Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Programa de Pós-Graduação em Engenharia Elétrica

Análise Multiescala da Propagação de Epidemias Utilizando Regressão por Processo Gaussiano

Bruno Cardoso Dantas

Campina Grande, Paraíba, Brasil

©Bruno Cardoso Dantas, 7 de março de 2025

Análise Multiescala da Propagação de Epidemias Utilizando Regressão por Processo Gaussiano

Bruno Cardoso Dantas

Tese de Doutorado submetida à Coordenadoria do Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Campina Grande - Campus de Campina Grande como parte dos requisitos necessários para obtenção do grau de Doutor em Ciências no Domínio da Engenharia Elétrica.

Área de Concentração: Processamento da Informação

Orientador: Wamberto José Lira de Queiroz
Orientador: Edmar Candeia Gurjão

Campina Grande, Paraíba, Brasil
Março de 2025

D192a Dantas, Bruno Cardoso.

Análise multiescala da propagação de epidemias utilizando Regressão por Processo Gaussiano / Bruno Cardoso Dantas. – Campina Grande, 2025.

137 f. : il. color.

Tese (Doutorado em Engenharia Elétrica) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2025.

"Orientação: Prof. Dr. Wamberto José Lira de Queiroz, Prof. Dr. Edmar Candeia Gurjão".

Referências.

1. Análise Multiescala. 2. Modelagem Computacional Aplicada à Dinâmica de Doenças Infectocontagiosas. 3. Regressão por Processo Gaussiano – COVID-19. 4. Modelos Epidemiológicos Orientados a Dados. I. Queiroz, Wamberto José Lira de. II. Gurjão, Edmar Candeia. III. Título.

CDU 519.254:616.9(043)

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECÁRIA MEIRE EMANUELA DA SILVA MELO CRB-15/568

Análise Multiescala da Propagação de Epidemias Utilizando Regressão por Processo Gaussiano

Raimundo Carlos Silvério Freire, DSc.
Presidente da Comissão e Examinador Interno

Wamberto José Lira de Queiroz, DSc.
Orientador

Edmar Candeia Gurjão, DSc.
Orientador

Danilo Freire de Souza Santos, DSc.
Examinador Interno

Eanes Torres Pereira, DSc.
Examinador Externo

Francisco Madeiro Bernardino Júnior, DSc.
Examinador Externo

Campina Grande - PB

Processo: 23096.012283/2025-82 Documento: 5287957

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
POS-GRADUACAO EM ENGENHARIA ELETRICA
Rua Aprigio Veloso, 882, - Bairro Universitario, Campina Grande/PB, CEP 58429-900

REGISTRO DE PRESENÇA E ASSINATURAS

1 - ATA DA DEFESA PARA CONCESSÃO DO GRAU DE DOUTOR EM CIÊNCIAS, NO DOMÍNIO DA ENGENHARIA ELÉTRICA, REALIZADA EM 07 DE MARÇO DE 2025
(Nº 395)

CANDIDATO(A): **BRUNO CARDOSO DANTAS.** COMISSÃO EXAMINADORA: RAIMUNDO CARLOS SILVÉRIO FREIRE, Dr. UFCG - Presidente da Comissão e Examinador Interno, WAMBERTO JOSÉ LIRA DE QUEIROZ, D.Sc., UFCG - Orientador, EDMAR CANDEIA GURJÃO, D.Sc., UFCG - Orientador, DANILO FREIRE DE SOUZA SANTOS, Dsc., UFCG - Examinador Interno, EANES TORRES PEREIRA, Dr., UFCG - Examinador Externo, FRANCISCO MADEIRO BERNARDINO JÚNIOR, D.Sc, UPE - Examinador Externo. TÍTULO DA TESE: Análise Multiescala da Propagação de Epidemias Utilizando Regressão por Processo Gaussiano e Deep Learning . ÁREA DE CONCENTRAÇÃO: Processamento da Informação. HORA DE INÍCIO: **14h00** - LOCAL: **Sala Virtual, conforme Art. 5º da PORTARIA SEI Nº 01/PRPG/UFCG/GPR, DE 09 DE MAIO DE 2022.** Em sessão pública, após exposição de cerca de 45 minutos, o(a) candidato(a) foi arguido(a) oralmente pelos membros da Comissão Examinadora, tendo demonstrado suficiência de conhecimento e capacidade de sistematização, no tema de sua tese, obtendo conceito APROVADO com modificações no texto, de acordo com as exigências da Comissão Examinadora, que deverão ser cumpridas no prazo de 30 dias. Face à aprovação, declara o presidente da Comissão, achar-se o examinado, legalmente habilitado a receber o Grau de Doutor em Ciências, no domínio da Engenharia Elétrica, cabendo a Universidade Federal de Campina Grande, como de direito, providenciar a expedição do Diploma, a que o(a) mesmo(a) faz jus. Na forma regulamentar, foi lavrada a presente ata, que é assinada por mim, Leandro Ferreira de Lima, e os membros da Comissão Examinadora. Campina Grande, 07 de Março de 2025.

LEANDRO FERREIRA DE LIMA
Secretário

RAIMUNDO CARLOS SILVÉRIO FREIRE, Dr. UFCG
Presidente da Comissão e Examinador Interno

WAMBERTO JOSÉ LIRA DE QUEIROZ, D.Sc., UFCG
Orientador

EDMAR CANDEIA GURJÃO, D.Sc., UFCG
Orientador

DANILO FREIRE DE SOUZA SANTOS, Dsc., UFCG
Examinador Interno

Processo: 23096.012283/2025-82 Documento: 5287957

EANES TORRES PEREIRA, Dr., UFCG
Examinador Externo

BRUNO CARDOSO DANTAS
Candidato

2 - APROVAÇÃO

2.1. Segue a presente Ata de Defesa de Tese de Doutorado do candidato **BRUNO CARDOSO DANTAS**, assinada eletronicamente pela Comissão Examinadora acima identificada.

2.2. No caso de examinadores externos que não possuam credenciamento de usuário externo ativo no SEI, para igual assinatura eletrônica, os examinadores internos signatários **certificam** que os examinadores externos acima identificados participaram da defesa da tese e tomaram conhecimento do teor deste documento.

Documento assinado eletronicamente por **LEANDRO FERREIRA DE LIMA**, **SECRETÁRIO (A)**, em 10/03/2025, às 10:21, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **EANES TORRES PEREIRA**, **PROFESSOR 3 GRAU**, em 10/03/2025, às 10:59, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **DANILO FREIRE DE SOUZA SANTOS**, **PROFESSOR(A) DO MAGISTERIO SUPERIOR**, em 10/03/2025, às 11:25, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **RAIMUNDO CARLOS SILVERIO FREIRE**, **PROFESSOR 3 GRAU**, em 10/03/2025, às 12:15, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **WAMBERTO JOSE LIRA DE QUEIROZ**, **PROFESSOR(A) DO MAGISTERIO SUPERIOR**, em 10/03/2025, às 13:20, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **EDMAR CANDEIA GURJAO**, **PROFESSOR 3 GRAU**, em 10/03/2025, às 13:39, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **Bruno Cardoso Dantas**, **Usuário Externo**, em 24/03/2025, às 09:55, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

A autenticidade deste documento pode ser conferida no site https://sei.ufcg.edu.br/autenticidade, informando o código verificador **5287957** e o código CRC **E01BD049**.

Processo: 23096.012283/2025-82 Documento: 5287957

Dedico este trabalho a meu querido pai,

João Dantas, *in memoriam*.

Agradecimentos

Agradeço, em primeiro lugar, a Deus por todas as oportunidades que tem me proporcionado ao longo da minha caminhada.

A minha família pelo apoio incondicional. A minha mãe, Maria Cardoso, por sempre me incentivar e acreditar em mim. E a meu pai, João Dantas, que infelizmente não está mais entre nós, por todo esforço feito para me proporcionar uma boa educação. Agradeço também às minhas irmãs Poliana e Mariana, com quem compartilho lembranças inestimáveis da infância.

Com todo o meu carinho e profunda gratidão, agradeço a minha namorada, Amanda Barbosa, por sua presença constante, apoio inabalável e amor incondicional, que foram essenciais ao longo desta jornada.

À Universidade Federal de Campina Grande, e a todo o seu corpo docente e de funcionários, por me acolher e dar suporte para o desenvolvimento deste trabalho.

Agradeço aos professores Wamberto e Edmar por todo o ensinamento, orientações, paciência, apoio, incentivo, oportunidade de trabalho e confiança em mim depositada ao longo dessa jornada.

Estendo minha gratidão a todos os professores com quem tive a honra de cruzar caminhos ao longo desses anos, pela generosa partilha de conhecimento e pelas experiências inestimáveis que proporcionaram. Em especial a meu grande amigo e mestre Hermann Atila Hrdlicka.

A todos os meus amigos, que de alguma forma se fizeram presentes e contribuiram para a minha formação. Agradeço em especial aos meus amigos Paulo Ricardo e Divya pela amizade e rede de apoio.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e à Coordenadoria de Pós-Graduação em Engenharia Elétrica (COPELE), pelo suporte e financiamento desta pesquisa.

"Você nunca fará nada neste mundo sem coragem. É a melhor qualidade da mente ao lado da honra."

Aristóteles.

Resumo

O crescimento exponencial do número de casos de infecção por COVID-19 impactou milhões de vidas ao redor do mundo, como relatado por diversos veículos de imprensa e plataformas de divulgação de dados estatísticos sobre a pandemia. Analisar esses dados pode contribuir para a previsão do comportamento da doença em diferentes escalas, auxiliando na tomada de decisão sobre medidas de contenção em diversos níveis territoriais. Neste contexto, este trabalho investigou a aplicação de diferentes abordagens de modelagem utilizadas em estudos epidemiológicos, com o objetivo de identificar a mais adequada à compreensão da dinâmica de propagação de epidemias em uma análise multiescala. Para isso, explorou-se o desempenho dos modelos compartimentais, do modelo de regressão aditiva (por meio do software *Prophet*) e da Regressão por Processo Gaussiano, sendo este último adotado como foco principal da pesquisa. Como principal contribuição, o estudo propõe o desenvolvimento de uma abordagem baseada em *deep learning* para otimizar o processo de seleção do *kernel* no modelo de Regressão por Processo Gaussiano, solucionando uma limitação conhecida na literatura e aprimorando sua capacidade preditiva e eficiência computacional.

Palavras-chave: Análise Multiescala. Modelagem Computacional Aplicada à Dinâmica de Doenças Infectocontagiosas. Regressão por Processo Gaussiano Aplicada à COVID-19. Modelos Epidemiológicos Orientados a Dados.

Abstract

The exponential growth in the number of COVID-19 infections has impacted millions of lives around the world, as reported by various news outlets and data dissemination platforms. Analyzing such data can support the prediction of disease behavior at different scales, aiding decision-making on containment measures across various territorial levels. In this context, this study investigates the application of different modeling approaches used in epidemiological research, aiming to identify the most suitable one for understanding the dynamics of epidemic spread in a multiscale analysis. To this end, the performance of compartmental models, additive regression models (using the Prophet software), and Gaussian Process Regression was evaluated, with Gaussian Process Regression being adopted as the main focus of the research. As its main contribution, the study proposes a deep learning-based approach to optimize the kernel selection process in Gaussian Process Regression models, addressing a well-known limitation in the literature and enhancing both predictive performance and computational efficiency.

Keywords: Multiscale analysis. Computational Modeling Applied to the Dynamics of Infectious Diseases. Gaussian Process Regression Applied to COVID-19. Data-Driven Epidemiological Models.

Sumário

Lista de abreviaturas, símbolos, siglas e acrônimos

Símbolos

$p(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N))$: Distribuição de probabilidade conjunta das variáveis aleatórias $f(\mathbf{x}_1)$ até $f(\mathbf{x}_N)$ 17

β : Coeficiente de transmissão 13

γ : Taxa de recuperação 13

$\rightarrow$: Transição de Estados 13

σ : Coeficiente de incubação 15

d/dt: Derivada 13

E: Exposto 14

I: Infectado 13

N: Tamanho da população 13

R: Removido 13

R^2: Coeficiente de Determinação 42

S: Suscetível 13

t: Tempo 11

Siglas

ANNs: Redes Neurais Artificiais (*Artificial Neural Networks*) 4

CNNs: Redes Neurais Convolucionais (*Convolutional Neural Networks*) 33

CoIM: Centro de Massa de Infecção (*Center of Infection Mass*) 33

DTs: Árvores de Decisão (*Decision Trees*) 4

GP: Processo Gaussiano (*Gaussian Process*) 17

GPR: Regressão por Processo Gaussiano (*Gaussian Process Regression*) 4

LML: Log-Verossimilhança Marginal (*Log Marginal Likelihood*) 42

LSTM: Memória de Longo-Curto Prazo (*Long Short-Term Memory*) 4

MSE: Erro Quadrático Médio (*Mean Squared Error*) 42

NAR: Autorregressiva Não linear (*Nonlinear Autoregressive*) 34

OMS: Organização Mundial da Saúde 1

RLIM: Modelo de Infeção Recursivo e Latente (*Recursive and Latent Infection Model*) 32

RNN: Rede Neural Recorrente (*Recurrent Neural Network*) 4

SRAG: Síndrome Respiratória Aguda Grave 4

STD: Desvio Padrão (*Standard Deviation*) 42

Acrônimos

ARIMA: Autorregressivo Integrado de Média Móvel (*Autoregressive Integrated Moving Average*) 33

COVID-19: COronaVIrus Disease 2019 1

ESPIN: Emergência de Saúde Pública de Importância Nacional 3

PRISMA: Principais Itens para Relatar Revisões Sistemáticas e Meta-análises (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) 30

SARIMA: Autorregressivo Integrado de Média Móvel Sazonal (*Seasonal Autoregressive Integrated Moving Average*) 33

SARS-CoV-2: Severe Acute Respiratory Syndrome Coronavirus 2 4

SEIR: Suscetível-Exposto-Infectado-Removido 3

SIR: Suscetível-Infectado-Removido 12

SIS: Suscetível-Infectado-Suscetível 3

Lista de Tabelas

Lista de Figuras

Parte I

Introdução

Capítulo 1

Introdução

Neste capítulo, são apresentadas a contextualização, a relevância e a motivação que norteiam este trabalho, bem como os objetivos geral e específicos da realização da pesquisa e as suas contribuições para a comunidade científica. Por fim, é apresentada a organização dos capítulos seguintes do texto.

1.1 Contextualização

Desde a constatação do surto da infecção por coronavírus (COVID-19), em dezembro de 2019, em Wuhan (China)[1–3] e sua reclassificação como pandemia pela Organização Mundial da Saúde (OMS)[4], já houve mais de 777 milhões de casos confirmados e, dentre esses, mais de 7 milhões de mortes em 216 países até Janeiro de 2025[5]. O Brasil atingiu 714 mil mortes em um total de 39,1 milhões de casos confirmados[6], tendo sido, entre março e maio de 2021, classificado como o epicentro global da pandemia. Vidas foram impactadas e, dentre outros efeitos danosos, surgiram incertezas no comércio global e nas cadeias de abastecimento[7]. Várias medidas (distanciamento social, medidas de barreira) e metodologias foram desenvolvidas para combater esse fenômeno[8, 9]. O aumento exponencial dessa enfermidade pôde também ser observado em extensões territoriais menores, como o estado Paraíba, por exemplo, com 729 mil casos confirmados e mais de 10 mil óbitos[10], e a cidade de Campina Grande - PB, com 42 mil casos confirmados e 1 mil óbitos[11].

A rápida disseminação global da COVID-19 mobilizou várias iniciativas científicas a

fim de compreender e atenuar os impactos da doença. Os períodos críticos da pandemia expuseram grandes desafios na previsão do número de casos, dificultando a tomada rápida de decisão em relação às medidas de contenção. Esse cenário tornou evidente as limitações dos modelos tradicionais e reforçou a necessidade de abordagens mais adaptáveis e eficazes. Neste sentido, os cientistas desempenham um papel crucial no combate ao coronavírus, seja por meio da pesquisa e desenvolvimento de vacinas ou da utilização de tecnologias avançadas para compreender o comportamento do vírus. As iniciativas que se concentram na análise de dados são particularmente úteis, pois podem ajudar a prever cenários, entender a frequência e a distribuição dos casos, e acompanhar a evolução da doença. Essas informações são fundamentais para a implementação eficaz de medidas preventivas.

A interpretação estatística pode ser eficaz para esclarecer suposições, fornecendo estimativas precisas dos parâmetros analisados, permitindo a extração de conhecimento e a detecção de padrões. Aliada ao processamento e gerenciamento adequados de dados, pode expor novos conhecimentos e facilitar a resposta a oportunidades e desafios em tempo hábil[12]. Na área da saúde, a exploração de informações acerca de levantamentos epidemiológicos permite evidenciar dados cruciais, identificar tendências, fazer previsões gerais e avaliar a incerteza e o erro nas previsões[8, 13].

A compreensão do comportamento de cenários baseada na análise de dados é uma prática empregada há bastante tempo. Há uma vasta literatura sobre modelagem matemática em epidemiologia[14], passando pelos modelos compartimentais[15], estabelecendo as bases da epidemiologia moderna, até as atuais abordagens orientadas a dados[16]. A modelagem epidemiológica é fundamental no monitoramento e controle de doenças infecciosas, dando suporte à criação de políticas públicas e à alocação de recursos de saúde. Modelos matemáticos e computacionais são usados há décadas para a previsão da propagação de doenças como gripe e sarampo, fornecendo informações fundamentais para o emprego de medidas de mitigação eficazes. Porém, a pandemia de COVID-19 evidenciou limitações nas estratégias utilizadas até então, principalmente diante da heterogeneidade da população atingida e da subnotificação de casos, reforçando a necessidade de abordagens mais robustas, capazes de capturar incertezas e variações nos padrões de disseminação da doença.

As estatísticas relativas aos número de disseminação da COVID-19 demonstraram não só a magnitude da pandemia, mas também evidenciaram a necessidade de métodos preditivos

capazes de orientar políticas de saúde e alocação de recursos eficientes, particularmente em cenários com grandes variações regionais. Assim, identifica-se a importância da análise de diferentes métodos para o estudo de epidemias e identificação de possíveis medidas de controle em um contexto multiescala.

1.2 Relevância e Motivação

A pandemia de COVID-19 exigiu habilidades para superar, de forma rápida e eficiente, os desafios que surgiram nos sistemas de saúde[17], cuja gestão demandou a melhor combinação dos recursos disponíveis, buscando o aprimoramento do funcionamento das organizações envolvidas no controle da doença[18]. A escala global da pandemia afetou substancialmente a saúde da população e a economia mundial. Com isso, muitos países, incluindo o Brasil, declararam Emergência de Saúde Pública de Importância Nacional (ESPIN), o que levou à adoção de medidas de contenção da disseminação do vírus[19, 20]. A capacidade de prever com precisão a evolução de uma pandemia é essencial para a aplicação de medidas preventivas. Neste sentido, dada a urgência da fortificação dos setores da saúde em um momento que obrigou o distanciamento social, iniciativas de Saúde Digital e Telemedicina foram amplamente desenvolvidas[17, 20] e foram fundamentais para superar as incertezas surgidas naquele período, permitindo aos pacientes assistência médica e declarações de qualidade de saúde diferenciadas[18].

Para o ambiente de Saúde Digital, a OMS considerou estratégias de reforço dos sistemas de saúde pela adoção de tecnologias emergentes[21]. As tecnologias desenvolvidas para controle e análise de dados relacionados à COVID-19, tanto nacional quanto internacionalmente, variaram desde aplicativos móveis para rastreio de contatos até sistemas de análise de dados para predição de tendências epidêmicas e cálculo de riscos, além de plataformas de telemedicina e ferramentas de diagnóstico assistido por computador[20].

Para o entendimento do comportamento da COVID-19, a maioria dos estudos realizados no Brasil utiliza modelos compartimentais: como os modelos SIS (Suscetível-Infectado-Suscetível)[22, 23], SEIR (Suscetível-Exposto-Infectado-Removido)[24] ou ainda suas variações que adicionam o estado de saúde de cada indivíduo[25]. A ideia central de modelagem compartimental é separar a população hospedeira do agente causador da doença em agru-

pamentos que indicam o estado de saúde de cada indivíduo com relação à doença em cada instante de tempo. No entanto, resultados imprecisos na previsão de cenários futuros do comportamento da doença podem ocorrer com esses modelos em nosso país, devido a sua dinâmica. Além da população brasileira ser heterogênea, os dados coletados são insuficientes e com baixos níveis de testes de diagnósticos para SARS-CoV-2[26]. Um exemplo disso é o banco de dados epidemiológico brasileiro SRAG (Síndrome Respiratória Aguda Grave) que apresenta um grande número de erros e inconsistências devido à inserção manual de dados.

Em alguns estudos relacionados a este trabalho[27–29], mostrou-se que considerar apenas as contagens de casos confirmados e de casos de mortes seria muito enganoso, uma vez que as tendências são afetadas pelo número de testes diários que, por sua vez, mudam de região para região, gerando incertezas nas projeções dos números de casos e óbitos, dificultando a adoção de medidas estratégicas e contribuindo para sobrecarga de hospitais e agravamento da crise sanitária. Além disso, relatar estatísticas em nível nacional não diz muito sobre a dinâmica da doença em nível regional[27].

Dessa forma, evidencia-se a importância de abordagens que integrem informações oriundas de diferentes escalas territoriais para compreender em mais detalhes a dinâmica de disseminação de doenças. A abordagem multiescala permite a identificação de padrões que podem não ser detectados em análises isoladas, possibilitando uma visão mais ampla do fenômeno.

À vista disso, ferramentas de aprendizado de máquina têm grande importância na gestão da COVID-19, permitindo análises de dados complexas, com precisão nas previsões, e facilitando a tomada de decisão acerca de medidas de contenção, tendo sido utilizadas em diversas aplicações que incluem sistemas de alerta precoce, análise de imagens de tomografia computadorizada e modelagem epidemiológica para entender e prever a propagação da doença[30]. Exemplos dessas ferramentas abrangem Redes Neurais Artificiais (ANNs, do inglês *Artificial Neural Networks*), modelos de Árvore de Decisão (DTs, do inglês *Decision Trees*) e Memória de Longo-Curto Prazo (LSTM, do inglês *Long Short-Term Memory*), que é um tipo de Rede Neural Recorrente (RNN, do inglês *Recurrent Neural Network*)[31–33]. Além dessas abordagens, a Regressão por Processo Gaussiano (GPR, do inglês *Gaussian Process Regression*) também se destaca como uma ferramenta robusta para modelagem epidemiológica, permitindo a captura de padrões complexos e a quantificação da incerteza nas

previsões.

Pelo exposto, verifica-se a necessidade da análise de técnicas que auxiliem na compreensão multiescala da dinâmica de doenças infecciosas e na avaliação de modelos epidemiológicos computacionais e de abordagens baseadas em dados. Nesse contexto, as propriedades estatísticas do modelo GPR permitem não só modelar a evolução temporal de epidemias com grande flexibilidade, mas também quantificar a incerteza associada às previsões, o que é fundamental para sustentar a tomada de decisão em cenários de crise sanitária. Dessa forma, as investigações realizadas ao longo desta pesquisa direcionaram o estudo para aplicação desse modelo às análises propostas. Portanto, esta tese tem como motivação explorar diferentes abordagens com potencial de explicar a propagação multiescala de epidemias, especialmente a de COVID-19, e, a partir disso, implementar o modelo GPR, fornecendo subsídios para a formulação de políticas públicas e estratégias eficientes de resposta a crises sanitárias.

Vale destacar ainda que, apesar da eficácia do modelo GPR, seu desempenho depende criticamente da escolha adequada do *kernel*, cuja definição ainda representa um desafio em aberto na literatura. Essa limitação motivou o desenvolvimento de uma abordagem baseada em *deep learning* para a seleção automatizada do *kernel*, com aplicação à análise multiescala da propagação da COVID-19 proposta nesta pesquisa.

1.3 Objetivos

1.3.1 Objetivo Geral

O objetivo geral desta pesquisa é investigar abordagens para a análise multiescala da dinâmica de doenças infectocontagiosas, com ênfase na proposição de um método automatizado de seleção de *kernel* em modelos GPR, utilizando técnicas de *deep learning*, com o intuito de aprimorar a eficiência preditiva da análise proposta. A pesquisa busca contribuir para a modelagem de surtos epidêmicos em diferentes escalas territoriais.

1.3.2 Objetivos Específicos

Para alcançar o objetivo principal, alguns objetivos específicos precisam ser contemplados:

1. Avaliar a aplicabilidade e desempenho de diferentes modelos para o estudo da evolução territorial de surtos epidêmicos, destacando as potencialidades do GPR e suas limitações relacionadas à escolha do *kernel*;

2. Propor e implementar um algoritmo computacional utilizando *deep learning* para a seleção automatizada de *kernels* em um modelo GPR;

3. Avaliar o desempenho do modelo GPR otimizado na análise da propagação da COVID-19 em diferentes escalas territoriais (local, regional e nacional);

4. Analisar a influência da granularidade dos dados na acurácia das previsões do modelo GPR, investigando a adequação de diferentes escalas para o estudo de epidemias

5. Implantar e validar o modelo proposto no contexto específico da propagação do vírus da COVID-19, comparando os resultados obtidos com diferentes configurações e abordagens.

1.4 Contribuição

Com base nos pontos destacados na Subseção 1.3, podem-se citar como principais contribuições desta pesquisa:

1. Avaliação crítica de diferentes modelos aplicados à dinâmica de epidemias em múltiplas escalas;

2. Aprimoramento de algoritmo computacional para o modelo de GPR, considerando sua aplicação à modelagem multiescala de surtos epidêmicos;

3. Proposição e validação de um algoritmo baseado em *deep learning* para a seleção automatizada de *kernels* em GPR, abordando uma limitação relevante e ainda em aberto na literatura de modelagem probabilística;

4. Implementação e aplicação do modelo proposto em um cenário real de propagação do vírus de COVID-19.

1.5 Organização do Texto

Este trabalho é dividido em 8 capítulos. No Capítulo 2, são apresentados os fundamentos teóricos que serviram como alicerce para o desenvolvimento do trabalho. No Capítulo 3, é feita uma revisão bibliográfica que lista os principais trabalhos relacionados com o tema desta pesquisa. No Capítulo 4, é apresentada a metodologia adotada para o desenvolvimento da pesquisa. No Capítulo 5, são apresentados os dados experimentais, de COVID-19, utilizados para a implementação dos algoritmos propostos. No Capítulo 6, é feita uma avaliação crítica dos modelos estudados para a determinação do melhor modelo para análises epidemiológicas e são apresentados os resultados e considerações acerca dessa avaliação. No Capítulo 7, é apresentada uma proposta de otimização de um modelo GPR, bem como o desenvolvimento de um modelo de *deep learning* para automatização da seleção do *kernel*. No Capítulo 8, o modelo GPR é aplicado aos dados de propagação da COVID-19, realizando uma análise no contexto multiescala. Por fim, no Capítulo 9, são apresentadas as conclusões da pesquisa, além de proposições para trabalhos futuros.

Parte II

Fundamentos

Capítulo 2

Fundamentos Teóricos

Neste capítulo, são apresentados os conceitos teóricos que embasam a modelagem epidemiológica aplicada nesta tese, definindo-se as principais abordagens utilizadas. Além disso, é introduzido o conceito de análise multiescala, que é fundamental para o entendimento das variações de disseminação de epidemias em diferentes níveis territoriais e/ou temporais. Esses fundamentos, juntamente com as formulações matemáticas envolvidas, são úteis para o desenvolvimento das questões investigadas ao longo deste trabalho.

2.1 Conceitos fundamentais

Definição 1. *Modelo* é um padrão, uma representação simplificada ou uma descrição feita de um sistema real, elaborada para demonstrar seu funcionamento e facilitar a análise de fenômenos complexos. Modelos podem assumir diversas formas — matemáticos, computacionais ou baseados em aprendizado — e são ferramentas imprescindíveis para a previsão e controle de processos naturais, sociais e tecnológicos[34–36].

Um modelo, portanto, exerce um papel fundamental na concepção de um conjunto de diretrizes para a representação da organização lógica de análise e exploração[34]. Nenhum modelo, porém, consegue sintetizar toda a realidade e, por isso, deve ser entendido como um instrumento de trabalho passível de reformulações e aprimoramentos, à medida que novos conhecimentos sobre o assunto são descobertos ou evidenciados pela sua aplicação prática[37].

Definição 2. *Modelos Epidemiológicos* são utilizados para compreender a propagação de doenças e identificar medidas para conter a disseminação de vírus e outras enfermidades[38].

As estimativas estatísticas derivadas desses modelos têm grande importância no combate a epidemias, auxiliando autoridades na definição de estratégias de atuação em diferentes cenários — como, por exemplo, o emprego do distanciamento social na pandemia de COVID-19 para reduzir o risco de contágio[38].

Definição 3. *Epidemiologia Matemática* aplica um conjunto de equações e modelos matemáticos que descrevem a interação entre a população e o ambiente, resultando em uma análise detalhada a respeito da dinâmica de doenças infecciosas[39]. Essa abordagem quantitativa permite prever tendências, avaliar medidas de controle e desenvolver políticas públicas fundamentadas em evidências científicas[39–41].

2.1.1 Classificação de Modelos Epidemiológicos

Os modelos epidemiológicos partem de duas hipóteses fundamentais:

1. **Determinística:** considera que a população é dividida em indivíduos suscetíveis, infectados e recuperados, que podem transitar entre esses estados. O modelo determinístico assume que, quando um indivíduo suscetível entra em contato com um infectado, está sujeito à infecção. Após combater o agente patógeno, o indivíduo se recupera e adquire imunidade permanente, passando para o compartimento dos recuperados. Essa estrutura, em que cada etapa específica da epidemia é atribuída a um subgrupo específico, assume que as interações entre indivíduos ocorrem de maneira homogênea e não considera variações aleatórias e, portanto, assume que o sistema evolui de maneira previsível a partir de taxas de infecção e recuperação fixas[39, 42–44];

2. **Estocástica:** incorpora a aleatoriedade inerente ao processo de transmissão, assumindo que cada indivíduo tem uma chance igual de entrar em contato com um infectado, mas reconhece que essa probabilidade pode variar de forma imprevisível devido a flutuações no comportamento e na rede de contatos. Essa abordagem é particularmente útil para populações pequenas ou para contextos em que as variações do risco de exposição e outras dinâmicas da doença são significativas, proporcionando uma análise mais realista da propagação da epidemia[39, 42–44].

Essas duas hipóteses servem como base para diversas estruturas de modelos epidemiológicos, cada uma com suas vantagens e limitações, e são escolhidas conforme as características dos dados e os objetivos da análise.

Diante dos conceitos apresentados, fica evidente que diferentes tipos de modelos podem ser empregados para a análise epidemiológica. De modo geral, esses modelos podem ser classificados em:

- **Modelos Matemáticos:** baseados em equações (por exemplo, equações diferenciais) para descrever a dinâmica de um sistema por meio de equações e fórmulas matemáticas. São comumente utilizados em modelagem epidemiológica para representar a evolução da transmissão de doenças, descrevendo a interação entre populações e o ambiente, permitindo prever tendências, avaliar medidas de controle e desenvolver políticas públicas baseadas em evidências[45].

- **Modelos Computacionais:** implementações de algoritmos computacionais e métodos numéricos para simulação e previsão, empregando técnicas estatísticas para tratar e interpretar dados[46].

- **Modelos Baseados em Inteligência Artificial (IA):** empregam técnicas de aprendizado de máquina para identificar padrões complexos a partir dos dados e realizar previsões ou tomar decisões[47, 48].

A escolha do modelo depende das características dos dados e dos objetivos da previsão, podendo, inclusive, ser adotada uma abordagem híbrida que combine elementos de diferentes categorias.

2.2 Modelos Compartimentais

Os modelos compartimentais são largamente utilizados na epidemiologia para representar a dinâmica de disseminação de doenças. Esses modelos dividem a população em compartimentos de acordo com o estado de saúde dos indivíduos ao longo do tempo ($\mathbf{t}$), da seguinte forma:

- Indivíduos **Suscetíveis** ($\mathbf{S(t)}$) podem contrair a doença em um dado momento. Inclui pessoas que não foram infectadas e, portanto, são suscetíveis à infecção em caso de contato com um indivíduo infectado. Sendo assim, o número de suscetíveis diminui à medida que os indivíduos são expostos ao agente infeccioso.

- Indivíduos **Infectados** ($\mathbf{I(t)}$) contraíram a doença em um dado momento e podem disseminá-la. Compreende, portanto, pessoas que estão infectadas pela doença no tempo t e podem transmiti-la aos suscetíveis. O número de infectados pode aumentar ou diminuir ao longo do tempo, de acordo com as medidas de controle praticadas e a taxa de transmissão da doença, que é a taxa de surgimento de novas infecções devido ao contato entre indivíduos suscetíveis e infectados.

- Indivíduos **Recuperados** ($\mathbf{R(t)}$) se recuperaram da doença e adquiriram imunidade, não podendo mais transmiti-la ou contraí-la. Normalmente, o número de recuperados aumenta com o passar do tempo. Considerando contextos em que a imunidade não é permanente, a classe $\mathbf{R(t)}$ pode ser denominada **Removidos**, incluindo indivíduos que se recuperaram e adquiriram imunidade transitória ou que foram removidos da população por outros motivos, como morte.

- Indivíduos **Expostos** ($\mathbf{E(t)}$) foram expostos à doença, mas ainda não são infectados ou capazes de transmitir a doença, ou seja, estão em uma fase latente da infecção, em que foram expostos ao agente infeccioso, mas ainda não apresentam sintomas ou não são contagiosos. Dessa forma, indivíduos expostos tendem a, com o tempo, evoluir para o estado infectado. Essa classe é considerada apenas em algumas variações dos modelos epidemiológicos.

As funções representativas dos modelos compartimentais são importantes para modelar a propagação de doenças em uma população, ajudando os estudiosos e tomadores de decisão a entender o curso de epidemias. Dentre os diversos tipos modelos compartimentais, os três mais utilizados no Brasil são[39, 43, 49, 50]:

1. **Modelo SIR (Suscetível-Infectado-Removido):** inclui o grupo de pessoas recuperadas da doença. Elas não são infectadas novamente porque o organismo já está

imune e, dessa forma, não transmitem a doença. A suposição básica deste tipo de modelo é que um indivíduo pode passar sucessivamente por estágios S, I e R e adquirir imunidade permanente ou morrer.

O fluxo desse modelo pode ser expresso como

$$\mathrm{S} \rightarrow \mathrm{I} \rightarrow \mathrm{R}. \tag{2.1}$$

Considerando uma população com uma quantidade N fixa de indivíduos, tal que $\mathrm{N} = \mathbf{S(t)} + \mathbf{I(t)} + \mathbf{R(t)}$[15], derivam-se as equações

$$\frac{d\mathbf{S(t)}}{dt} = -\frac{\beta \mathbf{S(t)I(t)}}{\mathrm{N}}, \tag{2.2}$$

$$\frac{d\mathbf{I(t)}}{dt} = \frac{\beta \mathbf{S(t)I(t)}}{\mathrm{N}} - \gamma \mathbf{I(t)}, \tag{2.3}$$

$$\frac{d\mathbf{R(t)}}{dt} = \gamma \mathbf{I(t)}, \tag{2.4}$$

em que β e γ representam, respectivamente, o coeficiente de transmissão e a taxa de recuperação.

Várias suposições foram feitas na formulação dessas equações: primeiro, uma pessoa qualquer da população deve contrair a doença com a mesma probabilidade β de infecção que as demais. Portanto, a pessoa infectada está em contato e é capaz de transmitir a doença para outras $\beta \mathrm{N} \mathbf{S(t)}$ por unidade de tempo, e a razão do contato entre a pessoa infectada e a pessoa suscetível é $\mathbf{S(t)}/\mathrm{N}$[51]. Para a segunda e terceira equações, deve-se considerar o número de pessoas que deixam a classe suscetíveis como sendo igual ao número de pessoas que entram para a classe infectados. No entanto, uma fração (γ, que representa a taxa de recuperação média/morte, ou $1/\gamma$, o período infeccioso médio) de infectados deixa esta classe por unidade de tempo para entrar na classe removidos. Estes processos, que ocorrem simultaneamente, são referidos como a Lei da Ação de Massa, uma ideia amplamente aceita de que a taxa de contato entre

dois grupos em uma população é proporcional ao tamanho de cada um dos grupos envolvidos. Finalmente, admite-se que a taxa de infeção e de recuperação é muito mais rápida do que a escala de tempo de nascimentos e mortes e, portanto, estes fatores são ignorados neste modelo[51].

2. **Modelo SIS (Suscetível-Infectado-Suscetível):** considera a possibilidade de vulnerabilidade mesmo após o indivíduo ser infectado e ter vencido a doença. De acordo com esse modelo, o indivíduo suscetível pode ser infectado ao entrar em contato com o patógeno e, após superar a infecção, volta a ser suscetível.

   O fluxo desse modelo pode ser representado por

$$\mathrm{S} \rightarrow \mathrm{I} \rightarrow \mathrm{S}. \tag{2.5}$$

   Esse modelo pode ser facilmente deduzido a partir do modelo SIR, bastando considerar que os indivíduos se recuperam sem imunidade à doença, isto é, uma vez recuperados, os indivíduos se tornam imediatamente suscetíveis.

   As equações diferenciais que descrevem o modelo são

$$\frac{d\mathbf{S}(\mathbf{t})}{dt} = -\frac{\beta \mathbf{S}(\mathbf{t})\mathbf{I}(\mathbf{t})}{\mathrm{N}} + \gamma \mathbf{I}(\mathbf{t}), \tag{2.6}$$

$$\frac{d\mathbf{I}(\mathbf{t})}{dt} = \frac{\beta \mathbf{S}(\mathbf{t})\mathbf{I}(\mathbf{t})}{\mathrm{N}} - \gamma \mathbf{I}(\mathbf{t}). \tag{2.7}$$

3. **Modelo SEIR (Suscetível-Exposto-Infectado-Removido):** considera uma fase latente ou exposta, comum em muitas doenças, durante a qual se diz que o indivíduo está infectado, mas não infeccioso.

   O fluxo desse modelo pode ser descrito como

$$\mathrm{S} \rightarrow \mathrm{E} \rightarrow \mathrm{I} \rightarrow \mathrm{R}. \tag{2.8}$$

   O número de indivíduos em cada compartimento pode ser denotado, respectivamente,

por $\mathbf{S(t)}$, $\mathbf{E(t)}$, $\mathbf{I(t)}$ e $\mathbf{R(t)}$, em que, considerando uma população fixa, tem-se: $\mathrm{N} = \mathbf{S(t)} + \mathbf{E(t)} + \mathbf{I(t)} + \mathbf{R(t)}$.

As equações diferenciais para esse modelo são

$$\frac{d\mathbf{S(t)}}{dt} = -\frac{\beta \mathbf{S(t)I(t)}}{\mathrm{N}}, \tag{2.9}$$

$$\frac{d\mathbf{E(t)}}{dt} = \frac{\beta \mathbf{S(t)I(t)}}{\mathrm{N}} - \sigma \mathbf{E(t)}, \tag{2.10}$$

$$\frac{d\mathbf{I(t)}}{dt} = \sigma \mathbf{E(t)} - \gamma \mathbf{I(t)}, \tag{2.11}$$

$$\frac{d\mathbf{R(t)}}{dt} = \gamma \mathbf{I(t)}. \tag{2.12}$$

em que σ representa o coeficiente de incubação, que indica a taxa com a qual os indivíduos passam do estado E para o estado I.

Para fundamentar as projeções dos modelos epidemiológicos, é importante compreender alguns conceitos-chave:

1. Tempo;
2. Número básico de reprodução: número médio de indivíduos suscetíveis à contaminação por um indivíduo infectado durante o período de infecção;
3. Taxa de transmissibilidade: velocidade com que a doença se espalha;
4. Taxa de propagação: depende das características biológicas do agente patogênico e é definida para prever quando esse agente em circulação na população persiste no modelo SIS;
5. Limiar de epidemia: considera que o agente patogênico pode se espalhar apenas se sua taxa de propagação exceder esse nível.

2.3 Modelo de Regressão Aditiva

O modelo de regressão aditiva é uma abordagem estatística para a previsão de séries temporais que considera múltiplas componentes, como tendências de longo prazo, sazonalidade e efeitos de eventos como feriados. Os dados são decompostos nessas componentes, que são modeladas separadamente e combinadas de forma aditiva para formar a previsão final, capturando o comportamento de longo prazo, modelando padrões repetitivos em intervalos regulares e absorvendo fenômenos como picos atípicos. A ideia geral é que cada componente atua de forma aditiva na variável dependente, oferecendo uma investigação mais precisa e uma compreensão mais instintiva das influências individuais para a série temporal[52].

Um exemplo prático dessa abordagem é o *Prophet*, uma ferramenta que possibilita a decomposição de séries temporais em componentes interoperáveis, facilitando a previsão e a análise sem a necessidade de que o usuário tenha entendimento profundo em estatística avançada, particularmente em cenários em que o comportamento da série temporal é afetado por múltiplos fatores.

Definição 4. *Prophet*[53] é um software de código aberto desenvolvido pela equipe *Core Data Science* do *Facebook*.

Está disponível no *PyPI*[54, 55] e no GitHub[56].

O *Prophet* é baseado em um modelo aditivo em que as tendências não lineares são ajustadas à sazonalidade anual, semanal e diária, além de aos efeitos de eventos especiais (como feriados). Projetado para ser robusto a dados ausentes e mudanças abruptas, esse software funciona melhor com séries temporais com fortes efeitos sazonais e várias temporadas de dados históricos. O *Prophet* é capaz de identificar variações bruscas, comuns em séries temporais reais, e fazer um ajuste adequado na tendência da série temporal[57].

2.4 Regressão por Processo Gaussiano

A Regressão por Processo Gaussiano é uma ferramenta de aprendizado de máquina[58] que possibilita gerar previsões sobre os dados incorporando conhecimento prévio[59, 60]. Do ponto de vista da teoria de aprendizado de máquina, esse modelo utiliza *Lazy Learning*[1]

[1] É um método de aprendizagem em que a generalização por trás da informação do treino é apenas feita quando uma questão é feita ao sistema, funcionando contrário à *Eager Learning* (Aprendizagem Ansiosa)[61].

