R^2_{ij}]\right) = \beta_0 + \alpha_i + \gamma_j + (\alpha\gamma)_{ij}, \tag{5.1}$$

em que $i = 1,2$ denota a topologia da rede e $j = 1,\ldots,5$ denota o método de regressão simbólica. O termo α_i representa o efeito principal da topologia da rede, γ_j representa o efeito principal do método de regressão simbólica e $(\alpha\gamma)_{ij}$ representa o efeito de interação entre a topologia da rede e o método de regressão simbólica.

Com base nos resultados obtidos, constatou-se que, embora alguns modelos tenham conseguido recuperar adequadamente a forma estrutural das equações, os algoritmos apresentaram maior dificuldade, estatisticamente significativa, na reconstrução dos processos epidêmicos simulados em redes do tipo Barabási–Albert model (BA).

Uma visualização dessa evidência pode ser observada na Figura 25. Em que o desempenho médio dos regressores, quando aplicados a dinâmicas propagadas em rede BA, aproxima-se do desempenho do modelo de Floresta Aleatória, indicando uma redução relativa da vantagem estrutural dos métodos simbólicos nesse cenário mais heterogêneo.

Figura 25 – Comparações pós-hoc permitiram avaliar diferenças médias, em relação ao R^2, de desempenho entre os métodos de regressão simbólica em relação à Floresta Aleatória (RF), separadamente por tipo de rede.



Os valores de MAE apresentam distribuição altamente assimétrica à direita, com magnitude variando de 10^{-18} a 10^{-1}, que compromete a estabilidade numérica de modelos baseados em distribuições positivas contínuas. Para contornar esse problema, aplicou-se uma transformação logarítmica da forma $\log(\text{MAE}+\delta)$, em que δ representa uma constante positiva de pequena magnitude, introduzida para evitar indefinições e atenuar a influência de valores próximos de zero. Essa transformação contribuiu para estabilizar a variância e reduzir a assimetria da distribuição dos resíduos.

Na sequência, foi ajustado um modelo linear, permitindo realizar inferência estatística sobre os efeitos da topologia da rede e do algoritmo de regressão simbólica. Os resultados indicaram ausência de evidências estatisticamente significativas de interação entre esses fatores, sugerindo que o erro produzido pelos algoritmos não depende da topologia da rede considerada.

De modo geral, os modelos estatísticos aplicados evidenciaram diferenças estatisticamente significativas no desempenho dos algoritmos quando avaliados pelo coeficiente de determinação (R^2). Esse comportamento era, em certa medida, esperado, uma vez que em redes BA a distribuição de graus apresenta alta heterogeneidade, caracterizada pela presença de nós altamente conectados (hubs). Tal estrutura favorece uma propagação mais rápida e não homogênea da epidemia, aumentando a complexidade dinâmica do sistema e, consequentemente, impondo maior desafio aos algoritmos de regressão simbólica na identificação da forma funcional subjacente. Em contraste, para erro absoluto médio (MAE), não foram observadas diferenças estatisticamente significativas.

5.4 Resultados e discussões

A análise dos algoritmos de regressão simbólica evidenciou a eficácia de três abordagens na modelagem de dinâmicas epidemiológicas em redes complexas, tanto na recuperação da expressão geradora dos dados quanto no desempenho numérico das predições. Destacaram-se o Symbolic Regression (SR), método de natureza evolutiva baseado em algoritmos genéticos, e dois métodos de caráter não convencional, PySINDy e PyKAN, fundamentados, respectivamente, em identificação esparsa de sistemas dinâmicos e em redes do tipo Kolmogorov–Arnold.

O PySR destaca-se por ser fundamentado em algoritmos genéticos e programação genética, combinando elevado poder computacional com relativa facilidade de uso. Em contrapartida, os modelos não convencionais foram concebidos para incorporar e, em muitos casos, exigir a interação humano-máquina, a qual desempenha papel central na condução da busca por representações simbólicas adequadas. Embora essa característica amplie significativamente a flexibilidade e a capacidade de modelagem, sua utilização tende a ser mais complexa, em função da robustez e do amplo conjunto de ferramentas disponíveis. Por essa razão, tais métodos mostram-se particularmente adequados para pesquisadores com conhecimento aprofundado em sistemas dinâmicos e dinâmicas complexas.

Nesse contexto, os algoritmos PySINDy e PyKAN podem ser compreendidos não apenas como métodos de regressão simbólica, mas também como ferramentas de modelagem orientadas por conhecimento, nas quais informações a priori podem ser incorporadas de forma explícita. Essa característica contribui tanto para a melhoria do desempenho preditivo quanto para o aumento da interpretabilidade das equações identificadas, aspecto especialmente relevante em aplicações epidemiológicas.

Embora o AI-Feynman apresente excelente desempenho em termos de métricas de erro, ele pode falhar na identificação de determinadas estruturas funcionais em razão de sua automatização interna. O algoritmo aplica múltiplos processos e transformações aos dados, o que frequentemente resulta em expressões excessivamente complexas. Além disso, em comparação com os métodos anteriormente discutidos, trata-se de uma abordagem relativamente mais lenta. Ainda assim, o AI-Feynman permite que a busca simbólica seja interrompida em qualquer iteração, possibilitando ao pesquisador obter expressões úteis sem a necessidade de execução completa do algoritmo. Sua estratégia de busca é guiada pelo princípio do Comprimento Mínimo de Descrição, que busca equilibrar a redução do erro com a simplicidade da expressão final, mas que pode levar ao insucesso quando uma dinâmica subjacente adequada não é identificada.

O GPLearn, por sua vez, apresentou desempenho inferior à maioria dos algoritmos avaliados, superando apenas o ODEFormer, que obteve resultados inconsistentes. Apesar de também se basear em algoritmos genéticos, o GPLearn não dispõe do mesmo poder computacional nem da flexibilidade observada no PySR, tornando-se pouco competitivo na presença das demais alternativas consideradas neste estudo.

Por fim, ao analisar os processos epidemiológicos sob diferentes topologias de rede, observou-se que a maioria dos modelos apresentou desempenho significativamente influenciado pela estrutura da rede quando avaliado pelo coeficiente de determinação (R^2). Esse resultado pode ser explicado pelo fato de que o R^2 depende da variância total da dinâmica, a qual pode variar conforme a topologia considerada.

Em contraste, ao utilizar a métrica de erro absoluto médio (MAE), baseada no erro ponto a ponto, não foram identificadas diferenças estatisticamente significativas entre os erros estimados pelos algoritmos nos diferentes contextos de rede. Tal achado indica que, embora a capacidade de explicação global da variância seja sensível à estrutura topológica, o erro médio local permanece relativamente estável entre os distintos cenários de propagação analisados.

A integração desses resultados com os testes de Wilcoxon permite inferir sobre a robustez dos métodos avaliados no cenário simulado. Para a maioria das dinâmicas recuperadas, não foram observadas diferenças estatisticamente significativas em comparação às dinâmicas originais, evidenciando a capacidade dos algoritmos de reproduzir adequadamente o comportamento estrutural subjacente dos sistemas estudados.

CAPÍTULO

6

CONCLUSÕES, LIMITAÇÕES E PROPOSTA DE PESQUISA

6.1 Conclusão

Esta pesquisa forneceu uma revisão tanto do contexto histórico da regressão simbólica quanto dos algoritmos de última geração atualmente utilizados por essa técnica emergente, ao mesmo tempo em que comparou vários deles na tarefa de recuperar as equações que regem a dinâmica não linear. Além disso, a análise propõe uma nova metodologia semi-supervisionada para inferência de equações diferenciais em sistemas dinâmicos epidemiológicos complexos, baseada em regressão simbólica. Entre todos os algoritmos de regressão simbólica disponíveis, GPLearn, AI-Feynman, PySINDy, PySR, PyKAN e ODEFormer foram selecionados para essas comparações detalhadas, usando dados sintéticos de sistemas físicos, biológicos e epidemiológicos, sendo este último uma aplicação ainda pouco explorada para esses modelos. Embora esses algoritmos tenham demonstrado alto desempenho em processos físicos, até este momento ainda não haviam sido validados em contextos epidemiológicos caracterizados por múltiplos processos dinâmicos acoplados que se propagam mediante diferentes meios. A análise identificou os algoritmos com maior capacidade preditiva quando aplicados a dinâmicas epidemiológicas em redes complexas. Os resultados revelam que a regressão simbólica, em especial as bibliotecas PySINDy, PySR e PyKAN apresentaram potencial promissor para modelagem de sistemas epidêmicos complexos, com destaque para a PySR.

