UNIVERSIDADE DO ESTADO DA BAHIA

DEPARTAMENTO DE CIÊNCIAS EXATAS E DA TERRA

CURSO DE GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

PEDRO JESUS DO ROSÁRIO

PROJETO HERMES: IDENTIFICAÇÃO DE ONDAS PARA PARA ANÁLISE DESCRITIVA E PREDITIVA DO QUADRO EPIDEMIOLÓGICO DA COVID-19

SALVADOR

2021

PEDRO JESUS DO ROSÁRIO

PROJETO HERMES: IDENTIFICAÇÃO DE ONDAS PARA PARA ANÁLISE DESCRITIVA E PREDITIVA DO QUADRO EPIDEMIOLÓGICO DA COVID-19

Monografia apresentada ao curso de Sistemas de Informação do Departamento de Ciências Exatas e da Terra da Universidade do Estado da Bahia - UNEB, como requisito à obtenção do grau de bacharel em Sistemas de Informação. Área de Concentração: Ciência da Computação

Orientador: Leandro S. Coelho de Souza

Coorientador: Diego Gervasio Frias Suarez

SALVADOR

2021

FICHA CATALOGRÁFICA
Sistema de Bibliotecas da UNEB

J58p Jesus do Rosário, Pedro

PROJETO HERMES: IDENTIFICAÇÃO DE ONDAS PARA PARA ANÁLISEDESCRITIVA E PREDITIVA DO QUADRO EPIDEMIOLÓGICO DA COVID-19 / Pedro Jesus do Rosário. - Salvador, 2021.
80 fls.

Orientador(a): Leandro S. Coelho de Souza.
Coorientador(a): Diego G. Friaz Suarez.
Inclui Referências
TCC (Graduação - Sistemas de Informação) - Universidade do Estado da Bahia. Departamento de Ciências Exatas e da Terra. Campus I. 2021.

1.Coronavirus. 2.Pandemia. 3.COVID-19. 4.Predição. 5.Regressão.

CDD: 004

PEDRO JESUS DO ROSÁRIO

PROJETO HERMES: IDENTIFICAÇÃO DE ONDAS PARA PARA ANÁLISE DESCRITIVA E PREDITIVA DO QUADRO EPIDEMIOLÓGICO DA COVID-19

Monografia apresentada ao curso de Sistemas de Informação do Departamento de Ciências Exatas e da Terra da Universidade do Estado da Bahia - UNEB, como requisito à obtenção do grau de bacharel em Sistemas de Informação. Área de Concentração: Ciência da Computação

Aprovada em:

BANCA EXAMINADORA

Leandro S. Coelho de Souza (Orientador)
Universidade do Estado da Bahia – UNEB

Diego Gervasio Frias Suarez (Coorientador)
Universidade do Estado da Bahia - UNEB

Dra. Marcia São Pedro Leal Souza
Diretora de Vigilância Epidemiológica de Secretaria da Saúde (SEBAB)

Hugo Saba Pereira Cardoso
Universidade do Estado da Bahia - UNEB

Magno Conceição das Mercês
Universidade do Estado da Bahia - UNEB

Vagner de Souza Fonseca
Organização Panamericana da Saúde - OMS

Walter Massa Ramalho
Organização Panamericana da Saúde - OMS

Silvano Barbosa de Oliveira
Organização Panamericana da Saúde - OMS

Dedico esta monografia à minha mãe, formada com honras nesta mesma instituição, e que embora hoje traduzida em uma imensurável saudade, seguiu sendo protagonista da minha motivação durante toda a minha formação, bem como a pessoa que acreditava no meu potencial até quando nem mesmo eu o fazia.

AGRADECIMENTOS

Gostaria de agradecer às seguintes pessoas:

Minha noiva Késia Luane, com quem pude contar com a companhia durante toda esta jornada, seja em tempos bons ou ruins.

Meus familiares, por todo o suporte que me permitiu por muito tempo focar, de forma exclusiva, nos meus compromissos acadêmicos, além de todo o suporte emocional a mim disponibilizado.

Alex Freire Spinola e Thiago Luís Armede, outrora colegas de turma e grupo para trabalhos acadêmicos, hoje valiosos amigos e colegas de profissão, aos quais possuo eterna dívida por todos os momentos de apoio tanto nas esferas acadêmicas quanto profissionais.