(aprendizagem preguiçosa)[61] e uma medida de correlação entre os dados observados (a função *kernel*) para prever uma variável em um instante futuro a partir de dados de treinamento.

A previsão não é apenas uma estimativa para essa variável, mas contém também a informação da incerteza[58,62]. Uma das principais vantagens do GPR é a sua capacidade de quantificar a incerteza das previsões, fornecendo intervalos de confiança.

Nesse modelo estatístico, as observações podem ocorrer em um domínio contínuo, por exemplo, tempo ou espaço[47,58].

Definição 5. *Regressão por Processo Gaussiano* (GP) é um modelo estatístico em que as observações ocorrem em um domínio contínuo. Pode-se assumir que cada ponto em algum espaço contínuo de entrada está associado com uma variável aleatória com distribuição normal[47]. Trata-se de um processo aleatório em que qualquer $\mathbf{x} \in \mathbb{R}^d$ é atribuído a uma variável aleatória $f(\mathbf{x})$ e a distribuição conjunta de um número finito dessas variáveis $p(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N))$ é gaussiana[47,58,63].

Matematicamente, pode-se usar a notação

$$p(\mathbf{f}|\mathbf{X}) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}, \mathbf{K}), \tag{2.13}$$

em que $\mathbf{f} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N))$, $\mathbf{x}_i$ representa o i-ésimo vetor coluna de dimensão d, $\boldsymbol{\mu} = (m(\mathbf{x}_1), \ldots, m(\mathbf{x}_N))$ é um vetor de médias das variáveis aleatórias do vetor e $\mathbf{K}$ representa a matriz de covariância de elementos $K_{ij} = \text{cov}(\mathbf{x}_i, \mathbf{x}_j) = \text{k}(\mathbf{x}_i, \mathbf{x}_j)$, em que $\text{k}(,)$ é a função de covariância denominada *kernel*[47]. A escolha dessa função é fundamental para determinar as propriedades de generalização do modelo GP[64].

2.4.1 Padrões de *Kernel*

Os *kernels* são funções de covariância que determinam a correlação entre os dados[47,65] e são fundamentais para o desempenho do GPR, pois codificam os pressupostos sobre o que está sendo aprendido, informando o tipo de dependência que pode haver entre duas variáveis aleatórias, com a suposição de fornecer valores altos para variáveis aleatórias próximas e baixos para variáveis aleatórias distantes[58,63,65].

Definição 6. *Kernel* é uma função de duas entradas $\mathbf{x}_i, \mathbf{x}_j$, em que $\mathbf{x}_i$ e $\mathbf{x}_j$ são vetores

coluna em um espaço euclidiano de dimensão d[64].

O *kernel* é uma parte fundamental do GPR porque determina como as observações estão relacionadas entre si, influenciando diretamente as previsões e a incerteza associada a elas. Dada essa importância, são exploradas, a seguir, as funções *kernel* padrão[2] mais consideradas na literatura.

Constant Kernel

O *Constant Kernel* é dado por

$$\mathrm{k}(\mathbf{x}_i, \mathbf{x}_j) = constant\_value \; \forall \; \mathbf{x}_i, \mathbf{x}_j. \tag{2.14}$$

Definição 7. *Constant Kernel* retorna um valor constante para qualquer par de entradas, tendo a função de adicionar uma constante à função de covariância do GP, o que influencia diretamente a média do modelo.

Pode ser utilizado como parte de um produto ou de uma soma de *kernels*, em que, respectivamente, dimensiona a magnitude do outro *kernel* ou modifica a média do GP.

Dot Product Kernel

O *Dot Product Kernel* é dado por

$$\mathrm{k}(\mathbf{x}_i, \mathbf{x}_j) = \sigma_0^2 + \mathbf{x}_i \cdot \mathbf{x}_j. \tag{2.15}$$

Definição 8. *Dot Product Kernel* é não-estacionário e pode ser obtido a partir de regressão linear. Assume-se que os coeficientes das variáveis de entrada, $\mathbf{x}_d (d = 1, ..., D)$, que correspondem aos pesos atribuídos a cada uma dessas variáveis, têm distribuição normal $\mathcal{N}(0, 1)$ e o termo de ajuste (bias) segue uma distribuição normal $\mathcal{N}(0, \sigma_0^2)$.

É invariante à rotação de coordenadas em torno da origem, mas não a translações. O *Dot Product Kernel* usualmente é combinado com a exponenciação, como $\mathrm{k}(\mathbf{x}_i, \mathbf{x}_j) = (\sigma_0^2 + \mathbf{x}_i \cdot \mathbf{x}_j)^p$. A Figura 2.1 mostra um exemplo com expoente $p = 2$.

Seu parâmetro é:

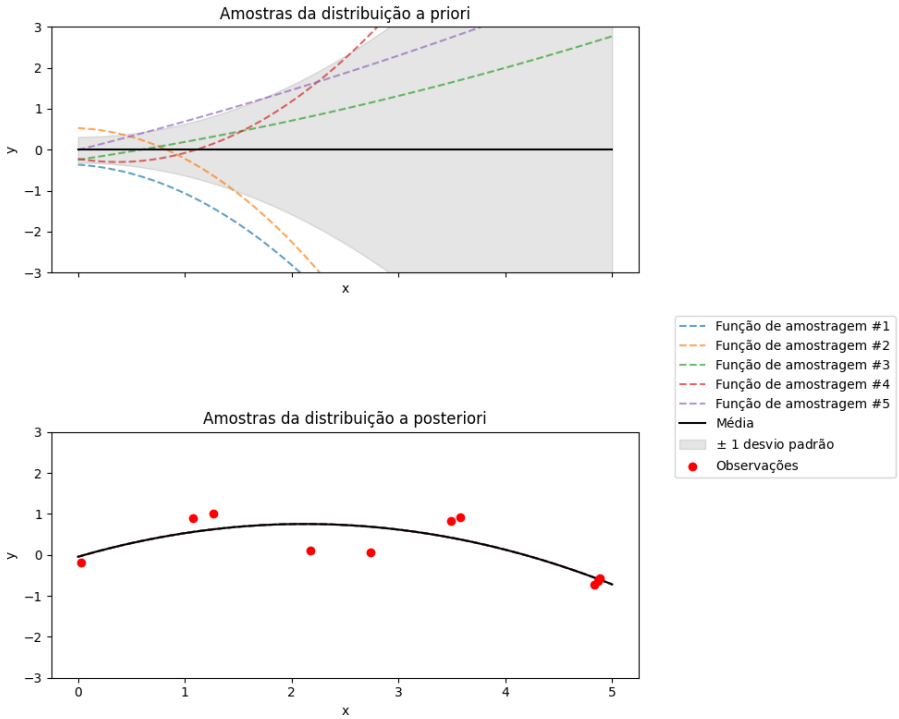
[2]O detalhamento dos *kernels* aqui listados está disponível na documentação da biblioteca *Scikit-learning*[66]

- σ_0^2: controla a não-homogeneidade do kernel

  $\sigma_0^2 = 0$: o *kernel* é chamado de linear homogêneo;

  caso contrário: o *kernel* é não homogêneo.

Figura 2.1: GP a priori e a posteriori com *Dot Product Kernel* com expoente 2.



Fonte: Scikit-learn scikit@2024

O exemplo ilustrado na Figura 2.1 mostra as distribuições a priori e a posteriori de um GPR utilizando um *Dot Product Kernel*. Para ambos os casos, são representados a média, o desvio padrão e cinco funções amostra. Da mesma forma, as Figuras 2.2 a 2.5 trazem essas representações para os *kernels* especificados. Os gráficos em cada uma das figuras ilustram como diferentes *kernels* afetam o comportamento do modelo GPR tanto antes (distribuição a priori) quanto depois (distribuição a posteriori) da observação dos dados.

A distribuição a priori representa possíveis funções que o modelo GPR considera razoáveis com base em determinado *kernel* e seus parâmetros. Já a distribuição a posteriori reflete a distribuição das funções a partir da observação de dados reais, capturando informações acerca dos dados observados e, assim, retratando tanto as suposições iniciais quanto as tendências dos dados.

As curvas coloridas representam possíveis funções amostra da distribuição do modelo GPR com base em cada um dos *kernels* especificados. Para as distribuições a posteriori, essas curvas são ajustadas em razão da observação dos dados (pontos vermelhos), simbolizando o ajuste da hipótese inicial para se alinhar à evidência observada e reduzir a incerteza. As Figuras 2.1 a 2.5 ilustram, portanto, como diferentes *kernels* afetam a modelagem da relação entre as variáveis em um modelo GPR e como o modelo incorpora conhecimento a partir dos dados, oferecendo previsões mais precisas.

Exp-Sine-Squared Kernel

O *Exp-Sine-Squared Kernel* é dado por

$$\mathrm{k}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{2\,\mathrm{sen}^2(\pi \mathrm{d}(\mathbf{x}_i, \mathbf{x}_j)/p)}{l^2}\right). \tag{2.16}$$

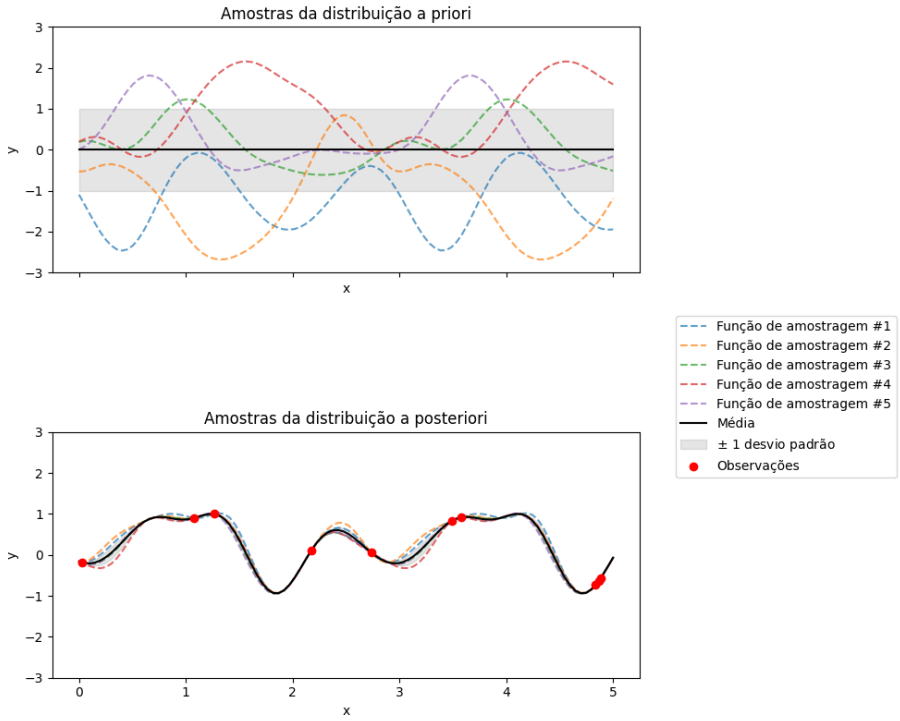
Definição 9. *Exp-Sine-Squared Kernel* permite modelar dados com comportamento periódico.

Seus parâmetros são:

- $l > 0$: escala de comprimento;
- $p > 0$: periodicidade;
- $\mathrm{d}(\cdot, \cdot)$: distância euclidiana.

A Figura 2.2 mostra o resultado do GP a priori e a posteriori para o *Exp-Sine-Squared Kernel*.

Figura 2.2: GP a priori e a posteriori com *Exp-Sine-Squared Kernel*.



Fonte: Scikit-learn scikit@2024

Radial Basis Function (RBF) *Kernel*

O RBF *Kernel* é dado por

$$\mathrm{k}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\mathrm{d}(\mathbf{x}_i, \mathbf{x}_j)^2}{2l^2}\right). \tag{2.17}$$

Definição 10. *RBF Kernel* é um kernel estacionário também chamado de "exponencial quadrático"(*squared exponential*).

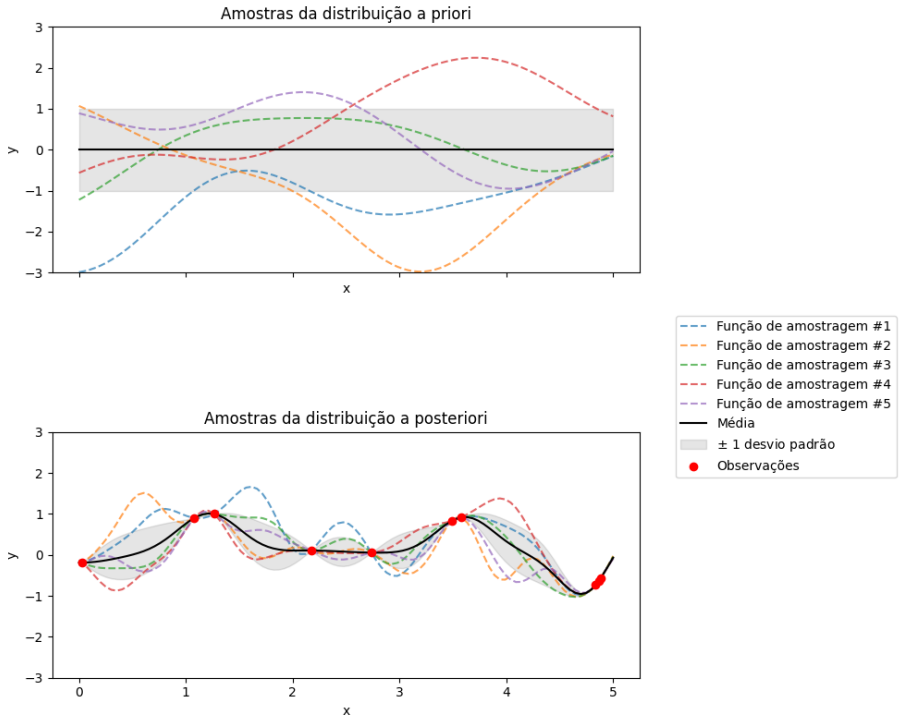
É infinitamente diferenciável, assim, GPs que utilizam esse *kernel* têm derivadas quadráticas médias de todas as ordens e, portanto, são muito suaves.

Seus parâmetros são:

- $l > 0$: escala de comprimento;
- $\mathrm{d}(\cdot, \cdot)$: distância euclidiana.

A Figura 2.3 ilusta as amostras das distribuições a priori e a posteriori de um GP com RBF *Kernel*.

Figura 2.3: GP a priori e a posteriori com RBF *Kernel*.



Fonte: Scikit-learn scikit@2024

Matérn Kernel

O *Matérn Kernel* é dado por

$$\mathrm{k}(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\frac{\sqrt{2\nu}}{l} d(\mathbf{x}_i, \mathbf{x}_j) \right)^{\nu} \mathrm{K}_{\nu} \left(\frac{\sqrt{2\nu}}{l} d(\mathbf{x}_i, \mathbf{x}_j) \right). \tag{2.18}$$

Definição 11. *Matérn Kernel* é um *kernel* estacionário que é uma generalização do *RBF Kernel*, com um parâmetro adicional ν.

Seus parâmetros são:

- ν: controla a suavidade da função resultante. Quanto menor o ν, menos suave é a função aproximada

  $\nu \to \infty$: o kernel resultante é equivalente ao RBF kernel;

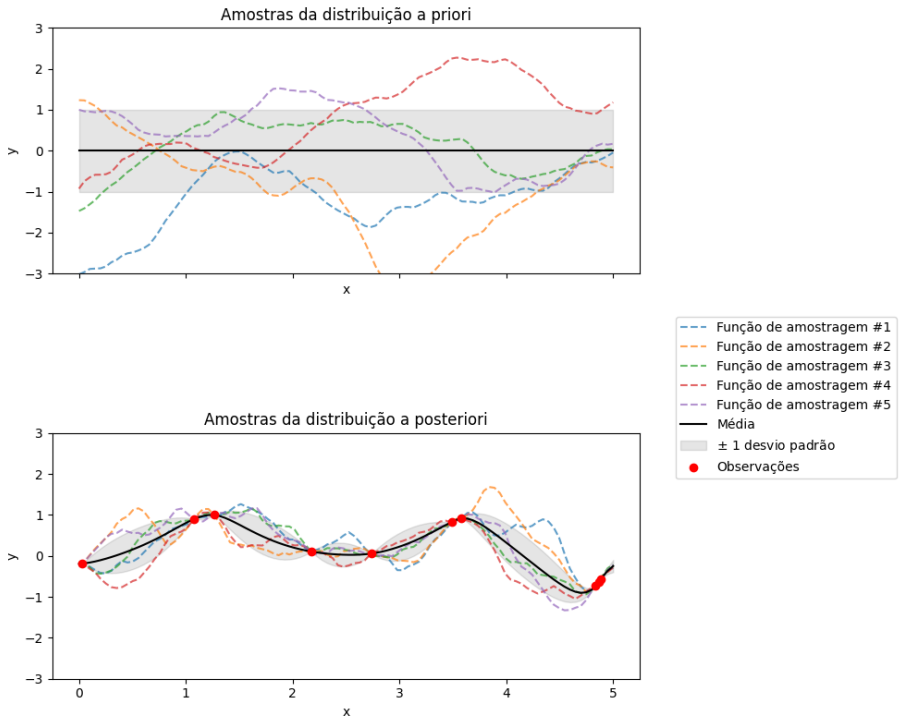
- $l > 0$: escala de comprimento;

- $d(\cdot, \cdot)$: distância euclidiana;

- $\mathrm{K}_\nu(\cdot)$: função de Bessel modificada de segunda espécie;
- $\Gamma(\cdot)$: função Gama.

A Figura 2.4 apresenta funções amostra a priori e a posteriori de um GP com *Matérn Kernel*.

Figura 2.4: GP a priori e a posteriori com *Matérn Kernel*.



Fonte: Scikit-learn scikit@2024

Rational Quadratic Kernel

O *Rational Quadratic Kernel* é dado por

$$\mathrm{k}(\mathbf{x}_i, \mathbf{x}_j) = \left(1 + \frac{d(\mathbf{x}_i, \mathbf{x}_j)^2}{2\alpha l^2}\right)^{-\alpha}. \tag{2.19}$$

Definição 12. *Rational Quadratic Kernel* pode ser visto como uma soma infinita de RBF Kernels com diferentes escalas de comprimento características.

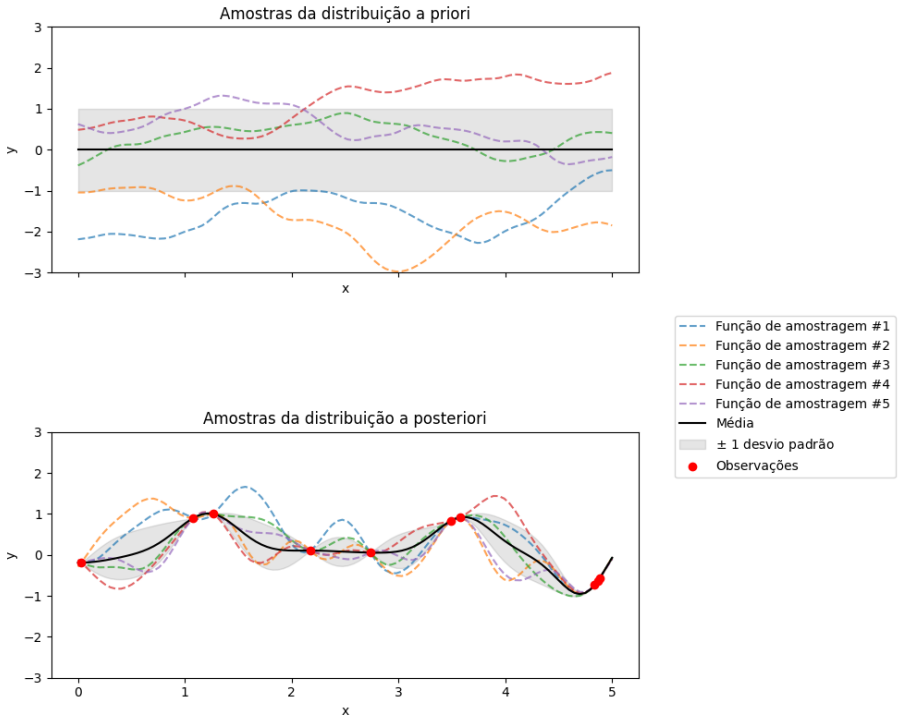
Seus parâmetros são:

- $l > 0$: escala de comprimento;
- $\alpha > 0$: mistura de escalas;

- $d(\cdot, \cdot)$: distância euclidiana.

A Figura 2.5 mostra um GP a priori e a posteriori com *Rational Quadratic Kernel*

Figura 2.5: GP a priori e a posteriori com *Rational Quadratic Kernel*.



Fonte: Scikit-learn scikit@2024

White Kernel

O *White Kernel* é dado por

$$\mathrm{k}(\mathbf{x}_i, \mathbf{x}_j) = noise\_level \text{ se } \mathbf{x}_i = \mathbf{x}_j, 0 \text{ caso contrário.} \tag{2.20}$$

Definição 13. *White Kernel* é utilizado, principalmente, como parte de uma soma de kernels, em que explica a componente de ruído do sinal.

Seu parâmetro é:

- *noise_level*: variância do ruído.

A seleção dos *kernels* é orientada por sua capacidade de capturar diferentes características dos dados. Combinações destes *kernels*, realizadas por meio de operações de soma e

produto, possibilitam a criação de funções de covariância adaptadas às particularidades dos dados.

2.5 Análise Multiescala

A análise multiescala é uma abordagem que investiga fenômenos em diferentes níveis de granularidade, seja temporal, espacial ou populacional. No contexto epidemiológico, essa análise é fundamental, pois a disseminação de uma doença pode variar substancialmente entre diferentes regiões e intervalos de tempo.

- A **escala temporal** permite identificar tanto tendências de longo prazo quanto variações rápidas.
- A **escala espacial** simplifica o entendimento das diferenças na transmissão e do impacto das intervenções em bairros, cidades, estados e países.
- A **escala populacional** considera as variações nos padrões de disseminação entre diferentes grupos populacionais.

Com essa abordagem, modelos como o GPR podem ser avaliados e ajustados para assimilar tanto tendências globais quanto particularidades locais. Isso permite revelar padrões e dinâmicas que não seriam visíveis em uma única escala, contribuindo, assim, para previsões mais precisas e informadas e apoiando decisões de saúde pública mais eficazes.

2.6 Considerações Finais

Neste capítulo, foram apresentados os fundamentos teóricos essenciais para a modelagem epidemiológica. Inicialmente, foram apresentados os conceitos fundamentais de modelo, estabelecendo a base para uma discussão mais aprofundada sobre as diferentes classificações de modelos que podem ser empregadas para a análise de epidemias.

Em seguida, foram discutidos os modelos compartimentais, que são amplamente utilizados para descrever a dinâmica de doenças infecciosas e oferecem uma estrutura clara e matemática bem definida para a análise da propagação de doenças, sendo particularmente

úteis para estudos teóricos e simulações com dados agregados. Esse tipo de modelo, porém, apresenta limitações, uma vez que supõe homogeneidade nas populações, podendo, assim, não refletir com a precisão a realidade da pandemia de COVID-19.

Ademais, foi apresentado o modelo de Regressão Aditiva, exemplificado pelo *Prophet*, um software desenvolvido pelo *Facebook* para a previsão de séries temporais e que é especialmente adequado para dados com evidente sazonalidade e tendências não lineares. No entanto, apesar de ser uma ferramenta poderosa para a previsão de curto prazo e ajustes de sazonalidades, sua aplicação em contextos epidemiológicos é limitada, uma vez que os mecanismos de transmissão de doenças não são explicitamente considerados.

Foi discutida, ainda, a Regressão por Processo Gaussiano, uma abordagem flexível e robusta para a modelagem e previsão de séries temporais, que se destaca pela capacidade de incorporar incertezas de maneira explícita e por utilizar funções *kernel* para capturar relações não lineares complexas nos dados. Discutiu-se a importância da escolha dos *kernels* e apresentaram-se seus principais tipos, destacando suas características.

Cada um dos modelos analisados oferece vantagens e desvantagens distintas para a aplicação no contexto de estudo. Os modelos compartimentais são intuitivos e matematicamente refinados, porém podem falhar em capturar heterogeneidades populacionais e complexidades inerentes de uma pandemia. O *Prophet* é eficaz na captura de sazonalidades e tendências, mas não é adequado para a modelagem de processos de transmissão de doenças. O GPR, no entanto, mostrou-se particularmente promissor em apreender diversidades populacionais e complexidades comuns em surtos epidêmicos, devido a sua abordagem flexível e capacidade de incorporar incertezas. Sendo assim, apesar de todos os modelos discutidos contribuírem para a análise de uma pandemia, o modelo GPR se destacou como a ferramenta mais promissora para os estudos subsequentes desta tese, uma vez que sua capacidade de capturar e modelar comportamentos complexos proporciona uma perspectiva única e detalhada sobre a disseminação da doença.

Finalmente, foi introduzido o conceito de análise multiescala, destacando sua relevância para a compreensão das variações na disseminação de doenças em diferentes níveis de granularidade.

Essa base teórica sustenta os métodos propostos neste trabalho e orienta a aplicação de técnicas avançadas para a previsão e controle de epidemias, contribuindo para o desenvolvi-

mento de políticas públicas mais eficazes. Neste sentido, os esforços da tese se direcionaram para a utilização do GPR como principal metodologia de análise, apoiando-se em suas vantagens para explorar a dinâmica da pandemia em diversas escalas, desde local até nacional.

Parte III

Estado da Arte

Capítulo 3

Revisão Bibliográfica

Neste capítulo, é apresentada uma revisão da literatura relevante para o desenvolvimento desta pesquisa, situando o trabalho atual no contexto de investigações anteriores. Este capítulo aborda os principais estudos, teorias e metodologias que norteiam esta tese.

3.1 Contextualização e Relevância

Estudar a dinâmica da transmissão epidemiológica é fundamental para o entendimento e o controle eficaz de doenças infecciosas. A análise da propagação de doenças em uma população fornece conhecimento acerca dos mecanismos de transmissão, bem como dos fatores que influenciam a disseminação, possibilitando a criação de estratégias eficazes para mitigar o impacto de epidemias. Essa compreensão é importante para uma resposta rápida a surtos em tempo real e, ainda, para o desenvolvimento de modelos preditivos e estratégias de intervenção baseadas em evidências.

Neste sentido, a revisão bibliográfica realizada situa o presente trabalho no contexto de pesquisas anteriores, além de identificar as principais teorias e metodologias que fundamentam a análise da dinâmica epidemiológica, permitindo detectar possíveis lacunas existentes e oportunidades para novas contribuições. Ao integrar e avaliar estudos anteriores, pretende-se estabelecer o contexto para as abordagens propostas nesta pesquisa, evidenciando a importância do entendimento da dinâmica da transmissão epidemiológica para o desenvolvimento de soluções eficazes.

São discutidos trabalhos relacionados aos modelos epidemiológicos e suas aplicações práticas, abrangendo desde modelos tradicionais até abordagens mais recentes. Além disso, é explorada também a literatura sobre métodos computacionais que são empregados no contexto de séries temporais, com ênfase em GPR. São revisadas, ainda, pesquisas acerca da disseminação de pandemias e da eficácia de modelos para entender a dinâmica de doenças em diferentes escalas territoriais.

3.2 Metodologias e Diretrizes para Revisão Sistemática

A fim de entender e aprofundar a análise de abordagens e modelos multiescala na dinâmica da transmissão epidemiológica, foi realizada uma busca extensa na literatura científica, visando identificar e selecionar os estudos mais relevantes em bases de dados como *Scielo*, *Pubmed*, *ScienceDirect* e *Springer*. As palavras-chave utilizadas na pesquisa incluíram termos como "Prediction"ou (OR) "Multi-Scale", combinados (AND) com "COVID-19", OR "COVID", OR "Infectious Diseases". O período coberto pela busca foi de 1º de janeiro de 2020 a 14 de dezembro de 2024.

A busca inicial resultou em 1.391 artigos (*Scielo*: 15 artigos – 1,07%; *Pubmed*: 94 artigos – 6,75%; *ScienceDirect*: 284 artigos – 20,41%; *Springer*: 998 artigos – 71,74%). Destes, foram selecionados 162 artigos que se enquadravam nos critérios estabelecidos. Os artigos selecionados foram categorizados nas seguintes áreas: modelos baseados em aprendizagem de máquina (29 artigos – 17,90%), modelos compartimentais SIR (16 artigos – 9,87%), modelos compartimentais SEIR (6 artigos – 3,70%), variações de modelos compartimentais (16 artigos – 9,87%), modelos matemáticos ou estatísticos (21 artigos – 12,96%), modelos baseados em aprendizado profundo (18 artigos – 11,12%), modelo ARIMA (10 artigos – 6,71%) e outros modelos (46 artigos – 28,3%).

A seleção dos artigos incluídos na revisão sistemática seguiu as diretrizes *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA)[67]. Os artigos selecionados foram aqueles relacionados à dinâmica da COVID-19 e baseados em modelos. Foram excluídos artigos que não estavam em inglês ou português, bem como aqueles não baseados em modelos. Artigos sobre a avaliação de fatores de risco por óbito de COVID-19, estratificação de risco para predição de disseminação e gravidade, modelos geoespaciais, modelos de

previsão e monitoramento de resultados de COVID-19 em hospitais, diagnóstico eficaz de pacientes, e comportamento humano individual em relação a medidas de controle, também foram excluídos.

A Tabela A.1, no Apêndice A, mostra mais detalhes sobre a metodologia PRISMA utilizada na revisão sistemática. Os artigos resultantes das buscas descritas acima estão disponíveis no diretório */literature-review-file* do repositório GitHub[68] desta tese.

3.3 Fundamentos da Modelagem Epidemiológica

A modelagem epidemiológica utiliza técnicas matemáticas e/ou computacionais que simulam o comportamento de epidemias ou pandemias, incluindo taxas de transmissão da doença, suscetibilidade da população e resultados de intervenções públicas, o que pode ser útil para a previsão da dinâmica futura de uma doença e a avaliação da eficácia de medidas de controle adotadas[69, 70].

Sendo assim, essa ferramenta é de fundamental importância, uma vez que permite prever como uma doença pode se disseminar em diversos cenários. Assim, é possível planejar, implementar e avaliar o impacto de estratégias de contenção. Também é possível comunicar de forma clara e transparente a população acerca da situação atual e futura de uma doença e da necessidade de atender aos protocolos indicados[71].

Os modelos também possibilitam que pesquisadores investiguem hipóteses e elaborem estratégias alternativas, oferecendo informações relevantes para a saúde pública que apoiam a tomada de decisões e o planejamento estratégico do controle de doenças infecciosas [70, 72, 73].

3.4 Abordagens de Modelagem Epidemiológica

3.4.1 Modelos Compartimentais

Os modelos compartimentais são amplamente utilizados na modelagem epidemiológica para entender e prever a dinâmica da transmissão de doenças infecciosas. Diversos estudos utilizaram modelos SIR, SEIR e SIS, bem como variações destes modelos, para investigar a dinâmica da pandemia de COVID-19 desde o seu surgimento. A estimação de parâmetros

desses modelos em diferentes fases da disseminação de uma doença infecciosa pode revelar variações na transmissibilidade, bem como identificar fraquezas em medidas de prevenção e controle das infecções, além de prever picos e pontos finais futuros da epidemia[74].

Postnikov[73] utilizou dados do Centro Europeu de Prevenção e Controle de Doenças para demonstrar que o modelo SIR pode reproduzir adequadamente a dinâmica epidêmica da COVID-19, sendo útil para estimativas preditivas e refletindo medidas preventivas por meio de desvios do modelo. Em *System inference for the spatio-temporal evolution of infectious diseases: Michigan in the time of COVID-19*[72], o modelo SIR clássico foi estendido para incluir novos parâmetros, considerando mudanças nos testes, quarentena e protocolos de tratamento, além de populações móveis, em Michigan. Já em *Prediction and mathematical analysis of the outbreak of coronavirus (COVID-19) in Bangladesh*[75], um modelo SIR modificado foi utilizado para analisar a disseminação da doença em Bangladesh. Além desses, um modelo SIR adaptado que incorpora subnotificações e respostas de saúde pública foi desenvolvido no Brasil, em um estudo que demonstra que picos de infecção, taxas de mortalidade e parâmetros de transmissão mudam sob cenários de dados subnotificados[71].

Um modelo SEIR foi desenvolvido por Garrido[76] para otimizar previsões relacionadas à hospitalização e admissão em UTIs devido à pandemia, em Granada, Espanha. Yu[77] propôs um Modelo de Infeção Recursivo e Latente (RLIM, do inglês *Recursive and Latent Infection Model*), baseado no modelo SEIR, que ajusta dados atuais de infecção e recuperação para prever casos futuros e identificar pontos de mudança na transmissão do vírus nos Estados Unidos.

Além disso, variações dos modelos compartimentais foram propostas para esse contexto. Mandal[78] introduziu uma classe de quarentena e medidas de interveção governamentais para prever tendências de curto prazo em estados altamente afetados da Índia com o chamado modelo SEQIR. Já o modelo SIPHERD analisou o impacto de *lockdowns* e taxas de testes, também na Índia[79]. Utilizando derivadas fracionárias, o modelo SEQUIHR foi utilizado para prever a disseminação da COVID-19 no Egito[80]. Em *Simulation and prediction of spread of COVID-19 in The Republic of Serbia by SEAIHRDS model of disease transmission*[81], foi proposto o modelo SEAIHRDS para simular a transmissão do vírus na Sérvia, incorporando cenários de vacinação.

3.4.2 Modelos Baseados em Técnicas de Inteligência Artificial

Apesar de os modelos compartimentais serem úteis para entender a tendência dinâmica de uma pandemia, eles são baseados em suposições matemáticas que podem não prever a real situação de disseminação da doença[82]. Assim, abordagens baseadas em técnicas de inteligência artificial passaram a ser utilizadas para entender o comportamento do vírus da COVID-19, prever a evolução da doença e avaliar o seu impacto. Diversos estudos demonstraram a eficácia dessas técnicas na predição de novos casos e análise do efeito de medidas de controle.

Wang[83] comparou a eficácia dos modelos Autorregressivo Integrado de Médias Móveis (ARIMA, do inglês *Autoregressive Integrated Moving Average*), Autorregressivo Integrado de Médias Móveis Sazonal (SARIMA, do inglês *Seasonal Autoregressive Integrated Moving Average*) e *Prophet* na previsão de casos de COVID-19 em países distintos, mostrando que cada modelo pode apresentar variações de desempenho dependendo do local de aplicação. Abbasimehr[84] propôs abordagens híbridas combinando LSTM, Redes Neurais Convolucionais (CNNs, do inglês *Convolutional Neural Networks*) e otimização Bayesiana para prever séries temporais de COVID-19. Técnicas de aprendizado de máquina foram utilizadas por Kuo[85] para construir um modelo de previsão de casos de COVID-19 baseado em dados demográficos, ambientais e de mobilidade em condados nos EUA, avaliando diferentes cenários de medidas de controle.

Modelos de regressão polinomial e logística foram aplicados por Amar[86] para prever o tamanho final da pandemia no Egito. Utilizando métodos de mineração de dados e aprendizado de máquina, Hirschprung[87] introduziu o conceito de Centro de Massa de Infecção (CoIM, do inglês *Center of Infection Mass*) para prever a disseminação do vírus na Europa Ocidental. Combinando dados relativos ao número de casos, mortes e recuperações com algoritmos de aprendizagem de máquina, Behnam[88] previu a tendência da doença no Irã. O crescimento global da COVID-19 foi previsto por Gothai[89] utilizando algoritmos de aprendizado de máquina supervisionado, e Guo[90] desenvolveu uma ANN para modelar os casos confirmados e de mortes globalmente.

Em *COVID-19 prediction using AI analytics for South Korea*[91], aprendizado de máquina e *deep learning* foram utilizados para analisar fatores demográficos e prever a so-

brevivência de pacientes infectados na Coreia do Sul. A tendência epidêmica na Rússia, no Peru e no Irã foi prevista por Wang[92], utilizando um modelo LSTM aprimorado com mecanismo de atualização contínua. Ketu[93] introduziu um modelo híbrido CNN-LSTM, Verma[82] comparou várias arquiteturas de redes neurais e Tomar[94] utilizou LSTM e ajuste de curva para prever a epidemia na Índia. Já em *Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM*[95], diferentes modelos de *deep learning* foram avaliados para prever casos confirmados, mortes e recuperações em dez países afetados pela pandemia. RNNs foram treinadas com dados limitados para fazer previsões de longo prazo para infecções diárias no Brasil[96].

Alzahrani[97] utilizou o modelo ARIMA para prever o número esperado de casos diários na Arábia Saudita nas quatro semanas seguintes. Uma rede neural autorregressiva não linear (NAR, do inglês *Nonlinear Autoregressive*) e o modelo ARIMA foram aplicados por Khan[98] e Swaraj[99] para prever o número de casos na Índia; o primeiro comparou a precisão dos modelos, enquanto o segundo propôs um modelo híbrido. Também na Índia, Singh[100] desenvolveu um modelo ARIMA para rever o número de casos em seis estados mais afetados e analisou a correlação entre o aumento da temperatura e os casos de COVID-19.

3.4.3 Modelos Matemáticos e Estatísticos

Outra abordagem para estudar a pandemia de COVID-19 é a utilização de modelos matemáticos e estatísticos que permitam prever a disseminação do vírus e orientar políticas públicas. Essa abordagem oferece uma forma de capturar a complexidade dos dados e realizar previsões baseadas em técnicas quantitativas. Diversos estudos exploraram a aplicação de modelos matemáticos e estatísticos para prever a propagação do vírus e otimizar estratégias de reposta à pandemia.

Uma discussão acerca de modelos preditivos existentes, como os modelos SEIR e SIR, é apresentada em *COVID-19 Prediction Models and Unexploited Data*[101], mostrando que esses modelos falharam em explorar incertezas cruciais e fatores inéditos. Este estudo aponta, então, a necessidade de empregar modelos matematicamente comprovados, orientados por dados e que ajustem parâmetros de forma dinâmica ao longo do tempo, a fim de melhorar a precisão das previsões.

AlArjani[102] revisou nove modelos matemáticos aplicados à previsão da dinâmica de transmissão do corona vírus, destacando a importância da modelagem matemática para prever áreas futuras de infecção, bem como para traçar planos governamentais, especialmente durante a segunda onda da pandemia.

Cruz[103] utilizou diferentes tipos de modelos de transmissão dinâmica para estimar o aumento de casos e medir os efeitos das intervenções de distanciamento social no estado de São Paulo. Utilizando apenas os casos confirmados fornecidos pelo relatório técnico diário COVID-19 MEXICO, Torrealba-Rodriguez[104] apresentou a modelagem e previsão de casos de COVID-19 no México por meio de modelos matemáticos e computacionais. Singhal[105] desenvolveu dois modelos para capturar a tendência e prever casos futuros, sendo um modelo matemático baseado em parâmetros da disseminação do vírus e o outro modelo não-paramétrico baseado no método de decomposição de Fourier. Rath[106] aplicou modelos de regressão linear e múltipla para prever a tendência de casos ativos na Índia, prevendo com sucesso casos ativos futuros. Gupta[107] desenvolveu e validou um modelo de predição de risco de mortalidade para pacientes hospitalizados com COVID-19 no sul dos Estados Unidos.

Uma abordagem estatística baseada na teoria dos registros foi utilizada por Khraibani[108] para prever a probabilidade de novos casos recordes no Líbano. Em *Learning delay dynamics for multivariate stochastic processes, with application to the prediction of the growth rate of COVID-19 cases in the United States*[109] é abordada a resolução de equações diferenciais aleatórias com atraso e a predição da trajetória de casos de COVID-19 nos Estados Unidos. Um modelo baseado em equações diferenciais ordinárias é utilizado em *Data-driven prediction of COVID-19 cases in Germany for decision making*[110] para prever a ocupação de leitos de UTI devido à COVID-19 na Alemanha.

Ak[111] desenvolveu uma estrutura computacional baseada em Processos Gaussianos para prever doenças infecciosas e mostrou que a formulação de GPs obteve melhores resultados do que algoritmos padrão de aprendizado de máquina em contextos de previsão temporal, espacial e espaço-temporal, o que mostra o potencial dessa abordagem para contribuir em questões de saúde pública.

3.5 Estudos de Caso Relevantes

Durante a averiguação dos 162 artigos selecionados, foram observadas quatro contribuições relevantes que realizam análise multiescala. Essas contribuições destacam-se por suas abordagens distintas e *insights* valiosos sobre a dinâmica de transmissão da COVID-19.

O artigo de Bouchnita[29] propõe um modelo multiescala que simula a dinâmica de transmissão da COVID-19 com o objetivo de descrever o movimento de agentes individuais utilizando um modelo de força social. Cada indivíduo no modelo pode estar em um dos seguintes estados: suscetível, infectado, em quarentena, imunizado ou falecido. O modelo considera mecanismos de transmissão direta e indireta e foi parametrizado para reproduzir a dinâmica inicial da doença na Itália. Os resultados mostram que, mesmo com as medidas de distanciamento social, situações de pânico aumentam o risco de transmissão de infecção em multidões. Além disso, a transmissão antes da apresentação de sintomas acelera o crescimento exponencial dos casos e a persistência do SARS-CoV-2 em superfícies duras é determinante para o número de casos no pico da epidemia. O estudo também revela que a restrição de movimento dos indivíduos achata a curva epidêmica e sugere que medidas mais rigorosas do que o isolamento e o *lockdown* foram usadas para controlar a epidemia em Wuhan, China.

Lopez[112] aplica um modelo SEIR modificado que incorpora efeitos de proporções variadas de contenção com o objetivo de avaliar a taxa de confinamento nas primeiras fases do surto epidêmico para avaliar os cenários que minimizam a incidência e também a mortalidade. Utilizando dados do início do primeiro pico da pandemia, o modelo projeta cenários para diferentes regiões. Resultados indicam que, sem intervenções, o número de infectados poderia ultrapassar 1,4 milhões em 27 de abril e que o aumento das medidas de isolamento poderia reduzir drasticamente o pico para cerca de 100 mil casos no início de abril. A análise também destaca a eficácia das intervenções de distanciamento social rigorosas na redução do número de infectados e mortes.