Adicionalmente, ficou constatado que a topologia da rede não exerce efeito no desempenho desses algoritmos quanto ao MAE, mas exerce efeito em relação ao $R2$, o que já era esperado. No entanto, a robustez dos mesmos, nos cenários simulados, mostrou-se surpreendente, demonstrando desempenho estatisticamente equivalente nas dinâmicas produzidas por diferentes topologias de rede. As equações aprendidas revelaram-se generalizáveis, independentemente das características estruturais específicas do meio de propagação, o que parece viabilizar sua

implementação em cenários reais de saúde pública.

Vale enfatizar que a abordagem adotada privilegiou a identificação da estrutura matemática intrínseca aos dados, e não o ajuste de formas funcionais genéricas. Em saúde pública, desvendar a arquitetura formal das dinâmicas e determinar parâmetros fundamentais, como taxas de transmissão e recuperação, constitui o cerne do processo analítico.

Embora o presente estudo tenha sido conduzido em cenários simulados e controlados, a metodologia proposta é, em princípio, extensível a contextos reais de saúde pública, nos quais a identificação da forma funcional subjacente pode contribuir para a interpretação mecanicista da dinâmica epidemiológica, o aprimoramento de modelos preditivos e o suporte à tomada de decisão baseada em evidências.

6.2 Limitação

Em experimentos totalmente controlados, como o da presente pesquisa, quando temos controle da geração dos dados e de sua derivada exata, alguns algoritmos de regressão simbólica se saem muito bem, mas quando o controle é parcial ou inexistente, isto é, temos dados provenientes de um processo físico conhecido e não temos a resposta, eles nem sempre apresentam o mesmo desempenho. Assim como ocorre com os dados do mundo real, os dados simulados também podem apresentar um elevado nível de ruído, uma característica inerente aos processos estocásticos que os geram. Esse ruído, onipresente em qualquer conjunto de dados oriundo de fenômenos aleatórios, não apenas dificulta a análise, mas também pode levar a inferências equivocadas.

BREIMAN defendia a importância de um conjunto diversificado de ferramentas para a análise de dados, argumentando que a ênfase metodológica deve ser orientada pelo problema e pelos dados em questão, não pelo modelo escolhido. Um dos principais desafios da análise moderna reside na ubiquidade de dados ruidosos, o que motivou o desenvolvimento de diversas técnicas robustas. Entre elas, destacam-se os modelos de regularização, como Ridge (HOERL; KENNARD, 1970) e Lasso (TIBSHIRANI, 1996), bem como otimizadores adaptativos como Adam (KINGMA; BA, 2014), projetado para lidar eficientemente com esse tipo de dado por meio de sua alta adaptabilidade e suavização de gradientes. Em uma linha diferente, mas igualmente importante, os algoritmos genéticos (AG), precursores do primeiro modelo de regressão simbólica da história (KOZA, 1992). Consequentemente, os atuais algoritmos de regressão simbólica contam com um vasto arsenal de técnicas, que inclui desde regressão esparsa (SILVA *et al.*, 2020), redes neurais (UDRESCU; TEGMARK, 2020; LIU *et al.*, 2024b; LIU *et al.*, 2024a) e algoritmos genéticos (STEPHENS, 2016; CRANMER, 2023), até suporte estatístico para a avaliação de desempenho.

É precisamente neste contexto de evolução metodológica e na necessidade de abordagens robustas e híbridas que se insere um novo projeto de investigação.

6.3 Proposta de pesquisa

Nossa nova proposta de pesquisa, intitulada “Uma Estratégia Híbrida Baseada em Regressão Simbólica para Modelagem de Sistemas Dinâmicos”, tem como objetivo investigar a modelagem e a inferência em sistemas complexos a partir de dados reais extraídos do DataSUS, com foco na aplicação e no aprimoramento de algoritmos de regressão simbólica fundamentados em técnicas previamente discutidas. Entre esses algoritmos destacam-se PySINDy, PySR e PyKAN.

O objetivo principal é elaborar uma metodologia híbrida de dois estágios, com diferentes modelos de regressão simbólica. Cada algoritmo dentro do estágio terá um papel distinto e complementar. O primeiro atuará como suavizador dos dados, produzindo derivadas aproximações contínuas e fornecendo uma representação simbólica não necessariamente interpretável, mas sim, muito suave e próxima da resposta real. O segundo, com base na previsão obtida no estágio anterior, deverá retornar a forma funcional interpretável da dinâmica. Tratar dados ruidosos por meio de metodologias híbridas pode ser algo enriquecedor quando lidamos com dados reais que se propagam sobre a rede de contatos altamente complexa, mas abordar essa metodologia híbrida como método de inferir a forma estrutural de dinâmicas subjacentes aos dados pode ser algo promissor.

Espera-se que a estratégia híbrida permita encontrar com maior eficiência a forma estrutural interpretável dos dados. Se bem-sucedida, essa metodologia pode estabelecer a regressão simbólica como uma das metodologias-chave na inferência de sistemas dinâmicos, buscando trazer conhecimento preciso da representação simbólica interpretável, fundamental para a descrição de processos dinâmicos de populações finitas, tanto em meio homogêneo quanto heterogêneo.

REFERÊNCIAS

ABDELLAOUI, I. A.; MEHRKANOON, S. Symbolic regression for scientific discovery: an application to wind speed forecasting. In: IEEE. **2021 IEEE Symposium Series on Computational Intelligence (SSCI)**. [S.l.], 2021. p. 01–08. Citado nas páginas 28 e 78.

ADAMIC, L. A. Zipf, power-laws, and pareto-a ranking tutorial. **Xerox Palo Alto Research Center, Palo Alto, CA, http://ginger. hpl. hp. com/shl/papers/ranking/ranking. html**, 2000. Citado na página 35.

ALBERT, R.; JEONG, H.; BARABÁSI, A.-L. Diameter of the world-wide web. **nature**, Nature Publishing Group UK London, v. 401, n. 6749, p. 130–131, 1999. Citado na página 35.

ALMEIDA, R. F. d. Pesquisa bibliográfica sobre aprendizado de máquina aplicado à condução veicular autônoma: Uma revisão. 2024. Citado na página 28.

ANNAS, S.; PRATAMA, M. I.; RIFANDI, M.; SANUSI, W.; SIDE, S. Stability analysis and numerical simulation of seir model for pandemic covid-19 spread in indonesia. **Chaos, solitons & fractals**, Elsevier, v. 139, p. 110072, 2020. Citado na página 46.

BARABÁSI, A.-L. Network science. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, The Royal Society Publishing, v. 371, n. 1987, p. 20120375, 2013. Citado nas páginas 34, 36 e 41.

BARABÁSI, A. L.; ALBERT, R. Emergence of scaling in random networks. In: **The structure and dynamics of networks**. [S.l.]: Princeton University Press, 2011. p. 349–352. Citado na página 75.

BARRAT, A.; BARTHELEMY, M.; VESPIGNANI, A. **Dynamical processes on complex networks**. [S.l.]: Cambridge university press, 2008. Citado nas páginas 36, 37, 48 e 77.

BARTHÉLEMY, M.; BARRAT, A.; PASTOR-SATORRAS, R.; VESPIGNANI, A. Velocity and hierarchical spread of epidemic outbreaks in scale-free networks. **Physical review letters**, APS, v. 92, n. 17, p. 178701, 2004. Citado na página 35.

BELÉNDEZ, A.; PASCUAL, C.; MÉNDEZ, D.; BELÉNDEZ, T.; NEIPP, C. Exact solution for the nonlinear pendulum. **Revista brasileira de ensino de física**, SciELO Brasil, v. 29, p. 645–648, 2007. Citado na página 43.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, p. 5–32, 2001. Citado na página 65.