Meus colegas de turma Cleber Del Rei, Yan Gabriel, Diego dos Santos, Henrique Pacheco, Cândido Júnior e Matheus Tanure, pessoas com as quais pude contar com auxílios que transcenderam os muros da academia, e que seguirão com o meu apresso por tempo indeterminado.

Diego Costa e Diógenes Sampaio, que, através de diferentes abordagens, me ofereceram o suporte e o encaminhamento profissional que me fizeram chegar ao ponto em que me encontro na minha carreira enquanto desenvolvedor de software.

Ao meu grande amigo e orientador Leandro Coelho, que hoje ocupa a cadeira de diretor do Departamento de Ciências Exatas e da Terra - DCET I (UNEB), e que me apresentou a iniciativa científica, sendo o canal para o meu primeiro contato com trabalhos científicos, que me fizeram me destacar na Jornada de Iniciação Científica, e foi de grande valor para a realização desta pesquisa.

Ao saudoso professor Jorge Farias, que se empenhava para extrair o melhor das minhas capacidades, e que, por meio de uma metodologia orientada a desafios, me fez perceber o quão longe eu poderia ir se acreditasse no meu potencial.

"O futuro pertence àqueles que se preparam hoje para ele."

(Malcom X)

RESUMO

O presente trabalho tem como objetivo apresentar um estudo sobre o uso de funções sigmoides, com foco na função de Richards e suas derivações, para modelar a trajetória da pandemia da COVID-19. A partir dos trabalhos já publicados com o objetivo de modelar epidemias e pandemias, é possível identificar a existência de uma quantidade considerável de pesquisas acerca da capacidade descritiva, isto é, a eficácia em representar a trajetória de uma determinada epidemia dentro de um intervalo de tempo definido (também conhecido como onda), de modelos baseados em funções de crescimento sigmoides. No entanto, poucos destes trabalhos possuem foco na COVID-19, tampouco propõem uma análise sobre a capacidade preditiva, isto é, a capacidade de projetar o cenário futuro de uma pandemia com base no seu histórico, dos modelos anteriormente citados. Das pesquisas que visam a realização de predições de quadros epidemiológicos, pôde-se provar que é possível projetar, em até 45 dias pra alguns casos (entenda-se casos como cidades, estados, países ou regiões), o cenário de um determinado surto de doença, por meio de regressões não-lineares baseadas em funções sigmoides. Desta forma, a presente pesquisa descreve a análise do poder de predição de um modelo baseado na função de Richards, por meio da construção de uma interface gráfica interativa, trazendo além dos resultados dos testes de usabilidade, uma análise da variação de eficácia do modelo em questão quando aplicado a diferentes fases da pandemia, bem como a diferentes regiões do mundo.

Palavras-chave: Coronavirus. Predição. Richards. Covid-19. Pandemia. Gompertz. Logística. Regressão.

ABSTRACT

This paper aims to present a study on the use of sigmoid functions, focusing on the Richards function and its derivations, to fit the trajectory of the COVID-19 pandemic. From the works already published with the objective of fit epidemic outbreaks, it is possible to identify the existence of a certain amount of research on descriptive capacity, that is the effectiveness in representing the trajectory of a specified epidemic within a specified time interval defined (a.k.a wave), of models based on sigmoid growth functions. However, few of these works focus on COVID-19, nor do they propose an analysis of a predictive capacity, that is the ability to project the future scenario of a pandemic based on its history, on the aforementioned models. From research that aimed at making predictions of epidemiological conditions, it could be proved that it is possible to project, in up to 45 days for some cases (understanding cases such as cities, states, countries or regions), the scenario of a specific disease, through nonlinear regressions based on sigmoid functions. Thus, this research analyzes the predictive power of a model based on the Richards function, through the construction of an interactive graphical interface, bringing, in addition to the results of usability tests, an analysis of the variation in the model's effectiveness in difference when applicable to different phases of the pandemic as well as to different regions of the world.