O estudo de Scarpone[113] apresenta uma abordagem multimétodo estruturada para analisar dados de incidência transversal dentro de uma estrutura de Análise Exploratória de Dados Espaciais na escala NUTS3 (condado) na Alemanha. A análise geoespacial, interpretação geográfica heurística, análise de aprendizado de máquina bayesiano e modelagem

aditiva generalizada foram utilizadas para avaliar associações entre taxas de incidência e 368 variáveis independentes, identificando variáveis preditoras importantes, como localização geográfica, densidade do ambiente construído e características socioeconômicas. Os resultados sugerem medidas de distanciamento social e redução de viagens desnecessárias como métodos eficazes para reduzir a contaminação.

Por fim, a pesquisa de Quaranta[114] realiza uma análise territorial multiescala da pandemia na Itália usando várias abordagens baseadas em dados. Uma regressão logística é empregada para capturar a evolução dos casos positivos em cada região e em todo o país, enquanto uma versão aprimorada de um modelo SIR é ajustada para as diferentes dinâmicas epidêmicas territoriais via algoritmo de evolução diferencial. Técnicas de *clustering* hierárquico e análise multidimensional são exploradas para revelar semelhanças e diferenças no desenvolvimento geográfico da epidemia. A combinação de identificações paramétricas e análises multiescala baseadas em dados fornece uma compreensão mais aprofundada da disseminação epidêmica não linear e espacialmente não uniforme na Itália, destacando-se como uma contribuição fundamental para o entendimento das dinâmicas regionais da COVID-19 e, assim, mostrando-se como o artigo mais aderente a este estudo.

3.6 Considerações Finais

Neste capítulo, foi apresentada uma revisão abrangente da literatura relevante, contextualizando o desenvolvimento desta pesquisa no panorama das investigações anteriores. Por meio da revisão dos principais estudos, teorias e metodologias, buscou-se fornecer uma base sólida para a compreensão dos conceitos e análises utilizados ao longo desta tese.

Inicialmente, a contextualização e relevância da revisão bibliográfica foram discutidas, destacando a importância de estudar a dinâmica da transmissão epidemiológica. Esse enfoque é crucial para o entendimento dos padrões de disseminação de doenças e para a elaboração de estratégias eficazes de controle e mitigação. Em seguida, foram detalhadas as metodologias e diretrizes adotadas para a realização da revisão sistemática, que seguiu as diretrizes PRISMA. Essa abordagem garantiu uma seleção criteriosa e uma análise detalhada dos estudos mais relevantes no campo.

Os fundamentos da modelagem epidemiológica foram então descritos, proporcionando

uma base teórica para as subsequentes discussões sobre as diferentes abordagens de modelagem epidemiológica. Foi realizada uma revisão da literatura que abrangeu desde modelos compartimentais até técnicas de inteligência artificial e modelos matemáticos e estatísticos, todos aplicados no contexto da pandemia de COVID-19.

Por fim, foram destacados os estudos de caso relevantes selecionados na revisão sistemática, com ênfase naqueles que realizaram análises multiescala. Dentre estes, o trabalho de Quaranta[114] merece uma menção especial devido à sua abordagem inovadora e ao aprofundamento no contexto multiescala da dinâmica de disseminação da COVID-19 na Itália, fornecendo *insights* valiosos para a utilização desse contexto em outros territórios.

Em síntese, este capítulo não apenas embasou teoricamente a pesquisa, mas também evidenciou a diversidade e a complexidade das abordagens de modelagem epidemiológica, ressaltando a importância de estudos multiescala na compreensão das dinâmicas de transmissão de doenças. Assim, a revisão realizada aqui fornece uma fundamentação robusta para as análises e discussões que serão apresentadas nos capítulos subsequentes desta tese.

Parte IV

Metodologia

Capítulo 4

Metodologia

Neste capítulo, é apresentada a metodologia adotada para o desenvolvimento desta pesquisa, justificando as escolhas das ferramentas, técnicas e modelos utilizados. O estudo foi conduzido considerando abordagens estatísticas e computacionais a fim de identificar, implementar e otimizar modelos capazes de prever a propagação de epidemias (particularmente, a de COVID-19) considerando diferentes níveis de granularidade territorial (análise multiescala).

4.1 Revisão Bibliográfica e Seleção de Modelos

Para identificar os modelos utilizados em pandemias anteriores e, especificamente, na COVID-19, foi realizada a revisão da literatura descrita em detalhes no Capítulo 3.

A partir da revisão, identificou-se que os modelos compartimentais tradicionais foram os primeiros a ser empregados para modelar a propagação de doenças infecciosas, motivo pelo qual eles foram considerados como o ponto de partida deste estudo. Porém, embora fundamentais para a compreensão da dinâmica de doenças infecciosas, estes modelos apresentam limitações significativas para lidar com a dinâmica e a heterogeneidade características da COVID-19, o que impulsionou a busca por modelos alternativos que pudessem ser aplicados à proposta desta pesquisa. Neste sentido, dois métodos foram identificados: o *Prophet* e o modelo GPR.

O *Prophet* é um modelo de previsão de séries temporais baseado em regressão aditiva,

que decompõe os dados em componentes de tendência, sazonalidade e eventos especiais. Sua aplicação foi considerada por se tratar de uma ferramenta consolidada, amplamente utilizada em aplicações industriais e acadêmicas, e, principalmente, por permitir modelar padrões sazonais e tendências de forma robusta e se ajustar a feriados e mudanças bruscas, que são características marcantes em séries temporais de dados epidemiológicos. A análise de resultados preliminares mostrou, no entanto, que, para previsões de longo prazo, o *Prophet* não captura adequadamente a complexidade e as flutuações dos dados reais, projetando um crescimento linear que não reflete a variação dos casos de COVID-19.

O modelo GPR, por sua vez, é uma ferramenta de aprendizado de máquina não paramétrica que modela incertezas de forma explícita, utilizando funções *kernel* para capturar relações não lineares entre os dados, o que torna esse modelo particularmente adequado para a modelagem da disseminação de doenças em diferentes escalas territoriais. Os registros de casos na pandemia, especialmente no início, apresentavam bastante subnotificação e variabilidade; o GPR é capaz de modelar as incertezas inerentes aos dados epidemiológicos, fornecendo intervalos de confiança robustos. Isso faz dele uma ferramenta poderosa para análise multiescala e, por isso, esse modelo foi selecionado como a abordagem principal deste estudo.

4.2 Análise Multiescala

A análise multiescala foi motivada pela variabilidade significativa identificada na taxa de reprodução do vírus em diferentes regiões, inclusive dentro da mesma cidade. Estudos demonstraram que a disseminação do vírus pode variar significativamente em diferentes níveis de granularidade, devido a fatores como densidade populacional, mobilidade e condições socioeconômicas. Essa análise permite a investigação da dinâmica de disseminação da doença, possibilitando identificar variações e padrões que podem ser mascarados em análises exclusivamente nacionais ou locais. Com isso, é possível oferecer subsídios para a implementação de medidas de contenção mais eficazes em níveis locais, regionais e nacionais, otimizando a alocação de recursos e a implementação de estratégias de controle direcionadas. Essa abordagem é fundamental para capturar as dinâmicas complexas da disseminação da doença, reconhecendo que diferentes contextos territoriais podem ter comportamentos

distintos. Neste sentido, esta pesquisa testou os modelos em níveis distintos (bairros, zonas, cidades, estados e país), permitindo avaliar o desempenho do GPR em cada escala e ajustar a abordagem conforme a disponibilidade e qualidade dos dados.

4.3 Otimização do Modelo GPR e Seleção Automática do *Kernel*

O GPR utiliza uma função *kernel* para determinar a correlação entre os dados e fazer previsões. Dado que a escolha do *kernel* afeta o desempenho do GPR de forma significativa, foi identificado um potencial de melhorias importantes no desempenho do modelo por meio da automatização da seleção do *kernel*. Assim, foi desenvolvido um método de seleção automática da melhor combinação de *kernels*.

O modelo proposto é uma rede neural *feedforward* com três camadas ocultas, treinada para identificar a combinação de *kernels* que melhor se ajusta aos dados, com base em métricas de desempenho. O processo envolve:

- **Geração de combinações de *kernels*:** *kernels* padrão são combinados por meio de soma e multiplicação para formar elementos que possam capturar diferentes características dos dados;
- **Cálculo de métricas de desempenho:** para cada combinação gerada, são calculadas métricas como Erro Quadrático Médio (MSE, do inglês *Mean Squared Error*), Log-Verossimilhança Marginal (LML, do inglês *Log Marginal Likelihood*), Desvio Padrão (STD, do inglês *Standard Deviation*) e Coeficiente de Determinação (R^2);
- **Biblioteca de *Kernels*:** cada combinação de *kernels* é associada a suas respectivas métricas e uma biblioteca de *kernels* é salva para a etapa posterior;
- **Critério de seleção:** a melhor combinação de *kernels* é selecionada em um modelo de *deep learning*, que é treinado utilizando as métricas da biblioteca de *kernels* para aprender padrões e relações entre elas, indicando a probabilidade de uma combinação de *kernels* ser a melhor, adotando um critério baseado na média ponderada dos R^2 de treinamento, global e teste;

- **Otimização iterativa:** após selecionado o melhor *kernel*, o modelo GPR é executado em múltiplas iterações e, em cada uma, as métricas são avaliadas e os hiperparâmetros do *kernel* são ajustados até que se atinja um critério de convergência ou um limite de iterações (*patience*), retornando o modelo com o melhor valor das métricas.

4.4 Considerações Finais

A metodologia adotada nesta pesquisa integra um mapeamento sistemático detalhado, a seleção de modelos com base nas limitações identificadas em métodos tradicionais e a implementação de técnicas de otimização, garantindo que a análise seja feita de forma robusta e fundamentada em critérios objetivos. Além disso, a análise multiescala permite uma melhor compreensão da dinâmica de disseminação da COVID-19 em diferentes níveis territoriais, auxiliando na tomada de decisão e formulação de políticas públicas de forma eficaz.

Parte V

Base de Dados

Capítulo 5

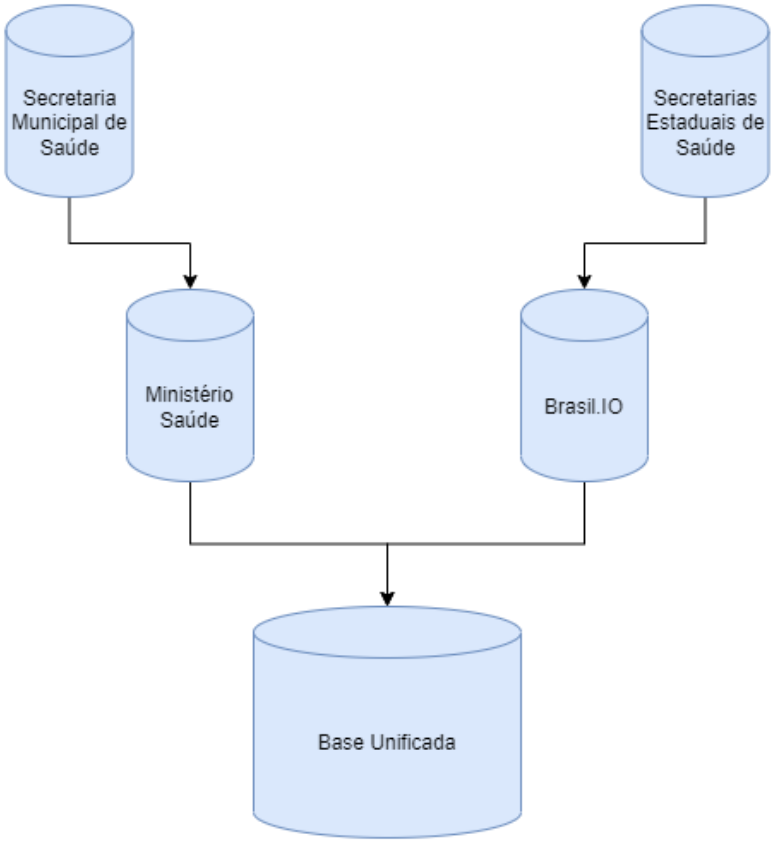
Dados Experimentais

Neste capítulo, é apresentada a estrutura da base de dados da COVID-19 utilizada nesta pesquisa, descrevendo os dados que guiam a investigação, detalhando suas fontes, os critérios de seleção e os métodos de coleta empregados. São exploradas as principais características dos dados, incluindo variáveis relevantes e sua organização. A análise exploratória dos dados também é discutida, destacando as abordagens utilizadas para preparar e entender a base de dados antes da aplicação dos modelos. Este capítulo proporciona uma compreensão detalhada do conjunto de dados experimentais que sustenta as análises e conclusões desta tese, garantindo a transparência e a robustez metodológica necessárias para a pesquisa.

5.1 Dados da COVID-19

Este estudo utiliza dados que apresentam números relacionados à COVID-19 no Brasil, disponibilizados pelo Ministério da Saúde[6], em conjunto com dados das Secretarias Estaduais de Saúde, disponibilizados por meio do Brasil.IO[115, 116], além de dados da Secretaria Municipal de Saúde da cidade de Campina Grande[11]. A base de dados completa está disponível no diretório */data* do repositório GitHub[68]. A estrutura de captação desses dados está exemplificada na Figura 5.1.

Figura 5.1: Estrutura da base de dados considerada neste estudo



Fonte: Próprio Autor.

Os conjuntos de dados epidêmicos nacionais e estaduais foram atualizados diariamente desde 25 de fevereiro de 2020 e, além dos dados totais de óbitos, contêm os dados de casos suspeitos, vacinados, recuperados e de testes realizados. Todos esses dados são atualizados por meio do portal Brasil.IO[115]. A Tabela 5.1 traz uma amostra desses conjuntos de dados.

Tabela 5.1: Amostra de Conjunto de Dados Nacionais da COVID-19

	Data	Total de Casos	Óbitos	Casos Suspeitos	Casos Recuperados	Testes Realizados	Vacinados	Vacinados Segunda Dose	Casos Ativos	Novos Vacinados	N. Vacinados Segunda Dose
...	...	...	...	...	...	...	...	...	...	...	...
465	2021-06-04	16852317	471191	6655234	14874923	49857892	48963900	22997889	1506203	642437	73860
466	2021-06-05	16913984	472861	6655234	14944069	49878944	49491413	23095537	1497054	527513	97648
467	2021-06-06	16954210	473735	6655234	15002817	49939952	49685501	23126008	1477658	194088	30471
468	2021-06-07	16990262	474785	6655234	15022649	50029606	50381104	23232975	1492828	695603	106967
469	2021-06-08	17042198	477027	6655234	15055747	50064972	51092883	23343758	1509424	711779	110783
...	...	...	...	...	...	...	...	...	...	...	...

Para a cidade de Campina Grande, a base de dados referente às notificações de casos

de COVID-19 foi iniciada em 2 de janeiro de 2020. A fim de unificar a análise para todas as bases de dados, definiu-se a data de início como 22 de março de 2020, que foi a primeira data em comum para todos os registros[1].

Além dos casos notificados no município, a prefeitura de Campina Grande também disponibiliza uma base de dados que apresenta um maior número de informações dos pacientes, sendo assim mais detalhada que os conjuntos de dados nacionais e estaduais, possibilitando uma análise mais específica em regiões territoriais menores. Essa base caracteriza cada paciente em relação ao estado de saúde e informações pessoais, incluindo seus dados geoespaciais. A Tabela 5.2 traz uma amostra dessa base de dados.

Tabela 5.2: Amostra de Conjunto de Dados Municipais da COVID-19 de Campina Grande

	...	Data do Teste	...	Sexo	...	CEP	...	Sintoma-Dor de Garganta	...	Data da Notificação	...
...	...	...	...	...	...	...	...	...	...	...	...
13	...	2021-06-27	...	Feminino	...	58.485-000	...	Sim	...	2021-06-28	...
14	...	2021-06-28	...	Masculino	...	58.400-000	...	Não	...	2021-06-28	...
15	...	2021-06-26	...	Feminino	...	58.411-000	...	Não	...	2021-06-28	...
16	...	2021-06-28	...	Feminino	...	58.400-000	...	Não	...	2021-06-28	...
17	...	2021-06-28	...	Feminino	...	58.780-000	...	Não	...	2021-06-28	...
...	...	...	...	...	...	...	...	...	...	...	...

Essa base de dados inclui, além dos dados constantes na Tabela 5.2: Estado da Notificação; Município da Notificação; Tipo de Teste; Resultado; Evolução Caso; Estado do Teste; Data de encerramento; Classificação Final; Resultado Totais; Resultado IgA; Resultado IgM; Resultado IgG; Teste Sorológico; Data do Teste (Sorológico); Estado de Residência; Tem CPF?; Estrangeiro; Município de Residência; Data de Nascimento; País de origem; É profissional de saúde?; CBO; Bairro; Raça/Cor; Profissional de Segurança; Etnia; Comunidade/Povo Tradicional?; Comunidade/Povo Tradicional; Sintoma-Dispneia; Sintoma-Febre; Sintoma-Tosse; Sintoma-Outros; Sintoma-Dor de Cabeça; Sintoma-Distúrbios Gustativos; Sintoma-Distúrbios Olfativos; Sintoma-Coriza; Sintoma-Assintomático; Condições-Doenças respiratórias crônicas descompensadas; Condições-Doenças cardíacas crônicas; Condições-Diabetes; Condições-Doenças renais crônicas em estágio avançado (graus 3, 4 ou 5); Condições-

[1]Para a unificação das datas, foi necessário alterar algumas bases de dados, desconsiderando alguns registros feitos antes da data inicial considerada. As bases de dados – originais e modificadas – estão disponíveis no diretório */data* do repositório GitHub[68].

Imunossupressão; Condições-Gestante; Condições-Portador de doenças cromossômicas ou estado de fragilidade imunológica; Condições-Puérpera (até 45 dias do parto); Condições-Obesidade; Data do início dos sintomas; Descrição do Sintoma.

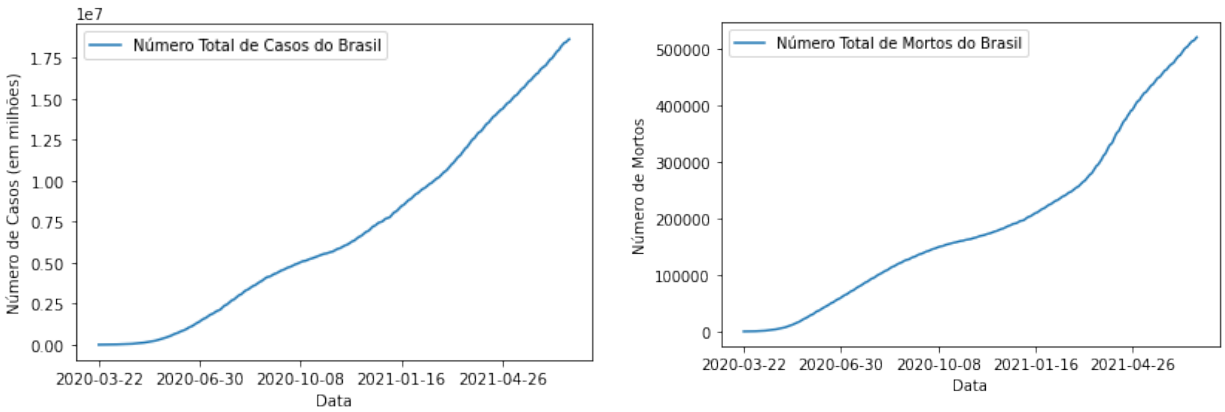
Ademais, bases de dados referentes a alguns setores censitários (bairros) de Campina Grande também foram coletadas. Em síntese, foram utilizadas várias bases de dados para compor a base unificada que é analisada neste estudo e está representada na Figura 5.1. A Tabela 5.3 descreve essas bases. A base de dados unificada pode ser acessada no diretório */data* do repositório GitHub[68].

Tabela 5.3: Exemplo de estrutura de base de dados unificada da COVID-19

	Título do Dataset	Descrição	Local de Origem
1	dataset_Brasil	Dados Totais da COVID-19 no Brasil	Brasil.IO
2	dataset_Paraiba	Dados Totais da COVID-19 na Paraíba	Brasil.IO
3	dataset_CampinaGrande_casos_acumulados	Dados Totais da COVID-19 em Campina Grande	SMS de Campina Grande
4	dataset_CampinaGrande_casos_diarios	Dados Diários da COVID-19 no Brasil	SMS de Campina Grande
5	dataset_Bodocongo	Dados Totais da COVID-19 no bairro Bodocongó	SMS de Campina Grande
6	dataset_Catole	Dados Totais da COVID-19 no bairro Catolé	SMS de Campina Grande
7	dataset_Malvinas	Dados Totais da COVID-19 no bairro Malvinas	SMS de Campina Grande

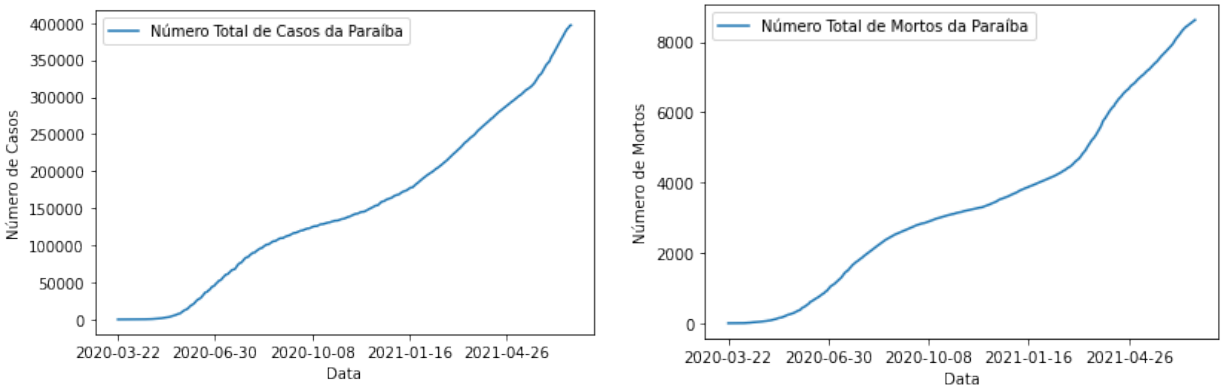
A partir das bases de dados analisadas, foram extraídos dados para se estudar o comportamento da doença em escalas nacional, estadual e municipal. Dessa extração, foram obtidos gráficos que relacionam os números de COVID-19 com o período de disseminação da doença nessas escalas. Os gráficos obtidos são mostrados nas Figuras 5.2 a 5.7.

Figura 5.2: Curvas de casos e óbitos - COVID-19 do Brasil



Fonte: Ministério da Saúde.

Figura 5.3: Curvas de casos e óbitos - COVID-19 da Paraíba

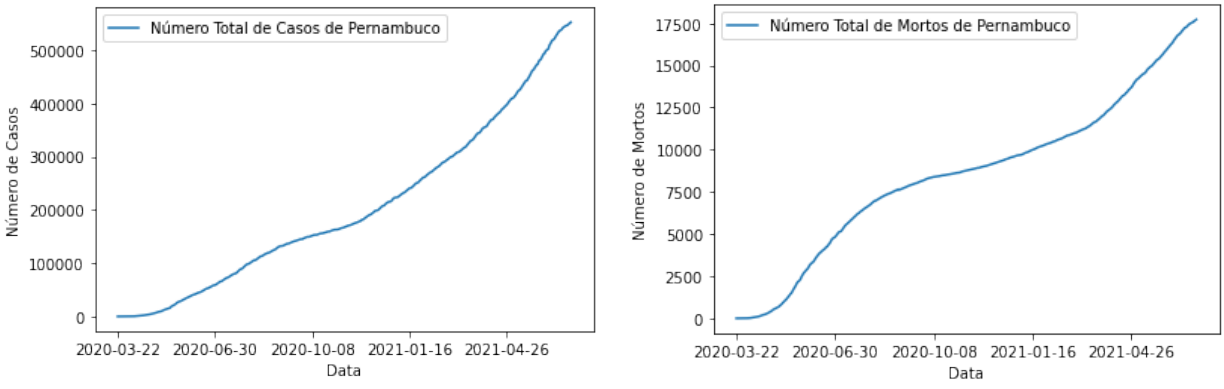


Fonte: Ministério da Saúde.

É possível observar que as curvas, tanto para os números de casos quanto para os de óbito, seguem a mesma tendência de crescimento quando se relaciona a escala nacional (Figura 5.2) com a estadual (Figura 5.3).

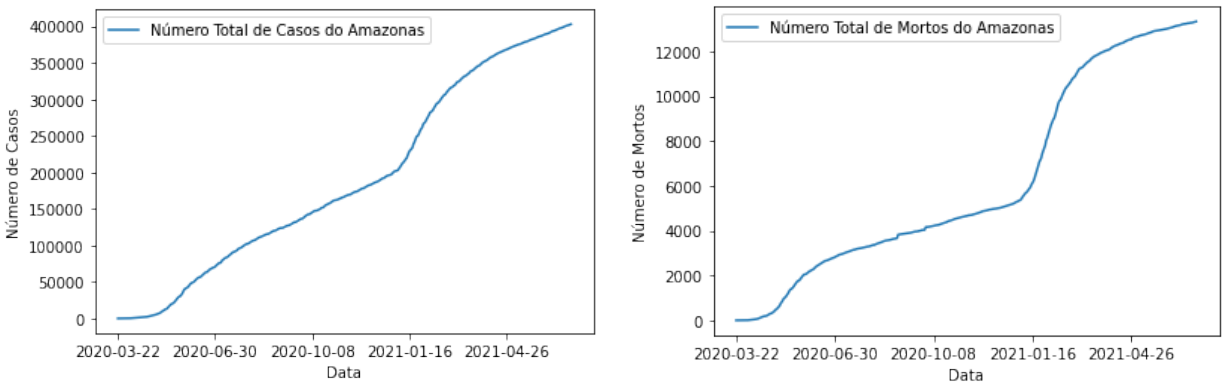
Para fins de comparação, além dos dados do estado da Paraíba, foram extraídos, da base de dados unificada, os números da COVID-19 nos estados de Pernambuco (Figura 5.4) e Amazonas (Figura 5.5), o que reforça a tendência de crescimento equivalente em diferentes escalas.

Figura 5.4: Curvas de casos e óbitos - COVID-19 de Pernambuco



Fonte: Ministério da Saúde.

Figura 5.5: Curvas de casos e óbitos - COVID-19 do Amazonas

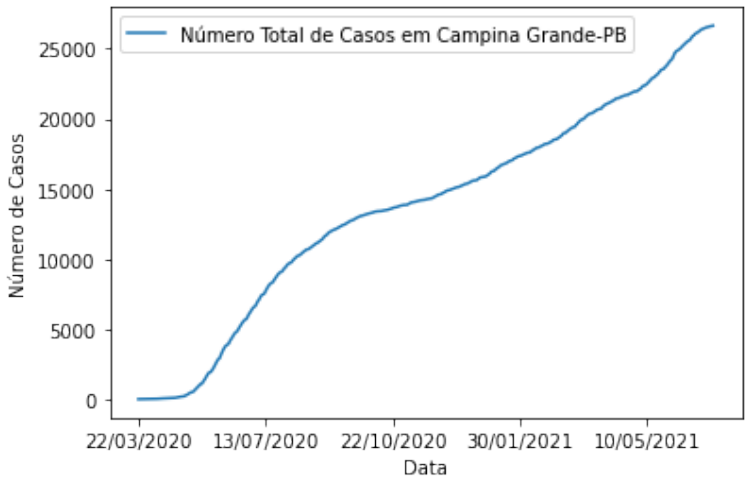


Fonte: Ministério da Saúde.

Vale ressaltar que, embora as curvas sigam essa tendência, alguns casos particulares são identificados quando é feita a análise dos gráficos. Como exemplo, pode-se citar o mês de janeiro de 2021, quando houve uma crise de falta de oxigênio em Manaus, o que provocou um colapso no sistema de saúde e levou a um crescimento mais acelerado no número de casos e mortes no estado do Amazonas (Figura 5.5), o que é evidenciado fazendo-se uma análise da curva desse estado nesse período em relação às demais apresentadas.

Em escala municipal, os dados dos números de casos resultam em uma curva, mostrada na Figura 5.6, que aponta para a mesma tendência de crescimento já observada em escalas estaduais e nacional.

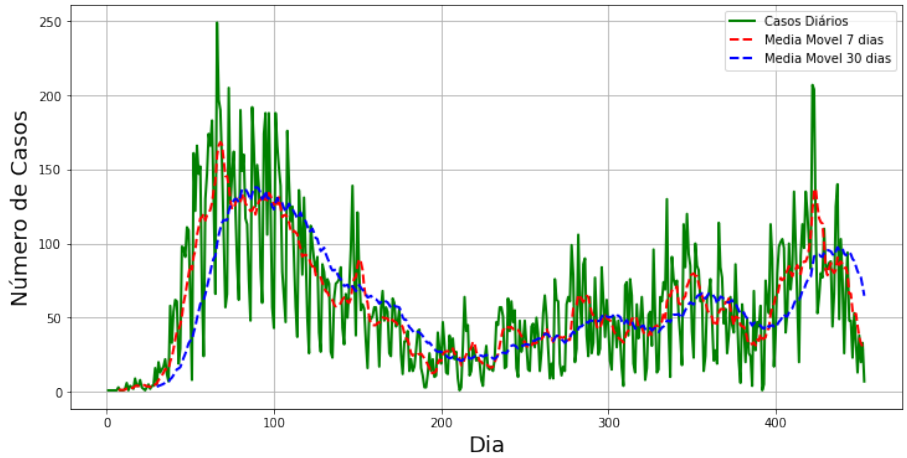
Figura 5.6: Curva de casos - COVID-19 de Campina Grande



Fonte: Secretária Municipal de Saúde de Campina Grande.

A partir dos dados referentes à cidade de Campina Grande, foram obtidas as médias móveis semanal e mensal dos casos diários, mostrada na Figura 5.7.

Figura 5.7: Média Móvel de Casos Diários - COVID-19 de Campina Grande



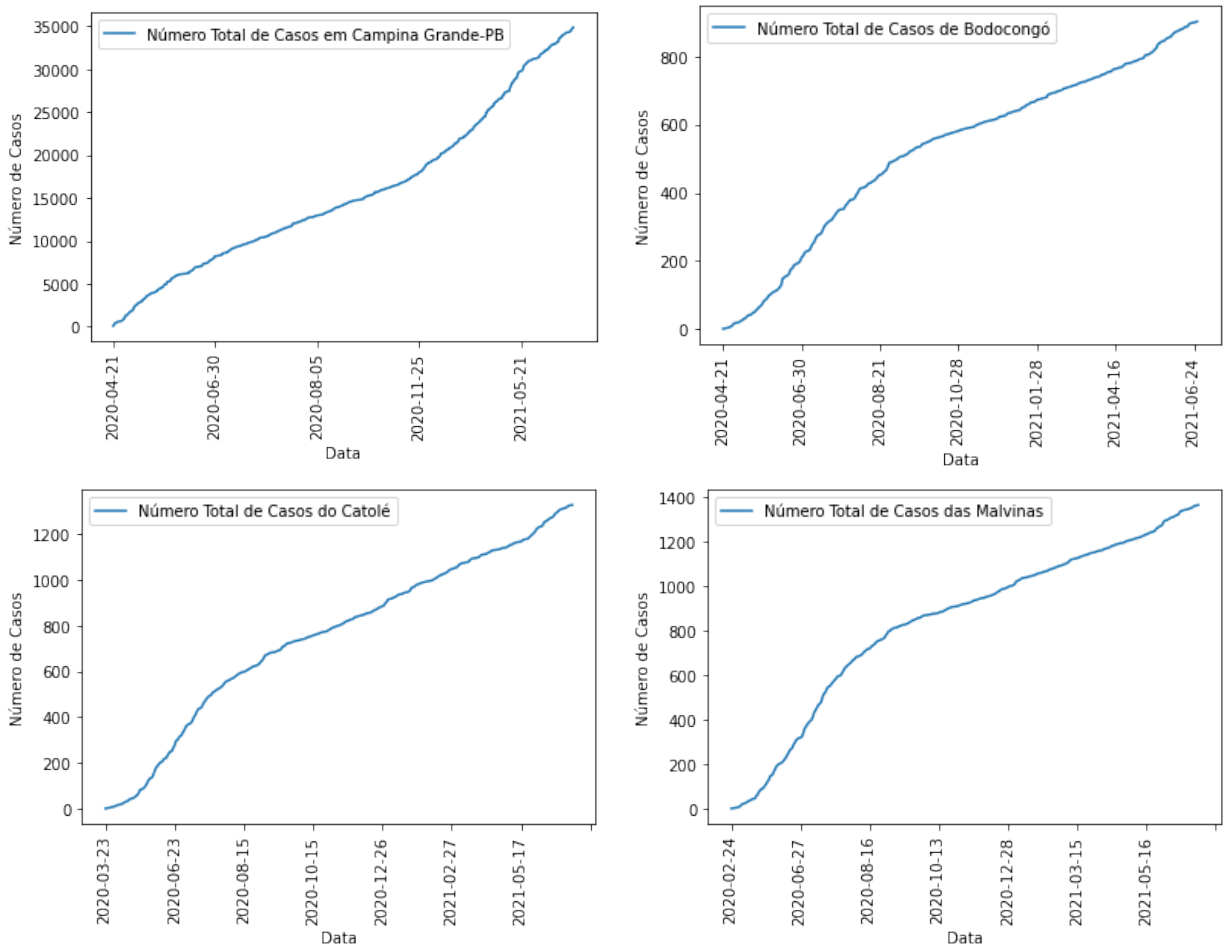
Fonte: Secretária Municipal de Saúde de Campina Grande.

A análise conjunta dos gráficos das Figuras 5.6 e 5.7 reforça a ideia de que, embora sigam uma mesma tendência, em determinados períodos, escalas menores de análise se destacam com um nível de crescimento mais acelerado que as demais, ressaltando a importância de se diminuir a granularidade do estudo para a obtenção de dados mais detalhados. Assim, foi vista a necessidade de se avaliar também a disseminação da doença em diferentes setores censitários da cidade.

Para a análise por bairros, não foi feito o filtro de datas utilizado para as demais bases de dados. Constatou-se que, nessa escala, alguns períodos de dias não constavam na base de dados coletada, o que pode ter ocorrido devido a subnotificações ou pelo fato de realmente não ter havido casos da doença naquele bairro nesses períodos.

A partir dos dados referentes aos setores censitários de Campina Grande, foram obtidos os gráficos que mostram a disseminação da COVID-19 em alguns dos seus bairros mais populosos. Esse gráficos são apresentados na Figura 5.8.

Figura 5.8: Análise das curvas de número de casos acumulados de COVID-19 por bairros de Campina Grande



Fonte: Secretaria Municipal de Saúde de Campina Grande.

Observando-se todas as curvas da Figura 5.8 em conjunto, fica clara a diferença no crescimento dos números da doença em bairros distintos e também de um bairro em relação à cidade. Isso mostra que alguns setores de uma mesma cidade têm crescimento mais acelerado que outros em determinados períodos, evidenciando que a análise numa granularidade maior pode dar uma falsa sensação de controle da doença, que pode estar se disseminando de forma mais acelerada em regiões menores.

Uma outra análise a ser feita é a proporção de casos em relação à população de cada bairro. Com dados do IBGE[2], constatou-se que, ao final do período analisado, o número de

[2]Os dados coletados estão disponíveis no diretório */data* do repositório GitHub[68].

casos de COVID-19 correspondia a cerca de 23% da população dos bairros de Bodocongó e Catolé, enquanto, nas Malvinas, a disseminação da doença foi equivalente a cerca de 12% da sua população.

Com as conclusões dessas duas análises, entende-se que a implantação de políticas públicas que estudassem não só a disseminação da doença no nível municipal, mas também em regiões menores dentro da cidade, avaliando a relação dessa disseminação com a dinâmica adotada em cada bairro, poderia ter resultado em um melhor controle dos casos e óbitos no contexto geral.

5.2 Considerações Finais

Neste capítulo, foram apresentados os dados experimentais utilizados nesta pesquisa, detalhando a estrutura da base de dados da COVID-19 consumida. A descrição das fontes, critérios de seleção e métodos de coleta assegurou a transparência e a robustez metodológica dos dados. As características principais dos dados, incluindo variáveis relevantes e sua organização, foram exploradas de forma a permitir a compreensão do material que sustenta as análises realizadas.

A análise exploratória dos dados proporcionou compreensões relevantes sobre a distribuição e o comportamento da COVID-19 em diferentes níveis de granularidade. Gráficos e análises críticas revelaram como a doença pode apresentar padrões distintos quando observada em escalas variadas, desde o nível nacional até bairros específicos dentro de uma cidade.

As conclusões derivadas dessa análise indicam que políticas públicas mais eficazes poderiam ter sido implementadas se tivessem considerado a disseminação da doença não apenas no nível municipal, mas também em sub-regiões menores, proporcionando uma resposta mais adequada à pandemia.

Portanto, este capítulo fornece conhecimentos consistentes acerca dos dados experimentais que serão utilizados para as análises e discussões seguintes. A compreensão desses dados é importante para o desenvolvimento de modelos e estratégias de controle da doença, assegurando que as conclusões desta tese sejam fundamentadas em uma análise detalhada e confiável.

Parte VI

Resultados

Capítulo 6

Avaliação Crítica de Modelos

As diferenças entre os diversos tipos de modelos utilizados para representar o estado atual de uma pandemia têm sido investigadas a fim de entender seus desempenhos neste contexto. Trabalhos recentes mostraram que os modelos podem ser aplicados em momentos distintos de disseminação de uma doença. Neste capítulo, são apresentados os resultados obtidos na análise de dados da COVID-19 utilizando algoritmos já existentes na literatura. Posteriormente, é feita a análise do desempenho desses algoritmos.

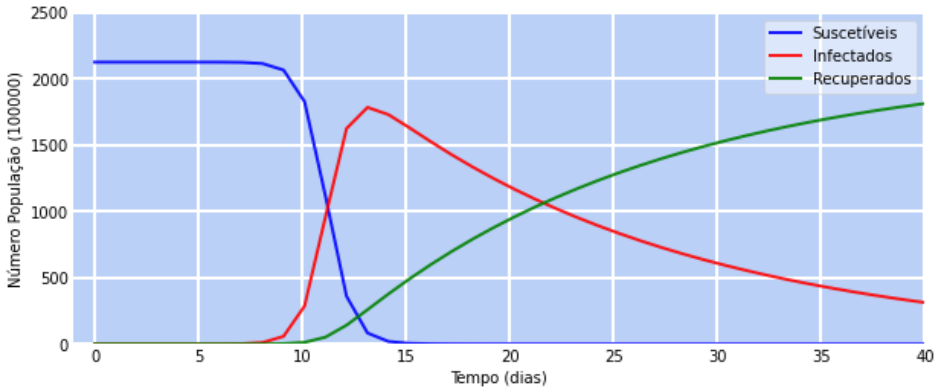
6.1 Modelos Compartimentais

Os primeiros modelos apresentados na literatura para estudos epidemiológicos e entendimento de disseminação de doenças infecciosas foram os modelos compartimentais. Eles representam uma proposta simples e direta para a compreensão do comportamento de epidemias. Sendo assim, para entendimento da propagação da COVID-19, buscou-se inicialmente avaliar a aplicação desses modelos a esse contexto. Para o estudo com base nos modelos compartimentais, as equações referentes a cada um deles foram aplicadas aos seus respectivos códigos, disponíveis no diretório */model-code* do repositório GitHub[68]. Os resultados obtidos são expostos a seguir.

6.1.1 Modelo SIR

As Equações 2.2 a 2.4, referentes ao Modelo SIR, foram aplicadas aos dados de cada uma das escalas analisas neste estudo. As curvas obtidas são apresentadas nas Figuras 6.1 a 6.8.

Figura 6.1: Curvas das funções que caracterizam o modelo SIR aplicadas ao Brasil para taxa de transmissão 1,78



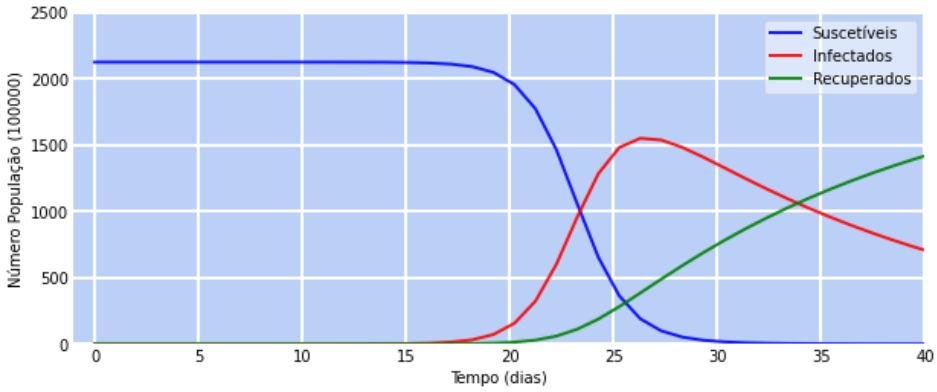
Fonte: Próprio Autor.

Uma das dificuldades encontradas na aplicação do modelo SIR foi a definição da Taxa de Transmissão da COVID-19, um parâmetro determinante para a precisão do modelo. A taxa de transmissão representa o número médio de pessoas para as quais cada indivíduo infectado transmite a doença a cada dia e é inferida por meio de modelos matemáticos epidemiológicos. Essa dificuldade ficou evidente na análise das curvas mostradas nas Figuras 6.1 a 6.3.

A fórmula disponível na literatura[117] para o cálculo da taxa de transmissão resultou em valores muito altos (entre 5 e 7) quando aplicada aos dados coletados das bases utilizadas, o que não condiz com o histórico conhecido da doença. Isso ressalta a complexidade na obtenção dessa taxa para cada período específico de análise, refletindo a necessidade de uma abordagem mais refinada para a estimativa da taxa de transmissão ao longo do tempo.