____. Statistical modeling: The two cultures. **Quality control and applied statistics**, Executive Sciences Institute, v. 48, n. 1, p. 81–82, 2003. Citado na página 88.

BRUNTON, S. L.; PROCTOR, J. L.; KUTZ, J. N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. **Proceedings of the national academy of sciences**, National Acad Sciences, v. 113, n. 15, p. 3932–3937, 2016. Citado nas páginas 27 e 28.

____. Sparse identification of nonlinear dynamics with control (sindyc). **IFAC-PapersOnLine**, Elsevier, v. 49, n. 18, p. 710–715, 2016. Citado na página 28.

CAMPS-VALLS, G.; GERHARDUS, A.; NINAD, U.; VARANDO, G.; MARTIUS, G.; BALAGUER-BALLESTER, E.; VINUESA, R.; DIAZ, E.; ZANNA, L.; RUNGE, J. Discovering causal relations and equations from data. **arXiv preprint arXiv:2305.13341**, 2023. Citado nas páginas 25 e 78.

CASTILLO, F. A.; VILLA, C. M. Symbolic regression in multicollinearity problems. In: **ACM Conferences**. New York, NY, USA: Association for Computing Machinery, 2005. p. 2207–2208. Citado na página 74.

CAVA, W. L.; ORZECHOWSKI, P.; BURLACU, B.; FRANCA, F. de; VIRGOLIN, M.; JIN, Y.; KOMMENDA, M.; MOORE, J. Contemporary Symbolic Regression Methods and their Relative Performance. **Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks**, v. 1, dez. 2021. Disponível em: <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/c0c7c76d30bd3dcaefc96f40275bdc0a-Abstract-round1.html>. Citado nas páginas 29, 48, 49, 66, 72 e 74.

CHAMPION, K.; ZHENG, P.; ARAVKIN, A. Y.; BRUNTON, S. L.; KUTZ, J. N. A unified sparse optimization framework to learn parsimonious physics-informed models from data. **IEEE Access**, IEEE, v. 8, p. 169259–169271, 2020. Citado na página 28.

CHAN, J. Y.-L.; LEOW, S. M. H.; BEA, K. T.; CHENG, W. K.; PHOONG, S. W.; HONG, Z.-W.; CHEN, Y.-L. Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. **Mathematics**, Multidisciplinary Digital Publishing Institute, v. 10, n. 8, p. 1283, abr. 2022. ISSN 2227-7390. Citado na página 74.

CHEN, Y.; ANGULO, M. T.; LIU, Y.-Y. Revealing complex ecological dynamics via symbolic regression. **BioEssays**, Wiley Online Library, v. 41, n. 12, p. 1900069, 2019. Citado na página 28.

CHUNG, F.; LU, L.; VU, V. The spectra of random graphs with given expected degrees, internet math. 2004. Citado nas páginas 29 e 77.

CONOVER, M.; RATKIEWICZ, J.; FRANCISCO, M.; GONÇALVES, B.; MENCZER, F.; FLAMMINI, A. Political polarization on twitter. In: **Proceedings of the international aaai conference on web and social media**. [S.l.: s.n.], 2011. v. 5, n. 1, p. 89–96. Citado na página 35.

CONOVER, W. J. **Practical Nonparametric Statistics, 3rd Edition**. Hoboken, NJ, USA: Wiley, 1999. 350 p. ISBN 978-0-471-16068-7. Disponível em: <https://www.wiley.com/en-us/Practical+Nonparametric+Statistics%2C+3rd+Edition-p-9780471160687>. Citado na página 65.

COSTA, L. M. C. da; MERCHAN-HAMANN, E. Pandemias de influenza e a estrutura sanitária brasileira: breve histórico e caracterização dos cenários. **Revista Pan-Amazônica de Saúde**, v. 7, n. 1, p. 15–15, 2016. Citado nas páginas 44 e 76.

CRANMER, M. Interpretable machine learning for science with pysr and symbolicregression. jl. **arXiv preprint arXiv:2305.01582**, 2023. Citado nas páginas 28, 29, 48, 54, 71, 78 e 88.

DARWIN, C. **The Origin of Species by Means of Natural Selection. Popular impression of the corrected copyright edition**. [S.l.]: John Murray, London, 1910. Citado na página 26.

D'ASCOLI, S.; BECKER, S.; MATHIS, A.; SCHWALLER, P.; KILBERTUS, N. ODEFormer: Symbolic Regression of Dynamical Systems with Transformers. **arXiv**, out. 2023. Citado nas páginas 56, 71, 73 e 78.

DIVEEV, A.; SHMALKO, E. **Machine Learning Control by Symbolic Regression**. [S.l.]: Springer, 2021. Citado na página 26.

DIZ-PITA, É.; OTERO-ESPINAR, M. V. Predator–Prey Models: A Review of Some Recent Advances. **Mathematics**, Multidisciplinary Digital Publishing Institute, v. 9, n. 15, p. 1783, jul. 2021. ISSN 2227-7390. Citado na página 43.

DOROGOVTSEV, S. N.; MENDES, J. F. **The Nature of Complex Networks**. [S.l.]: Oxford University Press, 2022. Citado na página 40.

DUBČÁKOVÁ, R. **Eureqa: software review**. [S.l.]: Springer, 2011. Citado na página 27.

ERDŐS, P.; RÉNYI, A. *et al.* On the evolution of random graphs. **Publ. math. inst. hung. acad. sci**, v. 5, n. 1, p. 17–60, 1960. Citado na página 75.

FALOUTSOS, M.; FALOUTSOS, P.; FALOUTSOS, C. On power-law relationships of the internet topology. **ACM SIGCOMM computer communication review**, ACM New York, NY, USA, v. 29, n. 4, p. 251–262, 1999. Citado na página 35.

Feynman. **The Feynman Lectures on Physics**. 2024. [Online; accessed 7. Nov. 2024]. Disponível em: <https://www.feynmanlectures.caltech.edu>. Citado na página 27.

FOSTER, D. V.; FOSTER, J. G.; GRASSBERGER, P.; PACZUSKI, M. Clustering drives assortativity and community structure in ensembles of networks. **Physical Review E**, APS, v. 84, n. 6, p. 066117, 2011. Citado na página 36.

GAEL, J. V.; SAATCI, Y.; TEH, Y. W.; GHAHRAMANI, Z. Beam sampling for the infinite hidden Markov model. In: **ACM Other conferences**. New York, NY, USA: Association for Computing Machinery, 2008. p. 1088–1095. Citado na página 56.

GILBERT, E. N. Random graphs. **The Annals of Mathematical Statistics**, JSTOR, v. 30, n. 4, p. 1141–1144, 1959. Citado nas páginas 40 e 75.

GOLBERG, D. E. Genetic algorithms in search, optimization, and machine learning. **Addion wesley**, v. 1989, n. 102, p. 36, 1989. Citado nas páginas 26 e 78.

GÓMEZ, S.; ARENAS, A.; BORGE-HOLTHOEFER, J.; MELONI, S.; MORENO, Y. Discrete-time markov chain approach to contact-based disease spreading in complex networks. **Europhysics Letters**, IOP Publishing, v. 89, n. 3, p. 38009, 2010. Citado nas páginas 29, 48, 75 e 77.

GUDETTI, J. P.; YAZDI, S. J. M.; BAQERSAD, J.; PETERS, D.; GHAMARI, M. **Data-Driven Modeling of Linear and Nonlinear Dynamic Systems for Noise and Vibration Applications**. [S.l.], 2023. Citado na página 28.

GUSTAFSON, S.; BURKE, E. K.; KRASNOGOR, N. On improving genetic programming for symbolic regression. v. 1, p. 912–919, 2005. Citado na página 26.

HAGBERG, A.; CONWAY, D. Networkx: Network analysis with python. **URL: https://networkx. github. io**, 2020. Citado na página 61.

HAKEN, H.; SAUERMANN, H. Nonlinear interaction of laser modes. **Z. Phys.**, Springer-Verlag, v. 173, n. 3, p. 261–275, jun. 1963. Citado na página 43.