Keywords: Coronavirus. Pandemic. Covid-19. Predictions. Richards. Gompertz. Logistic.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

2019-nCov	*2019 New Coronavirus*
COVID-19	*Coronavirus Desease 2019*
IEEE	*Institute of Electrical and Electronics Engineers*
OMS	Organização Mundial da Saúde
PIMAT	*Predições Inteligentes sobre Métodos Aplicados a séries Temporais*
RMSE	Raiz quadrada do erro-médio
R^2	R Quadrado
Sars-Cov-2	*Severe Acute Respiratory Syndrome Coronavirus 2*
SSE	Soma dos Quadrados dos Erros
WHO	*World Health Organization*

LISTA DE SÍMBOLOS

s	Grau de assimetria de assíntotas
k	Assíntota superior ou plateau
r	Taxa de crescimento do número de mortes
$N(t)$	Número de mortes acumuladas no dia t

SUMÁRIO

1 INTRODUÇÃO

Em dezembro de 2019, na cidade de Wuhan, situada na província chinesa de Hubei, um fato chamava a atenção da comunidade médica. Um surto de pneumonia passou a acometer os moradores locais. A pneumonia, no momento de origem desconhecida, se espalhava de forma rápida, e apesar de não apresentar grau de letalidade alto, representava sério risco ao sistema de saúde da China, bem como do mundo. O motivo do risco é que, por conta do vírus ser extremamente contagioso (1), este levava um número crescente de pessoas a buscar serviços médicos simultaneamente, o que poderia, em algum momento, representar uma demanda a qual os hospitais da região não estavam preparados para lidar.

Além do risco de sobrecarga no sistema de saúde, o alto número de infectados indica um potencial risco de surgimento de variantes de vírus em geral (cepas), uma vez que estes possuem a capacidade de se adaptar a novos hospedeiros (2), o que tornaria mais complexa a busca por formas de anular ou eliminar os efeitos provocados pelos mesmos.

Com o avanço do número de novos casos na China, no dia 31/12/2019, a Organização Mundial de Saúde (OMS) emitiu um alerta global sobre casos de pneumonia de origem desconhecida na China, e no dia seguinte é anunciado pelo governo chinês que a infecção teria começado em um grande mercado de frutos do mar, em Wuhan, que foi imediatamente fechado após o surgimento do surto. Apesar de classificado com foco em frutos do mar, neste mercado eram comercializadas carnes das mais variadas espécies de animais, sob condições sanitárias questionáveis, motivando portanto adoção da hipótese de que o paciente 0 (primeiro indivíduo a contrair a doença) teria sido infectado por meio da ingestão de algum alimento adquirido neste comércio (esta hipótese, apesar de possuir sólida fundamentação, ainda não foi comprovada. Conforme relatado em (3), em Outubro de 2021 a China iniciou uma análise de mais de 200 mil amostras de sangue, ainda com o intuito de se provar a origem da COVID-19).

Embora já fosse alarmante a situação em algumas províncias da China, os olhos do mundo se voltaram para esta doença no dia 07/01/2020, quando foi confirmado, por autoridades chinesas, que a pneumonia citada era causada por uma variante de um grupo de vírus já conhecida pela comunidade médica e científica. O grupo em questão é intitulado Coronavírus, pela sua

aparência similar à de uma coroa (a palavra *corona* significa "coroa"em alguns idiomas como italiano e espanhol) (4). Uma variante deste grupo de vírus, nomeada Betacoronavirus, já havia protagonizado, em 2002, uma epidemia de SARS (*Severe acute respiratory syndrome*) - uma doença respiratória grave caracterizada por febre, dor de cabeça, dores no corpo, tosse seca, hipóxia e geralmente pneumonia. Na época, a doença havia se espalhado em mais de 25 países, culminando em 8.098 casos confirmados e 774 mortes (5). Dada a necessidade de identificação desta nova variante do Coronavirus, o governo Chinês atribuiu ao vírus o nome *2019 New Coronavirus (2019-nCov)*, sendo posteriormente renomeado para *Severe Acute Respiratory Syndrome Coronavirus 2 (Sars-Cov-2)*, e a doença provocada pelo mesmo foi intitulada *Coronavirus Desease 2019 (COVID-19)* (6).