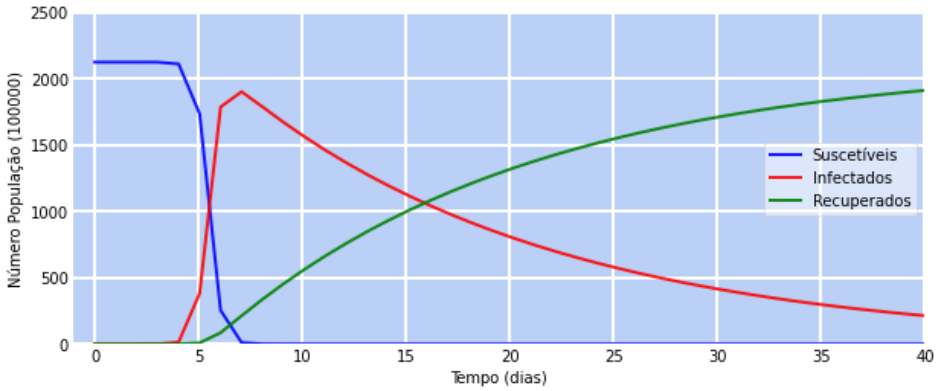
Para fins de análise, adotou-se uma taxa de transmissão de 1,78, baseada em dados históricos da COVID-19 no Brasil. No período inicial da pandemia, a taxa de transmissão no país variou de 1,31 a 2,61. A adoção desse valor resultou nos gráficos da Figura 6.1. Além disso, para avaliação da influência dessa taxa no Modelo SIR, foram arbitrados os valores de 0,89 e 3,56 e, com eles, foram obtidos os gráficos das Figuras 6.2 e 6.3.

Figura 6.2: Curvas das funções que caracterizam o modelo SIR aplicadas ao Brasil para taxa de transmissão 0,89



Fonte: Próprio Autor.

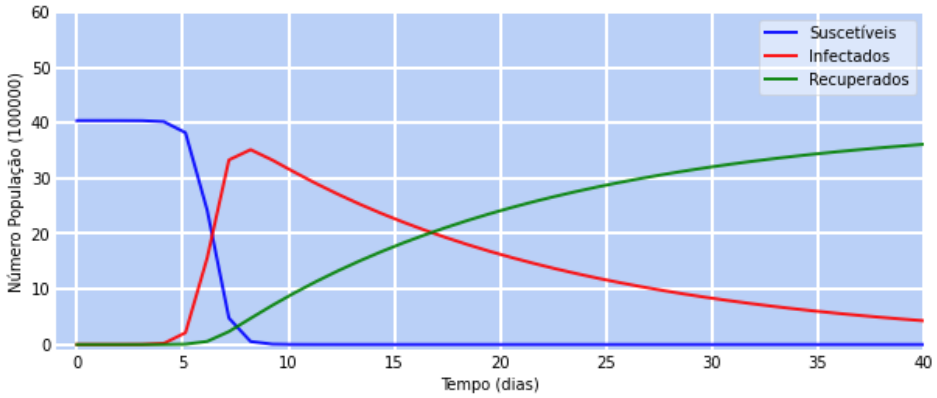
Figura 6.3: Curvas das funções que caracterizam o modelo SIR aplicadas ao Brasil para taxa de transmissão 3,56



Fonte: Próprio Autor.

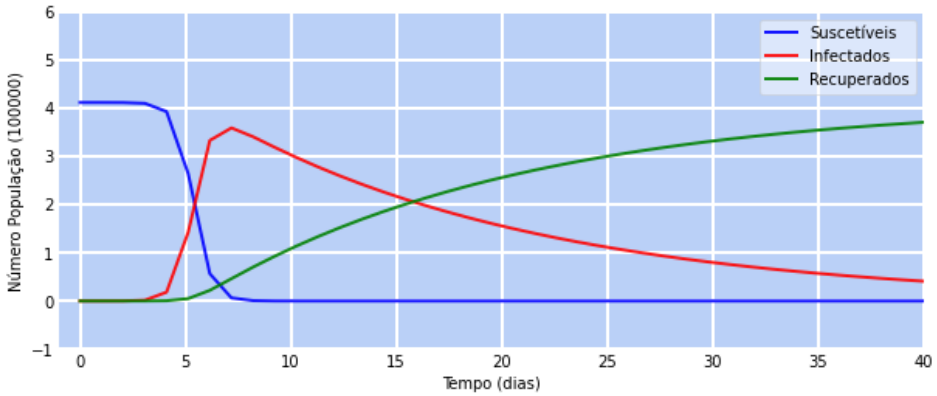
Também com base em dados históricos da pandemia, foi utilizada a Taxa de Transmissão de 2,47 para a Paraíba e para Campina Grande e seus bairros. Foi adotada a mesma taxa para essas escalas, pois não há a divulgação desse dado por cidade ou por regiões censitárias do município. As curvas obtidas para essas escalas são apresentadas nas Figuras 6.4 a 6.8.

Figura 6.4: Curvas das funções que caracterizam o modelo SIR aplicadas à Paraíba



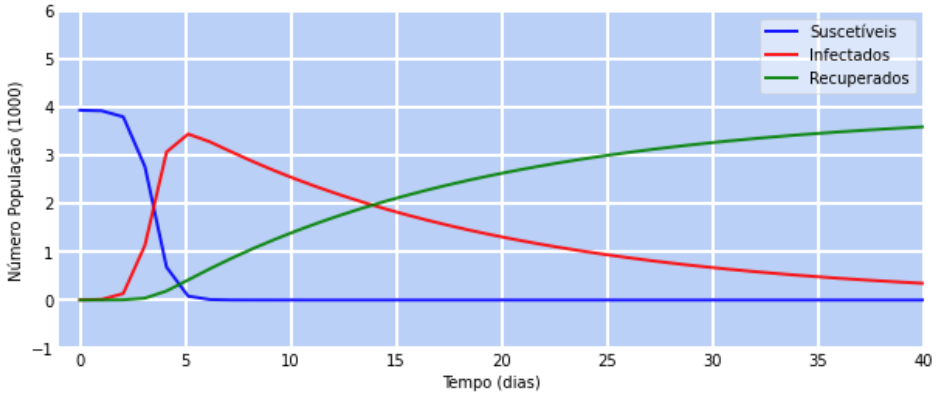
Fonte: Próprio Autor.

Figura 6.5: Curvas das funções que caracterizam o modelo SIR aplicadas a Campina Grande



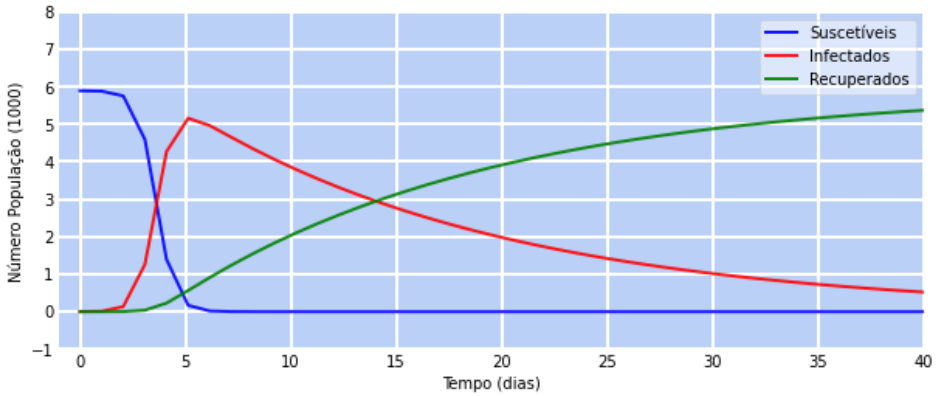
Fonte: Próprio Autor.

Figura 6.6: Curvas das funções que caracterizam o modelo SIR aplicadas ao Bairro Bodocongó



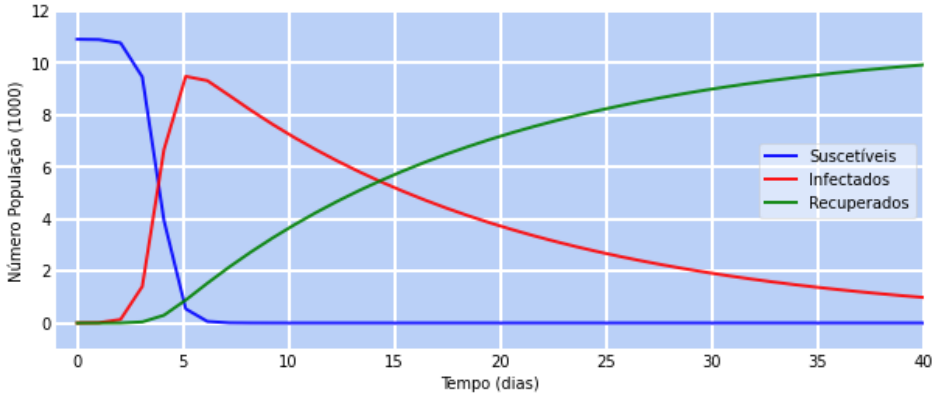
Fonte: Próprio Autor.

Figura 6.7: Curvas das funções que caracterizam o modelo SIR aplicadas ao Bairro Catolé



Fonte: Próprio Autor.

Figura 6.8: Curvas das funções que caracterizam o modelo SIR aplicadas ao Bairro Malvinas



Fonte: Próprio Autor.

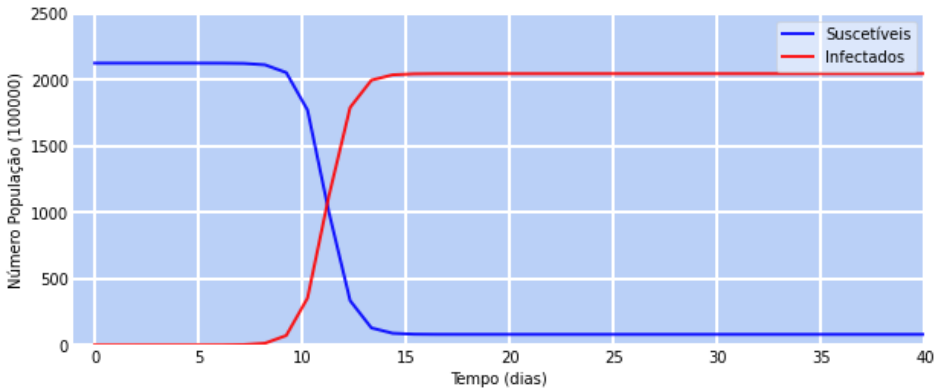
A adoção de uma mesma taxa de transmissão para essas regiões territoriais resultou em curvas com comportamentos semelhantes para diferentes níveis de granularidade. No entanto, conforme já discutido anteriormente, a disseminação da doença ocorre de formas distintas para um mesmo período quando se observam determinadas áreas dentro do mesmo município, por exemplo. Isso reforça a hipótese de que é importante diminuir a granularidade para se fazer o estudo do comportamento da doença.

6.1.2 Modelo SIS

A partir das Equações 2.6 e 2.7, referentes ao modelo SIS, foram obtidas as curvas das Figuras 6.9 a 6.14, para cada uma das escalas analisadas neste estudo. As taxas de

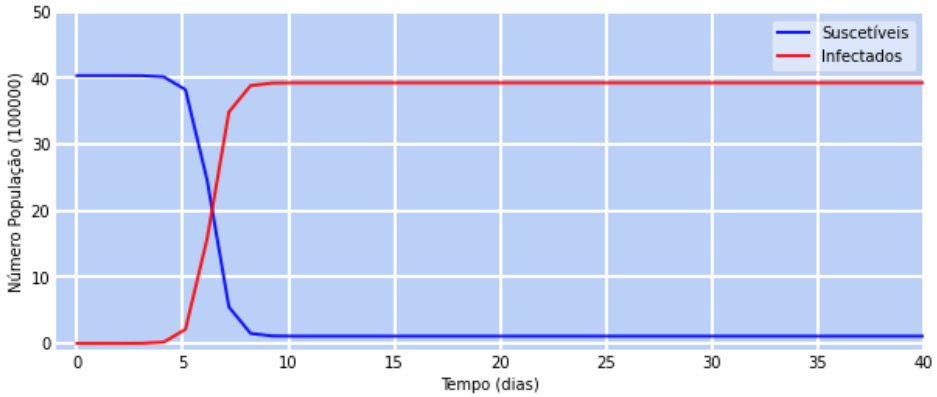
transmissão adotadas para o modelo SIS foram as mesmas utilizadas anteriormente.

Figura 6.9: Curvas das funções que caracterizam o modelo SIS aplicadas ao Brasil



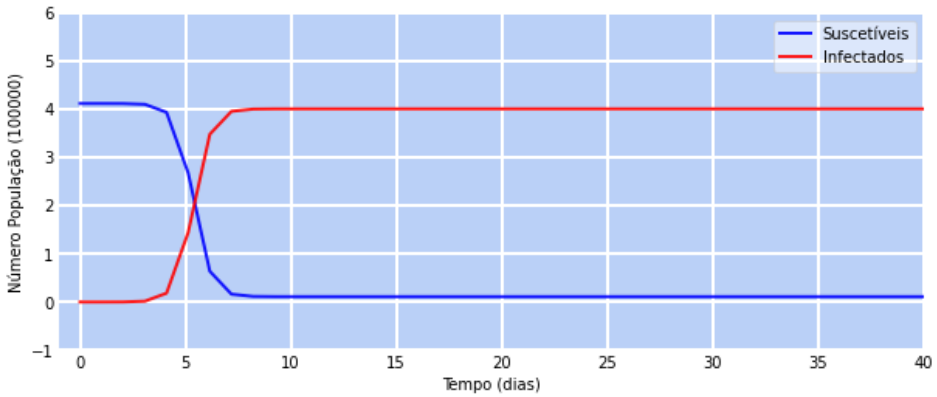
Fonte: Próprio Autor.

Figura 6.10: Curvas das funções que caracterizam o modelo SIS aplicadas à Paraíba



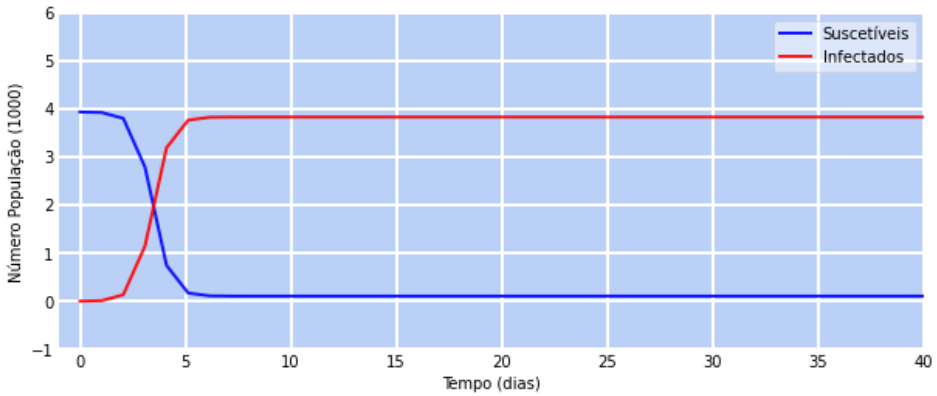
Fonte: Próprio Autor.

Figura 6.11: Curvas das funções que caracterizam o modelo SIS aplicadas a Campina Grande



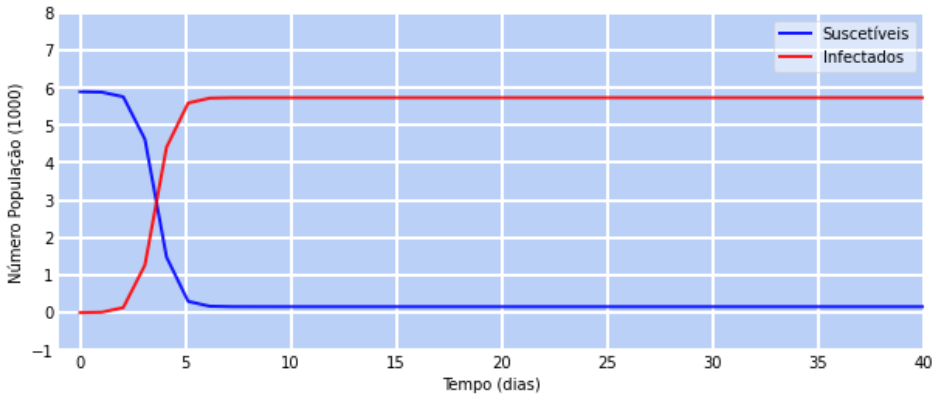
Fonte: Próprio Autor.

Figura 6.12: Curvas das funções que caracterizam o modelo SIS aplicadas ao Bairro Bodocongó



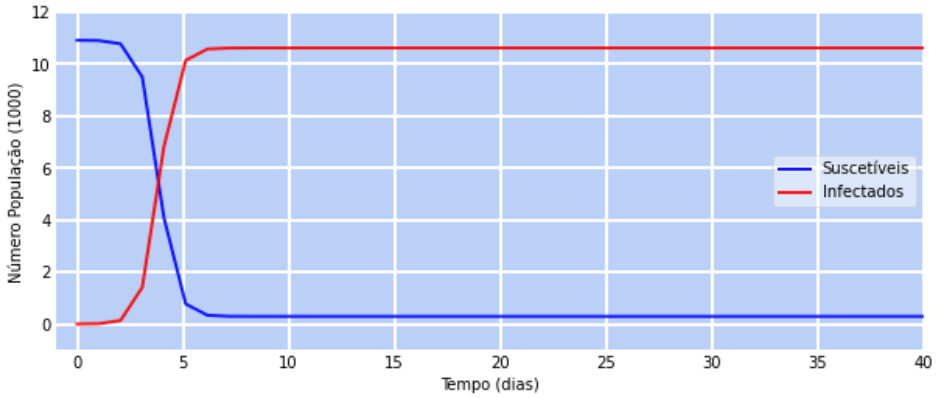
Fonte: Próprio Autor.

Figura 6.13: Curvas das funções que caracterizam o modelo SIS aplicadas ao Bairro Catolé



Fonte: Próprio Autor.

Figura 6.14: Curvas das funções que caracterizam o modelo SIS aplicadas ao Bairro Malvinas



Fonte: Próprio Autor.

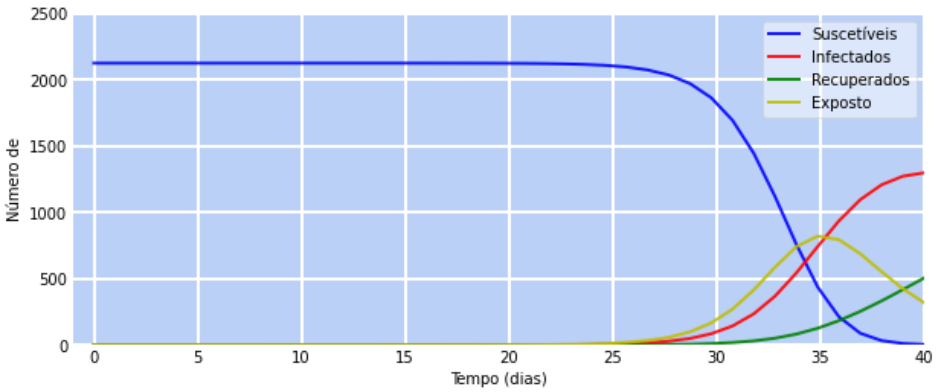
Conforme exposto anteriormente, para o modelo SIS, os indivíduos podem ser suscetíveis ou estar infectados, sendo todo não infectado considerado suscetível à doença. O decrescimento da curva de suscetíveis proporcional ao crescimento da curva de infectados, bem como a estabilização concomitante de ambas as curvas, deixa clara essa relação descrita pelo modelo.

Com todas essas curvas avaliadas paralelamente, tem-se que os resultados são iguais independentemente da região em que se aplica, mesmo para diferentes populações. O comportamento das curvas se repete apresentando apenas uma sutil diferença no dia em que essas curvas se cruzam. Sendo assim, o Modelo SIS mostra resultados que não são significativos para o estudo, portanto não se adequa para esta tese.

6.1.3 Modelo SEIR

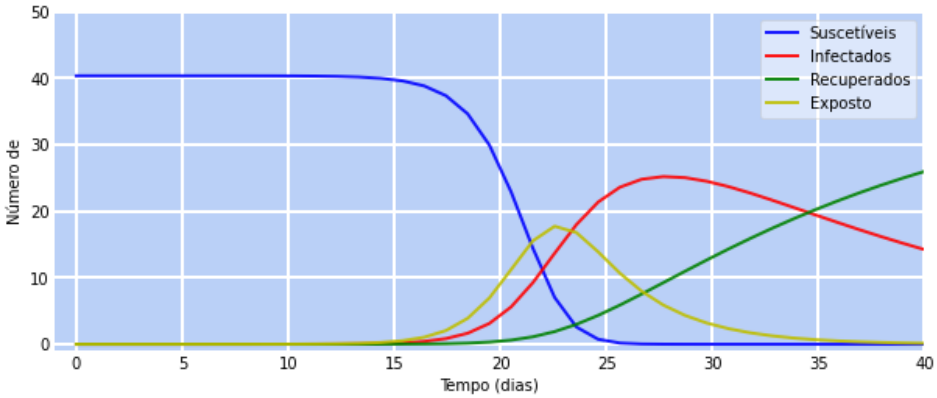
Para o modelo SEIR, adotou-se uma Taxa de Incubação de 0,3 e a População Exposta foi definida como 200 habitantes. A População Exposta Inicial considerada foi de 0 habitantes[117]. Para a Taxa de Transmissão, foram considerados os valores utilizados nos modelos anteriores, baseados em dados históricos. Com a aplicação desses dados às Equações 2.9 a 2.12, referentes a esse modelo, foram obtidas as curvas das Figuras 6.15 a 6.20, para cada uma das escalas analisadas neste estudo.

Figura 6.15: Curvas das funções que caracterizam o modelo SEIR aplicadas ao Brasil



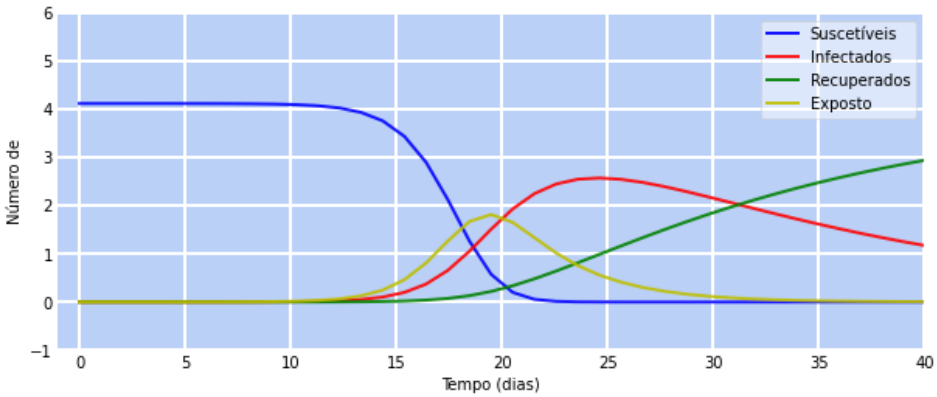
Fonte: Próprio Autor.

Figura 6.16: Curvas das funções que caracterizam o modelo SEIR aplicadas à Paraíba



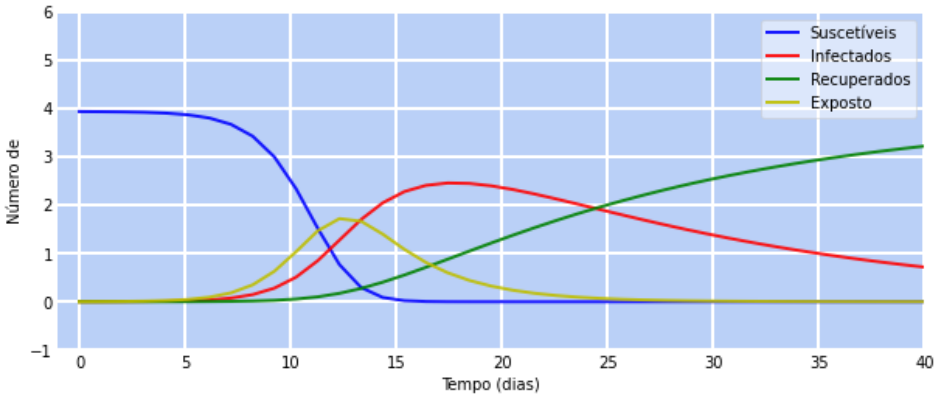
Fonte: Próprio Autor.

Figura 6.17: Curvas das funções que caracterizam o modelo SEIR aplicadas a Campina Grande



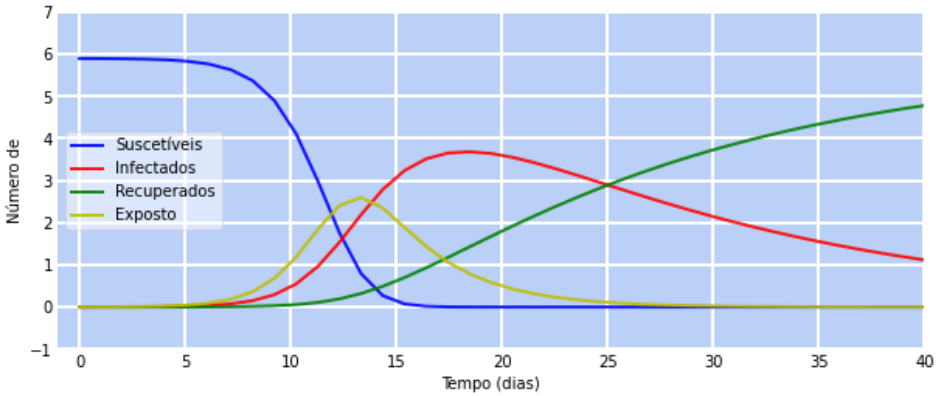
Fonte: Próprio Autor.

Figura 6.18: Curvas das funções que caracterizam o modelo SEIR aplicadas ao Bairro Bodocongó



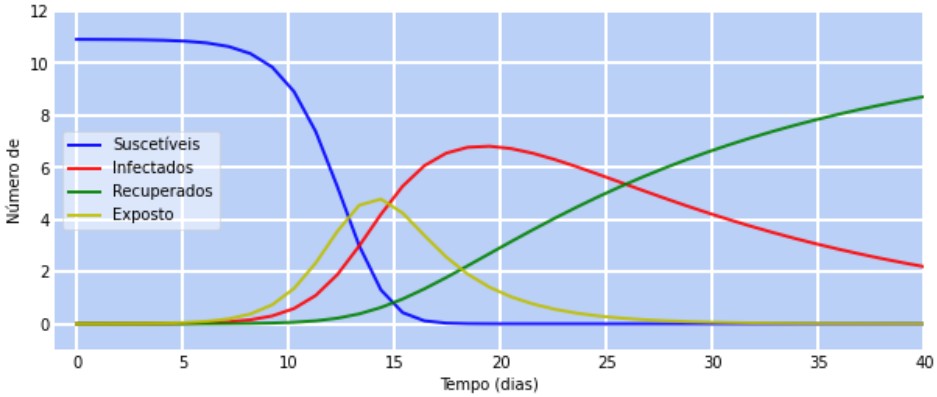
Fonte: Próprio Autor.

Figura 6.19: Curvas das funções que caracterizam o modelo SEIR aplicadas ao Bairro Catolé



Fonte: Próprio Autor.

Figura 6.20: Curvas das funções que caracterizam o modelo SEIR aplicadas ao Bairro Malvinas



Fonte: Próprio Autor.

Os valores aqui utilizados foram fixados para a análise em um determinado período, mas, dadas as características de disseminação de epidemias, é interessante que se considerem taxas dinâmicas, que possam ser extraídas de bases oficiais dos governos, que ainda não são disponíveis. Esse é um ponto a ser levado em consideração para estudos futuros que objetivem a utilização de modelos compartimentais, ou suas variações, para análise de dados epidemiológicos.

6.1.4 Considerações Sobre os Modelos Compartimentais

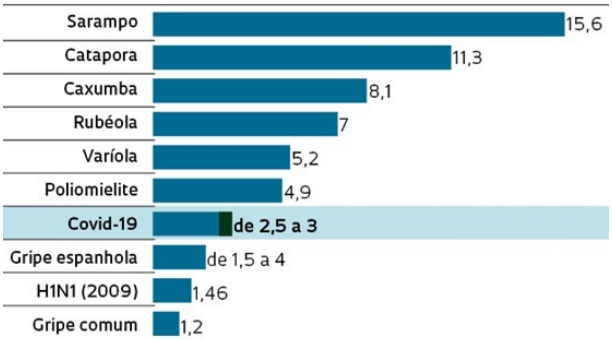
Com a análise da aplicação desses modelos, pode-se constatar que eles preveem valores mais precisos em escalas de tempo curtas, entre cinco e oito dias, porém, conforme a escala

de tempo cresce, erros numéricos se acumulam, provocando uma diminuição na acurácia dos valores obtidos. Assim, conclui-se que os modelos devem ser aplicados em pequenos intervalos de tempo, sendo reajustados ao fim desses curtos períodos, para que apresentem resultados mais satisfatórios para atender ao objetivo de estudo.

Além disso, conforme discutido na Subseção 5.1.1, a Taxa de Transmissão é um fator primordial para a precisão de todos os Modelos Compartimentais. Pequenas variações nesse dado podem resultar em um grande aumento na disseminação da doença e, computacionalmente, afetam o resultado de estudo sobre o comportamento dos modelos, conforme mostrado anteriormente. Dessa forma, é fundamental que esse dado seja determinado com a máxima exatidão possível, porém fatores como base de dados limitada, casos de subnotificação e falta de regularidade na divulgação de informações dificultam a determinação precisa dessa taxa. Essa determinação se torna mais precisa com o enriquecimento da base de dados analisada, uma vez que ela é inferida por meio de modelos matemáticos epidemiológicos que consideram fatores como o número de contato entre pessoas infectadas e suscetíveis, a probabilidade de transmissão por contato e a duração do período infeccioso.

Para definir a Taxa de Transmissão, deve-se inicialmente estabelecer o Número Básico de Reprodução da infecção, que determina o quanto um patógeno é infeccioso em um local em que nenhum indivíduo adquiriu imunidade a ele. Para o vírus causador da COVID-19, o Número Básico de Reprodução gira em torno de 2,5 e 3, ou seja, um infectado contamina de duas a três pessoas em média[118]. A Figura 6.21 mostra o Número Básico de Reprodução de diversas doenças infectocontagiosas.

Figura 6.21: Número Básico de Reprodução de Algumas Enfermidades Infectocontagiosas



Fonte: Zaparolli, Revista pesquisa FAPESP [118]

Para se obter o Número Básico de Reprodução, devem-se considerar fatores como: quantidade de indivíduos suscetíveis com que um infectado tem contato, o risco de transmissão em cada um desses contatos e o tempo médio em que o infectado transmite a doença, que, no caso da COVID-19, é cerca de dois dias antes da apresentação dos primeiros sintomas e dura até mais sete dias[118]. A Taxa de Transmissão, que é o número efetivo de reprodução da doença, é utilizada para analisar a taxa de disseminação ao longo do tempo e equivale ao Número Básico de Reprodução exposto às condições reais de evolução da doença. Devido a isso, essa taxa muda constantemente e depende dos valores do momento em que foi medida, mostrando a relação da sociedade com o agente infeccioso. Para ser usada de forma adequada no acompanhamento da evolução da doença, a Taxa de Transmissão requer informações precisas e constante atualização[118].

Ademais, é importante considerar que, em determinados períodos, a doença apresenta comportamentos diferentes dependendo do local de análise, tendo um crescimento mais acelerado em algumas áreas do que em outras de uma mesma região. Assim, adotar uma mesma Taxa de Transmissão para uma área territorial consideravelmente grande pode mascarar essas diferenças na disseminação da doença que são observadas quando se diminui a granularidade do estudo. Tendo em vista os pontos aqui levantados, conclui-se que a adoção de Taxas de Transmissão diferentes para cada região estudada, levando em consideração a dinâmica da doença específica para aquela área, pode trazer mais exatidão na aplicação dos modelos compartimentais, gerando resultados mais precisos para a análise.

Vale ressaltar, ainda, que os modelos compartimentais, por si só, apresentam limitações inerentes significativas, especialmente no contexto da COVID-19. Por exemplo, o modelo SIR não considera a possibilidade de reinfecção, um aspecto crítico no contexto da COVID-19, dado que a pandemia apresentou várias ondas com novas infecções e o surgimento de novas variantes, como a Delta e a Omicron. Esses eventos comprovam a não eficácia dos modelos compartimentais puros para capturar toda a complexidade da pandemia, principalmente em um cenário heterogêneo como o Brasil. Para aumentar a eficácia dos modelos compartimentais, estudos recentes têm adaptado esses modelos e os combinado com técnicas de aprendizagem de máquina, inteligência artificial e outros modelos matemáticos. Essas abordagens híbridas têm mostrado resultados mais promissores, pois conseguem incorporar variáveis adicionais e lidar melhor com a heterogeneidade dos dados e a variabilidade das

taxas de transmissão.

No caso do Brasil, um país com grande diversidade e diferenças regionais significativas, a aplicação de modelos compartimentais exige um monitoramento eficaz e contínuo das taxas de transmissão, o que é desafiador em contextos que apresentam subnotificação e inconsistências na divulgação de dados, já que o controle e o monitoramento da pandemia baseiam-se nessas taxas, que devem ser precisas e atualizadas regularmente para refletir as condições reais da população.

Por fim, destaca-se que, embora os modelos compartimentais ofereçam uma base importante para o estudo epidemiológico, sua aplicação efetiva na prática requer adaptações e complementações com outras técnicas para capturar a complexidade e dinamismo da disseminação da COVID-19.

6.2 Modelo de Regressão Aditiva

O *Prophet*, software utilizado para a execução do Modelo de Regressão Aditiva, é uma ferramenta para a produção de séries temporais em larga escala, reconhecida por sua eficácia em séries que apresentam fortes efeitos sazonais e em contexto em que um grande número de séries precisa ser modelado por analistas com diversos níveis de expertise em métodos estatísticos. Ele é especialmente útil para lidar com dados faltantes e mudanças bruscas na tendência de crescimento ou queda de casos. Sua principal motivação consiste na observação de que técnicas de previsão totalmente automáticas podem ser inflexíveis e difíceis de ajustar, enquanto analistas com profundo conhecimento de domínio frequentemente não possuem o conhecimento necessário em modelagem por meio de séries temporais[53].

Esse software se baseia em um modelo de série temporal que pode ser decomposto, que consiste em três componentes principais combinadas de forma aditiva: tendência, sazonalidade e feriados/eventos. O *Prophet* presume uma taxa de crescimento constante com identificação automática de pontos de inflexão e, por padrão, modela efeitos sazonais periódicos por meio de séries de Fourier, além de demandar a discriminação de todas as ocorrências de feriados, capturando as influências desses eventos em séries temporais que não seguem padrões periódicos constantes. O modelo é projetado com parâmetros intuitivos e de fácil interpretação, permitindo que analistas com conhecimento de domínio, mas sem expertise

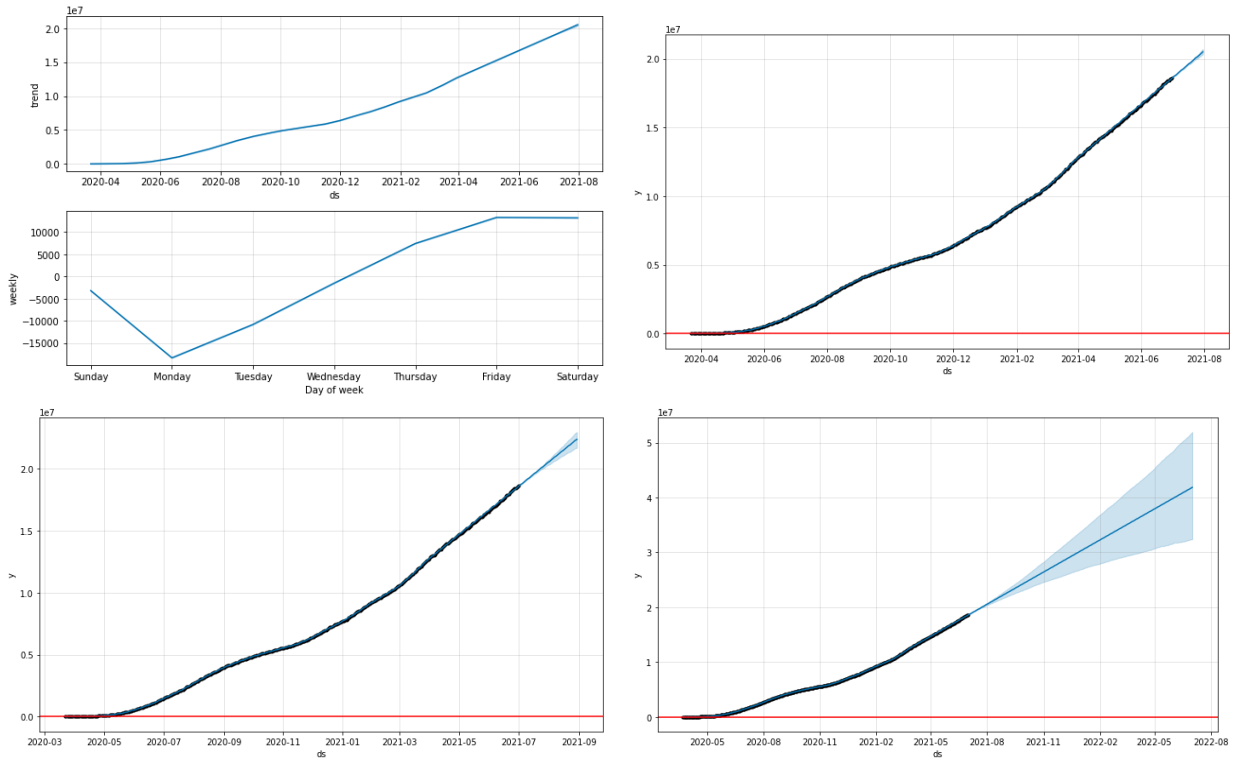
aprofundada em estatística de séries temporais, possam ajustar o modelo para incorporar seu conhecimento e expectativas.

Para o ajuste do modelo, é feita a maximização da estimativa a posteriori, integrando a incerteza sobre os parâmetros do modelo à incerteza das previsões, que são obtidas por meio da simulação de mudanças futuras na taxa de crescimento de acordo com dados históricos, considerando possíveis alterações de tendência. Ao enquadrar o problema de previsão como um exercício de ajuste de curva, o *Prophet* sacrifica algumas vantagens inferenciais de modelos generativos, que explicitamente modelam a dependência temporal nos dados. A análise do desempenho das previsões utiliza previsões históricas simuladas em vários pontos passados para simular a performance fora da amostra, ajudando a detectar erros.

Durante a pandemia de COVID-19, vários fatores contribuíram para a ausência de dados nas bases analisadas, como subnotificação e falta de divulgação de registros nos fins de semana e feriados. Além disso, o comportamento da população frente à pandemia ocasionou explosões de casos em determinados períodos, resultando em mudanças abruptas nos números registrados. Diante dessas características, a utilização do *Prophet* para analisar o comportamento dos casos da doença utilizando as bases de dados coletadas se mostrou apropriada.

As Figuras 6.22 a 6.27 ilustram o modelo de previsão do *Prophet* aplicado às bases de dados alvos deste estudo. O primeiro quadrante de cada figura apresenta as tendências (geral e semanal) da COVID-19 nas diferentes escalas estudas. Os demais quadrantes mostram, em preto, as curvas reais da doença e, em azul, as previsões do *Prophet* para períodos de 30 dias, 60 dias e 1 ano, respectivamente.

Figura 6.22: Curvas retornadas pelo *Prophet* para o número de casos acumulados de COVID-19 no Brasil

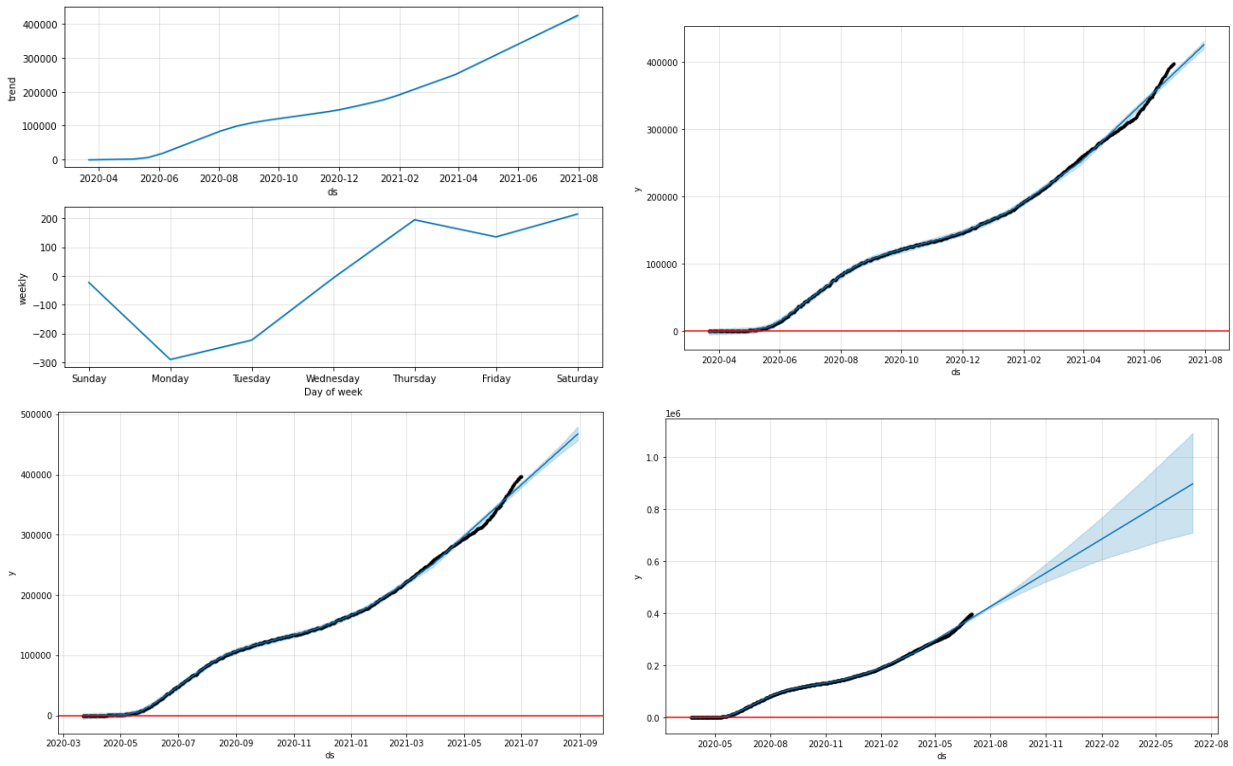


Fonte: Próprio Autor.