HOERL, A. E.; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. **Technometrics**, Taylor & Francis, v. 12, n. 1, p. 55–67, 1970. Citado na página 88.

HU, H.; YUAN, X.; HUANG, L.; HUANG, C. Global dynamics of an sirs model with demographics and transfer from infectious to susceptible on heterogeneous networks. **Math. Biosci. Eng**, v. 16, n. 5, p. 5729–5749, 2019. Citado na página 47.

HU, J.; CUI, J.; YANG, B. Learning interpretable network dynamics via universal neural symbolic regression. **Nature Communications**, Nature Publishing Group UK London, v. 16, n. 1, p. 6226, 2025. Citado na página 29.

KAPTANOGLU, A. A.; SILVA, B. M. de; FASEL, U.; KAHEMAN, K.; GOLDSCHMIDT, A. J.; CALLAHAM, J. L.; DELAHUNT, C. B.; NICOLAOU, Z. G.; CHAMPION, K.; LOISEAU, J.-C. *et al.* Pysindy: A comprehensive python package for robust sparse system identification. **arXiv preprint arXiv:2111.08481**, 2021. Citado nas páginas 28, 52, 53 e 78.

KAPTANOGLU, A. A.; SILVA, B. M. de; FASEL, U.; KAHEMAN, K.; GOLDSCHMIDT, A. J.; CALLAHAM, J.; DELAHUNT, C. B.; NICOLAOU, Z. G.; CHAMPION, K.; LOISEAU, J.-C.; KUTZ, J. N.; BRUNTON, S. L. Pysindy: A comprehensive python package for robust sparse system identification. **Journal of Open Source Software**, The Open Journal, v. 7, n. 69, p. 3994, 2022. Disponível em: <https://doi.org/10.21105/joss.03994>. Citado na página 71.

KEELING, M. J.; ROHANI, P. Introduction to simple epidemic models. In: **Modeling infectious diseases in humans and animals**. [S.l.]: Princeton University Press, 2011. p. 15–53. Citado na página 45.

KERMACK, W. O.; MCKENDRICK, A. G. A contribution to the mathematical theory of epidemics. **Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character**, The Royal Society London, v. 115, n. 772, p. 700–721, 1927. Citado nas páginas 45 e 77.

KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. **arXiv preprint arXiv:1412.6980**, 2014. Citado na página 88.

KISS, I. Z.; MILLER, J. C.; SIMON, P. L. Mean-field approximations for heterogeneous networks. In: **Mathematics of Epidemics on Networks: From Exact to Approximate Models**. [S.l.]: Springer, 2017. p. 165–205. Citado na página 48.

KISS, I. Z.; MILLER, J. S.; SIMON, P. **Mathematics of Epidemics on Networks: From Exact to Approximate Models**. [S.l.]: Springer International Publishing, 2017. Citado na página 61.

KIYANI, E.; SHUKLA, K.; KARNIADAKIS, G. E.; KARTTUNEN, M. A framework based on symbolic regression coupled with extended physics-informed neural networks for gray-box learning of equations of motion from data. **arXiv preprint arXiv:2305.10706**, 2023. Citado na página 28.

KOROLEV, I. Identification and estimation of the seird epidemic model for covid-19. **Journal of econometrics**, Elsevier, v. 220, n. 1, p. 63–85, 2021. Citado na página 47.

KORTE, C.; MILGRAM, S. Acquaintance networks between racial groups: Application of the small world method. **Journal of personality and social psychology**, American Psychological Association, v. 15, n. 2, p. 101, 1970. Citado na página 40.

KOZA, J. R. Genetic programming, on the programming of computers by means of natural selection. a bradford book. **MIT Press**, 1992. Citado nas páginas 26, 78 e 88.

LANDAJUELA, M.; LEE, C. S.; YANG, J.; GLATT, R.; SANTIAGO, C. P.; ARAVENA, I.; MUNDHENK, T.; MULCAHY, G.; PETERSEN, B. K. A unified framework for deep symbolic regression. **Advances in Neural Information Processing Systems**, v. 35, p. 33985–33998, 2022. Citado na página 29.

LANGLEY, P. Bacon: A production system that discovers empirical laws. In: CITESEER. **IJCAI**. [S.l.], 1977. p. 344. Citado nas páginas 26, 49 e 78.

LITTMAN, R. J. The plague of athens: epidemiology and paleopathology. **Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine: A Journal of Translational and Personalized Medicine**, Wiley Online Library, v. 76, n. 5, p. 456–467, 2009. Citado nas páginas 44 e 76.

LIU, Z.; MA, P.; WANG, Y.; MATUSIK, W.; TEGMARK, M. Kan 2.0: Kolmogorov-arnold networks meet science. **arXiv preprint arXiv:2408.10205**, 2024. Citado nas páginas 28, 55 e 88.

LIU, Z.; WANG, Y.; VAIDYA, S.; RUEHLE, F.; HALVERSON, J.; SOLJAČIĆ, M.; HOU, T. Y.; TEGMARK, M. Kan: Kolmogorov-arnold networks. **arXiv preprint arXiv:2404.19756**, 2024. Citado nas páginas 28, 55, 78 e 88.

LORENZ, E. N. Deterministic Nonperiodic Flow. **J. Atmos. Sci.**, American Meteorological Society, v. 20, n. 2, p. 130–141, mar. 1963. ISSN 0022-4928. Citado na página 43.

LUO, J.; YU, C. L. The application of symbolic regression on identifying implied volatility surface. **Mathematics**, MDPI, v. 11, n. 9, p. 2108, 2023. Citado na página 28.

MANGAN, N. M.; BRUNTON, S. L.; PROCTOR, J. L.; KUTZ, J. N. Inferring biological networks by sparse identification of nonlinear dynamics. **IEEE Transactions on Molecular, Biological and Multi-Scale Communications**, IEEE, v. 2, n. 1, p. 52–63, 2016. Citado na página 27.

MEYBORG, M. **The role of German universities in a system of joint knowledge generation and innovation. A social network analysis of publications and patents with a focus on the spatial dimension**. [S.l.]: KIT Scientific Publishing, 2014. Citado na página 38.

MILGRAM, S. The small world problem. **Psychology today**, New York, v. 2, n. 1, p. 60–67, 1967. Citado na página 40.

MILLER, B. N.; RANUM, D. L. **Resolução de Problemas com Algoritmos e Estruturas de Dados usando Python**. 2013. Disponível em: urlhttps://panda.ime.usp.br/panda/static/pythonds$_{pt}$/07 – $Grafos/StronglyConnectedComponents.html. Acessoem: 22demarode2023. Citadonapgina38$.

MILLER, J. C.; TING, T. Eon (epidemics on networks): a fast, flexible python package for simulation, analytic approximation, and analysis of epidemics on networks. **arXiv preprint arXiv:2001.02436**, 2020. Citado na página 61.

MISLOVE, A.; MARCON, M.; GUMMADI, K. P.; DRUSCHEL, P.; BHATTACHARJEE, B. Measurement and analysis of online social networks. In: **Proceedings of the 7th ACM SIGCOMM conference on Internet measurement**. [S.l.: s.n.], 2007. p. 29–42. Citado na página 39.

MIYAZAKI, M.; ISHIKAWA, K.-I.; NAKASHIMA, K.; SHIMIZU, H.; TAKAHASHI, T.; TAKAHASHI, N. Application of the symbolic regression program ai-feynman to psychology. **Frontiers in Artificial Intelligence**, Frontiers, v. 6, p. 1039438, 2023. Citado nas páginas 28 e 78.

MONTENEGRO, E. N.; BATISTA, E.; STROPPA, P. H. As epidemias e pandemias virais na história da humanidade-uma análise sistemática e biológica. **Rev Bras Cien Med Saúde**, p. 00–00, 2021. Citado nas páginas 44 e 76.

MURRAY, J. D. **Mathematical biology II: Spatial models and biomedical applications**. [S.l.]: Springer New York, 2001. v. 3. Citado nas páginas 44 e 76.

NEWMAN, M. E. The structure and function of complex networks. **SIAM review**, SIAM, v. 45, n. 2, p. 167–256, 2003. Citado nas páginas 33, 37 e 75.