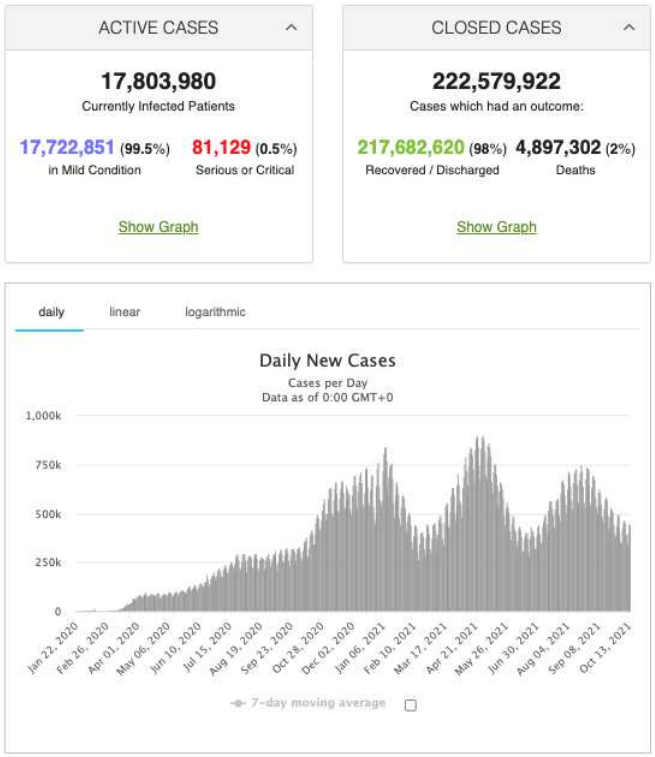
A COVID-19 seguiu sendo disseminada no território Chinês, que em 09/01/2020 registrou o primeiro óbito de uma pessoa por conta da doença. Nos dias seguintes, casos de COVID-19 foram identificados em outros países, como Tailândia (13/01/2020), Japão (16/01/2020) e Estados Unidos da América (21/01/2020) (7). Tal evolução levou a OMS a declarar, no dia 11/03/2020, a COVID-19 como uma pandemia. Na altura do fato, o número de casos da pneumonia fora do território chinês havia aumentado em 13 vezes em relação às ultimas duas semanas (7).

Nos meses seguintes, a pandemia em questão foi mote de uma série de eventos de grande impacto nos contextos políticos e econômicos em todo o globo, como sucessivos *circuit breaks* na bolsa de valores, atos de represália ao governo da China, pânico e estado de calamidade em diversas regiões do mundo (8), aumento do índice de violência doméstica (9), ansiedade, depressão, suicídio (10), etc. Concomitantemente, houve esforços no cenário médico e científico, como a busca por eventuais tratamentos farmacológicos, preventivos ou corretivos, além da urgência no desenvolvimento de vacina, levantamento de bases de dados epidemiológicos, e da definição de protocolos de segurança sanitária, com a finalidade de interromper o desenvolvimento da doença. Dada a ausência de soluções, os governos foram motivados a adotar uma série de medidas restritivas para suavizar a curva de crescimento do número de pessoas infectadas, de modo a permitir que os sistemas de saúde sejam capazes de lidar com a demanda por leitos hospitalares (11).

Em Outubro de 2021, o impacto da pandemia ainda apresenta variação considerável entre nações. Neste período de pandemia (passados quase 2 anos desde o seu início) já havia a

disponibilidade de diferentes vacinas para a COVID-19, e alguns países já vacinaram parcial ou integralmente a sua população. Apesar da existência de diferentes vacinas para a doença em questão, estas ainda carecem de estudos e aprimoramentos técnicos/científicos uma vez que pouco se avançou em questões importantes como duração do tempo de imunização, abrangência quanto à proteção contra variantes do vírus, possíveis efeitos colaterais e segurança em grupos de risco como imunossuprimidos, grupo de jovens e crianças, refletindo inclusive no quantitativo de doses necessárias para garantir um segurança mínima. Os dados coletados pela *World Health Organization* (WHO) revelam que, na data 23/10/2021, aproximadamente 50% da população mundial já havia recebido no mínimo a primeira dose da vacina, e 37.3% foram classificados como totalmente vacinados, estando a maior parte desta população distribuída nos continentes: América do Norte, América do Sul, Europa, Ásia e Oceania (12). No entanto, analisando o quadro de uma perspectiva global, o mundo ainda não venceu a pandemia. Como mostra a Figura 1, é possível notar que na data, registram-se aproximadamente 500 mil novos casos, e 17.8 milhões de pessoas enfrentando a doença no momento. Além disso, registra-se também em (7) uma média de aproximadamente 8 mil mortes diárias, e um total de 4.8 milhões de mortes em decorrência da COVID-19.

Figura 1 – Gráfico do registro de casos diários no intervalo de 22/01/2020 a 15/10/2021.