Na Figura 6.22, referente às previsões para o Brasil, observa-se que, a partir da previsão para 60 dias, o modelo já apresenta ruído. Para a previsão de 1 ano, esse ruído se intensifica, indicando uma certa instabilidade no modelo.

Outro ponto relevante é que, embora siga a tendência de crescimento observada nos dados reais, o resultado da previsão não é totalmente satisfatório, pois o *Prophet* projeta um crescimento linear, enquanto os casos reais mostram flutuações que não são capturadas pelo modelo.

Figura 6.23: Curvas retornadas pelo *Prophet* para o número de casos acumulados da Paraíba



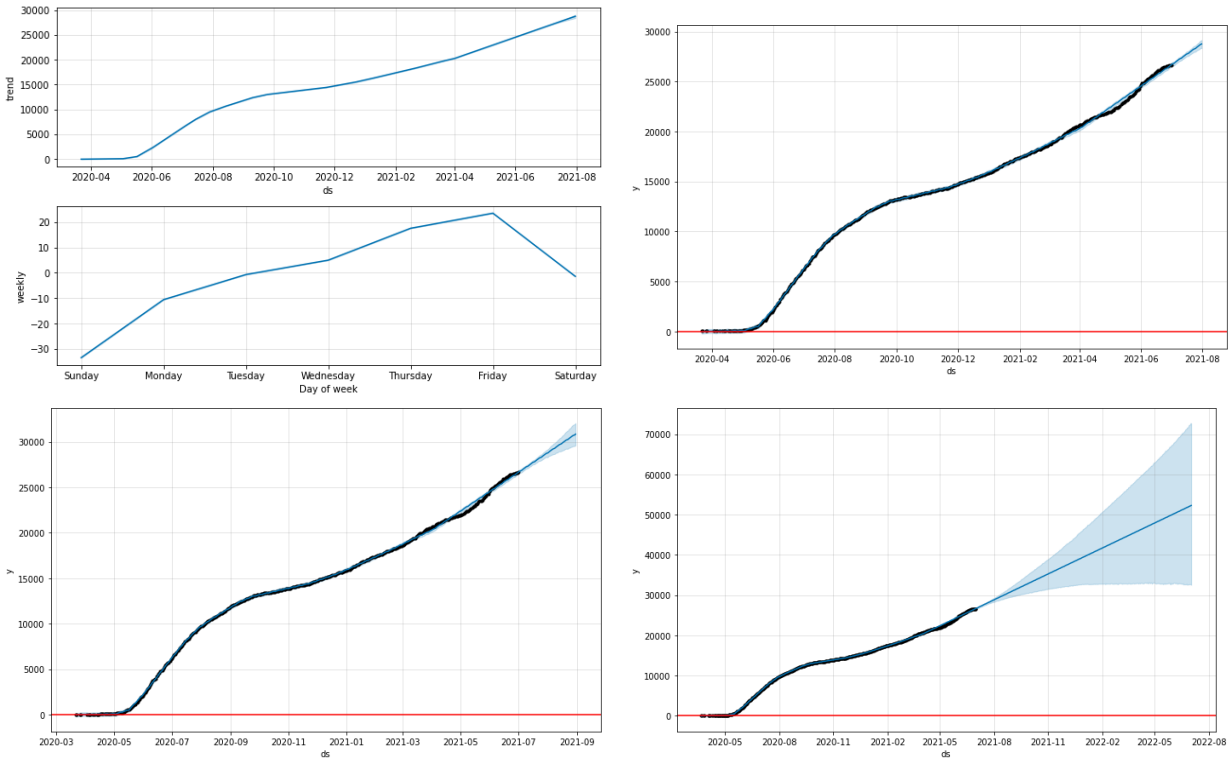
Fonte: Próprio Autor.

A análise das curvas obtidas para a Paraíba, mostradas na Figura 6.23, reforça essa observação: mesmo diante de flutuações mais acentuadas nos dados reais, a previsão do *Prophet* continua a seguir uma tendência linear.

Essa conclusão também se aplica aos resultados referentes a Campina Grande (Figura 6.24) e aos bairros da cidade (Figuras 6.25 a 6.27), em que se observa o mesmo comportamento do modelo de previsão do *Prophet*, independentemente das variações nos números de casos nas curvas reais dessas regiões.

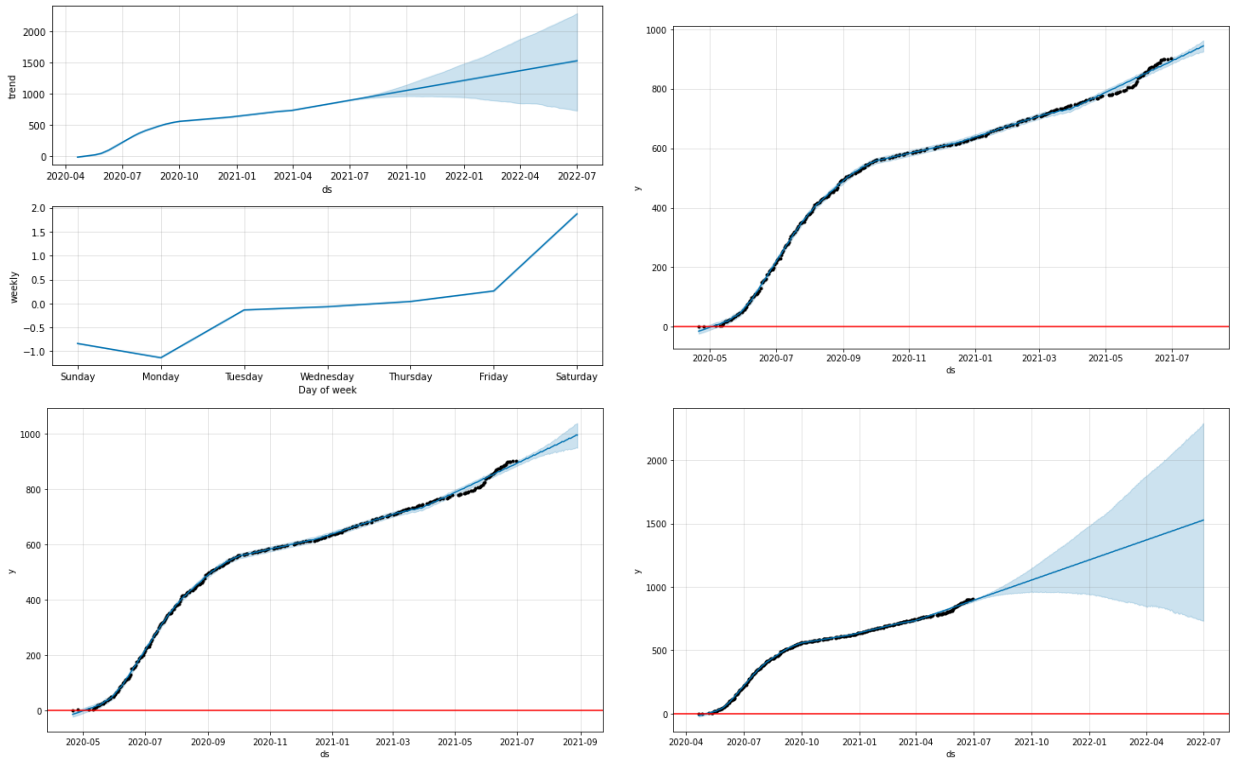
Com isso, apesar de sua capacidade de lidar com dados faltantes e mudanças abruptas no comportamento dos dados e da obtenção de previsões razoáveis para períodos mais curtos, observa-se que, para previsões de longo prazo, o *Prophet* pode não capturar adequadamente a complexidade e as flutuações dos dados reais. Para trabalhos futuros que pretendam utilizar este modelo para análise de dados epidemiológicos, uma perspectiva a ser considerada seria a combinação do *Prophet* com outros modelos que se ajustem melhor a não-linearidades e variações sazonais, podendo proporcionar previsões mais acertadas.

Figura 6.24: Curvas retornadas pelo *Prophet* para o número de casos acumulados de Campina Grande



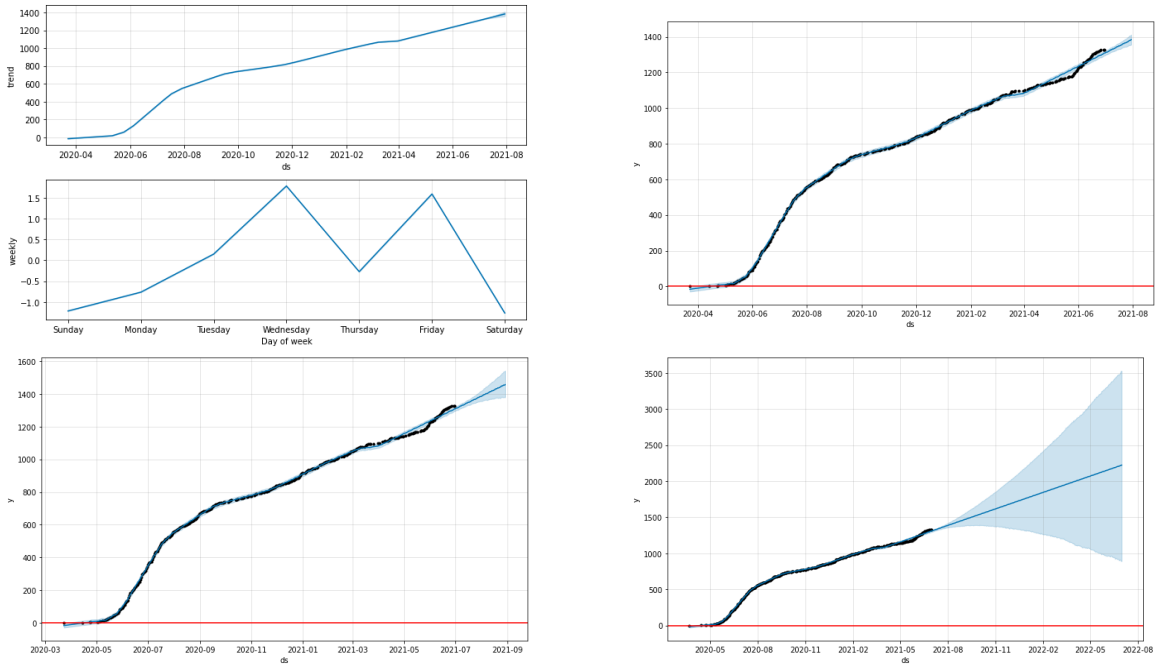
Fonte: Próprio Autor.

Figura 6.25: Curvas retornadas pelo *Prophet* para o número de casos acumulados do bairro Bodocongó



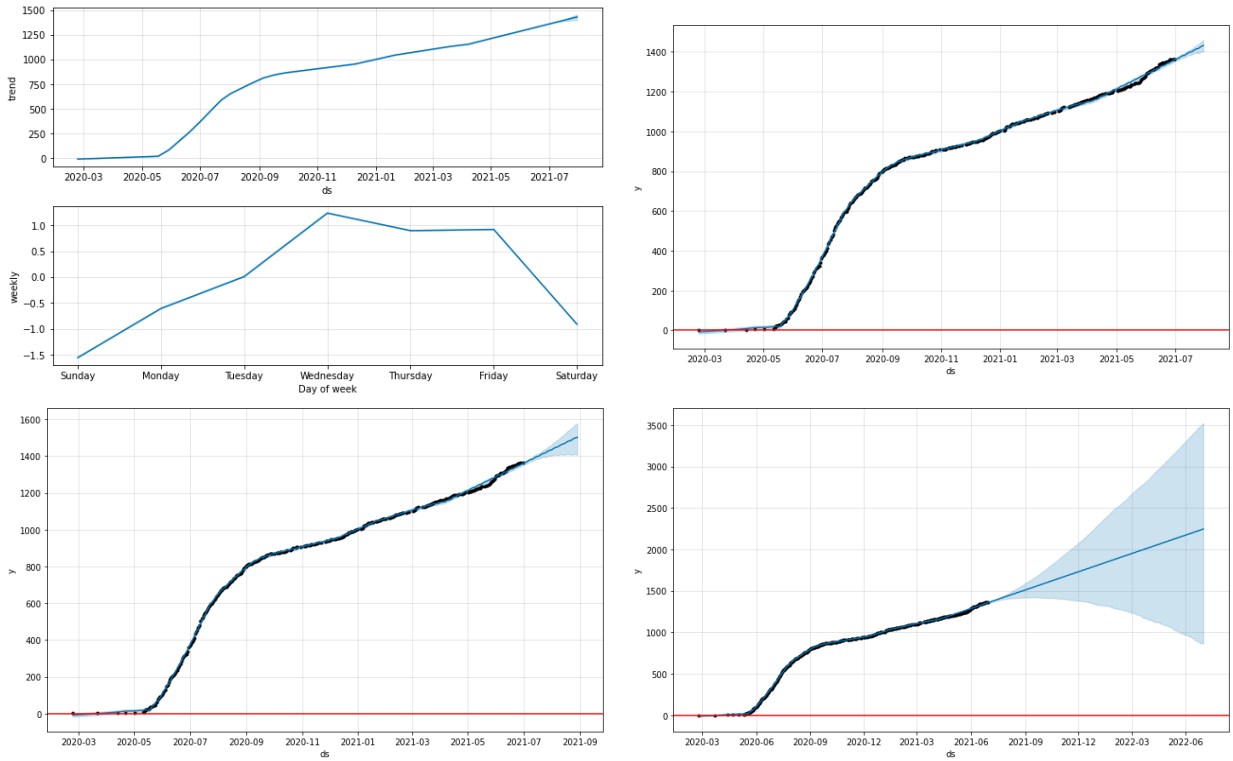
Fonte: Próprio Autor.

Figura 6.26: Curvas retornadas pelo *Prophet* para o número de casos acumulados do bairro Catolé



Fonte: Próprio Autor.

Figura 6.27: Curvas retornadas pelo *Prophet* para o número de casos acumulados do bairro Malvinas



Fonte: Próprio Autor.

6.2.1 Considerações Sobre o Modelo de Regressão Aditiva

A ferramenta *Prophet*, desenvolvida pelo *Facebook*, foi inicialmente considerada para a previsão de séries temporais relativas à COVID-19 devido à sua capacidade de lidar com dados faltantes e mudanças bruscas nas tendências de crescimento ou queda de casos, que são características comuns em dados epidemiológicos como os da pandemia de COVID-19. Diferentemente de modelos de previsão puramente automáticos que podem operar como uma "caixa-preta", o *Prophet* é apresentado como uma abordagem que combina modelos configuráveis com a intervenção do analista-no-circuito. Seu *design* modular, baseado em uma decomposição da série temporal em componentes de tendência, sazonalidade e feriados/eventos, visa oferecer parâmetros intuitivos que podem ser ajustados com base no conhecimento de domínio, mesmo por analistas sem especialização em modelagem de séries temporais.

Apesar dessas vantagens intrínsecas ao *design* do *Prophet*, é importante destacar algumas limitações observadas na sua aplicação em contextos epidemiológicos. Embora o modelo demonstre robustez para previsões de curto prazo e ajuste de sazonalidades, ele não considera explicitamente os mecanismos de transmissão de doenças, o que é um aspecto primordial quando se trata de modelos epidemiológicos, em que a dinâmica da transmissão pode afetar significativamente a precisão das previsões.

O *Prophet* foi utilizado como uma abordagem alternativa para analisar o comportamento da COVID-19 nas bases de dados coletadas. Os resultados mostraram que, embora o modelo tenha acompanhado a tendência geral de crescimento observada nos dados reais, ele tende a projetar um crescimento linear que não captura adequadamente as flutuações e complexidades dos dados reais. Este comportamento foi observado nas previsões para todas as escalas estudadas.

A capacidade do *Prophet* de lidar com dados faltantes e mudanças abruptas permitiu a obtenção de previsões razoáveis para períodos mais curtos. No entanto, para previsões de longo prazo, a linearidade das previsões introduz ruído e instabilidade, resultando em previsões menos precisas. Esse comportamento foi evidenciado pela análise das previsões para 60 dias e 1 ano, em que o ruído e a divergência em relação aos dados reais se tornaram mais pronunciados.

Fatores inerentes à formulação do *Prophet* podem justificar sua inadequação ao contexto estudado nesta pesquisa. A dinâmica de propagação de uma doença infecciosa é influenciada por uma complexa interação de fatores biológicos, comportamentais e sociais que não são diretamente modelados pela estrutura do *Prophet*, que enquadra a previsão como um problema de ajuste de curva no tempo. Além disso, sua capacidade de capturar mudanças abruptas e não-lineares impulsionadas por dinâmicas epidemiológicas complexas, como o surgimento de novas variantes ou o impacto de intervenções de saúde pública, mostrou-se limitada.

O modelo base do *Prophet* modela a série temporal principalmente em função do tempo e de eventos definidos pelo usuário, não integrando de forma nativa variáveis exógenas fundamentais em epidemiologia, como taxas de vacinação, níveis de testagem, medidas de distanciamento social ou mobilidade populacional, que exercem um impacto direto e significativo na evolução da pandemia.

Em resumo, o *Prophet* representa uma abordagem pragmática para a previsão em escala, combinando um modelo flexível e ajustável com ferramentas de avaliação automatizada para otimizar o uso do conhecimento de domínio dos analistas, mesmo com experiência limitada em modelagem por séries temporais. Sua arquitetura modular e a ênfase na interpretabilidade dos parâmetros o tornam uma ferramenta valiosa para organizações que necessitam gerar um grande volume de previsões confiáveis e de alta qualidade. Contudo, é importante considerar a sua natureza de "ajuste de curva", que pode limitar a inferência sobre a estrutura temporal subjacente aos dados. Adicionalmente, a estimativa da incerteza da tendência assume uma continuidade nos padrões de mudança da taxa de crescimento, o que pode não se sustentar em cenários de rupturas estruturais significativas.

Além disso, vale ressaltar que a busca na literatura revelou a ausência de artigos relevantes que utilizem o *Prophet* especificamente para análise de dados epidemiológicos de COVID-19, indicando que, embora os resultados obtidos sugiram que essa ferramenta possa ser valiosa para análises preliminares de dados com sazonalidade, essa abordagem é pouco explorada nesse contexto.

Para trabalhos futuros, é recomendável considerar a combinação do *Prophet* com outros modelos que capturem melhor as não-linearidades e variações sazonais dos dados epidemiológicos. Modelos híbridos ou o ajuste de hiperparâmetros específicos do *Prophet* podem proporcionar previsões mais acuradas, especialmente em cenários de longo prazo.

Finalmente, o uso do *Prophet* para prever a evolução dos casos de COVID-19 nas diversas escalas geográficas analisadas proporcionou uma visão inicial valiosa, mas indicou a necessidade de abordagens complementares e mais especializadas para obter previsões mais confiáveis e informativas. A exploração de modelos híbridos e o ajuste fino dos parâmetros do *Prophet* são caminhos promissores para futuras análises.

6.3 Modelo de Regressão por Processo Gaussiano

O modelo GPR utiliza informações conhecidas para prever o comportamento futuro dos dados. Esse modelo é treinado utilizando dados históricos, como datas e números de casos, que são extraídos diretamente da base de dados e devem estar formatados da maneira adequada para garantir a captura correta das informações necessárias. Em linhas gerais,

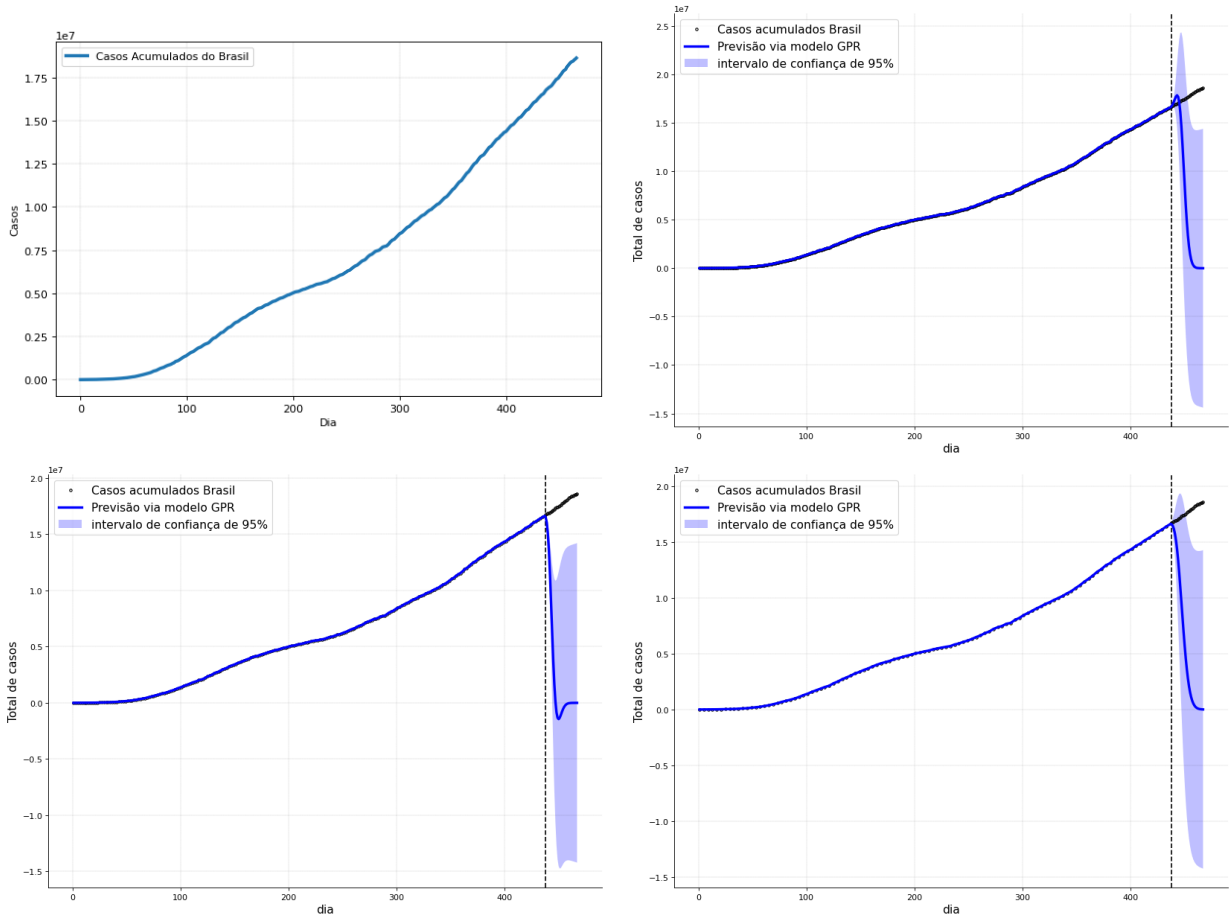
o GPR se baseia nas relações entre os dados para realizar suas previsões; diferentemente dos Modelos Compartimentais, que utilizam parâmetros como a taxa de reprodução e a população para fazer inferências.

Além da quantidade de amostras de treinamento, o fator primordial do modelo GPR é o *kernel*, que é essencial para capturar o comportamento da série temporal. Para uma análise inicial, adotou-se uma composição padrão do *kernel*[1] para todas as aplicações do modelo. O *kernel* e suas diferentes composições serão discutidos em um tópico mais adiante, mas, para o entendimento inicial do modelo GPR, escolheu-se um *kernel* que, em uma breve avaliação, mostrou-se adequado para as diferentes escalas estudadas.

As Figuras 6.28 a 6.33 ilustram a aplicação do modelo GPR a cada uma das escalas analisadas neste estudo. Em todas essas figuras, o primeiro quadrante representa a curva real de casos acumulados para a escala dada. Nos demais quadrantes, alguns dados são suprimidos do fim da base de dados: o segundo quadrante exclui os últimos 30 dias, o terceiro quadrante representa 50% dos dados, e o quarto quadrante contém apenas 25% dos dados totais. A única variação feita para a obtenção de cada uma das curvas nesses quadrantes é o número de amostras oferecidas ao modelo GPR, que utiliza apenas os dados acumulados para gerar seus resultados.

[1]Para todos os casos dessa primeira análise, utilizou-se o *kernel* codificado como $kernel = ConstantKernel() * RBF() * DotProduct(sigma_0 = 0) * ConstantKernel(constant_value = 0.005)$

Figura 6.28: Curvas fornecidas pelo modelo GPR para o número de casos acumulados de COVID-19 no Brasil



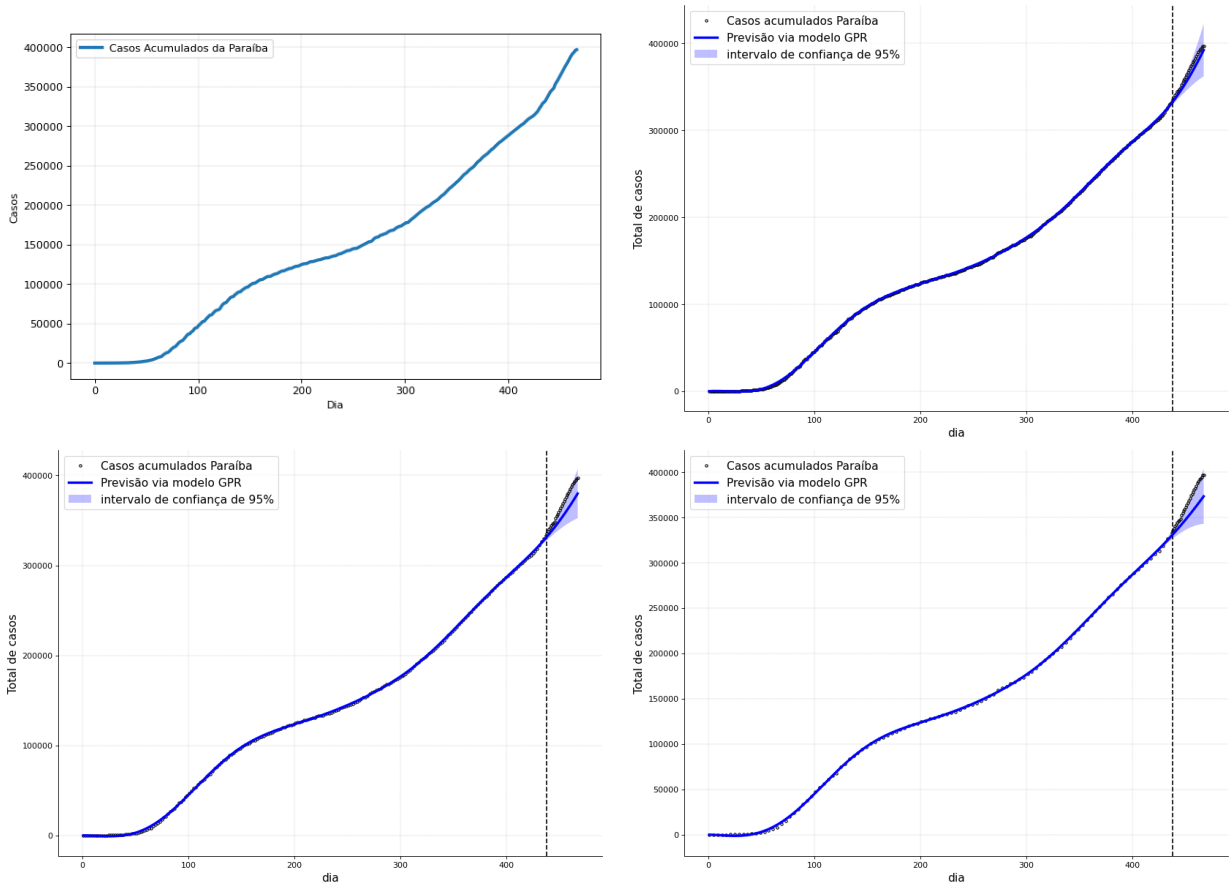
Fonte: Próprio Autor.

A análise dos resultados da aplicação do Modelo GPR aos dados do Brasil mostra que a exclusão de 30 dias da base de dados já introduz considerável ruído no intervalo de confiança. Um intervalo de confiança muito amplo indica a introdução de muito ruído e a dificuldade de fazer uma previsão que condiga com a realidade. Isso sugere que o *kernel* utilizado não foi adequado para captar a variação temporal dos dados brasileiros, podendo não ser capaz de modelar eficientemente a complexidade da série temporal.

Em suma, o modelo não se adequou ao estudo em relação ao Brasil. Observando as curvas, percebe-se que a previsão não segue a tendência dos casos reais: enquanto o número de casos acumulados está em crescimento, o modelo prevê uma queda abrupta. Nos segundo e quarto quadrantes da Figura 6.28, o intervalo de confiança coincide brevemente com os

casos reais, mas está, na maior parte do tempo, muito distante da realidade. Nesse caso, seria indicado avaliar a possibilidade de utilização de um outro *kernel* que melhor se ajuste à variação dos dados brasileiros, reduzindo o ruído e melhorando a precisão das previsões.

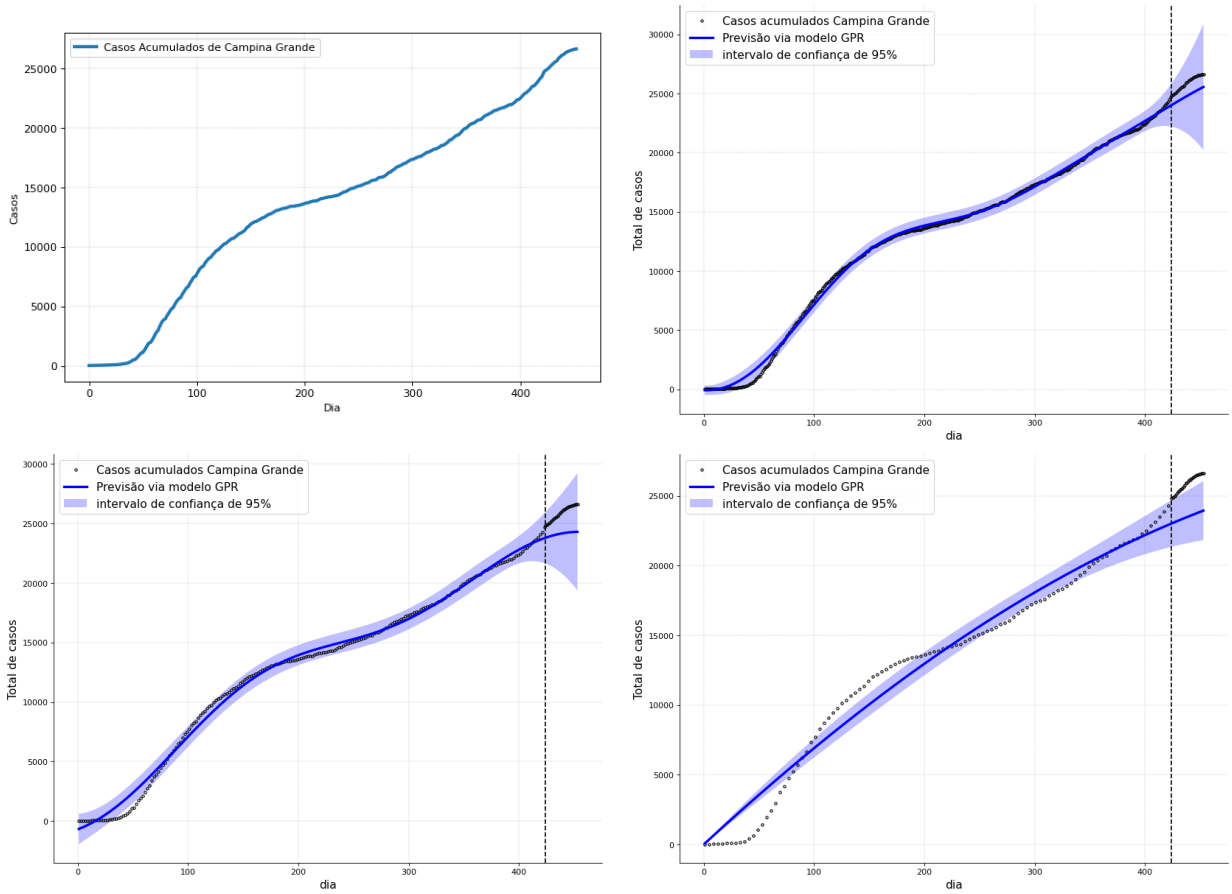
Figura 6.29: Curvas fornecidas pelo modelo GPR para o número de casos acumulados de COVID-19 na Paraíba



Fonte: Próprio Autor.

Para a aplicação do Modelo GPR à base de dados da Paraíba, os mesmos parâmetros, estrutura do modelo e *kernel* foram utilizados. Observando a Figura 6.29, percebe-se que o GPR se adequou melhor nesse caso, ajustando-se melhor à quantidade de dados referentes ao estado. A exclusão dos últimos 30 dias da base de dados da Paraíba resulta em uma previsão via GPR que praticamente se iguala ao caso real. Conforme mais dados são suprimidos, adiciona-se mais ruído, e o modelo começa a se distanciar da realidade, mas ainda segue a tendência de crescimento real dentro do intervalo de confiança.

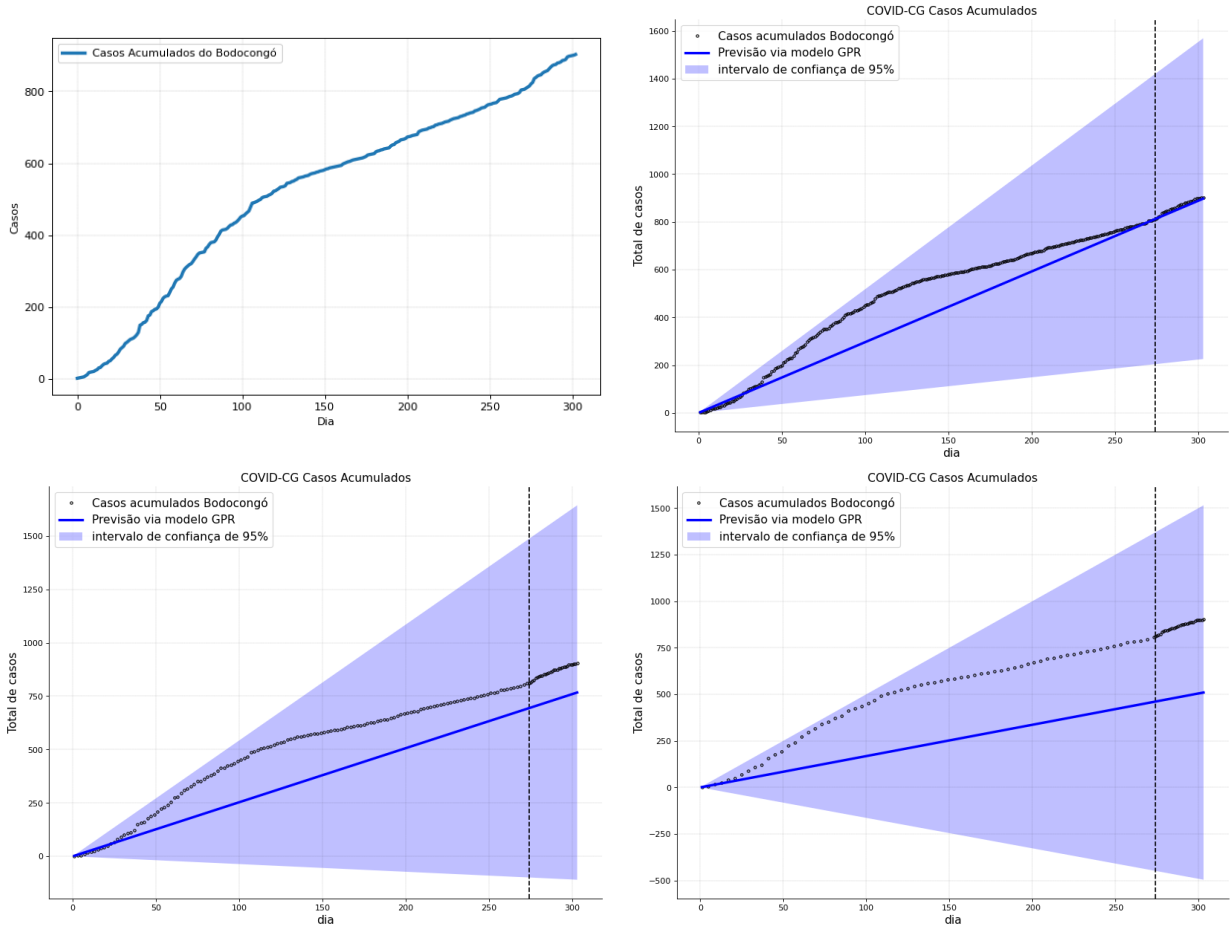
Figura 6.30: Curvas fornecidas pelo modelo GPR para o número de casos acumulados de COVID-19 em Campina Grande



Fonte: Próprio Autor.

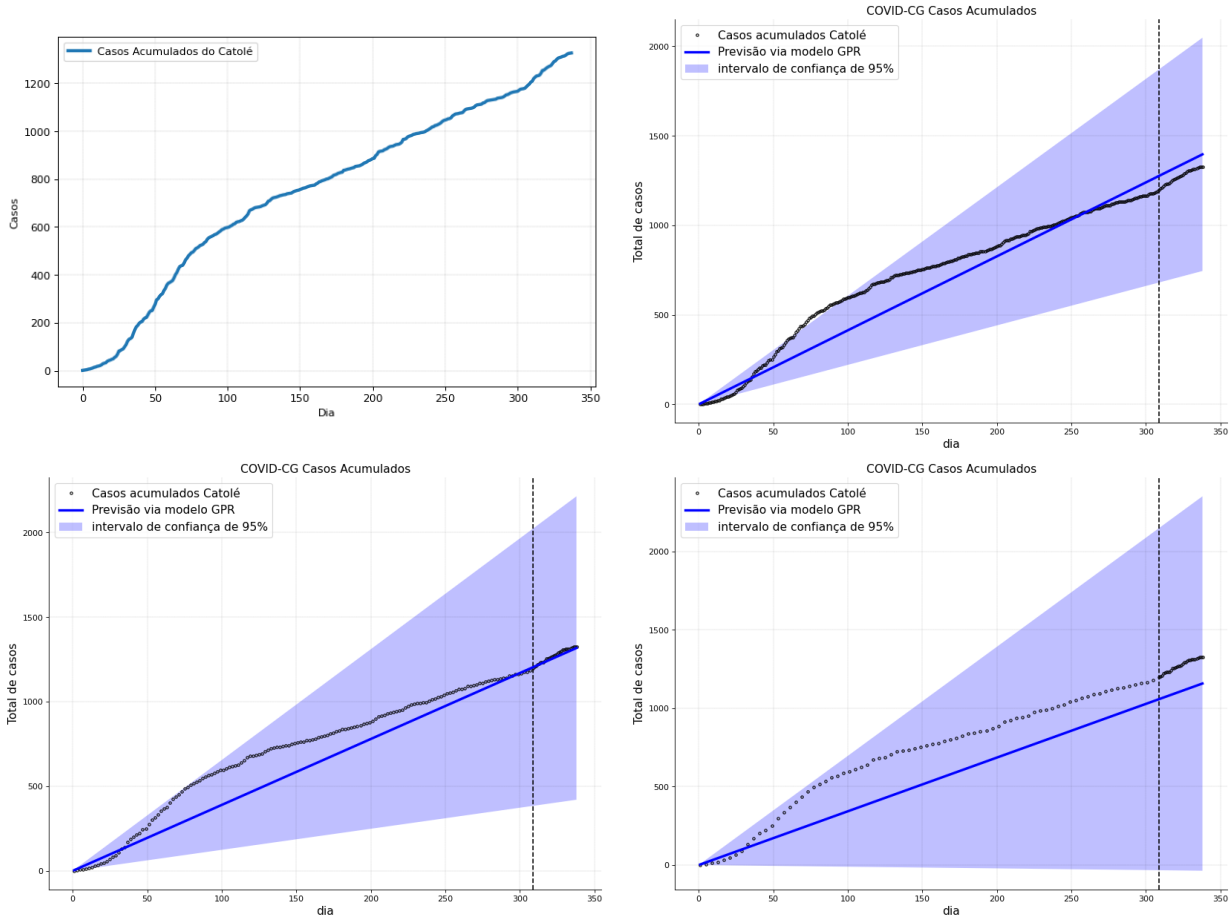
Enquanto a previsão via modelo GPR permanece dentro do intervalo de confiança, o resultado do modelo pode ser considerado satisfatório. Analisando a Figura 6.30, percebe-se que, conforme são excluídos dados da base, o modelo se distancia consideravelmente do caso real e, a partir do terceiro quadrante, começa a sair do intervalo de confiança em alguns pontos. Isso indica que, quando a base de dados fica muito pequena, o modelo já não responde de forma adequada. O intervalo de confiança largo sugere que o *kernel* padrão adotado talvez não seja o mais adequado para capturar a variação temporal dos dados com menos pontos.