_____. Power laws, pareto distributions and zipf's law. **Contemporary physics**, Taylor & Francis, v. 46, n. 5, p. 323–351, 2005. Citado na página 35.

OKE, M.; OGUNMILORO, O.; AKINWUMI, C.; RAJI, R. Mathematical modeling and stability analysis of a sirv epidemic model with non-linear force of infection and treatment. **Communications in Mathematics and Applications**, RGN Publications, v. 10, n. 4, p. 717, 2019. Citado na página 47.

PAIVA, G. M.; PIMENTEL, S. P.; ALVARENGA, B. P. de; MARRA, E. G. Regressão simbólica aplicada na previsão de irradiância solar intra-diária na cidade de goiânia (brasil). In: **Anais Congresso Brasileiro de Energia Solar-CBENS**. [S.l.: s.n.], 2018. Citado na página 78.

PAPKE, L. E.; WOOLDRIDGE, J. M. Econometric methods for fractional response variables with an application to 401 (k) plan participation rates. **Journal of applied econometrics**, Wiley Online Library, v. 11, n. 6, p. 619–632, 1996. Citado na página 83.

PASTOR-SATORRAS, R.; CASTELLANO, C.; MIEGHEM, P. V.; VESPIGNANI, A. Epidemic processes in complex networks. **Reviews of modern physics**, APS, v. 87, n. 3, p. 925, 2015. Citado na página 40.

PASTOR-SATORRAS, R.; VESPIGNANI, A. Epidemic spreading in scale-free networks. **Physical review letters**, APS, v. 86, n. 14, p. 3200, 2001. Citado na página 75.

QUADE, M.; ABEL, M.; KUTZ, J. N.; BRUNTON, S. L. Sparse identification of nonlinear dynamics for rapid model recovery. **Chaos: An Interdisciplinary Journal of Nonlinear Science**, AIP Publishing LLC, v. 28, n. 6, p. 063116, 2018. Citado na página 52.

QUADE, M.; ABEL, M.; SHAFI, K.; NIVEN, R. K.; NOACK, B. R. Prediction of dynamical systems by symbolic regression. **Physical Review E**, APS, v. 94, n. 1, p. 012214, 2016. Citado na página 78.

RAGHAV, S. S.; KUMAR, S. T.; BALAJI, R.; SANJAY, M.; SHUNMUGA, C. Interactive Symbolic Regression - A Study on Noise Sensitivity and Extrapolation Accuracy. In: **ACM Conferences**. New York, NY, USA: Association for Computing Machinery, 2024. p. 2076–2082. Citado na página 71.

RODRIGUES, F. A. Network centrality: an introduction. **A mathematical modeling approach from nonlinear dynamics to complex systems**, Springer, p. 177–196, 2019. Citado nas páginas 38 e 39.

ROTHLAUF, F. Representations for genetic and evolutionary algorithms. In: **Representations for Genetic and Evolutionary Algorithms**. [S.l.]: Springer, 2006. p. 9–32. Citado na página 49.

RUDY, S. H.; BRUNTON, S. L.; PROCTOR, J. L.; KUTZ, J. N. Data-driven discovery of partial differential equations. **Science advances**, American Association for the Advancement of Science, v. 3, n. 4, p. e1602614, 2017. Citado na página 27.

SAHOO, S.; LAMPERT, C.; MARTIUS, G. Learning equations for extrapolation and control. In: PMLR. **International Conference on Machine Learning**. [S.l.], 2018. p. 4442–4450. Citado na página 27.

SCHAEFFER, H.; MCCALLA, S. G. Sparse model selection via integral terms. **Physical Review E**, APS, v. 96, n. 2, p. 023302, 2017. Citado na página 61.

SCHMIDT, M.; LIPSON, H. Distilling free-form natural laws from experimental data. **science**, American Association for the Advancement of Science, v. 324, n. 5923, p. 81–85, 2009. Citado na página 26.

____. Symbolic regression of implicit equations. In: **Genetic programming theory and practice VII**. [S.l.]: Springer, 2009. p. 73–85. Citado na página 48.

____. Eureqa (version 0.98 beta). **Nutonian Inc., Boston MA**, 2014. Citado na página 29.

SILVA, B. M. D.; CHAMPION, K.; QUADE, M.; LOISEAU, J.-C.; KUTZ, J. N.; BRUNTON, S. L. Pysindy: a python package for the sparse identification of nonlinear dynamics from data. **arXiv preprint arXiv:2004.08424**, 2020. Citado nas páginas 28, 52, 78 e 88.

SIVANANDAM, S.; DEEPA, S. **Genetic algorithms**. [S.l.]: Springer, 2008. 15–37 p. Citado nas páginas 26 e 49.

STANISLAWSKA, K.; KRAWIEC, K.; KUNDZEWICZ, Z. W. Modeling global temperature changes with genetic programming. **Computers & Mathematics with Applications**, Elsevier, v. 64, n. 12, p. 3717–3728, 2012. Citado na página 28.

STEPHENS, T. **Genetic Programming in Python, with a scikit-learn inspired API: gplearn**. 2016. Citado nas páginas 27, 78 e 88.

THING, M. E.; KOKSBANG, S. M. cp3-bench: A tool for benchmarking symbolic regression algorithms tested with cosmology. **arXiv preprint arXiv:2406.15531**, 2024. Citado na página 29.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society Series B: Statistical Methodology**, Oxford University Press, v. 58, n. 1, p. 267–288, 1996. Citado na página 88.

UDRESCU, S.-M.; TEGMARK, M. Ai feynman: A physics-inspired method for symbolic regression. **Science Advances**, American Association for the Advancement of Science, v. 6, n. 16, p. eaay2631, 2020. Citado nas páginas 27, 29, 50, 78 e 88.

UGANDER, J.; KARRER, B.; BACKSTROM, L.; MARLOW, C. The anatomy of the facebook social graph. **arXiv preprint arXiv:1111.4503**, 2011. Citado na página 39.

VANNESCHI, L.; CASTELLI, M.; SILVA, S. Measuring bloat, overfitting and functional complexity in genetic programming. In: **Proceedings of the 12th annual conference on Genetic and evolutionary computation**. [S.l.: s.n.], 2010. p. 877–884. Citado na página 69.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L.; POLOSUKHIN, I. Attention Is All You Need. **arXiv**, jun. 2017. Citado na página 56.

VESPIGNANI, A. Modelling dynamical processes in complex socio-technical systems. **Nature physics**, Nature Publishing Group UK London, v. 8, n. 1, p. 32–39, 2012. Citado na página 33.

VIRGOLIN, M.; PISSIS, S. P. Symbolic regression is NP-hard. **Transactions on Machine Learning Research**, 2022. ISSN 2835-8856. Disponível em: <https://openreview.net/forum?id=LTiaPxqe2e>. Citado na página 74.

WANG, Y.; WAGNER, N.; RONDINELLI, J. M. Symbolic regression in materials science. **MRS Communications**, Cambridge University Press, v. 9, n. 3, p. 793–805, 2019. Citado na página 28.

WANGERSKY, P. J. Lotka-volterra population models. **Annual Review of Ecology and Systematics**, Annual Reviews, v. 9, p. 189–218, 1978. ISSN 00664162. Disponível em: <http://www.jstor.org/stable/2096748>. Citado na página 43.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world'networks. **nature**, Nature Publishing Group, v. 393, n. 6684, p. 440–442, 1998. Citado nas páginas 17, 40 e 41.

WONG, K. W.; CRANMER, M. Automated discovery of interpretable gravitational-wave population models. **arXiv preprint arXiv:2207.12409**, 2022. Citado na página 28.

ZACHARY, W. W. An information flow model for conflict and fission in small groups. **Journal of anthropological research**, University of New Mexico, v. 33, n. 4, p. 452–473, 1977. Citado na página 39.

ZHENG, P.; ASKHAM, T.; BRUNTON, S. L.; KUTZ, J. N.; ARAVKIN, A. Y. A unified framework for sparse relaxed regularized regression: Sr3. **IEEE Access**, IEEE, v. 7, p. 1404–1423, 2018. Citado na página 27.