Fonte: WORLDOMETERS, 2021. (7)

Tendo em vista os fatos, nota-se que, apesar vacinas, do avanço da medicina relacionada ao acolhimento e tratamento e das diversas medidas restritivas adotas em todo o globo durante o ano de 2020 e 2021, o fim da pandemia e o retorno às atividades de forma regular ainda não é uma realidade em muitos países (13). A campanha global de vacinação atualmente ainda não conseguiu alcançar todas as regiões do mundo. Países do continente Africano bem como do Oriente Médio ainda apresentam taxas de vacinação baixas quando comparados a países do continente europeu, por exemplo (12). Além disso, a possibilidade do surgimento de novas variantes do vírus, bem como teorias sobre o período de imunização das vacinas, deixam o mundo em constante alerta, haja visto que estas podem ocasionar novas ondas epidemiológicas, em países que já haviam controlado a pandemia, como por exemplo Israel, que em Junho de 2021 chegou a registrar 0 casos diários da doença, porém em Agosto do mesmo ano passou pela sua maior onda epidêmica, chegando a registrar 20.000 casos diários em Setembro do mesmo ano (7). A situação da COVID-19 ainda é considerada grave, devido a diferentes razões, como: Conflitos de interesse político-econômico, a natureza questionável de algumas medidas preventivas, casos de displicência quanto ao cumprimento dos protocolos de segurança sanitária instaurados (1), etc. Como consequência, surge a necessidade de avaliação constante da efetividade dos protocolos de segurança (politicas públicas de restrição de circulação, distanciamento social, etc), do subsidio à área da saúde, especificamente a área epidemiológica com ferramentas que permitam a analise local, regional e global e, investimento em pesquisa para tratamentos farmacológicos e produção de vacinas.

Além da divergência de opiniões acerca da efetividade de alguns protocolos de restrição adotados, da eficiência e eficácia das vacinas, e de outras ações e combate a COVID-19, ainda não é possível precisar o termino da pandemia, caso a relação de contágio e de medidas de combate/prevenção adotadas permaneça a mesma.

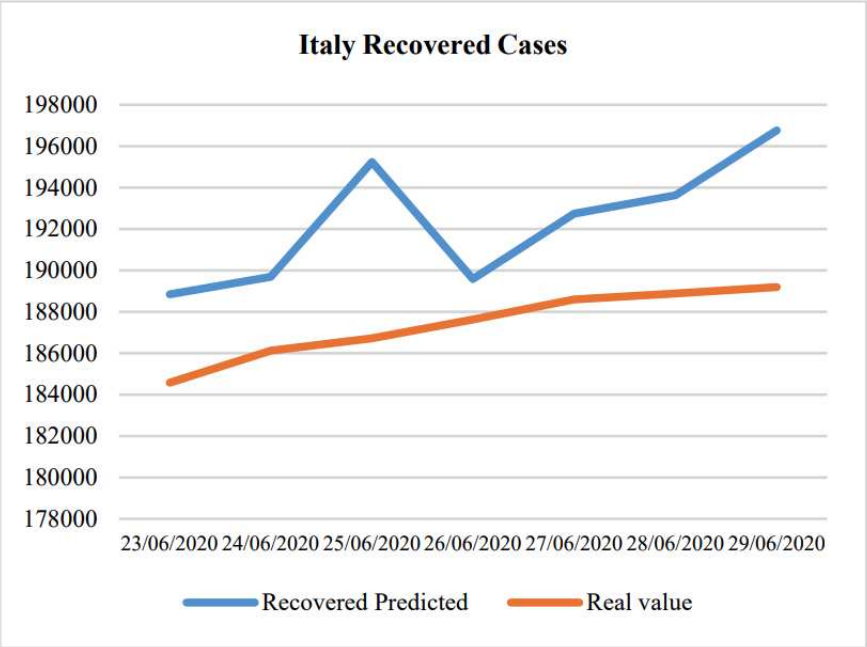
Partindo do fato de que o mundo ainda carece de uma visão acerca da efetividade das estratégias adotadas, de mais ferramentas para analisar a real situação da doença causada pelo Sars-Cov-2, bem como de estimar o cenário futuro da pandemia, fornecendo insumos para uma tomada de decisão por parte de lideranças sanitárias, públicas, e científicas, foi realizado o presente trabalho. O principal objetivo deste, é avaliar, de trabalhos já desenvolvidos pela comunidade científica, e dos principais trabalhos desenvolvidos pelo grupo de pesquisa *Predições Inteligentes sobre Métodos Aplicados a séries Temporais* (PIMAT), quais técnicas se mostram