Figura 6.31: Curvas fornecidas pelo modelo GPR para o número de casos acumulados de COVID-19 no bairro Bodocongó, em Campina Grande



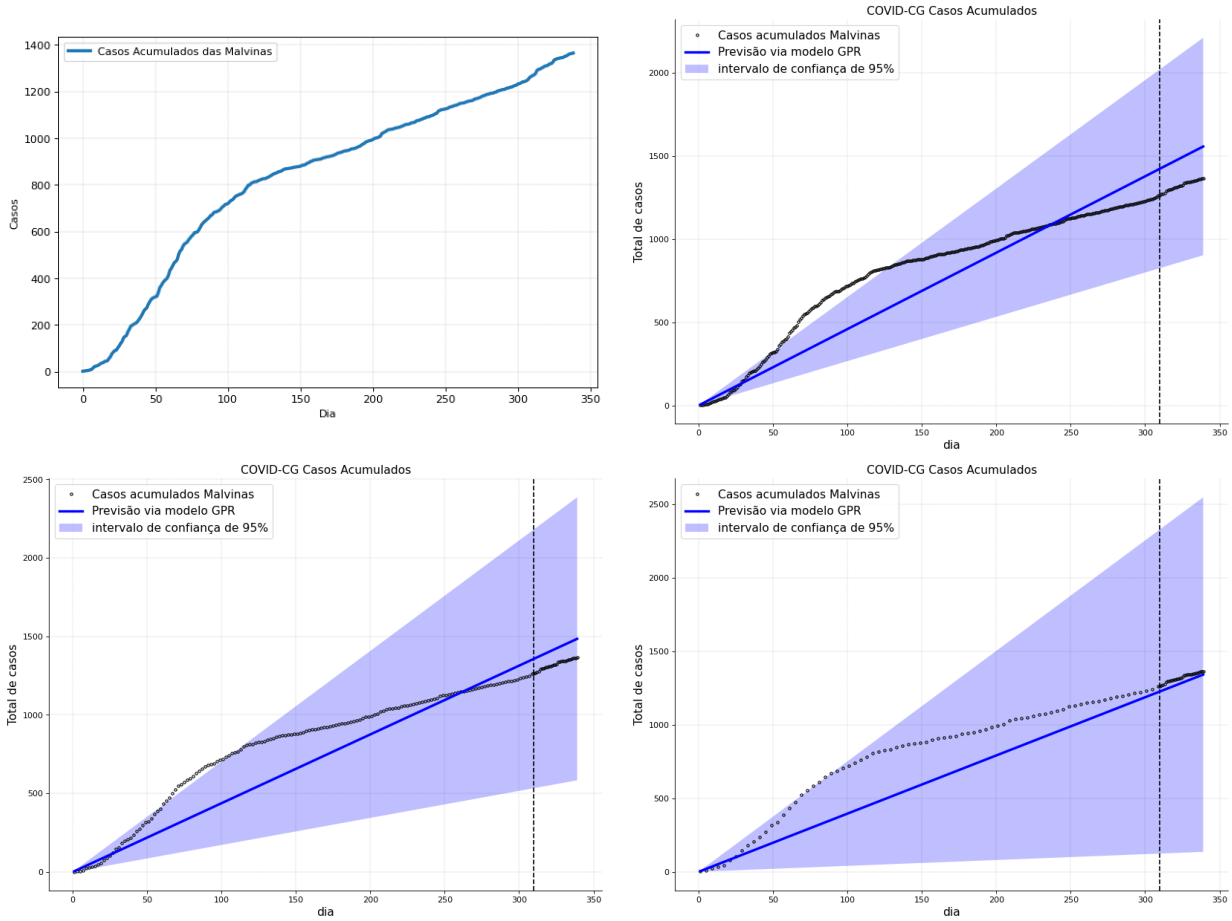
Fonte: Próprio Autor.

Figura 6.32: Curvas fornecidas pelo modelo GPR para o número de casos acumulados de COVID-19 no bairro Catolé, em Campina Grande



Fonte: Próprio Autor.

Figura 6.33: Curvas fornecidas pelo modelo GPR para o número de casos acumulados de COVID-19 no bairro Malvinas, em Campina Grande



Fonte: Próprio Autor.

Observando as Figuras 6.31 a 6.33, vê-se que, mesmo as curvas de casos estando, em sua maioria, dentro do intervalo de confiança, a previsão do GPR se distancia consideravelmente dos casos reais. Isso se deve ao intervalo de confiança muito amplo, indicando a presença de muito ruído na base de dados. Nesse caso, é necessário considerar que a aplicação do *kernel* utilizado não é adequada. Avaliar a utilização de outros *kernels* ou a combinação de múltiplos *kernels* pode ser uma abordagem mais eficiente para reduzir o ruído e melhorar a precisão das previsões, ajustando melhor o modelo às variações específicas de cada base de dados.

Portanto, a escolha do *kernel* é fundamental para o desempenho do modelo GPR. *Kernels* inadequados podem levar a previsões imprecisas e intervalos de confiança excessivamente

amplos, que não refletem a realidade dos dados. Assim, a seleção e combinação cuidadosa dos *kernels* são essenciais para melhorar a adequação do modelo e a confiabilidade das previsões.

6.3.1 Análise do *Kernel*

O *kernel* é um fator essencial para o funcionamento ótimo do modelo GPR. Fundamentalmente, ele determina o grau de correlação entre os dados, capturando a estrutura subjacente e as características da série temporal. A escolha de um *kernel* para o problema em questão é complexa e requer um conhecimento aprofundado sobre os dados e as características gerais de tendência que se deseja modelar[58, 63].

Para modelar as diferentes características do conjunto de dados deste estudo, foi utilizada uma função de covariância (*kernel*) que é uma combinação de outros *kernels* já disponíveis, ajustando seus hiperparâmetros com base nos dados estudados. Esse processo é importante para que o modelo consiga prever o comportamento futuro da pandemia de maneira precisa.

Um *kernel* bem ajustado ao modelo deve levar em consideração diversas características acerca dos dados. O *kernel* deve capturar as tendências de crescimento ou declínio dos casos ao longo do tempo e identificar se esse comportamento se dá em escala linear, exponencial ou em outro padrão, além de modelar variações periódicas, como picos, flutuações e irregularidades, o que é comum em dados epidemiológicos. Considerando dados de séries temporais que podem ter lacunas ou intervalos ausentes, o *kernel* também deve ser capaz de lidar com dados incompletos.

No contexto deste estudo, o *kernel* utilizado na análise inicial foi uma composição padrão que, em uma avaliação preliminar, mostrou-se adequado para as diferentes escalas analisadas. Porém, como observado, o desempenho do modelo variou de forma significativa dependendo da base de dados específica, o que sugere que o *kernel* padrão pode não ser a melhor escolha para todas as escalas e contextos. Portanto, uma etapa de reavaliação do *kernel* é fundamental, a fim de se identificar um *kernel* que melhor modele sazonalidades ou flutuações de curto prazo, podendo melhorar de forma significativa a precisão das previsões.

À vista do exposto, a escolha do *kernel* é essencial para o desempenho do modelo GPR e a confiabilidade das previsões. Um *kernel* inadequado pode resultar em previsões imprecisas

e intervalos de confiança excessivamente amplos.

6.3.2 Considerações Sobre o Modelo GPR

A análise do modelo GPR destacou vários pontos importantes sobre sua aplicação para previsões em epidemias, reforçando a importância de escolhas metodológicas cuidadosas, em especial a seleção adequada do *kernel*, e avaliando a eficácia do modelo no contexto estudado.

Em primeiro lugar, a aplicação do modelo GPR se mostrou uma abordagem promissora para prever o comportamento de séries temporais epidemiológicas. Sua capacidade de incorporar dados diretamente da base, sem a necessidade de hipóteses adicionais sobre a estrutura da população ou taxas de reprodução, diferencia-o, por exemplo, de modelos compartimentais tradicionais, tornando o GPR uma ferramenta flexível e poderosa para lidar com a variabilidade e incerteza inerentes a dados epidemiológicos.

No entanto, a eficácia do modelo GPR está intrinsecamente ligada à escolha do *kernel*, que é responsável por determinar como os dados são correlacionados e, portanto, como as previsões são geradas. A análise evidenciou que a utilização de um *kernel* padrão para todas as escalas de dados nem sempre é adequada, pois diferentes conjuntos de dados podem apresentar comportamentos distintos que exigem ajustes específicos.

A reavaliação do *kernel* é uma etapa fundamental no processo de modelagem. A seleção criteriosa de um *kernel* que capture as nuances da série temporal, como tendências de crescimento ou declínio, sazonalidades e flutuações de curto prazo, pode melhorar significativamente a precisão das previsões. Além disso, a combinação de múltiplos *kernels* ou o ajuste de hiperparâmetros pode proporcionar uma modelagem mais robusta e adaptativa às variações específicas dos dados.

No contexto deste estudo, conclui-se que a qualidade da base de dados é também um fator essencial para o desempenho adequado do modelo GPR. Observou-se que a exclusão de dados de diferentes períodos de tempo, bem como o estudo em diferentes níveis de granularidade, influenciou de maneira significativa os resultados do modelo GPR. Em alguns casos, o modelo conseguiu manter previsões dentro do intervalo de confiança mesmo com a exclusão de dados, indicando uma maior adequação do *kernel* utilizado. Já em outros casos,

a previsão se mostrou menos precisa, sugerindo a necessidade de um *kernel* mais adequado.

A importância da escolha do *kernel* é ainda mais evidente quando se considera que *kernels* inadequados podem levar a previsões imprecisas e intervalos de confiança amplos, que não refletem a realidade dos dados. Isso ressalta a necessidade de um processo de avaliação e ajuste, visando sempre a melhoria da adequação do modelo às características específicas da série temporal analisada.

Em suma, o modelo GPR apresenta um potencial significativo para a previsão de séries temporais em epidemias, oferecendo uma abordagem flexível e adaptativa. No entanto, sua eficácia está intimamente ligada à seleção adequada do *kernel*, que deve ser ajustado às particularidades dos dados em análise. O processo de reavaliação e ajuste contínuo do *kernel* é, portanto, indispensável para garantir a precisão e confiabilidade das previsões, contribuindo para uma melhor compreensão e modelagem das dinâmicas epidemiológicas.

6.4 Resultados

A avaliação comparativa entre os diversos modelos discutidos ao longo deste capítulo permitiu identificar as vantagens e limitações de cada abordagem, contribuindo para uma compreensão mais abrangente acerca da modelagem de dados epidemiológicos.

Os modelos compartimentais são bastante utilizados em epidemiologia para prever a dinâmica de doenças infecciosas. Esses modelos se baseiam em parâmetros como taxas de transmissão e recuperação e dividem a população em compartimentos distintos. Apesar de oferecerem uma estrutura teórica sólida, a precisão desses modelos depende fortemente da disponibilidade e exatidão dos parâmetros estimados. Além disso, eles podem ser menos flexíveis em lidar com dados irregulares ou lacunas, e a suposição de homogeneidade dentro dos compartimentos pode não refletir adequadamente a realidade de uma pandemia em diferentes regiões.

O *Prophet* oferece uma abordagem alternativa para a previsão de casos de COVID-19. Projetado para lidar com dados sazonais, feriados e tendências de longo prazo, ele é relativamente simples de implementar e interpretar. Porém pode apresentar limitações significativas quando aplicado a dados complexos e não lineares, comuns em epidemias. Sua capacidade de capturar variações temporais e padrões sazonais é limitada, o que pode levar

a previsões imprecisas em contextos dinâmicos.

Diante disso, a pesquisa culminou na aplicação do modelo GPR, que se destacou em relação aos modelos mencionados anteriormente por sua capacidade de incorporar variabilidade e incerteza dos dados de forma flexível. O GPR utiliza funções de covariância (*kernel*) para capturar a estrutura latente dos dados, permitindo uma modelagem mais robusta e adaptativa. A escolha e ajuste adequados do *kernel* são fundamentais para o desempenho do GPR, e a possibilidade de combinar múltiplos *kernels* ou ajustar hiperparâmetros proporciona uma vantagem significativa na captura de diferentes características da série temporal.

6.5 Considerações Finais

Após a análise crítica dos modelos compartimentais e do *Prophet* realizada neste capítulo, os resultados obtidos evidenciaram a necessidade de um modelo mais robusto e flexível, capaz de capturar as complexidades intrínsecas aos dados epidemiológicos. Nesse contexto, a pesquisa foi direcionada para a utilização do modelo GPR, que se distingue pela capacidade de incorporar incertezas explicitamente e pela utilização de funções *kernel*, que capturam relações não lineares complexas nos dados. Essas características tornam o GPR especialmente adequado para análises em diferentes escalas territoriais, desde locais até nacionais.

O GPR se mostrou particularmente promissor na possibilidade de contribuições científicas, como propostas de otimização do modelo e do método de seleção do *kernel*. Apesar da complexidade do modelo e da dificuldade em trabalhar com ele, sua aplicação apresentou resultados consistentes, sugerindo que melhorias contínuas na análise podem trazer benefícios significativos. A pesquisa de Quaranta[114], que propôs a análise da COVID-19 em múltiplas escalas geográficas, ofereceu *insights* relevantes sobre a disseminação da doença em contextos variados, o que foi fundamental para se buscar entender como o GPR pode ser adaptado para essas diferentes escalas.

Além disso, a pesquisa[114] contribuiu para a avaliação de como a análise em diferentes níveis de granularidade afeta o diagnóstico geral do comportamento de uma pandemia. Os resultados sugerem que a diminuição da granularidade pode ser válida, permitindo a

aplicação de políticas públicas mais eficazes e direcionadas de acordo com a realidade de cada região. O modelo GPR, com sua flexibilidade e precisão, revela-se uma ferramenta ideal para a previsão de epidemias, facilitando uma resposta mais eficiente às crises de saúde pública.

Em conclusão, a utilização do modelo GPR, ajustado de forma criteriosa por meio da seleção apropriada de *kernels*, mostrou-se superior em vários aspectos no contexto deste estudo. Sua aplicação oferece uma visão detalhada e confiável do comportamento de pandemias em diferentes escalas, contribuindo de forma expressiva para a formulação de políticas públicas baseadas em dados. O GPR não apenas melhora a precisão das previsões epidemiológicas, mas também proporciona uma compreensão mais profunda das dinâmicas da doença, possibilitando a implementação de medidas preventivas e de controle mais eficazes.

A concretização desta pesquisa visa a avaliar como a análise em diferentes níveis de granularidade afeta o diagnóstico geral do comportamento de uma pandemia e, assim, comprovar se é válido diminuir a granularidade para que sejam aplicadas políticas públicas de acordo com cada região. Essa abordagem pode ser determinante para otimizar recursos e estratégias de saúde pública, adaptando-as às necessidades específicas de cada localidade e, dessa forma, mitigando de maneira mais eficaz os impactos das pandemias.

Capítulo 7

Otimização do Modelo de Regressão por Processo Gaussiano

Neste capítulo, é apresentado, de forma mais detalhada, o modelo de Regressão por Processo Gaussiano, com uma proposta de um processo de otimização. O foco principal é aprimorar o mecanismo de identificação e seleção do *kernel* por meio de uma abordagem baseada em *deep learning*. Além disso, o capítulo descreve um procedimento iterativo de otimização para alcançar o melhor desempenho do modelo GPR.

7.1 Características do modelo

O GPR, conforme já discutido anteriormente, é uma técnica de aprendizagem de máquina baseada em *kernel*, fundamentada na teoria Bayesiana, que realiza regressão utilizando conhecimentos prévios sobre a distribuição dos dados ou padrões subjacentes para fazer previsões sobre novos dados[58]. Esse modelo se destaca por sua capacidade de incorporar incertezas explicitamente, o que o torna especialmente útil em cenários em que os dados são escassos ou apresentam alta variabilidade.

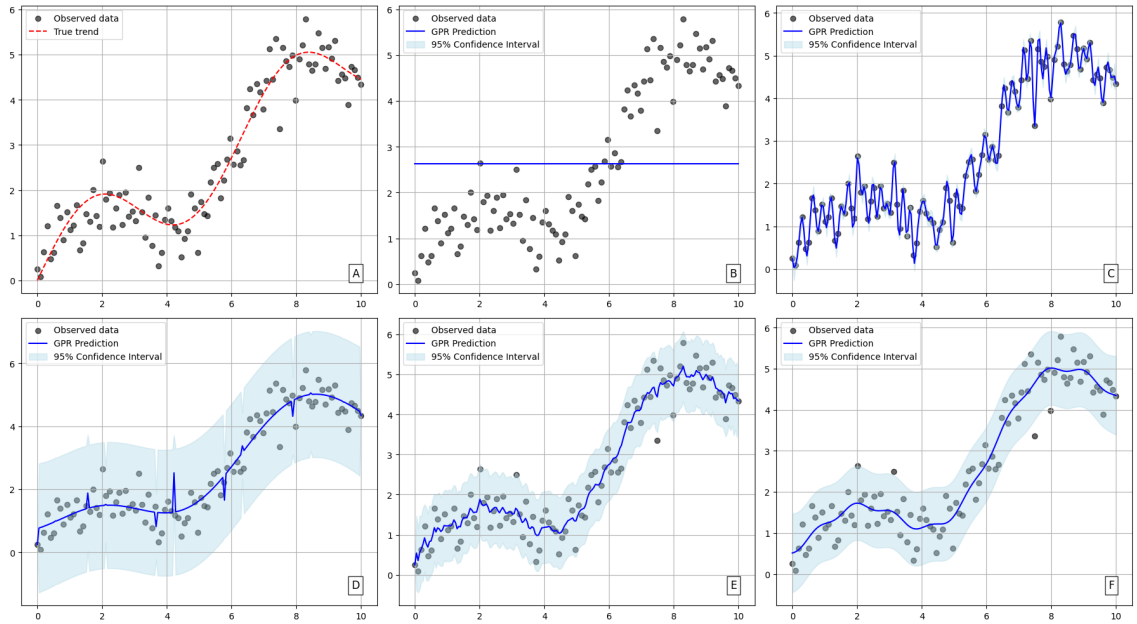
A essência do GPR está na definição da média e da covariância do GP. A média representa a tendência central dos dados, enquanto a covariância, determinada pela função *kernel*, mede a correlação entre os dados observados. A função *kernel* é fundamental para prever uma variável com base nos dados de treinamento em momentos futuros, capturando

a estrutura latente dos dados e permitindo modelar relações não lineares complexas.

Existem diferentes tipos de *kernel*, cada um com características específicas que se ajustam de maneiras distintas aos dados. A escolha e a combinação adequadas de *kernels* são fundamentais para o desempenho do GPR. Uma dificuldade inerente à utilização de modelos GPR é a seleção de um *kernel* adequado para o problema em questão, pois cada *kernel* se ajusta de forma diferente aos dados, afetando diretamente as previsões e o desempenho do modelo. Não há um *kernel* universalmente ótimo; portanto, o ajuste e o refinamento são necessários para escolher o *kernel* mais apropriado para cada aplicação.

A Figura 7.1 ilustra um processo iterativo de construção do *kernel* para modelar uma série temporal com GPR, apresentando uma visão geral do comportamento do modelo ao longo de cada uma das etapas, desde a série temporal original (A) até os resultados das predições (B — F) com a adição progressiva de *kernels* à composição. Cada passo introduz um novo componente ao *kernel*, aprimorando a capacidade do modelo de capturar diferentes padrões presentes nos dados.

Figura 7.1: Comparação da evolução do modelo GPR com diferentes composições de *kernels*.
(A) Série temporal original com dados observados e tendência real;
(B) Predições do GPR com *kernel* inicial: 1^2 (*Kernel* Constante);
(C) Adição do segundo *kernel*: $1^2 + \text{RBF}()$;
(D) Adição do terceiro *kernel*: $1^2 + \text{RBF}() + \text{ExpSineSquared}()$;
(E) Adição do quarto *kernel*: $1^2 + \text{RBF}() + \text{ExpSineSquared}() + \text{RationalQuadratic}()$;
(F) Modelo final, com adição do quinto *kernel*:
$1^2 + \text{RBF}() + \text{ExpSineSquared}() + \text{RationalQuadratic}() + \text{WhiteKernel}()$.



Fonte: Próprio Autor.

A Figura 7.1 exemplifica como a combinação de *kernels* simples pode melhorar o desempenho do modelo GPR ao modelar séries temporais com diferentes características. A série temporal (A) gerada para este exemplo combina uma função seno com uma tendência linear e ruído. Para aplicação do modelo GPR, inicia-se com um único *kernel* (B) e, a cada iteração, um novo *kernel* é adicionado à combinação, o modelo é ajustado aos dados e a predição é realizada. Cada gráfico mostra o impacto de adicionar um *kernel* à combinação e, no último gráfico, tem-se a predição final com o *kernel* completo, que captura bem as características dos dados.

No primeiro passo (B), um *kernel* constante fornece uma predição constante (linha azul horizontal) sem nenhuma variação, ainda não sendo capaz de capturar nenhuma estrutura relevante da série temporal. A incerteza (faixa azul) é mínima, pois o modelo basicamente considera que os dados são homogêneos.

No passo 2 (C), um *kernel* RBF é adicionado, permitindo que o modelo capture vari-

ações suaves nos dados, e a curva de predições começa a se ajustar à tendência geral da série temporal, embora ainda não incorpore estruturas mais complexas, por exemplo, periodicidade ou oscilações bruscas. A incerteza começa a crescer nas regiões em que os pontos apresentam maior dispersão. Apesar de a linha de predição parecer seguir os dados reais e o intervalo de confiança ser estreito, deve-se observar que o modelo não está generalizando bem, mas sim se ajustando ao ruído presente nos dados, e pode não capturar corretamente padrões subjacentes. Dessa forma, o modelo pode falhar em previsões para novos dados, pois se restringe muito às flutuações locais, que podem ser apenas ruído.

Em seguida (D), a adição de um *kernel Exp-Sine-Squared* permite a modelagem de padrões periódicos nos dados, tornando a predição mais refinada ao capturar a estrutura cíclica da série temporal de forma mais eficaz. Isso destaca a importância de levar em consideração componentes periódicos para séries com sazonalidade. O aumento da incerteza em algumas áreas indica que há variações que ainda não foram completamente modeladas.

O próximo passo (E) adiciona um *kernel Rational Quadratic*, que captura variações em múltiplas escalas, permitindo modelar flutuações com diferentes níveis de granularidade e tornando a curva ajustada mais flexível. A incerteza começa a diminuir em algumas regiões, mostrando que o modelo agora está representando melhor as características dos dados.

Finalmente (F), um *White Kernel* é adicionado para modelar o ruído presente nos dados, permitindo que o modelo reconheça e diferencie padrões estruturais de variações aleatórias. Com isso, tem-se a predição final, que se aproxima muito mais da estrutura real dos dados, e a incerteza ajustada de forma mais realista. O conjunto de *kernels* utilizado possibilita que o modelo aprenda não somente tendências globais, mas também variações locais e ruídos intrínsecos à série temporal.

Vale destacar que, além da soma, os *kernels* em GPR podem ser combinados por multiplicação (podendo apresentar as duas operações em uma mesma combinação), e cada operação impacta de forma diferente a modelagem dos dados. A soma modela de forma independente cada característica dos dados, tendo como resultado a superposição de cada contribuição. A multiplicação estabelece uma dependência entre os componentes, de forma que um *kernel* só tem efeito se o outro também for significativo. No exemplo apresentado, utilizou-se uma soma de *kernels* para facilitar a interpretação, permitindo avaliar a contribuição individual de cada *kernel*, e evitar que o modelo se tornasse muito complexo.

O exemplo descrito tornou evidente a importância da combinação de *kernels* ao mostrar que nenhum *kernel* isolado é suficiente para capturar todos os padrões identificados na série temporal. Isso evidencia como a escolha e combinação de kernels impactam diretamente o desempenho do modelo GPR e reforça a necessidade de um método sistemático para selecionar *kernels* adequados.

Pode-se concluir, portanto, que um dos pontos alvo de melhorias no modelo GPR é a otimização do processo de identificação e escolha do *kernel*. Este processo envolve a exploração de múltiplas combinações de *kernels* e o ajuste de hiperparâmetros para maximizar a precisão e a robustez do modelo. Diante do exposto, esta tese propõe um método eficiente para automatizar e aprimorar a seleção de *kernels*, garantindo que o GPR possa ser aplicado de maneira eficaz em diferentes contextos e escalas territoriais.

Em resumo, o modelo GPR se destaca por sua flexibilidade e capacidade de incorporar incertezas, o que é especialmente relevante na modelagem de dados epidemiológicos. Neste capítulo, objetiva-se aprimorar o modelo, focando na otimização do processo de identificação e escolha do *kernel*, de modo a proporcionar previsões mais precisas e confiáveis.

7.2 Proposta de aquisição automatizada de *kernel*

Até o momento, os trabalhos disponíveis na literatura[119–121] descrevem uma seleção não automatizada do *kernel*, em que o melhor *kernel* é escolhido por meio de uma análise do desempenho do modelo, com um conhecimento prévio dos dados, e aplicado para resolver problemas específicos. Este tipo de abordagem pode exigir um grande esforço e apresenta limitações para reprodutibilidade e aplicação em outros contextos. Pensando na necessidade de se executar um modelo GPR em um contexto em que não se tenha o conhecimento prévio dos dados ou que apresente dificuldades na seleção do *kernel* (composto ou não), propõe-se um modelo de *deep learning* que seleciona, computacionalmente, o melhor *kernel* a partir de um conjunto de dados. O método proposto destaca-se pela escolha automática do *kernel* a partir do conjunto de dados, podendo ser aplicado a diversos contextos e problemas variados, selecionando um *kernel* adequado de forma eficaz.

Escolher o melhor *kernel* para um determinado conjunto de dados em um modelo GPR envolve não só o conhecimento teórico sobre os tipos de *kernels*, mas também a avaliação de

desempenho do modelo. Para a seleção do *kernel*, é importante entender as características dos dados e os diversos *kernels* que podem ser usados. Além dos *kernels* comuns, a combinação de *kernels*, com soma e multiplicação, pode ser uma ferramenta importante para capturar melhor a característica dos dados. Após a seleção de um *kernel* ou combinação de *kernels*, é importante avaliar o desempenho do modelo utilizando métricas como MSE, STD e R^2.

Para o desenvolvimento do modelo aqui proposto, foram utilizados códigos *Python*, cujo processo, dividido em três etapas principais, é descrito nesta seção. Todo o material desenvolvido está disponível no diretório */Gaussian_Regressions_Process_Active* do repositório GitHub[68] desta tese.

7.2.1 GenK: geração e avaliação de *kernels*

Dada a importância da combinação de múltiplos *kernels* para modelar padrões complexos em dados, o primeiro módulo do modelo proposto automatiza o processo de geração e avaliação de composições de *kernels*. O código referente ao módulo GenK está disponível em */Gaussian_Regressions_Process_Active/1_-_GenK.ipynb*, no GitHub.

O processo do módulo GenK pode ser dividido em três etapas principais:

- Geração de combinações e permutações de *kernels* usando operações de soma e produto;
- Processamento da série temporal e ajuste do modelo GPR para cada *kernel* gerado;
- Cálculo de métricas para avaliar o desempenho de cada *kernel*.

Geração de *kernels*

A etapa inicial do módulo GenK é a geração de composições de *kernels* utilizando as operações de soma e multiplicação. A operação de soma combina *kernels* de forma aditiva, permitindo que diferentes componentes capturem características dos dados, enquanto a operação de multiplicação combina *kernels* de forma multiplicativa, permitindo modelar interações mais complexas entre os padrões.

As composições são geradas pela execução das funções **generate_combinations_with_repetition** e **generate_permutations_with_repetition**, que criam expressões com múltiplos *kernels* considerando duas abordagens:

1. **Combinações** são utilizadas quando apenas uma operação (soma OU multiplicação) está presente, uma vez que a ordem dos elementos não altera o resultado ($k_1+k_2+k_3 = k_3 + k_1 + k_2$, por exemplo);

2. **Permutações** devem ser consideradas quando ambas as operações aparecem em uma única expressão, já que, neste caso, a ordem dos operadores influencia o resultado ($k_1 + k_2 \times k_3 \neq k_3 + k_1 \times k_2$, por exemplo).

Para ambos os casos, é permitida a repetição de elementos ($k_1 + k_2 + k_2$ e $k_1 + k_1 \times k_2$, por exemplo).

Matematicamente, as combinações são geradas utilizando a lógica de combinação com repetição, que é um tipo de agrupamento da análise combinatória, cuja quantidade de termos gerados é dada por

$$\mathrm{C}_p^n = \frac{(n+p-1)!}{p!(n-1)!}, \tag{7.1}$$

em que n é o número de elementos distintos a serem combinados e p é o número de elementos em cada combinação. A combinação com repetição de n elementos tomados p a p representa todos os agrupamentos que podem ser formados escolhendo p entre n, sendo que um mesmo elemento pode se repetir.

Já para as permutações com repetição, considerando p posições a serem ocupadas por *kernels* que podem se repetir, o número total de permutações com repetição é dado por

$$\text{Permutação com repetição} = n^p, \tag{7.2}$$

em que n é o número de elementos (*kernels*) e p, o número de posições.

E o número de combinações possíveis de operações entre os *kernels* é dado por:

$$\text{Combinação de operações} = m^{p-1}, \tag{7.3}$$

em que m é o número de operações (neste caso, 2) e $p-1$, o número de operadores necessários para combinar p *kernels*.

Finalmente, o número total de permutações possíveis com operações diferentes é dado por

$$\text{Total} = n^p \times m^{p-1}. \tag{7.4}$$

A implementação dessas abordagens resulta em composições como

$$\begin{gathered}
\mathrm{k}_1 + \mathrm{k}_2, \\
\mathrm{k}_1 \times \mathrm{k}_2, \\
\mathrm{k}_1 + \mathrm{k}_2 \times \mathrm{k}_3, \\
\mathrm{k}_1 + \mathrm{k}_2 + \cdots + \mathrm{k}_n, \\
\mathrm{k}_1 \times \mathrm{k}_2 + \cdots \times \mathrm{k}_n,
\end{gathered}$$

em que k_i são *kernels*, com $i = 1, \ldots, 7$.

As combinações geradas consideram as duas abordagens descritas e combinam desde dois até sete *kernels*, utilizando operações de soma e multiplicação, de acordo com a documentação da biblioteca *scikit-learn*[66]. Vale destacar que a multiplicação atende também à combinação de exponenciação, uma vez que, por permitir a repetição de elementos, há casos como, por exemplo, $k \times k$ (equivalente a k^2) e a combinação exponencial de *kernels* representa apenas um *kernel* base (k) e um parâmetro escalar (p) combinados por k^p, não havendo a possibilidade de um *kernel* elevado a outro (${k_1}^{k_2}$, por exemplo).

Deve-se atentar, porém, para questões relativas à complexidade do modelo. Embora a documentação[66] não apresente um limite estrito para o número de *kernels* a serem combinados, na prática, percebe-se que, à medida que se adiciona mais *kernels* à combinação, a complexidade computacional do modelo aumenta e, além disso, combinações muito complexas de *kernels* podem levar ao *overfitting*, especialmente se o número de dados de treinamento for relativamente pequeno em comparação à complexidade do *kernel*, uma vez que os parâmetros do modelo dependem dos dados com os quais são treinados e seus valores são definidos por meio de uma iteração do algoritmo de treinamento e dos dados de treinamento. Combinações formadas por muitos *kernels* podem, ainda, afetar a interpreta-

bilidade do modelo e o entendimento de como cada parte do *kernel* está contribuindo para a predição.

Considerando o que foi colado, neste estudo, optou-se pela utilização efetiva apenas das combinações formadas por até quatro *kernels*. Foram testadas combinações maiores, bem como com diferentes tamanhos de conjuntos de dados, comprovando, porém, os pontos levantados acima.

A abordagem descrita permite a criação sistemática de uma ampla variedade de composições de *kernels* que são então armazenadas para posterior avaliação.

Métricas de desempenho

Após gerar todas as composições de *kernels*, é necessário calcular métricas de desempenho para cada uma delas. Para isso, um conjunto de dados de série temporal é carregado e dividido em treino e teste.

Cada uma das combinações de *kernels* é, então, avaliada por meio de métricas de desempenho do modelo GPR, que é ajustado aos dados de treino e avaliado no conjunto de teste. As métricas utilizadas incluem:

- MSE, que mede a discrepância entre os valores preditos e os valores reais, dimensionando a desigualdade entre eles. É dado por

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{7.5}$$

  em que y_i são os valores reais, $\hat{y}_i$ são as previsões do modelo e n é o número de observações.

- LML, que mede a verossimilhança dos dados sob o modelo GPR e é uma medida da adequação do modelo aos dados observados. É dado por

$$\log p(\mathbf{y}|\mathbf{X},\theta) = -\frac{1}{2}\mathbf{y}^T\mathbf{K}^{-1}\mathbf{y} - \frac{1}{2}\log\det(\mathbf{K}) - \frac{n}{2}\log 2\pi, \tag{7.6}$$

  em que $\mathbf{y}$ é o vetor de observações de saída e n é o número de observações; $\mathbf{X}$ é a matriz de dados de entrada do modelo; $\mathbf{K}$ é a matriz de covariância avaliada nos pontos $\mathbf{X}$,

que depende dos hiperparâmetros θ; $\mathbf{K}^{-1}$ é a matriz de covariância inversa e $\det(\mathbf{K})$ é o determinante da matriz de covariância[58].

- STD, que mede a incerteza associada às previsões, indicando a variabilidade dos erros das previsões em relação à média. É dado por

$$\text{STD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(e_i - \overline{e})^2}, \tag{7.7}$$

em que e_i são os erros e $\overline{e}$ é a média dos erros.

- $\mathbf{R^2}$, que mede o quanto da variabilidade dos dados é explicada pelo modelo. Um $\mathbf{R^2}$ próximo de 1 indica que o modelo explica bem a variabilidade dos dados e, quanto mais próximo de zero, mostra que o modelo não explica bem essa variabilidade. É dado por

$$\text{R}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \tag{7.8}$$

em que $\overline{y}$ é a média dos valores reais.

Essas métricas são importantes na avaliação do desempenho de modelos de regressão, ajustando o modelo para realizar previsões mais confiáveis.

A função **Metrics_to_performance** no código *1_-_Genk.ipynb* avalia cada combinação de *kernel* gerada, resultando em um conjunto de métricas que possibilita a comparação do desempenho de diferentes composições. Os valores são então armazenados em uma biblioteca de *kernels* para futura seleção do melhor *kernel*.

Nesta etapa, testou-se também a utilização de menos métricas para a formação da biblioteca de *kernels*, porém, o enriquecimento da biblioteca, levando a essa composição final, mostrou-se mais eficiente na seleção do *kernel* para o contexto desta aplicação. Para outras aplicações, pode-se avaliar a utilização de diferentes métricas, avaliando o desempenho do modelo para o dado contexto. Essa etapa é eficiente para avaliar uma grande quantidade de combinações de *kernels*, garantindo que os resultados sejam salvos de forma organizada para análises futuras.

7.2.2 BestK: modelo de *deep learning* para seleção do melhor *kernel*

Após a geração e avaliação das combinações de *kernels*, o módulo BestK é introduzido para selecionar a melhor composição de *kernel* para a modelagem dos dados. Para isso, um modelo de deep learning é treinado para identificar a composição de *kernel* com melhor desempenho, considerando um conjunto de métricas de avaliação. O código referente ao módulo BestK está disponível em */Gaussian_Regressions_Process_Active/2_-_BestK.ipynb*, no GitHub.

O processo do módulo BestK pode ser dividido em três etapas principais:

- Pré-processamento dos dados de desempenho dos *kernels*;
- Treinamento do modelo de *deep learning*;
- Seleção e armazenamento do melhor *kernel*.

Preparação dos dados

A primeira etapa do módulo BestK consiste no carregamento da biblioteca de *kernels* gerada na fase anterior, contendo as composições de *kernels* e suas respectivas métricas de desempenho. Essas métricas são selecionadas e armazenadas em uma variável que servirá como entrada para o modelo de *deep learning*. Para determinar a melhor composição de *kernel*, o modelo considera os valores dos R^2 de treinamento (R2_rt), de teste (R2_test) e do conjunto completo (R2). A escolha é feita com base na média ponderada desses três coeficientes, garantindo um equilíbrio entre a capacidade de generalização do *kernel* e sua precisão nos diferentes conjuntos de dados. Os pesos atribuídos a cada coeficiente podem ser ajustados conforme a necessidade do modelo. A composição de *kernel* com o maior valor da média ponderada é considerada a melhor e o modelo a identifica atribuindo o valor 1 à variável-alvo. Assim, o problema se torna um caso de classificação binária.

Outras abordagens foram testadas para esta etapa, como, por exemplo, a seleção da composição de *kernel* com base no menor MSE ou no maior LML. Mas a seleção com base no *kernel* que maximiza a média ponderada dos coeficientes de determinação (R^2), que indicam a capacidade do modelo de explicar a variabilidade dos dados, mostrou-se o critério

mais representativo da capacidade preditiva do modelo, resultando em melhor desempenho do GPR.

Com os dados processados, são definidos os conjuntos de entrada (X) e saída (y) a serem utilizados para treinar o modelo de *deep learning*. O conjunto de dados é então dividido em treino e teste, garantido que o modelo seja treinado e avaliado de forma imparcial.

Arquitetura do modelo

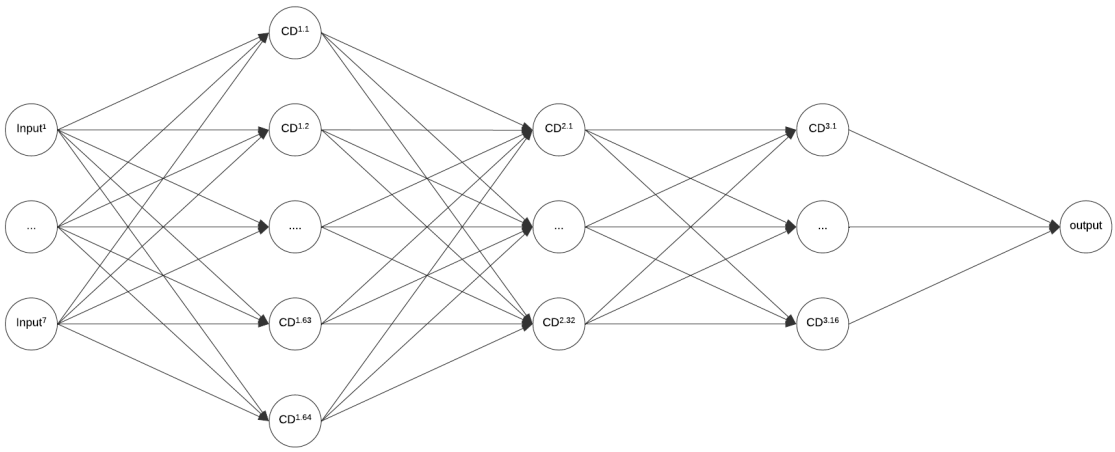
O modelo de *deep learning* utilizado para a seleção do melhor *kernel* é uma rede neural *feedforward* composta por três camadas ocultas totalmente conectadas. O modelo recebe como entrada o vetor de métricas de cada *kernel* e é treinado para aprender a relação entre essas métricas e a melhor composição, retornando um valor de probabilidade indicando se aquele *kernel* é a melhor opção.

O modelo é definido pela seguinte arquitetura:

- Camada de entrada, que recebe um vetor com os valores das métricas. Essa camada tem dimensionalidade igual ao número das métricas (neste caso, 6 neurônios), que são usadas como *features*;
- Três camadas ocultas densas totalmente conectadas, com 64, 32 e 16 neurônios, respectivamente, e ativação ReLU, para introduzir não-linearidade no modelo e refinar os padrões aprendidos a partir das métricas de entrada, capturando relações complexas entre elas;
- Camada de saída, com um único neurônio com função de ativação *sigmoid*, para classificação binária. Essa camada gera uma probabilidade que indica a probabilidade de uma combinação de *kernels* ser a melhor para o conjunto de dados.

A Figura 7.2 representa a rede neural criada por meio de um grafo em que cada nó representa os neurônios da camada de entrada (*Input* na figura), das camadas densas ocultas (CD^1 a CD^3) e da camada de saída (*Output*). As arestas conectam as camadas, representando o fluxo de informação entre os neurônios de uma camada para a seguinte.

Figura 7.2: Grafo Referente à Arquitetura do Modelo de *Deep Learning*



Fonte: Próprio Autor.

A rede foi treinada com função de perda *binary_ crossentropy*, otimizador Adam, 100 épocas, tamanho de *batch* 8 e 20% do conjunto de treino usado para validação, permitindo monitorar o desempenho do modelo durante o treinamento.

Seleção do melhor *kernel*

Os dados de entrada do modelo são as métricas de desempenho da biblioteca de *kernels* e a variável-alvo indica se uma combinação de *kernels* é a melhor. O modelo é treinado e usado para prever a probabilidade de que cada combinação de *kernels* seja a melhor. O índice do *kernel* com maior probabilidade de ser o melhor para o conjunto de dados é identificado usando a função argmax() e a composição correspondente é recuperada do conjunto de dados original.

A configuração do *kernel* selecionado como melhor é armazenada para reutilização em futuras análises. A abordagem do módulo BestK permite a identificação da melhor composição de *kernel*, garantindo que a escolha seja baseada em métricas quantitativas de desempenho. Esse módulo complementa a fase anterior e fornece a melhor configuração de *kernel* para ser utilizada em modelos GPR.

7.2.3 GaussianR: modelo GPR com o *kernel* selecionado

Após a seleção do melhor *kernel*, o módulo GaussianR é responsável por aplicar esse *kernel* em um modelo GPR para realizar previsões sobre a série temporal. Esse módulo garante que a modelagem dos dados seja realizada com a melhor configuração de *kernel* encontrada, maximizando a precisão das previsões. O código referente ao módulo GaussianR está disponível em */Gaussian_Regressions_Process_Active/3_-_GaussianR.ipynb*, no GitHub.

O processo do módulo GaussianR pode ser dividido em três etapas principais:

- Carregamento e preparação dos dados;
- Treinamento do modelo GPR com o *kernel* selecionado;
- Avaliação e visualização dos resultados.

Preparação dos dados

O módulo GaussianR inicia com o carregamento do *kernel* selecionado anteriormente e dos dados históricos que serão utilizados para o treinamento e teste do modelo GPR. A série temporal é normalizada e dividida em conjuntos de treino e teste, garantindo que o modelo GPR seja avaliado corretamente. Essa divisão garante que o modelo seja treinado em uma parte dos dados históricos e avaliado com os últimos registros da série.

Execução do modelo GPR

Com os dados devidamente preparados, o modelo GPR é definido utilizando o *kernel* selecionado e treinado para aprender a dinâmica da série temporal e fazer previsões futuras.

O modelo é ajustado utilizando os dados de treinamento e suas previsões são avaliadas com base nas métricas MSE, R^2, STD e LML.

Otimização iterativa do modelo GPR

Para encontrar o melhor modelo, múltiplas iterações do GPR são executadas, monitorando a melhora do MSE. Esta etapa foi incluída porque, com o prosseguimento das

investigações realizadas ao longo desta pesquisa, observou-se que, ao se executar múltiplas vezes e sem nenhuma alteração o modelo GPR para um mesmo conjunto de dados, as métricas apresentadas mostravam variações nos seus valores. Sendo assim, percebeu-se que, em vários casos, o resultado da primeira execução não apresentava os melhores valores de métricas para o modelo.