APÊNDICE

A

APÊNDICE A

A.1 Parâmetros utilizados para gerar dados sintéticos

As tabelas 11 e 12, mencionadas no Capítulo 3 na Seção 3.1, que descrevem os parâmetros, condições iniciais e intervalos de tempo utilizados para gerar dados sintéticos. A função SOLVE_IVP da biblioteca `SciPy` foi utilizada para integrar numericamente estas equações ao longo do tempo. A seleção destes parâmetros foi feita caso a caso, com a intenção de representar um cenário realista para cada sistema.

Quanto às configurações do algoritmo SR, Tabela 13 e Tabela 14 listam os parâmetros empregados para os sistemas pendular não linear, Lotka-Volterra e Lorenz, e os modelos epidemiológicos, nessa ordem. Quaisquer outros parâmetros disponíveis que não foram mencionados foram mantidos como padrão.

Tabela 11 – Parâmetros utilizados para resolver o atrator de Lorenz, o pêndulo não linear e a dinâmica predador-presa (Lotka-Volterra).

Parameters	Nonlinear pendulum	Lotka-Volterra	Lorenz
Condições iniciais	$\omega = 0$ $\theta = 45$	$u = 20$ $v = 5$	$x_0 = 0.6$ $y_0 = 2.0$ $z_0 = 1.0$
Coefficients	g = 9.8 l = 1.0	$\alpha = 2.0$ $\beta = 0.5$ $\gamma = 1.0$ $\delta = 0.375$	$\sigma = 2.0$ $\rho = 1.0$ $\beta = 2.6$
Janela de simulação	[0,5]	[0,7.5]	[0,5]
Tamanho do intervalo de tempo	2E-3	1E-1	2E-3

Tabela 12 – Parâmetros utilizados para gerar dados nos modelos epidêmicos compartimentais escolhidos (Seção 3.1). Os tamanhos dos compartimentos são listados como frações do número total de indivíduos.

Parâmetros	SIS	SIR	SIRV	SIRS	SEIR	SEIRD
Tamanho inicial do compartimento	$S_0 = 0.99$ $I_0 = 0.01$	$S_0 = 0.99$ $I_0 = 0.01$ $R_0 = 0$	$S_0 = 0.94$ $I_0 = 0.01$ $R_0 = 0$ $V_0 = 0.05$	$S_0 = 0.99$ $I_0 = 0.01$ $R_0 = 0$	$S_0 = 0.8$ $E_0 = 0.1$ $I_0 = 0.1$ $R_0 = 0$	$S_0 = 0.99$ $E_0 = 0$ $I_0 = 0.01$ $R_0 = 0$ $D_0 = 0$
Coeficientes	$\beta = 0.3$ $\gamma = 0.1$	$\beta = 0.5$ $\gamma = 0.1$	$\beta = 0.5$ $\gamma = 0.1$ $\varepsilon = 0.5$	$\beta = 0.5$ $\gamma = 0.1$ $\delta = 0.2$	$\beta = 0.5$ $\sigma = 0.5$ $\gamma = 0.1$	$\beta = 0.5$ $\sigma = 0.2$ $\gamma = 0.1$ $\mu = 0.1$
Janela de simulação	[0, 100]	[0,75]	[0, 35]	[0,60]	[0,80]	[0,120]
Tamanho do intervalo de tempo	1E-1	1E-1	1E-1	1E-1	1E-1	1E-1

Tabela 13 – Parâmetros empregados pelos modelos de regressão para obter uma expressão simbólica para um determinado sistema, que são os mesmos definidos na Seção 4.

SR modelo	Parâmetros	**Sistema dinâmico**		
		Lorenz	Pêndulo não linear	Lotka-Volterra
GPLearn	population_size	5000	5000	5000
	generations	50	100	50
	tournament_size	50	100	50
	stopping_criteria	0.01	0.01	0.01
	p_crossover	0.7	0.7	0.6
	p_subtree_mutation	0.2	0.2	0.2
	p_hoist_mutation	0.01	0.01	0.01
	p_point_mutation	0.09	0.09	0.09
	init_depth	$(2,6)$	$(2,2)$	$(8,9)$
	parsimony_coefficient	0.001	0.009	0.001
	function_sett	–	$[\times,+,-,\div,\sin]$	$[\times,+,-]$
	random_state	0	0	0
AI-Feynman	BF_try_time	[30,60,30]	60s	60s
	BF_ops_file_type	7ops	14ops	7ops
	polyfit_deg	2	1	2
	NN_epochs	[40,40,1000]	600	[100,300]
PySINDy	library_functions	Custom [1]	Fourier(n_frequencies = 1) Polynomial(degree=1)	Polynomial (degree=3)
	feature_library	x,y,z	θ,ω	u,v
	differentiation_method		FiniteDifference	FiniteDifference (order=2)
	Optimizer *optmizer*	*STLSQ*	*SR*3	*SR*3
	Optimizer *threshold*	0.2	0.4	0.6
	Optimizer *alpha*	$1E-4$	–	$1E-4$
	Optimizer *normalize_columns*	*Fals*	–	*True*
	Optimizer *thresholder*	–	$'l1'$	–
	Optimizer ν	–	–	1
	Optimizer *tol*	–	–	$1E-6$
PySR	Population	[30, 30, 30]	[1, 30]	
	N interation	[30,30,30]	[30,30]	[10,50,800]
	Binary-Operator	[[-, +, *],[-, +, *],[-, +, *]]	[[*, +],[+, *]]	[[*, +],[-, *]]
	Unary-Operator	[[ ],[ ],[ ]]	[[ ],["sin"]]	[[ ],[ ]]
	Nested_constraints	no	no	no
PyKAN	Topologia de rede	x : *kanpiler* y : *kanpiler* z : *kanpiler*	$\theta : [1,1]$ $\omega : [1,1]$	U : *kanpiler* V : *kanpiler*
	seed	0	12	0
	λ	1E-25	1E-3	1E-35
	steps	50	30	60
	λ_coef	1E-19	1e-15	1E-2

Tabela 14 – Parâmetros empregados por cada modelo de regressão e conjunto de dados, em que cada valor se refere ao parâmetro usado para um determinado compartimento do modelo epidêmico da análise da Seção 4.