eficientes em descrever a trajetória da pandemia da COVID-19, com base em critérios de eficiência (isto é, a relação entre taxa de assertividade e custo computacional), e construir uma ferramenta capaz de fornecer uma estimativa do número novos casos diários, bem como casos acumulados da doença causada pelo novo Coronavirus. O foco deste trabalho se traduz em validar a hipótese de que funções sigmoides (como as funções de Richards, Gompertz e Logística) podem ser utilizadas tanto para modelar, quanto para prever o cenário futuro da pandemia da COVID-19, bem como provar a possibilidade de identificação dinâmica da ocorrência de múltiplas ondas em uma série temporal epidemiológica.

Isto posto, foi realizada uma pesquisa nas bases do grupo PIMAT bem como em artigos médicos na plataforma *PubMed*, e em artigos de cunho matemático e estatístico na plataforma *Institute of Electrical and Electronics Engineers* (IEEE), tendo como principal meta realizar uma revisão de literatura, isto é, elencar trabalhos com objetivo similar (modelagem de pandemias e epidemias) afim de se obter insumos que viabilizassem uma pesquisa assertiva. Desta etapa, foram identificados trabalhos que se propunham a modelar a pandemia por meio de diversas abordagens, como em (14) e (15), que fizeram uso de técnicas de *Deep Learning* - Aprendizagem de máquina profundo, para modelar a curva de crescimento do número de recuperados da COVID-19 em países como Itália, Estados Unidos e Índia.

Tanto em (14) quanto em (15), foram verificados resultados promissores, como taxas de erro abaixo de 3% nas predições realizadas pelo modelo desenvolvido quando comparadas aos valores reais, presentes nos *datasets (bases de dados)*, como se observa nos seguintes gráficos:

Figura 2 – Comparação entre número oficial de recuperados na Itália, e as predições do modelo desenvolvido.

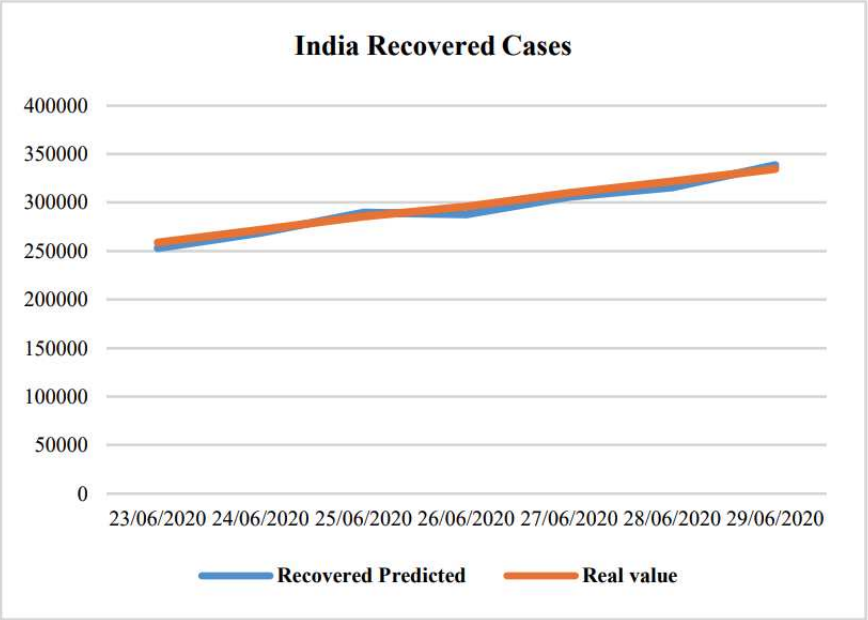


Fonte: KDAYEM, 2021 (14)