Isso se dá devido à natureza estocástica do treinamento do GPR e às variações decorrentes das diferentes inicializações. Com isso, identificou-se uma oportunidade de contribuição para o modelo, propondo um método de execução múltipla do GPR e avaliação das métricas obtidas, a fim de se identificar os melhores resultados e, consequentemente, obter as previsões mais acertadas do modelo GPR.

O método proposto introduz uma forma de encontrar o melhor modelo GPR com base nas suas métricas de avaliação e seu processo consiste em:

1. **Execução do Modelo GPR:** o modelo é treinado várias vezes para cada combinação de *kernel*.

2. **Avaliação das Métricas:** em cada iteração, são registradas as métricas de desempenho.

3. **Comparação e Seleção:** se o desempenho melhorar, o melhor modelo é atualizado; caso contrário, um contador de "*patience*"é incrementado.

4. **Ajuste dos Hiperparâmetros:** em cada uma dessas iterações, o modelo GPR é treinado com os dados fornecidos e, ao final das iterações, o melhor modelo encontrado é retornado.

5. **Convergência:** o ciclo iterativo continua até que o número de iterações consecutivas sem melhoria atinja o valor determinado de "*patience*".

Com isso, é possível encontrar o modelo que otimiza o valor das métricas e, consequentemente, os parâmetros que melhor se ajustam aos dados. Após o treinamento, o modelo com melhor desempenho é selecionado e suas métricas são exibidas.

Visualização de resultados

Para visualizar os resultados, o modelo GPR é utilizado para gerar previsões sobre a série temporal, incluindo intervalos de confiança de 95%. Os dados reais, as previsões e os intervalos de confiança do modelo são plotados com a execução da função **plot_gpr**, permitindo validar a eficácia do modelo GPR com o *kernel* selecionado, garantindo previsões confiáveis para a série temporal.

A abordagem do módulo GaussianR garante a aplicação do *kernel* ideal ao modelo GPR, maximizando a precisão das previsões. Esse módulo finaliza o pipeline de modelagem da série temporal, utilizando a configuração de *kernel* encontrada na fase anterior e aplicando técnicas de avaliação para validar a qualidade dos resultados.

7.2.4 Incremento com *Active Learning*

A fim de otimizar o processo de geração de combinações de *kernels* e cálculo das métricas de desempenho, foi incorporada a abordagem de *active learning*, uma técnica de aprendizado de máquina que busca melhorar a eficiência do treinamento do modelo ao selecionar, de forma inteligente, um subconjunto representativo dos dados em vez de processar todo o conjunto disponível.

No contexto do modelo proposto, a aplicação direta dessa técnica permitiu a seleção de uma amostra diversificada e representativa das possíveis combinações de *kernels*, reduzindo significativamente o tempo de processamento sem comprometer a qualidade dos modelos gerados. O processo de *active learning* no módulo GenK pode ser dividido em três etapas principais:

- Categorização e estruturação das combinações de *kernels*;
- Amostragem inteligente baseada em clusterização;
- Geração e avaliação das métricas de desempenho.

Estruturação das combinações de *kernel*

Inicialmente, todas as possíveis combinações de *kernels* são geradas utilizando operadores matemáticos de soma e multiplicação, conforme descrito na Seção 7.2.1. Essas combina-

ções são então categorizadas em três grupos principais, de acordo com sua complexidade estrutural:

- Somas simples contêm apenas operações de soma entre os *kernels*;
- Multiplicações simples contêm apenas operações de multiplicação;
- Combinações complexas contêm tanto soma quanto multiplicação.

A separação dessas categorias permite um controle mais refinado sobre a diversidade das amostras utilizadas para o treinamento do modelo.

Após a estruturação, um subconjunto referente a uma porcentagem de todas as combinações é selecionado aleatoriamente como ponto de partida para a amostragem baseada em *active learning*. Esse subconjunto representa um espaço inicial de busca otimizado para o processo de seleção.

Amostragem inteligente por meio de *clustering*

Para que a amostra selecionada represente a diversidade de combinações, foi utilizada uma técnica de clusterização baseada em K-Means, que agrupa as combinações em *clusters*, garantindo que a seleção final apresente distribuições equilibradas entre as três categorias. Essa etapa é realizada utilizando a representação vetorial das expressões dos *kernels* por meio do método **CountVectorizer**, que transforma as combinações em vetores numéricos antes da aplicação do algoritmo K-Means. A partir da clusterização, uma nova amostragem balanceada é selecionada.

Com isso, a amostra utilizada nos experimentos contém uma diversidade adequada de combinações, reduzindo a redundância e maximizando a eficiência da avaliação dos *kernels*.

Avaliação de desempenho dos *kernels* selecionados

O processo de preparação dos dados e avaliação de desempenho dos *kernels* segue os mesmos passos já descritos anteriormente, agora, porém, utilizando apenas a amostra selecionada com o *active learning*.

Com essa abordagem, há uma redução significativa no tempo necessário para avaliar as diferentes combinações de *kernels*, sem comprometer a qualidade dos modelos ajustados.

Impacto e benefícios do *active learning*

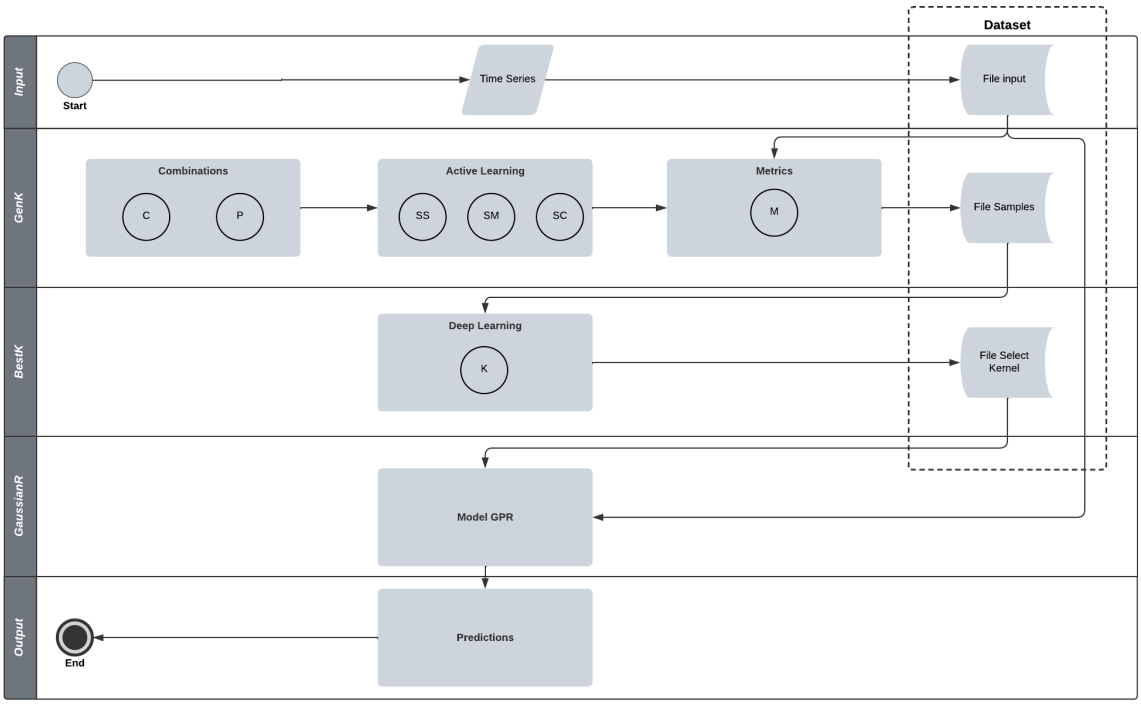
A inclusão de *active learning* no processo de seleção de *kernels* trouxe benefícios ao modelo proposto, tais como:

- Redução do espaço de busca, uma vez que o modelo trabalha apenas com um subconjunto representativo ao invés de avaliar todas as possíveis combinações;
- Eficiência computacional, pois o tempo de execução do processo foi reduzido, permitindo a avaliação de mais experimentos em menos tempo;
- Diversidade nos *kernels* testados, já que a clusterização garante que a amostra final contém diferentes tipos de *kernels*, melhorando a qualidade da seleção final.

Essa abordagem possibilitou uma otimização na escolha do *kernel* ideal para a modelagem da série temporal, garantindo um modelo final mais preciso e eficiente.

A Figura 7.3 ilustra todo o fluxo do processo do modelo proposto, descrito na Seção 7.2.

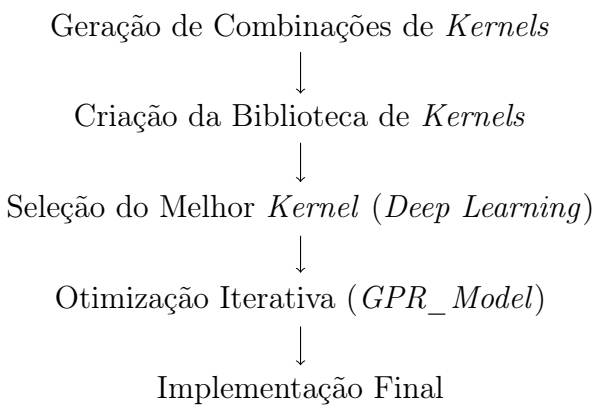
Figura 7.3: Fluxo do processo do modelo proposto



Fonte: Próprio Autor.

7.3 Proposta do novo algoritmo

O novo algoritmo proposto, otimizado para o contexto do estudo, foi desenvolvido com base no que foi discutido nas seções anteriores. O algoritmo objetiva a automatização da seleção do melhor *kernel* para o modelo GPR e, ainda, melhorar de forma iterativa o desempenho do modelo. O processo do novo algoritmo pode ser ilustrado pelo fluxograma:



Inicialmente, o algoritmo realiza a seleção do *kernel* com o modelo de *deep learning*, utilizando as várias combinações de *kernels* geradas e avaliadas conforme descrito anteriormente. O *kernel* selecionado é utilizado na função *GPR_Model* e, em seguida, o método de otimização iterativa é aplicado para encontrar a máxima precisão do modelo GPR.

A capacidade de identificar e selecionar automaticamente o melhor *kernel* para um dado conjunto de dados melhora a reprodutibilidade e a eficiência do processo de modelagem. Além disso, a otimização iterativa possibilita maior refinamento do modelo, proporcionando mais confiabilidade das previsões.

7.4 Resultados

A partir da aplicação do novo algoritmo proposto, obteve-se uma identificação eficaz da melhor combinação de *kernel* para o conjunto de dados específico, avaliando as métricas de desempenho de cada combinação no conjunto de dados analisado, e, em seguida, a utilização do *kernel* no Modelo GPR selecionado e a otimização iterativa do modelo, permitiram a execução de um modelo com melhor precisão de previsões.

A análise dos resultados revelou que, com essa abordagem, o modelo final tem melhor desempenho, o que se evidencia pela entrega de valores reduzidos das métricas MSE e STD e aumentados de R^2. Em relação ao modelo anterior, o novo algoritmo proporcionou uma abordagem mais eficiente de seleção do *kernel* e melhoria na confiabilidade das previsões.

A eficiência do algoritmo será demonstrada no Capítulo 7 a seguir, com a sua aplicação no contexto de análise multiescala da disseminação da COVID-19, evidenciando a capacidade do modelo de se adaptar a diferentes conjuntos de dados.

7.5 Considerações Finais

O estudo desenvolvido contribui de forma significativa para a área de modelagem com GPR com a proposição de um novo algoritmo para a seleção de *kernels*. A abordagem automatizada e iterativa apresentada melhora a eficiência do processo, reduzindo a necessidade de intervenção "manual"na escolha do *kernel* e oferecendo uma solução mais robusta para a aplicação do modelo GPR em contextos diversos.

A capacidade do novo algoritmo de adaptar-se a diferentes conjuntos de dados e contextos demonstra sua versatilidade e aplicabilidade em uma ampla gama de cenários. Além disso, a otimização iterativa assegura que o modelo seja o mais adequado para as necessidades intrínsecas a cada aplicação.

Em relação ao modelo original, esta proposta resulta em previsões mais confiáveis do modelo GPR, uma vez que seleciona de forma automatizada o melhor *kernel* dentro de uma extensa lista de combinações possíveis, bem como reduz os erros médios das previsões por meio do processo de otimização iterativa. Dessa forma, o algoritmo proposto oferece um mecanismo relevante para a comunidade científica no contexto de GPR, trazendo contribuições para pesquisas futuras e aplicações práticas.

Capítulo 8

Análise Multiescala da Propagação da COVID-19

Neste capítulo, são apresentados e discutidos os resultados obtidos com a aplicação do modelo GPR em diferentes escalas para previsões acerca da disseminação da COVID-19. Além disso, este capítulo investiga a influência da escolha do *kernel* na modelagem preditiva, considerando diferentes abordagens para a seleção do *kernel*, demonstrando as dificuldades e limitações da escolha manual, bem como a eficácia da otimização baseada em *deep learning*. A análise abrange desde previsões em nível local até projeções em escala nacional, buscando identificar padrões de propagação e variações na acurácia das previsões conforme o nível de granularidade dos dados. São exploradas as implicações de cada escala na compreensão do comportamento do surto epidêmico e na eficiência do modelo. Por fim, é discutido qual nível de granularidade se mostra mais adequado para investigar a evolução da pandemia de COVID-19 e para a adoção de medidas preventivas, com base nos resultados obtidos e nas características intrínsecas dos dados analisados.

8.1 Comparação de Métodos para Seleção do *Kernel*

Para avaliar os impactos da escolha do *kernel* na modelagem preditiva via GPR, foram testadas três abordagens distintas:

1. **Escolha aleatória de *kernels*:** seleção direta de um *kernel* da biblioteca de *kernels*,

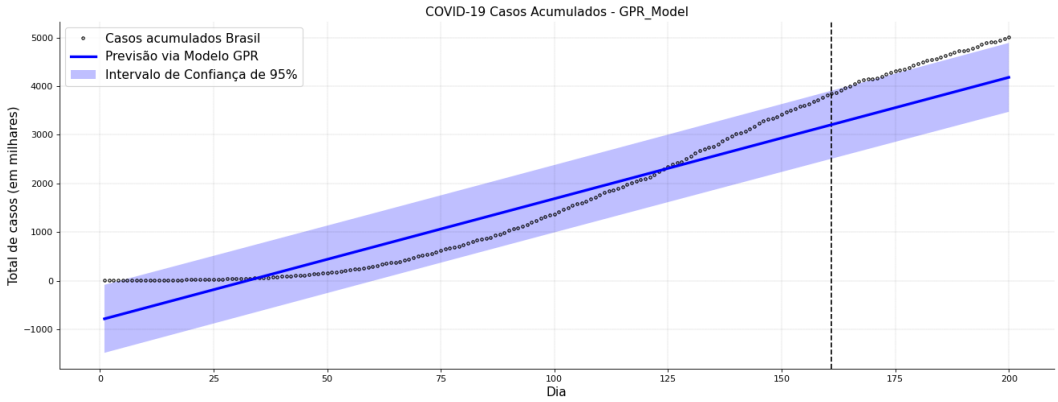
sem critérios definidos;

2. **Composição manual de *kernels*:** construção incremental de composições de *kernels*, adicionando progressivamente novas funções para acompanhar o ajuste;

3. **Otimização baseada em *deep learning*:** abordagem automatizada para identificar a melhor composição de *kernels*.

Inicialmente, um *kernel* foi arbitrariamente selecionado na biblioteca de *kernels* e aplicado ao modelo GPR. O resultado desta abordagem é apresentado na Figura 8.1.

Figura 8.1: Previsões via GPR com *kernel* aleatório



Fonte: Próprio Autor.

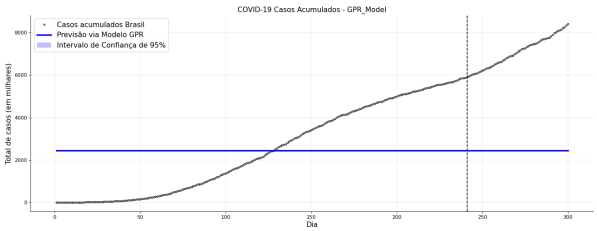
Os resultados apresentados na Figura 8.1 deixam claro que a composição de *kernels* escolhida aleatoriamente não captura bem as características dos dados reais. A previsão apresenta uma estrutura linear com intervalo de confiança muito amplo e desalinhado com a série temporal. Com isso, fica clara a necessidade de esforço e conhecimentos especializados para a seleção manual do *kernel* adequado.

Na próxima etapa de avaliação, foi realizada a construção incremental de uma composição de *kernels*, a fim de acompanhar o desempenho do modelo de forma progressiva. A Figura 8.2 evidencia a dificuldade em definir quais *kernels* em uma composição capturam cada característica dos dados, bem como a forma de combiná-los.

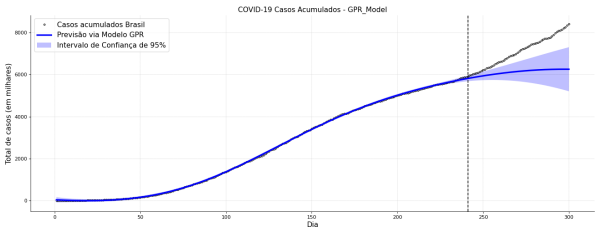
Por fim, foi testada a abordagem automatizada baseada em *deep learning* para identificar a melhor composição de *kernels* e a previsão obtida é mostrada na Figura 8.3

Figura 8.2: Previsões do Modelo GPR com construção incremental do *kernel*

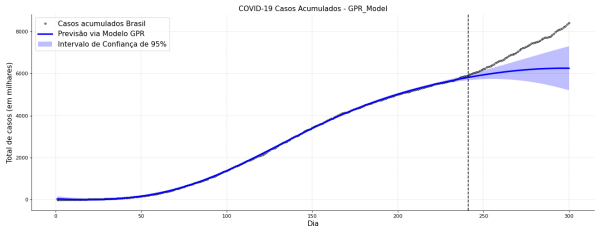
(a) 1^2 (Kernel Constante)



(b) 1^2 + RationalQuadratic()

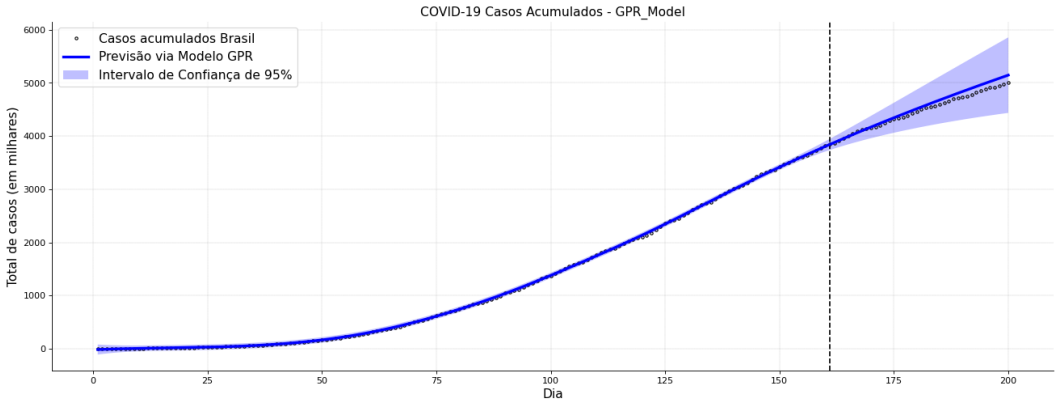


(c) 1^2 + RationalQuadratic() × ExpSineSquared()



Fonte: Próprio Autor.

Figura 8.3: Previsões via GPR com *kernel* selecionado por *deep learning*



Fonte: Próprio Autor.

O resultado obtido com a última abordagem avaliada mostra um bom desempenho do modelo GPR em capturar o comportamento dos dados, comprovando a eficácia do modelo proposto para a seleção automatizada de *kernels*.

Para uma comparação mais detalhada entre as três abordagens consideradas, a Tabela 8.1 apresenta as métricas de desempenho obtidas em cada caso.

A análise das métricas obtidas ressalta o desempenho do método baseado em *deep learning* sobre os demais, com diminuição do MSE e o STD e, principalmente, aumento dos R^2 em todos os intervalos avaliados.

Tabela 8.1: Métricas obtidas em cada abordagem

	MSE	**STD**	**R^2 train**	**R^2**	**R^2 test**	**LML**
Seleção aleatória	$1{,}37 \times 10^5$	$3{,}69 \times 10^5$	0,90681	0,92159	-4,17939	-46,07603
Composição manual	$5{,}42 \times 10^3$	$2{,}33 \times 10^2$	0,99987	0,96276	-1,46259	310,02668
Seleção via *deep learning*	$1{,}47 \times 10^2$	$1{,}21 \times 10^2$	0,99989	0,99955	0,94763	193,91415

Os resultados indicam que a seleção aleatória frequentemente leva a um desempenho inferior, conforme evidenciado pelas métricas de erro e pelas projeções dos gráficos. A composição manual, embora possa melhorar os resultados, é um processo custoso e altamente dependente do conhecimento do modelador. Em contraste, a abordagem baseada em *deep learning* demonstrou ganhos significativos, conforme ilustrado nos gráficos apresentados nas Figuras 8.1 a 8.3 e detalhado na Tabela 8.1.

Esses resultados ressaltam a importância de um processo sistemático para definição da estrutura do *kernel* e reforçam a necessidade de métodos que combinam *machine learning* e estatística para aprimorar previsões epidemiológicas. Diante disso, o método baseado em *deep learning* se mostrou eficaz e adequado para ser utilizado nas análises subsequentes desta tese.

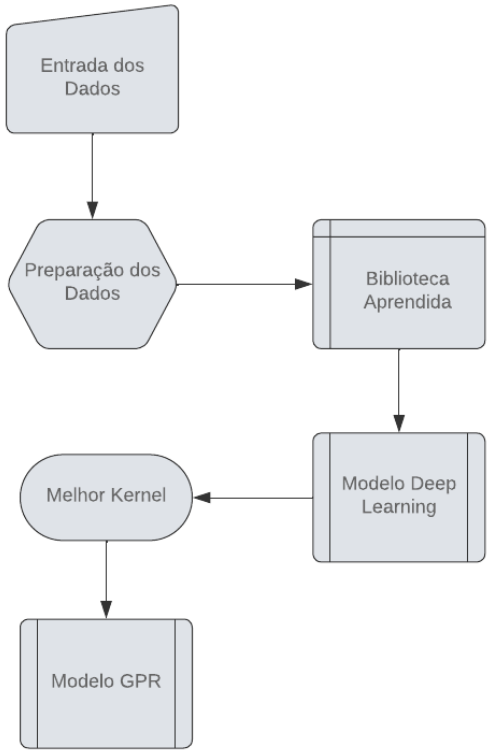
8.2 Resultados da Análise Multiescala

Para realizar a análise multiescala da disseminação do vírus de COVID-19, o modelo GPR foi aplicado em diferentes níveis de granularidade, desde bairros da cidade de Campina Grande até Brasil.

Inicialmente, foi feito o processo de seleção do melhor *kernel* para cada nível de granularidade a ser estudado, utilizando o modelo de *deep learning* descrito na Seção 7.2. Além disso, os parâmetros do modelo dependem dos dados e das características da base de dados, portanto, as métricas foram calculadas considerando as bases de dados completas e, ainda, fatias com quantidades menores de dados, gerando, por fim, uma biblioteca de *kernels* para cada um dos casos.

As bibliotecas geradas foram aplicadas ao modelo de *deep learning* para obtenção do melhor *kernel*, que foi, então, utilizado como o *kernel* do modelo GPR para fazer as previsões em cada uma das bases de dados. Este fluxo é ilustrado pela Figura 8.4.

Figura 8.4: Fluxograma para Aplicação do Modelo GPR



Fonte: Próprio Autor.

Considerando a obtenção do melhor *kernel* para cada conjunto de dados, o próximo passo é a realização das previsões com o modelo GPR e os resultados obtidos para cada nível de granularidade analisado são apresentados e discutidos nas seções a seguir.

8.2.1 Bairros

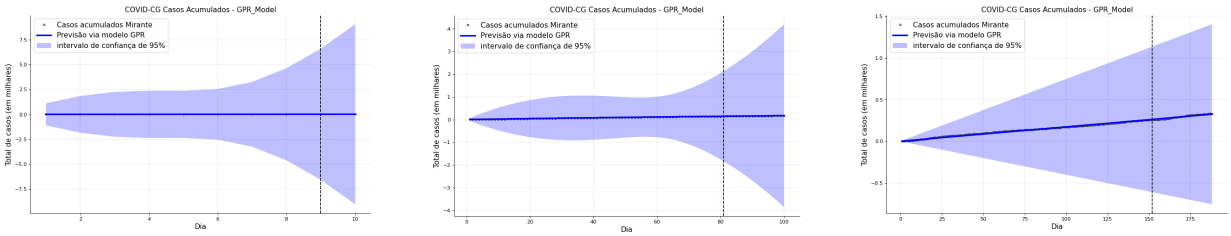
A granularidade mais fina analisada é representada por bairros da cidade de Campina Grande, Paraíba. Os dados coletados, referentes aos casos confirmados de COVID-19 acumulados em alguns bairros, foram utilizados no modelo GPR. Os resultados obtidos para três intervalos de tempo são mostrados na Figura 8.5.

Analisando os gráficos da Figura 8.5, é possível perceber uma dificuldade do GPR de se ajustar a esse nível de escala.

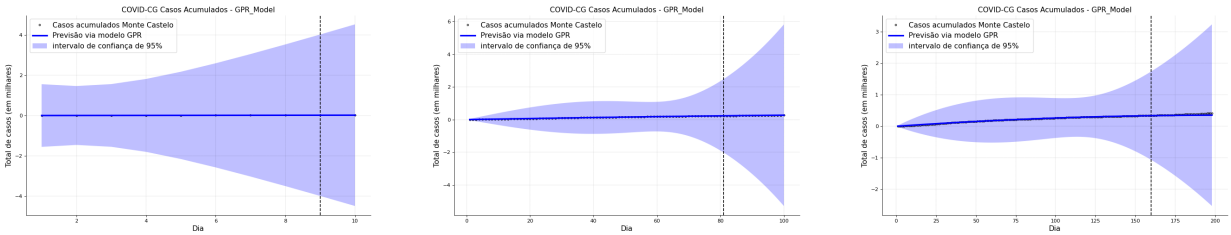
Um ponto a se destacar é que, além de a escolha do *kernel* ser crucial para o seu desempenho, o modelo GPR é, conforme já discutido, sensível à qualidade da base de dados.

Figura 8.5: Previsões do Modelo GPR para os Bairros Mirante e Monte Castelo

(a) 10 dias (b) 100 dias (c) 188 dias



(d) 10 dias (e) 100 dias (f) 198 dias



Fonte: Próprio Autor.

Observa-se que, com poucos dados, apesar de as previsões parecerem seguir a tendência dos dados reais, o intervalo de confiança é muito amplo, o que oferece pouca confiabilidade sobre os resultados gerados. O ajuste das predições aos dados reais, nesse caso, pode se dever ao fato de que, com um número pequeno de dados de treinamento, o modelo se ajusta facilmente aos pontos de dados específicos em vez de aprender a tendência geral.

Observando os gráficos das Figuras 8.5c e 8.5f, nota-se, ainda, que, conforme a quantidade de dados oferecidos ao GPR aumenta, o modelo parece começar a se ajustar melhor e oferecer mais confiabilidade nas suas previsões, embora ainda distante do ideal. Sendo assim, foi identificada a necessidade de aumentar a granularidade e realizar novas previsões.

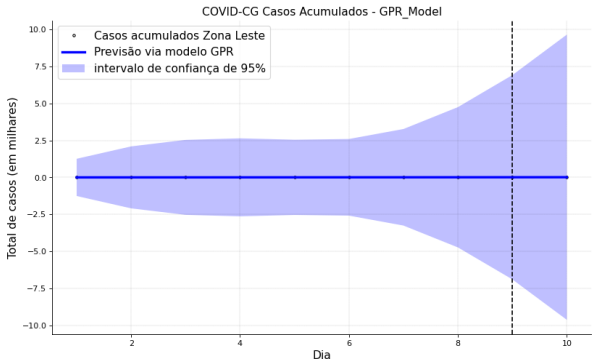
8.2.2 Zona

Percebida a necessidade de utilizar uma granularidade maior, foram coletados dados referentes a um agrupamento de bairros (zona). A Figura 8.6 ilustra os resultados alcançados com a aplicação do modelo GPR aos dados da Zona Leste da cidade de Campina Grande.

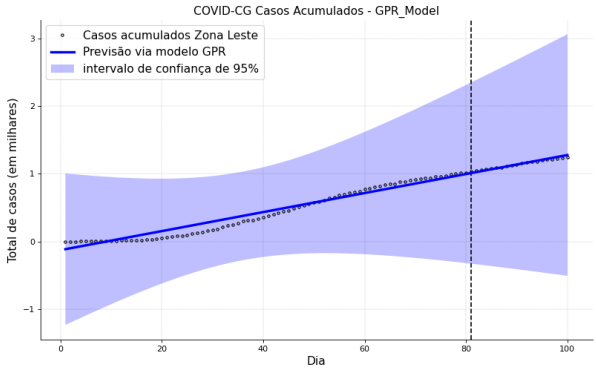
A análise dos gráficos da Figura 8.6 reforça que a qualidade da base de dados é determinante para o desempenho do GPR. Com o aumento da granularidade de bairro para

Figura 8.6: Previsões do Modelo GPR para a Zona Leste

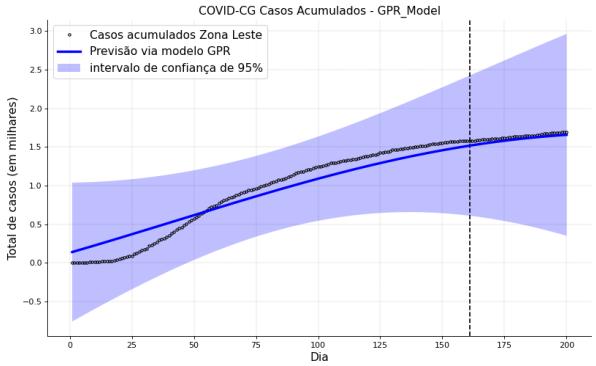
(a) 10 dias



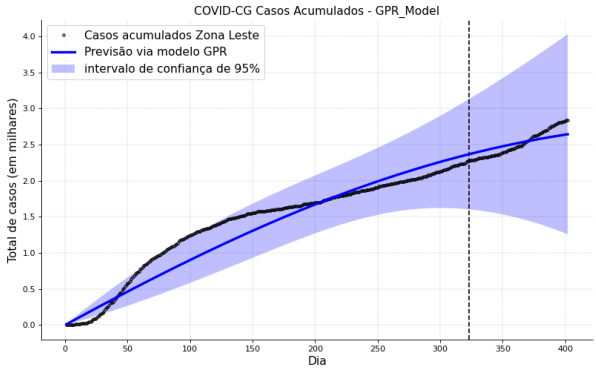
(b) 100 dias



(c) 200 dias



(d) 402 dias



Fonte: Próprio Autor.

zona, percebe-se uma discreta melhoria nas previsões do GPR que, apesar de ainda não se ajustar fortemente, já descreve melhor a tendência de crescimento dos dados. Além disso, o intervalo de confiança também apresenta um melhor ajuste. Mais especificamente, a Figura 8.6d evidencia a melhoria nas previsões e no ajuste do intervalo de confiança com o enriquecimento da base de dados oferecida ao modelo GPR.

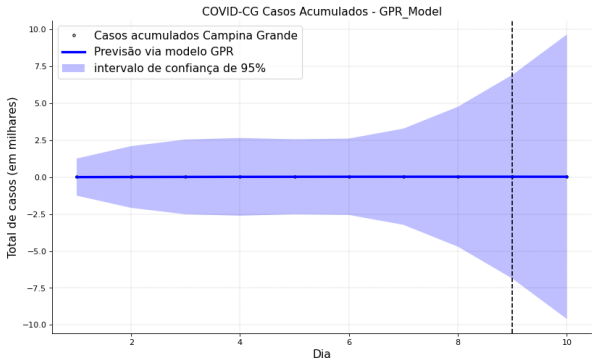
Mesmo com o notável aprimoramento do desempenho, os resultados obtidos ainda não são considerados os mais adequados. Com esse nível de escala, as previsões do modelo GPR continuam não sendo muito precisas. Dessa forma, aumentar ainda mais a granularidade pode proporcionar um melhor desempenho do modelo.

8.2.3 Cidade

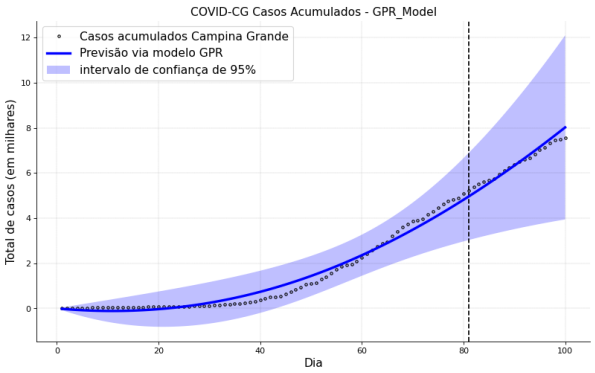
Diante do exposto, considerou-se então a escala de dados referentes à cidade de Campina Grande. Os gráficos da Figura 8.7 representam as previsões do modelo GPR para este nível de granularidade.

Figura 8.7: Previsões do Modelo GPR para a Campina Grande

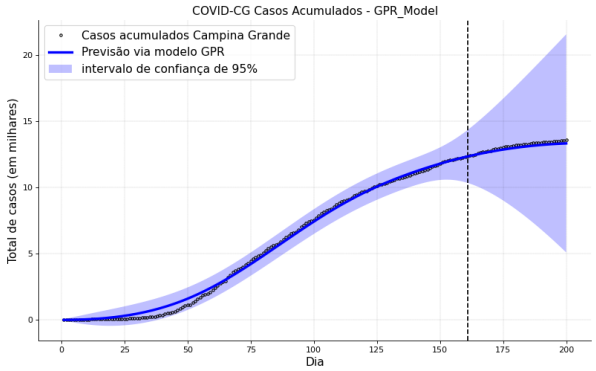
(a) 10 dias



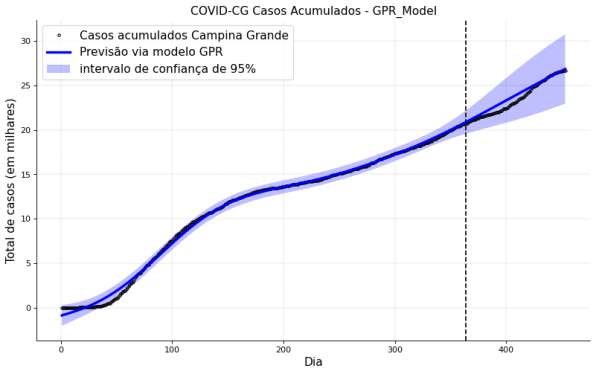
(b) 100 dias



(c) 200 dias



(d) 453 dias



Fonte: Próprio Autor.

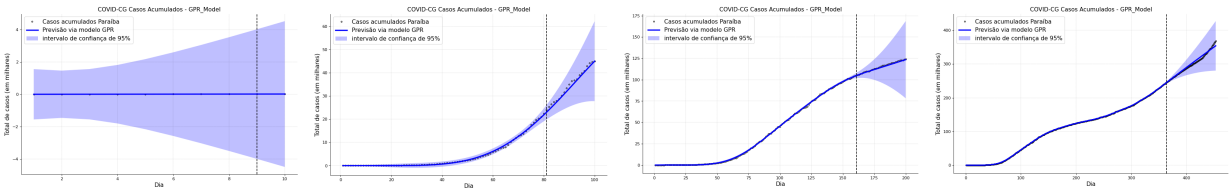
Os resultados ilustrados na Figura 8.7 deixam evidente a melhora no desempenho do modelo GPR para a escala municipal em comparação às escalas menores estudadas. Neste caso, percebe-se que as previsões do modelo são satisfatórias e seguem a tendência dos dados reais com um bom intervalo de confiança desde sua avaliação para 100 dias (Figura 8.7b). Com o aumento da base de dados (Figuras 8.7c e 8.7d), o modelo tem um desempenho ainda mais eficaz.

8.2.4 Estados

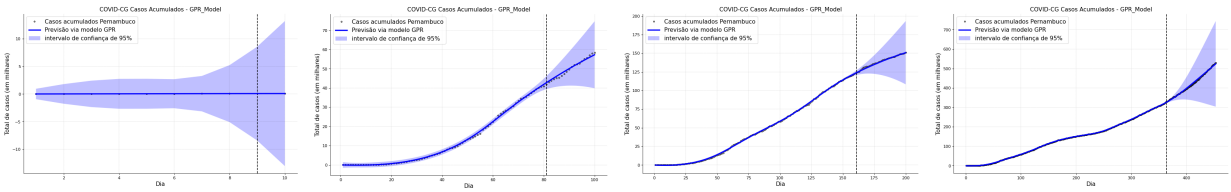
Com a identificação da escala municipal como a escala territorial mínima em que o modelo GPR faz previsões satisfatórias, aumentou-se a granularidade a fim de se reforçar a influência da qualidade da base de dados na precisão dessas previsões. Assim, foram coletados dados estaduais e a Figura 8.8 mostra os resultados obtidos com a aplicação do modelo GPR para Paraíba (8.8a – 8.8d), Pernambuco (8.8e – 8.8h) e Amazonas (8.8i – 8.8l).

Figura 8.8: Previsões do Modelo GPR para Paraíba, Pernambuco e Amazonas

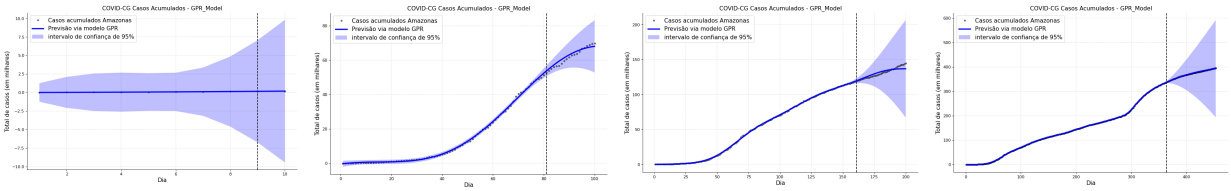
(a) 10 dias (b) 100 dias (c) 200 dias (d) 453 dias



(e) 10 dias (f) 100 dias (g) 200 dias (h) 453 dias



(i) 10 dias (j) 100 dias (k) 200 dias (l) 453 dias



Fonte: Próprio Autor.

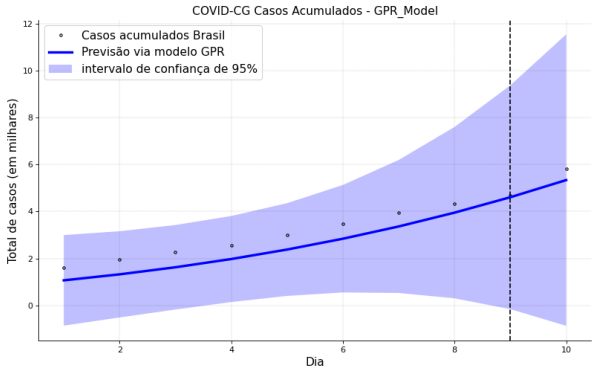
A observação dos gráficos mostrados na Figura 8.8 deixa claro que a qualidade e a precisão das previsões de um modelo GPR dependem diretamente da qualidade do conjunto de dados. De forma semelhante ao que foi percebido anteriormente, as previsões para as fatias de apenas 10 dias das bases de dados apresentam muita incerteza, dado o intervalo de confiança muito amplo, além de a linha da previsão ser rígida. A partir dos 100 dias, o modelo captura bem a tendência dos dados, seguindo a melhora no desempenho observada para a escala municipal. Para os demais casos, o aumento do desempenho continua evidente.

8.2.5 País

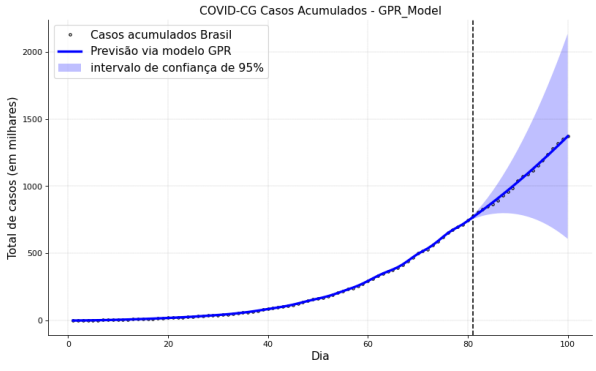
Para uma última análise, considerou-se a escala nacional. A Figura 8.9 apresenta as previsões do modelo GPR para a disseminação da COVID-19 no Brasil.

Figura 8.9: Previsões do Modelo GPR para o Brasil

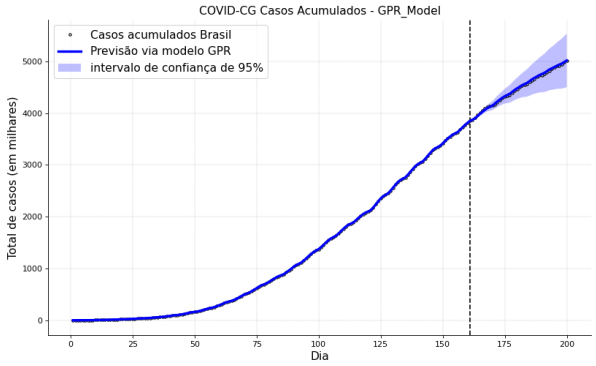
(a) 10 dias



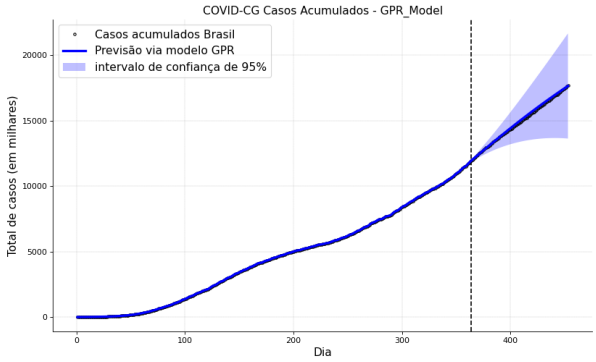
(b) 100 dias



(c) 200 dias



(d) 453 dias



Fonte: Próprio Autor.