SR model	Parameters	**Modelos epidemiológicos**					
		SIR	SIS	SEIR	SEIRD	SIRV	SIRS
GPLearn	population_size	10000	5000	[1E4, 5E4, 1E4, 1E3]	[1E4, 6E4, 6E4, 2E3, 2E3]	[2E3, 6.5E2, 2E3, 2E3]	[1E4, 8E4, 1E4]
	generations	100	20	[2E3, 5E3, 2E3, 2E3]	[1E2, 6E3, 6E3, 3.5E2, 3.5E2]	[3.5E2, 5E2, 3.5E2,3.5E2]	[5E2,1E3,6E2]
	tournament_size	100	20	[2E3, 5E3, 2E3, 2E3]	[1E2, 6E3, -, 3.5E2, 3.5E2]	[3.5E2, 1.5E2, 3.5E2,3.5E2]	[5E2, 1E3, 5E2]
	stopping_criteria	0.01	0.01	[1E-2, 1E-3, 1E-4, 1E-2]	[1E-3, 1E-2, 1E-2, 1E-3, 1E-3]	[1E-2, 1E-2, 1E-2, 1E-3]	1E-2
	p_crossover	0.7	0.6	0.7	[0.7,0.6,0.7,0.7,0.7]	0.7	0.7
	p_subtree_mutation	0.1	0.2	0.1	[0.1,0.2,0.2,0.1,0.1]	0.1	0.1
	p_hoist_mutation	0.05	0.01	[0.05, 0.05, 0.1, 0.1]	[0.1,0.01,0.01,0.1,0.1]	0.1	0.1
	p_point_mutation	0.1	0.08	0.1	[0.1,0.09,0.09,0.1,0.1]	0.1	0.1
	init_depth	(3,6)	(2,3)	[(2,5),(3,3),(2,3),(0,5)]	[(2,2),(3,4),(2,3),(1,1),(1,1)]	[(2,4),(3,3),(2,4)]	[(3,4),(3,3),(3,4)]
	parsimony_coefficient	0.7	0.01	[0.05, 0.001, 0.7, 0.05]	[0.05, 0.01, 0.01, 0.01, 0.01]	0.01	0.01
	function_sett	$[\times,+]$	$[\times,+,-,\div]$	[None, $[\times,+]$, $[\times,+]$], $[\times,+]$]	[[$\times,-$], [$\times,-$], [$\times,-$], [$\times,+,-$], [$\times,+,-$]]	[[$\times,-$], [$\times,-$], [$\times,-$], [$\times,+$]]	[[$\times,-,\neg$], [$\times,-$], [$\times,-,\neg$]]
	random_state	0	0	0	0	0	0
AI-Feynman	BF_try_time	[60s, 60s, 60s]	60s	[240, 300, 3600, 60]	[3600, 60, 60, 60, 30]	[300,300,3600,60]	[3600,60,600]
	BF_ops_file_type	7ops	7ops	7ops	7ops	7ops	7ops
	polyfit_deg	2	2	2	[2,2,2,2,1]	2	2
	NN_epochs	[400,400,4000]	400	[600,4000,600,600]	[1000,400,200,600, 600]	[4000,4000,600,1000]	[600,600,600]
PySINDy	Function	$x+y$,$x*y$	$x+y$,$x*y$	x,$x*y$	x,$x*y$	x,$x*y$	x,$x*y$
	Optimizer	STLSQ	STLSQ	STLSQ	OMP	OMP	STLSQ
	Threshold	0.6	0.6	1E-4	no	1E-4	1E-3
	α	1E-4	1E-4	1E-3	no	1E-3	1E-4
	n_nonzero_coefs	no	no	no	2	2	no
PySR	Population	[50,50,50]	[50,50]	[50,50,50, 50]	[50,50,50,50,50]	[50,50,50,50]	[50, 50, 50]
	model_selection	"best"	"best"	"best"	"best"	"best"	"best"
	Paralelism	"serial"	"serial"	"serial"	"serial"	"serial"	"serial"
	Random_state	42	42	42	42	42	42
	deterministic	"True"	"True"	"True"	"True"	"True"	"True"
	Maxsize	10	10	10	10	10	10
	ninteration	[1E3, 1E3,1E3]	[1E3,1E3]	[1E3,1E3,1E3,1E3]	[1E3,1E3,1E3,1E3,1E3]	[1E3,1E3,1E3,1E3]	[1E3,1E3,1E3]
	Binary-Operator	[+, *, -]	[+, *, -]	[+, *, -]	[+, *, -]	[[+, *, -]	[+, *, -]
	Unary-Operator	[[ ],[ ],[ ]]	[[ ],[ ]]	[[ ],[ ],[ ],[ ]]	[[ ],[ ],[ ],[ ]]	[[ ],[ ],[ ],[ ]]	
PyKAN	Network topology	[kanpiler, kanpiler, [1,1]]	[kanpiler, kanpiler]	[kanpiler, kanpiler, kanpiler, [1,1]]	[kanpiler, kanpiler, kanpiler, [1,1], [1,1]]	[kanpiler, kanpiler, [1,1], [1,1]]	[kanpiler, kanpiler, [2,1]]
	seed	0	12		0	12	12
	λ	[1E-15, 1E-15, 1E-4]	[1E-45,1E-25]	[1.5E-5, 1, 1E-15, 1E-15]	[1, 1, 1E-7, 1E-15, 1E-15]	[1E-3, 1E-3, 1E-15, 1E-15]	[100, 1E-10, 1E-15]
	steps	[50, 50,50]	[60,60]	[50,50,50,50]	[70,70,5,50,50]	[100,20,50,50]	[500,100,50]
	λ_coef	[1E-5,1E-5,1E-3]	[1E-2,1E-12]	[150, 80, 1, 1]	[1E-2, 1E-2, 1E-15, 1, 1]	[1E-15, 1E-15, 1,1]	[35,1,1]

A.2 Parâmetros utilizados pelos modelos de regressão simbólica para gerar dados sintéticos em diferentes meios de propagação.

Tabela 15 – Parâmetros empregados por cada modelo de regressão e conjunto de dados, onde cada valor se refere ao parâmetro utilizado para um determinado compartimento do modelo epidêmico na rede ER, seguindo a ordem da sigla. A tabela destaca os sistemas identificados pelo modelo SR especificado em cada linha. O tempo é medido em segundos.

SR	Parameters	Sistema dinâmico					
		SIR	**SIS**	**SEIR**	**SEIRD**	**SIRV**	**SIRS**
GPLearn	population_size	650	1E2	[1E3, 1E3, 5E3, 5E3]	[1E3, 1E3, 1E3, 1E3, 1E3]	[1E3, 1E3, 1E3, 1E3]	950
	generations	150	350	[5E2, 5E2, 1E2, 1E2]	[2E2, 4E2, 4E2, 1E2, 2E2]	[5E1, 5E2, 5E1,3.5E1]	350
	tournament_size	150	350	[5E2, 5E2, 1E2, 1E2]	[2E2, 4E2, 4E2, 1E2, 2E2]	[5E1, 5E2, 5E1,3.5E1]	350
	stopping_criteria	1E-2	1E-2	[1E-2, 1E-3, 1E-4, 1E-4]	[1E-2, 1E-2, 1E-3, 1E-4, 1E-4]	[1E-2, 1E-2, 1E-2, 1E-2]	1E-2
	p_crossover	0.7	0.7	[8E-1,0.7, 0.7,0.7]	[0.8,0.7,0.7,0.7,0.7]	0.7	0.7
	p_subtree_mutation	0.1	0.1	[0.05, 0.1, 0.1, 0.1]	[5E-2,5E-2,1E-1,1E-1,1E-1]	0.1	0.1
	p_hoist_mutation	0.1	0.1	[0.01, 0.05, 0.05, 0.05]	[1E-1,1E-1,5E-2,5E-2,5E-2]	0.1	0.1
	p_point_mutation	0.1	0.1	0.1	[5E-2,1E-1,1E-1,1E-1,1E-1]	0.1	0.1
	init_depth	none	none	[none,(2,5),(2,4),(2,4)]	[(),(),(2,4),(2,4),(2,4)]	[(),(),(2,4),(2,4)]	[(3,4),(3,3),(3,4)]
	parsimony_coefficient	0.01	1E-03	[1E-2, 1E-4, 1E-3, 1E-3]	[1E-3, 1E-3, 1E-3, 1E-3, 1E-3]	.01	0.01
	function_sett	$[\times,+]$	$[\times,+]$	$[[\times,+],[],[]],[\times,+]]$	[[], [], [], [], []]	$[[\times,+],[\times,+],[\times,-],[\times,-]]$	$[[\times,+],[\times,+],[\times,+]]$
	random_state	0	0	0	0	0	0
AI-Feynman	BF_try_time	[60,60,30]	[60, 60]	[60, 60, 60, 60]	[60, 60, 120, 60, 30]	[60,60,60,60]	[60,60,60]
	BF_ops_file_type	7ops	7ops	7ops	7ops	7ops	7ops
	polyfit_deg	2	2	2	[1,2,2,2,1]	2	2
	NN_epochs	[600,600,600]	[300,300]	[1000,1000,600,1000]	[600,600,900,600, 600]	[600,600,600,600]	[600,600,600]
PySINDy	Function			$x, x*y$	$x, x*y$	$x, x*y$	$x, x*y$
	Optimizer			STLSQ	STLSQ	STLSQ	STLSQ
	Threshold			1E-4	1E-4	1E-4	1E-3
	α			1E-3	1E-3	1E-3	1E-4
PySR	Population	[32,10,1]	[300,300]	[32,10,20, 1]	[1,10,10,1,1]	[30,10,1,10]	[30, 10, 32]
	N interation	[10,150, 20]	[100,150]	[10,150,150,20]	[100,150,150,20,20]	[10,150,20,150]	[10,150,20]
	Binary-Operator	[[*],[*, -],[*]]	[+, *]	[[*],[*, -],[*, -],[*]]	[[*][*, -],[*, -],[*], [*]]	[[-, *],[*, -],[*],[*]]	[[-, *],[*, -],[-, *]]
	Unary-Operator	[ ]	[ ]	[[ ],[ ],[ ],[ ]]	[[ ],[ ],[ ],[ ]]	[[ ],[ ],[ ],[ ]]	[[ ],[ ],[ ]]
PyKAN	Network topology	[kanpiler, kanpiler, [1,1]]	[kanpiler, kanpiler]	[kanpiler, kanpiler, kanpiler, [1,1]]	[kanpiler, kanpiler, kanpiler, [1,1], [1,1]]	[kanpiler, kanpiler, [1,1], [1,1]]	[kanpiler, kanpiler, [2,1]]
	λ	[1E-15, 1E-15, 1E-4]	[1E-45, 1E-45]	[1E.5-05, 1E-15, 1E-15, 1E-15]	[1.5E-09, 1E-07, 1E-07, 1E-15, 1E-15]	[1E-03, 1E-03, 1E-15]	[1E-2, 1E-10, 1E-15]
	steps	[50,50,50]	[60,60]	[50,70,50,50]	[100,5,5,50,50]	[100,20,50,50]	[500,100,50]
	λ_coef	[1E-05, 1E-05, 1E-03]	[1E-02, 1E-12]	[1, 1E-05, 1, 1]	[1.5E-09, 1E-07, 1E-07, 1E-15, 1E-15]	[1E-15, 1E-15,1,1]	[35,1,1]
ODEFormer	beam_size	50	50	50	50	50	50
	beam_Temperature	0.1	0.1	0.1	0.1	0.1	0.1