No entanto, os autores obtiveram tais retornos ao treinar o modelo com janelas de tempo e *datasets* específicos, não ficando provado no artigo que o modelo desempenharia de igual maneira para dados das demais regiões do mundo. Além disso, fica evidenciado no trabalho o poder computacional demandado pela solução, mesmo o processamento tendo sido feito sobre dados diários e acumulados de um intervalo de 50 dias. De maneira concomitante, já existem ferramentas publicadas que se propõem a fornecer projeções dos dados da COVID-19 com base em modelos de aprendizagem de máquina, como exemplificado em (16), o que colocaria em risco o valor científico da pesquisa caso a ferramenta a ser construída fizesse uso de modelo similar. Às características já citadas, soma-se o fato de que, do que se foi pesquisado sobre soluções envolvendo aprendizagem de máquina, como redes neurais, não foram encontradas evidências de que, passada a etapa de treinamento do modelo, seria possível identificar de que maneira o mesmo estrutura o seu processo de decisão, de modo que o seu funcionamento se assemelha ao modelo de caixa-preta. Portanto, concluiu-se que, embora existam trabalhos que apresentem altas taxas de assertividade, ao considerar o objetivos da pesquisa, o uso de *Deep Learning* não traduziria a melhor solução.

Outros trabalhos buscaram modelar a trajetória de pandemias e epidemias com métodos alternativos ao uso de *Machine Learning*, como na pesquisa realizada em (17), no

Figura 3 – Comparação entre número oficial de recuperados na Índia, e as predições do modelo desenvolvido.



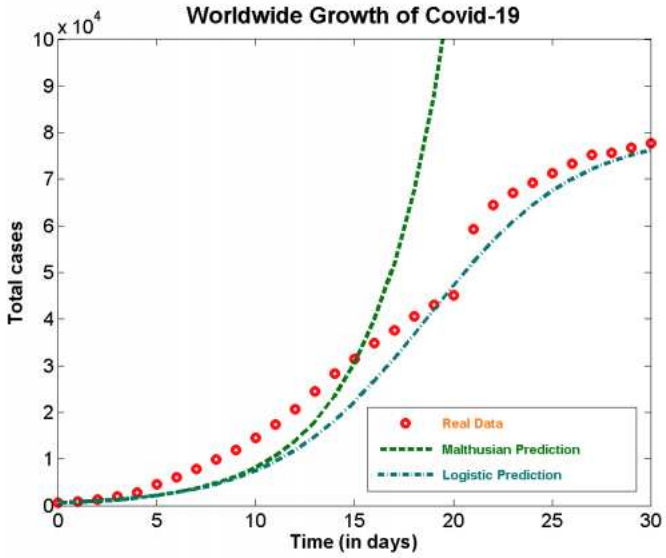
Fonte: KDAYEM, 2021 (14)

qual os autores buscaram analisar o comportamento das funções de crescimento Malthusiana e Logística para modelar o comportamento da pandemia da COVID-19. Os autores confirmaram a melhor eficiência da função Logística para descrever a pandemia quando comparada à função Malthusiana. No entanto, o trabalho diverge da proposta da presente pesquisa nos seguintes aspectos:

1. **Análise da pandemia sob uma perspectiva global:** Os autores utilizaram dados da pandemia de maneira global, não se preocupando com as particularidades de desenvolvimento da pandemia em cada pais/região, como é possível verificar na figura 4:
2. **Foram desconsideradas as particularidades das funções**: os autores se limitaram a comparar duas funções de crescimento, sendo uma exponencial (Malthusiana) e outra sigmoide (Logística), além de não levarem em consideração a característica função Logística de ser, obrigatoriamente, simétrica. O trabalho induz que a função logística apresenta melhores resultados do que a função Malthusiana. No entanto, a comparação traduz um cenário de pouco valor, haja visto que, não foram encontrados estudos que comprovem que a curva de crescimento de pandemias segue, obrigatoriamente, um modelo simétrico.

3. **Ausência de análise sob uma ótica estatística aos resultados**: os autores concluem no

Figura 4 – Comparação entre número oficial de casos acumulados dos primeiros 30 dias de pandemia, e a modelagem por regressão não linear utilizando as funções *Malthusian* e *Logistic*



.