A Figura 8.9a mostra que, para o caso da escala nacional, já se consegue observar uma tendência mínima de o GPR capturar o comportamento dos dados. Com o aumento dos dados, o GPR se mostra eficaz em aprender a tendência de crescimento.

Os resultados observados, em todos os níveis de granularidade, reforçam que, com menos dados, o modelo tem menos informações acerca da estrutura latente da base de dados e que a qualidade da base de dados é, juntamente com o *kernel*, um fator determinante para o bom desempenho do modelo.

8.3 Considerações Finais

A análise multiescala da propagação da COVID-19 com utilização do modelo GPR possibilitou verificar a influência dos níveis de granularidade e da qualidade da base de dados na precisão das previsões obtidas. Os resultados mostraram que aplicar o GPR em granularidades muito finas, como bairros, pode não fornecer bom desempenho. Para esses casos, foram obtidas previsões rígidas e com intervalos de confiança muito largos, independentemente do tamanho da base de dados, o que pode se dever à natureza do modelo, que tende a apresentar menor capacidade de generalização ao tentar capturar padrões em dados escassos. O aumento paulatino da granularidade evidenciou que o modelo tende a se ajustar a escalas maiores e que o seu desempenho depende também da base de dados.

Essas observações podem ser explicadas pela estrutura de covariância do modelo GPR. Em escalas menores e com pouca quantidade de dados disponível, o *kernel* pode ter dificuldade em capturar adequadamente as correlações entre os dados. Com isso, o modelo tende a subestimar a variabilidade dos dados e maximizar a incerteza, o que leva a previsões menos confiáveis, evidentes pelos intervalos de confiança amplos.

A escala municipal se destacou como o menor nível de granularidade com bons resultados, a partir do qual o GPR se mostrou eficiente em capturar a tendência dos dados reais, além de ter apresentado intervalos de confiança com bom ajuste. Sendo assim, essa escala apresentou um melhor balanço entre a necessidade de apreender particularidades locais e a robustez necessária para gerar previsões confiáveis. A melhora no desempenho se manteve na análise de escalas maiores (estadual e nacional). Em escalas maiores, a base de dados tende a ser mais robusta, facilitando a captura de tendências latentes pelo modelo que, assim, pode oferecer previsões mais confiáveis.

Conclui-se, assim, que selecionar uma escala apropriada para análise e previsão de surtos epidêmicos é importante, principalmente em contextos que possam apresentar problemas na coleta de dados e heterogeneidade populacional, como é o caso da pandemia de COVID-19 no Brasil.

Parte VII

Conclusão

Capítulo 9

Conclusões e sugestões para trabalhos futuros

Neste capítulo, é feita uma síntese das principais conclusões obtidas ao longo da pesquisa, destacando as contribuições do trabalho para o entendimento da disseminação da COVID-19 e a eficácia do modelo GPR na análise multiescala, bem como para a otimização do desempenho do modelo. Também são apresentadas sugestões para trabalhos futuros, com o intuito de aprimorar os métodos e abordagens utilizados e explorar novas direções de pesquisa que possam contribuir para a evolução do estudo realizado.

9.1 Conclusões

Nesta tese, investigou-se o comportamento da disseminação da COVID-19 utilizando o modelo GPR aplicado em um contexto multiescala que considerou diferentes níveis de granularidade dos dados – desde bairros até a escala nacional. A pesquisa buscou compreender como a variação da escala territorial impacta a acurácia das previsões sobre dados de doenças infectocontagiosas, contribuindo para a análise da dinâmica dessas doenças e subsidiando a tomada de decisão acerca de medidas de controle.

A partir do reconhecimento de que, embora o GPR apresente alto potencial para a modelagem de fenômenos complexos, sua eficácia depende criticamente da escolha adequada do *kernel*, o estudo se concentrou no desenvolvimento de um método automatizado de seleção

de *kernel* utilizando técnicas de *deep learning*, motivado por esta lacuna.

Inicialmente, foi realizada uma análise comparativa entre os modelos compartimentais, o modelo de regressão aditiva (*Prophet*) e o GPR, ressaltando a necessidade de um modelo robusto para capturar as complexidades inerentes aos dados de surtos epidêmicos. Nesse contexto, o GPR foi escolhido por se mostrar promissor para análises detalhadas em diferentes escalas territoriais.

A aplicação do GPR em uma análise multiescala da propagação da COVID-19 permitiu investigar como diferentes níveis de granularidade afetam a precisão das previsões. Os resultados indicaram que o GPR apresenta melhor desempenho em escalas maiores, como municipal, em relação às escalas menores, como a de bairros. Esse comportamento reflete características matemáticas intrínsecas do modelo, que tende a se ajustar melhor quando há um maior volume de dados e menor variabilidade local. Além disso, o tipo de *kernel* utilizado influencia diretamente a capacidade do modelo de correlacionar os dados e melhorar as previsões, mesmo em cenários com volume limitado de dados.

A utilização do modelo GPR representou uma oportunidade concreta de contribuição ao seu aprimoramento por meio da automatização do processo de seleção de *kernels* e do desenvolvimento de uma metodologia para sua otimização iterativa. Essas intervenções proporcionaram não apenas um aumento na capacidade preditiva do modelo, mas também ofereceram um mecanismo mais robusto e adaptável a diferentes áreas que demandam previsões baseadas em dados complexos. Conclui-se que o GPR é um modelo relevante para estudos voltados à propagação de epidemias e análise multiescala, podendo oferecer subsídios importantes para a elaboração de políticas de saúde pública.

Assim sendo, as contribuições desta tese são duas. Inicialmente, a análise multiescala salientou a importância da escolha adequada da escala territorial para a modelagem da propagação de epidemias. Além disso, os resultados obtidos com a proposta de otimização do modelo GPR demonstraram que a abordagem desenvolvida, aplicada ao contexto da pandemia de COVID-19, foi capaz de aprimorar o desempenho preditivo do modelo em diferentes escalas territoriais, ao mesmo tempo em que reduziu a complexidade do processo de modelagem. Com isso, esta tese contribui não apenas para a área de modelagem em saúde pública, mas também para o avanço metodológico dos modelos probabilísticos baseados em GPR, reforçando a importância e aplicabilidade das técnicas propostas durante a realização

desta pesquisa.

9.2 Sugestões para trabalhos futuros

Os resultados obtidos ao longo da pesquisa levam a oportunidades para o desenvolvimento de trabalhos futuros que proporcionem progressões nas abordagens aqui propostas.

O modelo de *deeep learning* proposto para a seleção de *kernels* se mostrou eficaz, porém é passível de aprimoramento no que se refere à eficiência computacional. A criação de uma biblioteca de *kernels* individual para cada base de dados mostrou bons resultados, mas essa metodologia pode ser melhorada a fim de se obter maior generalização. Sendo assim, uma proposição para investigação em potencial seria avaliar a aplicação de técnicas como *transfer learning* para explorar a possibilidade de uma biblioteca de *kernels* gerada para uma determinada base de dados ser utilizada em outras bases de dados. Dessa forma, a fase de treinamento do modelo de *deep learning* poderia ser acelerada e o algoritmo se tornaria mais flexível a diferentes contextos, sendo expandido em termos de eficiência e aplicabilidade.

Outra sugestão de estudo futuro diz respeito à adaptação do modelo GPR para outros cenários de doenças infecciosas, com padrões de propagação distintos, ou em regiões geográficas com características diferentes da estudada neste trabalho, a fim de verificar a versatilidade do modelo em contexto diferentes dos que foram considerados nesta pesquisa. Propõe-se ainda a integração do GPR a sistemas de monitoramento contínuo de dados epidemiológicos, para uma análise de dados em tempo real, criando um sistema de previsão dinâmico, melhorando o tempo hábil para a tomada de decisão em políticas de saúde pública.

O desenvolvimento dessas sugestões pode, além de melhorar o desempenho do modelo GPR, estender sua aplicabilidade a cenários diversos, aumentando as contribuições acerca do entendimento da disseminação de epidemias e da formulação de medidas de controle eficazes. Dessa forma, as investigações realizadas nesta tese podem ser o ponto de partida para novas pesquisas que continuarão a contribuir para a modelagem epidemiológica e, consequentemente, para a saúde pública.

Referências bibliográficas

1 D. G. Chen; X. Chen; J. K. Chen. Reconstructing and forecasting the COVID-19 epidemic in the United States using a 5-parameter logistic growth model. *Global Health Research And Policy*, v. 5, n. 25, p. 1–7, Mai 2020. doi.org/10.1186/s41256-020-00152-5.

2 J. Farooq; M. A. Bazaz. A novel adaptive deep learning model of Covid-19 with focus on mortality reduction strategies. *Chaos, Solitons Fractals*, v. 138, n. 1, p. 110148, Set 2020. doi.org/10.1016/j.chaos.2020.110148.

3 C. X. Deng. The global battle against SARS-CoV-2 and COVID-19. *International Journal of Biological Sciences*, v. 16, n. 10, p. 1676, Mai 2020. doi.org/10.7150/ijbs.45587.

4 N. Jebril. World health organization declared a pandemic public health menace: a systematic review of the coronavirus disease 2019 "COVID-19". *Available at SSRN 3566298*, v. 1, n. 2, p. 12, Abr 2020. doi.org/10.2139/ssrn.3566298.

5 World Health Organization. *WHO coronavirus (COVID-19) dashboard*. 2024. https://covid19.who.int/, acesso em 22 jun. 2024.

6 Ministério da Saúde. Secretaria de Vigilância em Saúde. *Painel Coronavírus*. 2024. https://covid.saude.gov.br/, acesso em 28 nov. 2024.

7 F. K. Ayittey et al. Economic impacts of Wuhan 2019-nCoV on China and the world. *Journal of Medical Virology*, v. 92, n. 5, p. 473, Fev 2020. doi.org/10.1002/jmv.25706.

8 F. Brauer. Mathematical epidemiology: past, present, and future. *Infectious Disease Modelling*, v. 2, n. 2, p. 113–127, Mai 2017. doi.org/10.1016/j.idm.2017.02.001.

9 M. A. M. T. Baldé. Fitting SIR model to COVID-19 pandemic data and comparative forecasting with machine learning. *medRxiv*, v. 1, n. 1, p. 1–20, Mar 2020. doi.org/10.1101/2020.04.26.20081042.

10 SES, Secretaria Estadual de Saúde da Paraíba. *Painel de Monitoramento*. 2024. https://paraiba.pb.gov.br/diretas/saude/coronavirus/dados-epidemiologicos-covid, acesso em 19 out. 2024.

11 SMSCG, Secretaria Municipal de Saúde de Campina Grande. *Casos de COVID-19*. 2024. https://campinagrande.pb.gov.br/coronavirus/, acesso em 21 jan. 2025.

12 J. Chen et al. Big data challenge: a data management perspective. *Frontiers of Computer Science*, v. 7, p. 157–164, Abr 2013. doi.org/10.1007/s11704-013-3903-7.

13 H. W. Hethcote; J. W. Van Ark. Epidemiological models for heterogeneous populations: proportionate mixing, parameter estimation, and immunization programs. *Mathematical Biosciences*, v. 84, n. 1, p. 85–118, Mai 1987. doi.org/10.1016/0025-5564(87)90044-7.

14 D. Bernoulli; D. Chapelle. Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l'inoculation pour la prévenir. *HAL*, v. 1, n. 1, p. 1–32, Mai 2023. https://inria.hal.science/hal-04100467.

15 W. O. Kermack; A. G. McKendrick. 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*, v. 115, n. 772, p. 700–721, Jan 1927. doi.org/10.1098/rspa.1927.0118.

16 N. E. Huang et al. Herd immunity vs suppressed equilibrium in COVID-19 pandemic: different goals require different models for tracking. *medRxiv*, Mai 2020. doi.org/10.1101/2020.03.28.20046177.

17 E. Santoro. Information technology and digital health to support health in the time of COVID-19. *Recenti Progressi in Medicina*, v. 111, n. 7, p. 393–397, Jul 2020. doi.org/10.1701/3407.33919.

18 C. Pernencar et al. Systematic mapping of digital health apps – a methodological proposal based on the World Health Organization classification of interventions. *Digital Health*, v. 8, p. 1–14, Dez 2022. doi.org/10.1177/20552076221129071.

19 B. A. B. de Souza Filho; É. F. Tritany. COVID-19: importância das novas tecnologias para a prática de atividades físicas como estratégia de saúde pública. *Cadernos de Saúde Pública*, v. 36, n. 5, p. 1–5, Ago 2020. doi.org/10.1590/0102-311X00054420.

20 I. C. Celuppi et al. Uma análise sobre o desenvolvimento de tecnologias digitais em saúde para o enfrentamento da COVID-19 no Brasil e no mundo. *Cadernos de Saúde Pública*, v. 37, n. 3, p. 1–12, Mar 2021. doi.org/10.1590/0102-311X00243220.

21 R. Silveira et al. GISSA intelligent chatbot experience – how effective was the interaction between pregnant women and a chatbot during the COVID-19 pandemic? *Procedia Computer Science*, v. 219, p. 1271–1278, Nov 2023. doi.org/10.1016/j.procs.2023.01.411.

22 F. S. Costa; I. J. L. de Sousa; J. A. R. Santos. SIR model applied in dynamics of COVID-19 contagion in São Luís-MA, Brazil. *International Journal of Modeling, Simulation, and Scientific Computing*, v. 12, n. 3, p. 2141003, Mar 2021. doi.org/10.1142/S1793962321410038.

23 A. G. M. Neves; G. Guerrero. Predicting the evolution of the COVID-19 epidemic with the A-SIR model: Lombardy, Italy and São Paulo state, Brazil. *Physica D: Nonlinear Phenomena*, v. 413, n. 1, p. e132693, Dez 2020. doi.org/10.1016/j.physd.2020.132693.

24 S. Afonso; J. Azevedo; M. Pinheiro. Epidemic analysis of COVID-19 in Brazil by a generalized SEIR model. *arXiv Preprint arXiv:2005.11420*, v. 1, n. 1, p. 11420, Mai 2020. doi.org/10.48550/arXiv.2005.11420.

25 A. S. Freitas; L. S. Silva; S. S. L. Sandes. New SIR model used in the projection of COVID-19 cases in Brazil. *medRxiv*, v. 1, n. 1, p. 32056, Mai 2020. doi.org/10.1101/2020.04.26.20080218.

26 S. Dana et al. Brazilian modeling of COVID-19 (BRAM-COD): a Bayesian Monte Carlo approach for COVID-19 spread in a limited data set context. *medRxiv*, v. 1, n. 1, p. 1–42, Mai 2020. doi.org/10.1101/2020.04.29.20081174.

27 K. Bartoszek et al. Are official confirmed cases and fatalities counts good enough to study the COVID-19 pandemic dynamics? a critical assessment through the case of Italy. *Nonlinear Dynamics*, v. 101, n. 3, p. 1951–1979, Jun 2020. doi.org/10.1007/s11071-020-05761-w.

28 D. H. Sousa. *Avaliação da eficiência do controlo do contágio e do tratamento médico à COVID-19 em países da OCDE utilizando a análise envoltória de dados*. Dissertação (Dissertação de Mestrado) — Universidade do Algarve – UAlgFE, Portugal, Fev 2021. https://sapientia.ualg.pt/handle/10400.1/18234.

29 A. Bouchnita; A. Jebrane. A hybrid multi-scale model of COVID-19 transmission dynamics to assess the potential of non-pharmaceutical interventions. *Chaos, Solitons Fractals*, v. 138, n. 1, p. 109941, Set 2020. doi.org/10.1016/j.chaos.2020.109941.

30 S. Solayman et al. Automatic COVID-19 prediction using explainable machine learning techniques. *International Journal of Cognitive Computing in Engineering*, v. 4, p. 36–46, Jun 2023. doi.org/10.1016/j.ijcce.2023.01.003.

31 C. Lv et al. Innovative applications of artificial intelligence during the COVID-19 pandemic. *Infectious Medicine*, v. 3, p. 100095, Mar 2024. doi.org/10.1016/j.imj.2024.100095.

32 A. Dairi et al. Comparative study of machine learning methods for COVID-19 transmission forecasting. *Journal of Biomedical Informatics*, v. 118, p. 103791, Jun 2021. doi.org/10.1016/j.jbi.2021.103791.

33 M. U. Tariq; S. B. Ismail. AI-powered COVID-19 forecasting: a comprehensive comparison of advanced deep learning methods. *Osong Public Health and Research Perspectives*, v. 15, n. 2, p. 115–136, Abr 2024. doi.org/10.24171/j.phrp.2023.0287.

34 R. A. S. Ramalho. Bibframe: modelo de dados interligados para bibliotecas. *Informação Informação*, v. 21, n. 2, p. 292–306, Dez 2016. doi.org/10.5433/1981-8920.2016v21n2p292.

35 A. Gouveia Jr. O conceito de modelo e sua utilização nas ciências do comportamento: breves notas introdutórias. *Estudos de Psicologia (Campinas)*, v. 16, p. 13–16, Abr 1999. doi.org/10.1590/S0103-166X1999000100002.

36 N. Bacaër et al. *Matemática e epidemias*. [S.l.: s.n.], 2022. https://hal.science/hal-03873865.

37 A. Ruffino-Netto; G. R. Arantes. Modelo matemático para estimar impacto epidemiológico da vacinação BCG. *Revista de Saúde Pública*, v. 11, n. 1, p. 502–509, Dez 1977. doi.org/10.1590/S0034-89101977000400007.

38 M. F. Lima-Costa; S. M. Barreto. Tipos de estudos epidemiológicos: conceitos básicos e aplicações na área do envelhecimento. *Epidemiologia e Serviços de Saúde*, v. 12, n. 4, p. 189–201, Dez 2023. doi.org/10.5123/S1679-49742003000400003.

39 M. H. R. Luiz. *Modelos matemáticos em epidemiologia*. Dissertação (Dissertação) — Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de Rio Claro, Nov 2012. https://repositorio.unesp.br/handle/11449/94348.

40 Y. Cheng et al. Evaluating the risk for usutu virus circulation in Europe: comparison of environmental niche models and epidemiological models. *International Journal of Health Geographics*, v. 17, n. 1, p. 1–14, Out 2018. doi.org/10.1186/s12942-018-0155-7.

41 E. G. J. Owusu-Ansah et al. Probabilistic modeling for an integrated temporary acquired immunity with norovirus epidemiological data. *Infectious Disease Modelling*, v. 4, n. 1, p. 99–114, Mai 2019. doi.org/10.1016/j.idm.2019.04.005.

42 H. W. Hethcote. The mathematics of infectious diseases. *SIAM Review*, v. 42, n. 4, p. 599–653, Mai 2000. doi.org/10.1137/S0036144500371907.

43 E. G. Nepomuceno; D. F. Resende; M. J. Lacerda. A survey of the individual-based model applied in biomedical and epidemiology. *Journal of Biomedical Research and Reviews*, v. 1, n. 1, p. 11–24, Fev 2019. doi.org/10.48550/arXiv.1902.02784.

44 E. G. Nepomuceno; R. H. C. Takahashi; L. A. Aguirre. Individual-based model (IBM): an alternative framework for epidemiological compartment models. *Brazilian Journal of Biometrics*, v. 34, n. 1, p. 133–162, Mai 2016. https://biometria.ufla.br/index.php/BBJ/article/view/95.

45 F. Brauer; C. Castillo-Chavez. *Mathematical models in population biology and epidemiology*. [S.l.]: Springer, 2012.

46 G. James et al. *An introduction to statistical learning*. [S.l.]: Springer, 2013. doi.org/10.1007/978-1-4614-7138-7.

47 M. Svensén; C. M. Bishop. *Pattern recognition and machine learning*. [S.l.]: Springer, Cham, 2007.

48 T. Hastie; R. Tibshirani; J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. [S.l.]: Springer, 2017. doi.org/10.1007/978-0-387-84858-7.

49 E. G. Nepomuceno. *Dinâmica, modelagem e controle de epidemias*. Dissertação (Dissertação de Mestrado) — Universidade Federal de Minas Gerais – UFMG, Belo Horizonte – MG, Dez 2005. https://www.ppgee.ufmg.br/defesas/534D.PDF.

50 C. M. R. Franco. Modelos matemáticos em epidemiologia aplicação: evolução epidêmica da COVID-19 no Brasil e no estado da Paraíba. *CES UFCG*, v. 1, n. 1, p. 1–22, Mai 2020. http://www.ces.ufcg.edu.br/portal/phocadownload/userupload/COVID-19_MODELO_SIR.pdf.

51 F. Brauer; C. Castillo-Chávez. Discrete population models. *Springer*, v. 40, n. 1, p. 51–94, Mai 2001. doi.org/10.1007/978-1-4757-3516-1_2.

52 L. P. de Lima. *Modelos aditivos generalizados: aplicação a um estudo epidemiológico ambiental.* Tese (Dissertação de Mestrado) — Universidade de São Paulo, São Paulo, Brasil, Abr 2001. https://www.researchgate.net/publication/34009692_Modelos_aditivos_generalizados_aplicacao_a_um_estudo_epidemiologico_ambiental.

53 S. J. Taylor; B. Letham. Forecasting at scale. *The American Statistician*, Taylor & Francis, v. 72, n. 1, p. 37–45, Jul 2018. doi.org/10.1080/00031305.2017.1380080.

54 Python Software Foundation. *Automatic forecasting procedure.* 2023. https://pypi.org/project/prophet/, acesso em 25 jan. 2025.

55 Facebook Open Source. *Quick start / Prophet.* 2023. https://facebook.github.io/prophet/docs/quick$_s tart.html, acessoem 12 jan.$2025.

56 Facebook Open Source. *Prophet: automatic forecasting procedure.* 2023. https://github.com/facebook/prophet, acesso em 22 jan. 2025.

57 Prophet. *Previsão em escala.* 2024. https://facebook.github.io/prophet/, acesso em 21 jan. 2025.

58 C. E. Rasmussen. *Gaussian processes in machine learning.* [S.l.]: Springer Berlin Heidelberg, 2004.

59 A. Kapoor et al. Gaussian processes for object categorization. *International Journal of Computer Vision*, v. 88, n. 2, p. 169–188, Jul 2010. doi.org/10.1007/s11263-009-0268-3.

60 H.-C. Kim; J. Lee. Clustering based on Gaussian processes. *Neural Computation*, v. 19, n. 11, p. 3088–3107, Nov 2007. doi.org/10.1162/neco.2007.19.11.3088.

61 E. Lewinson. *Python for finance cookbook – second edition: over 80 powerful recipes for effective financial data analysis.* [S.l.]: Packt Publishing, 2022. 2nd ed.

62 M. Ebden. Gaussian processes: a quick introduction. *arXiv Preprint arXiv:1505.02965*, v. 2, n. 02965, p. 1505, Ago 2015. https://arxiv.org/pdf/1505.02965.pdf.

63 C. M. Bishop. *Pattern recognition and machine learning.* [S.l.]: Springer, 2006.

64 D. Duvenaud. *Automatic model construction with Gaussian processes.* Tese (Tese de Doutorado) — University of Cambridge, Cambridge – Inglaterra, Jun 2014. http://www.cs.toronto.edu/ duvenaud/thesis.pdf.

65 K. P. Murphy. *Machine learning: a probabilistic perspective.* [S.l.]: MIT Press, 2012.

66 scikit-learn developers. *Gaussian processes*. 2024. https://scikit-learn.org/stable/modules/gaussian_process.html#kernels-for-gaussian-processes, acesso em 28 jan. 2025.

67 M. Regona et al. Opportunities and adoption challenges of AI in the construction industry: a PRISMA review. *Journal of Open Innovation: Technology, Market, and Complexity*, v. 8, n. 1, Fev 2022. doi.org/10.3390/joitmc8010045.

68 B. C. Dantas. *Repositório tese*. 2024. https://github.com/bcdmodelos, acesso em 11 jan. 2025.

69 L. de R. Alvarenga. *Modelagem de epidemias através de modelos baseados em indivíduos*. Dissertação (Dissertação de Mestrado) — Universidade Federal de Minas Gerais – UFMG, Belo Horizonte, MG, Set 2008. https://repositorio.ufmg.br/handle/1843/RHCT-7JXPK4.

70 F. Saleem et al. Machine learning, deep learning, and mathematical models to analyze forecasting and epidemiology of COVID-19: a systematic literature review. *International Journal of Environmental Research and Public Health*, v. 19, n. 9, p. 5099, Abr 2022. doi.org/10.3390/ijerph19095099.

71 S. B. Bastos et al. The COVID-19 (SARS-CoV-2) uncertainty tripod in Brazil: assessments on model-based predictions with large under-reporting. *Alexandria Engineering Journal*, v. 60, n. 5, p. 4363–4380, Out 2021. doi.org/10.1016/j.aej.2021.03.004.

72 Z. Wang et al. System inference for the spatio-temporal evolution of infectious diseases: Michigan in the time of COVID-19. *Computational Mechanics*, v. 66, p. 1153–1176, Ago 2020. doi.org/10.1007/s00466-020-01894-2.

73 E. B. Postnikov. Estimation of COVID-19 dynamics "on a back-of-envelope": does the simplest SIR model provide quantitative parameters and predictions? *Chaos, Solitons Fractals*, v. 135, n. 1, p. 109841, Jun 2020. doi.org/10.1016/j.chaos.2020.109841.

74 W. Ding; Q.-G. Wang; J.-X. Zhang. Analysis and prediction of COVID-19 epidemic in South Africa. *ISA Transactions*, v. 124, p. 182–190, Mai 2022. doi.org/10.1016/j.isatra.2021.01.050.

75 P. Shahrear; S. M. S. Rahman; M. M. H. Nahid. Prediction and mathematical analysis of the outbreak of coronavirus (COVID-19) in Bangladesh. *Results in Applied Mathematics*, v. 10, p. 100145, Mai 2021. doi.org/10.1016/j.rinam.2021.100145.

76 J. M. Garrido et al. Mathematical model optimized for prediction and health care planning for COVID-19. *Medicina Intensiva (English Edition)*, v. 46, n. 5, p. 248–258, Mai 2022. doi.org/10.1016/j.medine.2022.02.020.

77 X. Yu et al. RLIM: a recursive and latent infection model for the prediction of US COVID-19 infections and turning points. *Nonlinear Dynamics*, v. 106, n. 2, p. 1397–1410, Mai 2021. doi.org/10.1007/s11071-021-06520-1.

78 M. Mandal et al. A model based study on the dynamics of COVID-19: prediction and control. *Chaos, Solitons & Fractals*, Elsevier, v. 136, p. 109889, Jul 2020. doi.org/10.1016/j.chaos.2020.109889.

79 A. Mahajan; N. A. Sivadas; R. Solanki. An epidemic model SIPHERD and its application for prediction of the spread of COVID-19 infection in India. *Chaos, Solitons & Fractals*, Elsevier, v. 140, p. 110156, Nov 2020. doi.org/10.1016/j.chaos.2020.110156.

80 W. E. Raslan. Fractional mathematical modeling for epidemic prediction of COVID-19 in Egypt. *Ain Shams Engineering Journal*, Elsevier, v. 12, n. 3, p. 3057–3062, Set 2021. doi.org/10.1016/j.asej.2020.10.027.

81 S. Stanojevic et al. Simulation and prediction of spread of COVID-19 in the Republic of Serbia by SEAIHRDS model of disease transmission. *Microbial Risk Analysis*, Elsevier, v. 18, p. 100161, Ago 2021. doi.org/10.1016/j.mran.2021.100161.

82 H. Verma; S. Mandal; A. Gupta. Temporal deep learning architecture for prediction of COVID-19 cases in India. *Expert Systems with Applications*, Elsevier, v. 195, p. 116611, Jun 2022. doi.org/10.1016/j.eswa.2022.116611.

83 Y. Wang et al. Prediction and analysis of COVID-19 daily new cases and cumulative cases: time series forecasting and machine learning models. *BMC Infectious Diseases*, Springer, v. 22, n. 1, p. 495, Mai 2022. doi.org/10.1186/s12879-022-07472-6.

84 H. Abbasimehr; R. Paki. Prediction of COVID-19 confirmed cases combining deep learning methods and Bayesian optimization. *Chaos, Solitons Fractals*, Elsevier, v. 142, p. 110511, Jan 2021. doi.org/10.1016/j.chaos.2020.110511.

85 C.-P. Kuo; J. S. Fu. Evaluating the impact of mobility on COVID-19 pandemic with machine learning hybrid predictions. *Science of The Total Environment*, Elsevier, v. 758, n. 1, p. 144151, Mai 2021. doi.org/10.1016/j.scitotenv.2020.144151.

86 L. A. Amar; A. A. Taha; M. Y. Mohamed. Prediction of the final size for COVID-19 epidemic using machine learning: a case study of Egypt. *Infectious Disease Modelling*, Elsevier, v. 5, p. 622–634, Abr 2020. doi.org/10.1016/j.idm.2020.08.008.

87 R. S. Hirschprung; C. Hajaj. Prediction model for the spread of the COVID-19 outbreak in the global environment. *Heliyon*, Elsevier, v. 7, n. 7, Jun 2021. doi.org/10.1016/j.heliyon.2021.e07416.

88 A. Behnam; R. Jahanmahin. A data analytics approach for COVID-19 spread and end prediction (with a case study in Iran). *Modeling Earth Systems and Environment*, Springer, v. 8, n. 1, p. 579–589, Jan 2022. doi.org/10.1007/s40808-021-01086-8.

89 E. Gothai et al. Prediction of COVID-19 growth and trend using machine learning approach. *Materials Today: Proceedings*, Elsevier, v. 81, p. 597–601, Mai 2023. doi.org/10.1016/j.matpr.2021.04.051.

90 Q. Guo; Z. He. Prediction of the confirmed cases and deaths of global COVID-19 using artificial intelligence. *Environmental Science and Pollution Research*, Springer, v. 28, n. 1, p. 11672–11682, Jan 2021. doi.org/10.1007/s11356-020-11930-6.

91 A. Sinha; M. Rathi. COVID-19 prediction using AI analytics for South Korea. *Applied Intelligence*, Springer, p. 1–19, Jan 2021. doi.org/10.1007/s10489-021-02352-z.

92 P. Wang et al. Time series prediction for the epidemic trends of COVID-19 using the improved LSTM deep learning method: case studies in Russia, Peru and Iran. *Chaos, Solitons & Fractals*, Elsevier, v. 140, p. 110214, Jan 2020. doi.org/10.1016/j.chaos.2020.110214.

93 S. Ketu; P. K. Mishra. Retracted article: India perspective: CNN-LSTM hybrid deep learning model-based COVID-19 prediction and current status of medical resource availability. *Soft Computing*, Springer, v. 26, n. 2, p. 645–664, Nov 2022. doi.org/10.1007/s00500-021-06490-x.

94 A. Tomar; N. Gupta. Prediction for the spread of COVID-19 in India and effectiveness of preventive measures. *Science of The Total Environment*, Elsevier, v. 728, p. 138762, Ago 2020. doi.org/10.1016/j.scitotenv.2020.138762.

95 F. Shahid; A. Zameer; M. Muneeb. Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. *Chaos, Solitons Fractals*, Elsevier, v. 140, p. 110212, Jan 2020. doi.org/10.1016/j.chaos.2020.110212.

96 M. Hawas. Generated time-series prediction data of COVID-19's daily infections in Brazil by using recurrent neural networks. *Data in Brief*, Elsevier, v. 32, p. 106175, Out 2020. doi.org/10.1016/j.dib.2020.106175.

97 S. I. Alzahrani; I. A. Aljamaan; E. A. Al-Fakih. Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions. *Journal of Infection and Public Health*, Elsevier, v. 13, n. 7, p. 914–919, Out 2020. doi.org/10.1016/j.jiph.2020.06.001.

98 F. M. Khan; R. Gupta. ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in India. *Journal of Safety Science and Resilience*, Elsevier, v. 1, n. 1, p. 12–18, Set 2020. doi.org/10.1016/j.jnlssr.2020.06.007.

99 A. Swaraj et al. Implementation of stacking based ARIMA model for prediction of COVID-19 cases in India. *Journal of Biomedical Informatics*, Elsevier, v. 121, p. 103887, Set 2021. doi.org/10.1016/j.jbi.2021.103887.

100 S. Singh; K. S. Parmar; J. Kaur. Prediction of COVID-19 pervasiveness in six major affected states of India and two-stage variation with temperature. *Air Quality, Atmosphere & Health*, Springer, v. 14, p. 2079–2090, Set 2021. doi.org/10.1007/s11869-021-01075-x.

101 K. C. Santosh. COVID-19 prediction models and unexploited data. *Journal of Medical Systems*, Springer, v. 44, n. 170, Ago 2020. doi.org/10.1007/s10916-020-01645-z.

102 A. AlArjani; M. T. Nasseef; S. M. Kamal. Application of mathematical modeling in prediction of COVID-19 transmission dynamics. *Arab Journal of Science and Engineering*, Springer, v. 447, p. 10163–10186, Jan 2022. doi.org/10.1007/s13369-021-06419-4.

103 P. A. da Cruz; L. C. Crema-Cruz; F. S. Campos. Modeling transmission dynamics of severe acute respiratory syndrome coronavirus 2 in São Paulo, Brazil. *Revista da Sociedade Brasileira de Medicina Tropical*, SciELO, v. 54, p. e05532020, Jan 2021. doi.org/10.1590/0037-8682-0553-2020.

104 O. Torrealba-Rodriguez; R. A. Conde-Gutiérrez; A. L. Hernández-Javier. Modeling and prediction of COVID-19 in Mexico applying mathematical and computational models. *Chaos, Solitons & Fractals*, Elsevier, v. 138, p. 109946, Ago 2020. doi.org/10.1016/j.chaos.2020.109946.

105 A. Singhal et al. Modeling and prediction of COVID-19 pandemic using Gaussian mixture model. *Chaos, Solitons Fractals*, Elsevier, v. 138, p. 110023, Set 2020. doi.org/10.1016/j.chaos.2020.110023.

106 S. Rath; A. Tripathy; A. R. Tripathy. Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, Elsevier, v. 14, n. 5, p. 1467–1474, Out 2020. doi.org/10.1016/j.dsx.2020.07.045.

107 A. Gupta et al. Development and validation of a multivariable risk prediction model for COVID-19 mortality in the southern United States. *Mayo Clinic Proceedings*, Elsevier, v. 96, n. 12, p. 3030–3041, Dez 2021. doi.org/10.1016/j.mayocp.2021.09.002.

108 Z. Khraibani et al. Application of records theory on the COVID-19 pandemic in Lebanon: prediction and prevention. *Epidemiology and Infection*, Cambridge, v. 148, p. e192, Ago 2020. doi.org/10.1017/S0950268820001909.

109 P. Dubey et al. Learning delay dynamics for multivariate stochastic processes, with application to the prediction of the growth rate of COVID-19 cases in the United States. *Journal of Mathematical Analysis and Applications*, Elsevier, v. 514, n. 2, p. 125677, Out 2022. doi.org/10.1016/j.jmaa.2021.125677.

110 L. Refisch; F. Lorenz; T. Riedlinger. Data-driven prediction of COVID-19 cases in Germany for decision making. *BMC Medical Research Methodology*, Springer, v. 22, n. 116, Abr 2022. doi.org/10.1186/s12874-022-01579-9.

111 Ç. Ak et al. Spatiotemporal prediction of infectious diseases using structured Gaussian processes with application to Crimean–Congo hemorrhagic fever. *PLoS Neglected Tropical Diseases*, Public Library of Science, v. 12, n. 8, p. e0006737, Ago 2018. doi.org/10.1371/journal.pntd.0006737.

112 L. López; X. Rodó. A modified SEIR model to predict the COVID-19 outbreak in Spain and Italy: simulating control scenarios and multi-scale epidemics. *Results in Physics*, Elsevier, v. 21, n. 1, p. 103746, Fev 2021. doi.org/10.1016/j.rinp.2020.103746.

113 C. Scarpone et al. A multimethod approach for county-scale geospatial analysis of emerging infectious diseases: a cross-sectional case study of COVID-19 incidence in Germany. *International Journal of Health Geographics*, BioMed Central, v. 19, n. 1, p. 1–17, Ago 2020. doi.org/10.1186/s12942-020-00225-1.

114 G. Quaranta et al. Understanding COVID-19 nonlinear multi-scale dynamic spreading in Italy. *Nonlinear Dynamics*, Springer, v. 101, n. 3, p. 1583–1619, Set 2020. doi.org/10.1007/s11071-020-05902-1.

115 Brasil.io. *O Brasil em dados libertos*. 2024. https://brasil.io/home/, acesso em 16 jan. 2025.

116 W. Cota. Monitoring the number of COVID-19 cases and deaths in Brazil at municipal and federative units level. *SciELO Preprints*, SciELO Preprints, n. 1, Jul 2020. doi.org/10.1590/SciELOPreprints.362.

117 C. Starling et al. 404. COVID-19 normality rate: Criteria for optimal time to return to in-person learning. *Open Forum Infectious Diseases*, Oxford Academic, v. 8, n. 1, p. S303–S304, Dez 2021. doi.org/10.1093/ofid/ofab466.605.

118 D. Zaparolli. O desafio de calcular o R. *Revista Pesquisa FAPESP*, FAPESP, v. 293, p. 47, Jul 2022. https://revistapesquisa.fapesp.br/o-desafio-de-calcular-o-r/.

119 A. B. Abdessalem et al. Automatic kernel selection for Gaussian processes regression with approximate Bayesian computation and sequential Monte Carlo. *Frontiers in Built Environment*, Frontiers Media SA, v. 3, p. 52, Jun 2017. doi.org/10.3389/fbuil.2017.00052.

120 J.-L. Akian et al. Learning "best" kernels from data in Gaussian process regression. with application to aerodynamics. *Journal of Computational Physics*, Elsevier, v. 470, p. 111595, Dez 2022. doi.org/10.1016/j.jcp.2022.111595.

121 Y. Pan et al. Evaluation of Gaussian process regression kernel functions for improving groundwater prediction. *Journal of Hydrology*, Elsevier, v. 603, p. 126960, Dez 2021. doi.org/10.1016/j.jhydrol.2021.126960.

Apêndice A

Protocolo de Revisão Sistemática – Metodologia PRISMA

Tabela A.1: Protocolo de Revisão Sistemática

1	PALAVRAS-CHAVE		
		Predição	Prediction
		Múltiplas Escalas	Multi-Scale
		COVID	COVID
		Doenças Infecto Contagiosas	Infectious Diseases
2	IDIOMAS DOS ARTIGOS		
		Português	Inglês
3	STRING DE BUSCA		
		(previsão OR "múltiplas escalas") AND (COVID OR "Doenças Infecciosas")	
		(prediction OR "multiple scales"OR "multi-scale") AND (COVID OR "Infectious Diseases")	
4	FONTES DE BUSCA		
		Scielo – https://www.scielo.br/	
		Pubmed – https://pubmed.ncbi.nlm.nih.gov/	
		ScienceDirect – https://www.sciencedirect.com/	
		Springer – https://www.springer.com/br	
5	OBJETIVOS		

6	CRITÉRIOS DE INCLUSÃO	
		6.1 Revisão formal
		6.1.1 O documento completo está disponível
		6.1.2 Artigos publicados a partir de 2020
		6.1.3 Artigos publicados em periódicos
		6.1.4 Artigos publicados em revistas científicas
		6.1.5 Artigos publicos em congresso
		6.2 Literatura cinzenta e padronização
		6.2.1 A organização editorial e/ou os autores são reconhecidos na área
		6.2.2 A fonte possui metodologia e objetivos bem definidos
		6.2.3 É algo que enriquece ou acrescenta algo único na pesquisa
7	CRITÉRIOS DE EXCLUSÃO	
		7.1 Revisão formal
		7.1.1 Pôster
		7.1.2 Artigos resumidos
		7.1.3 Capítulos de livros
		7.1.4 O documento está duplicado
		7.2 Literatura cinzenta e padronização
		7.2.1 A metodologia não está claramente definida
		7.2.2 Existe conflito de interesses
		7.2.3 A data não está claramente indicada
		7.2.4 Fontes formais não foram vinculadas ou discutidas
8	PROCEDIMENTO DE SELEÇÃO, AVALIAÇÃO DE QUALIDADE E *SNOWBALLING*	

	8.1 Etapa 1	O procedimento de seleção será iniciado pela leitura de títulos e resumos de documentos e aplicação de critérios de inclusão e exclusão. Quando necessário, a conclusão será também analisada para aumentar a confiança na seleção. Cada documento vai ser avaliado pelo pesquisador. Posteriomente, o supervisor da pesquisa irá realizar a revisão das decisões com base na aplicação da estatística Cohen's Kappa.
	8.2 Etapa 2	Os documentos resultantes da primeira seleção serão analisados por completo para verificar se é possível extrair dados com base no formulário de extração de dados. Os documentos que possibilitam extração serão mantidos.
	8.3 Etapa 3	A técnica *snowballing* (*backwards and forwards*) será utilizada para reforçar a confiança na seleção de artigos. A técnica será aplicada em documentos selecionados para avaliação de qualidade.
	8.4 Etapa 4	Ao finalizar a primeira filtragem de documentos, atributos de qualidade serão analisados. A qualidade refere-se à medida em que o estudo minimiza o viés[1] e maximiza a validade interna[2] e externa[3].
9	SÍNTESE DE DADOS	
		Tabulação de informações extraídas dos documentos com base nas questões de pesquisa. Além disso, tabelas serão usadas para destacar similaridades e diferenças identificadas nos resultados (identificar se são homogêneos ou heterogêneos).

[1]Uma tendência a produzir resultados que se afastam sistematicamente dos resultados "verdadeiros". Resultados imparciais são internamente válidos.

[2]A extensão em que o desenho e a condução do estudo podem evitar erros sistemáticos. A validade interna é um pré-requisito para a validade externa.

[3]A extensão em que os efeitos observados no estudo são aplicáveis fora do estudo.