Tabela 16 – Parâmetros empregados por cada modelo de regressão e conjunto de dados, onde cada valor se refere ao parâmetro utilizado para um determinado compartimento do modelo epidêmico na rede BA, seguindo a ordem da sigla. A tabela destaca os sistemas identificados pelo modelo SR especificado em cada linha. O tempo é medido em segundos.

SR	Parâmetros	Sistema dinâmico					
		SIR	**SIS**	**SEIR**	**SEIRD**	**SIRV**	**SIRS**
GPLearn	population_size	650	1E2	[1E3, 1E3, 5E3, 5E3]	[1E3, 1E3, 1E3, 1E3, 1E3]	[1E3, 1E3, 1E3, 1E3]	950
	generations	150	350	[5E2, 5E2, 1E2, 1E2]	[2E2, 4E2, 4E2, 1E2, 2E2]	[5E1, 5E2, 5E1,3.5E1]	350
	tournament_size	150	350	[5E2, 5E2, 1E2, 1E2]	[2E2, 4E2, 4E2, 1E2, 2E2]	[5E1, 5E2, 5E1,3.5E1]	350
	stopping_criteria	1E-2	1E-2	[1E-2, 1E-3, 1E-4, 1E-4]	[1E-2, 1E-2, 1E-3, 1E-4, 1E-4]	[1E-2, 1E-2, 1E-2, 1E-2]	1E-2
	p_crossover	0.7	0.7	[8E-1,0.7, 0.7,0.7]	[0.8,0.7,0.7,0.7,0.7]	0.7	0.7
	p_subtree_mutation	0.1	0.1	[0.05, 0.1, 0.1, 0.1]	[5E-2,5E-2,1E-1,1E-1,1E-1]	0.1	0.1
	p_hoist_mutation	0.1	0.1	[0.01, 0.05, 0.05, 0.05]	[1E-1,1E-1,5E-2,5E-2,5E-2]	0.1	0.1
	p_point_mutation	0.1	0.1	0.1	[5E-2,1E-1,1E-1,1E-1,1E-1]	0.1	0.1
	init_depth	none	none	[none,(2,5),(2,4),(2,4)]	[(),(),(2,4),(2,4),(2,4)]	[(),(),(2,4),(2,4)]	[(3,4),(3,3),(3,4)]
	parsimony_coefficient	0.01	1E-03	[1E-2, 1E-4, 1E-3, 1E-3]	[1E-3, 1E-3, 1E-3, 1E-3, 1E-3]	0.01	0.01
	function_sett	$[\times,+]$	$[\times,+]$	$[[\times,+],[],[]],[\times,+]]$	[[], [], [], [], []]	$[[\times,+],[\times,+],[\times,-],[\times,-]]$	$[[\times,+],[\times,+],[\times,+]]$
	random_state	0	0	0	0	0	0
AI-Feynman	BF_try_time	[60,30,30]	[60, 60]	[60, 60, 60, 60]	[60, 60, 120, 60, 30]	[60,60,60,60]	[60,60,60]
	BF_ops_file_type	7ops	7ops	7ops	7ops	7ops	7ops
	polyfit_deg	2	2	[2,2,none,1]	[2,2,2,2,1]	[2,2,2,1]	2
	NN_epochs	[600,600,100]	[600, 600]	[1000,1000,600,1000]	[600,600,600,600, 600]	[600,600,600,600]	[600,600,600]
PySINDy	Function			$x, x*y$	$x, x*y$	$x, x*y$	$x, x*y$
	Optimizer			STLSQ	STLSQ	STLSQ	STLSQ
	Threshold			1E-4	1E-4	1E-4	1E-3
	α			1E-3	1E-3	1E-3	1E-4
PySR	Population	[32,10,1]	[300,300]	[32,10,20, 1]	[1,10,10,1,1]	[30,10,1,10]	[30, 10, 32]
	N interation	[10,150, 20]	[100,150]	[10,150,150,20]	[100,150,150,20,20]	[10,150,20,150]	[10,150,20]
	Binary-Operator	[[*],[*, -],[*]]	[+, *]	[[*],[*, -],[*, -],[*]]	[[*][*, -],[*, -],[*], [*]]	[[-, *],[*, -],[*],[*]]	[[-, *],[*, -],[-, *]]
	Unary-Operator	[ ]	[ ]	[[ ],[ ],[ ],[ ]]	[[ ],[ ],[ ],[ ]]	[[ ],[ ],[ ],[ ]]	[[ ],[ ],[ ]]
PyKAN	Network topology	[kanpiler, kanpiler, [1,1]]	[kanpiler, kanpiler]	[kanpiler, kanpiler, kanpiler, [1,1]]	[kanpiler, kanpiler, kanpiler, [1,1], [1,1]]	[kanpiler, kanpiler, [1,1], [1,1]]	[kanpiler, kanpiler, [2,1]]
	λ	[1E-15, 1E-15, 1E-4]	[1E-45, 1E-45]	[1E.5-05, 1E-15, 1E-15, 1E-15]	[1.5E-09, 1E-07, 1E-07, 1E-15, 1E-15]	[1E-03, 1E-03, 1E-15]	[1E-2, 1E-10, 1E-15]
	steps	[50,50,50]	[60,60]	[50,70,50,50]	[100,5,5,50,50]	[100,20,50,50]	[500,100,50]
	λ_coef	[1E-05, 1E-05, 1E-03]	[1E-02, 1E-12]	[1, 1E-05, 1, 1]	[1.5E-09, 1E-07, 1E-07, 1E-15, 1E-15]	[1E-15, 1E-15,1,1]	[35,1,1]
ODEFormer	beam_size	50	50	50	50	50	50
	beam_Temperature	0.1	0.1	0.1	0.1	0.1	0.1

ANEXO

A

REPOSITÓRIOS DE CÓDIGO E FERRAMENTAS

Beatriz Regina Brum. **SR_in_complex_dynamics**: Página do github com o experimento de simulação de sistemas dinâmicos epidêmicos em meios heterogêneos. GitHub, 2025. Disponível em: [<https://github.com/BeatrizReginaBrum/SR-in-complex-dynamics.git>].

Luiza Lober. **review_symb_regression**: xperimento de simulação de sistemas dinâmicos epidêmicos em meios heterogêneos. GitHub, 2025. Disponível em: <https://github.com/luizalober/review_symb_regression.git>.

Beatriz Regina Brum. **SR_in_networks_ER_BA**: Página do github com o experimento de simulação de sistemas dinâmicos epidêmicos em meios heterogêneos. GitHub, 2025. Disponível em: [<https://github.com/BeatrizReginaBrum/SR_in_networks_ER_BA.git>].


ICMC
USP
SÃO CARLOS