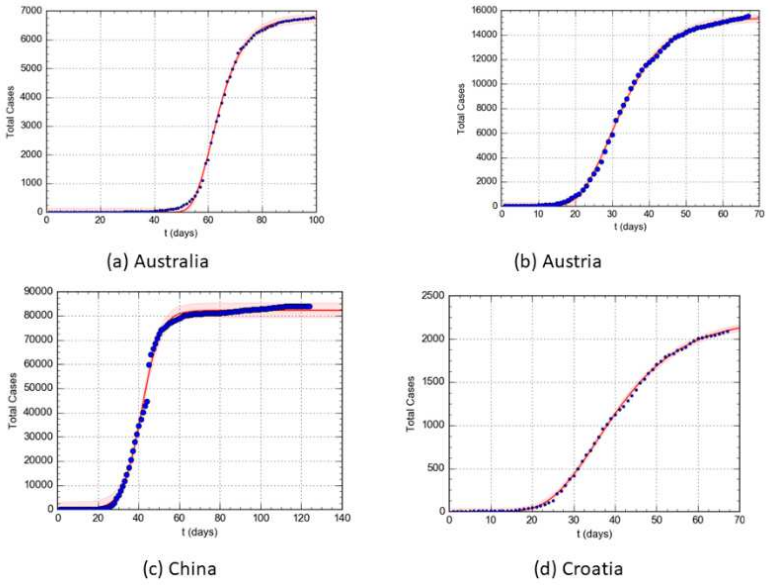
Fonte: ISLAM, 2020 (17)

artigo que função Logística se mostrou a mais indicada (dentre as funções analisadas) para modelar a pandemia, mas não apresentam dados estatísticos para o modelo utilizado - como o coeficiente de determinação, Raiz quadrada do erro-médio (RMSE), Soma dos Quadrados dos Erros (SSE) ou outros necessários para endossar os resultados.

4. **Conclusão excessivamente ampla**: Apesar de estudarem a capacidade descritiva das funções utilizadas os autores não apresentam direcionamento conclusivo pertinente para contribuir com a questão da COVID-19.

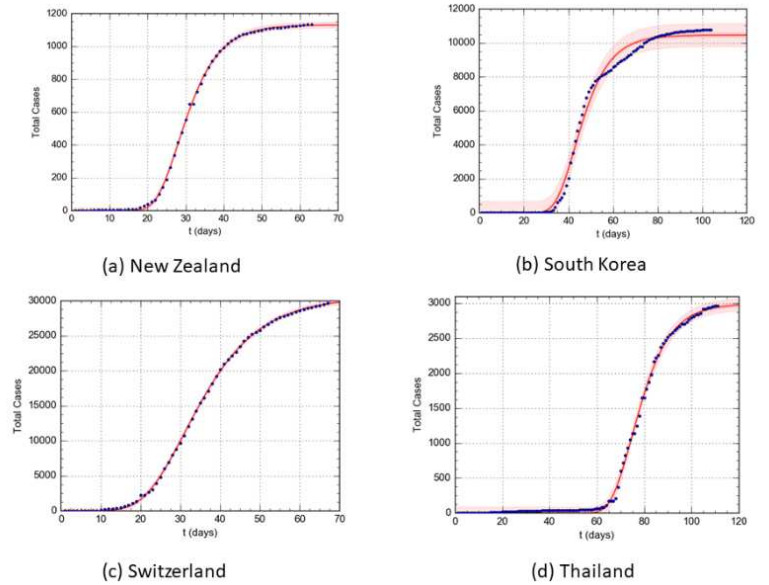
Outra pesquisa que se propôs a modelar a pandemia por meio de regressão não-linear com base em funções sigmoides pode ser encontrada no trabalho realizado em (18), no qual os autores propõem um método para determinar quando e como é possível usar o ajuste não linear das funções de crescimento sigmoide (Gompertz e Logística) para estimar a evolução dos casos COVID-19 ao longo do tempo. Os autores testaram a eficiência do método proposto em oito países e comprovaram a eficiência preditiva através do método gerando importante contribuição cientifica, como demonstrado nas imagens 5 e 6, onde os dados analisadas estão representados pelos pontos azuis e a linha vermelha representa a função sigmoide (entre Gompertz e Logística) que melhor se ajustou às medições:

Figura 5 – Modelagem para os casos da Austrália, Áustria, China e Croácia



Fonte: MELO, 2020 (18)

Figura 6 – Modelagem para os casos da Tailândia, Suíça, Coreia do Sul e Nova Zelândia



Fonte: MELO, 2020 (18)

Contudo, o referido trabalho difere da presente proposta, uma vez que não insere na sua análise o uso da função de Richards - função sigmoide com proposta mais generalista, da qual se derivam tanto a função Logística quanto a função de Gompertz. Além disso, não há em (18) uma descrição da eficiência e limitações dos modelos apresentados no que diz respeito à capacidade preditiva, sendo portanto a análise limitada à capacidade descritiva dos mesmos. A

isso, soma-se o fato de que os autores não levaram em consideração um cenário com múltiplas ondas, o que é tratado na presente pesquisa.

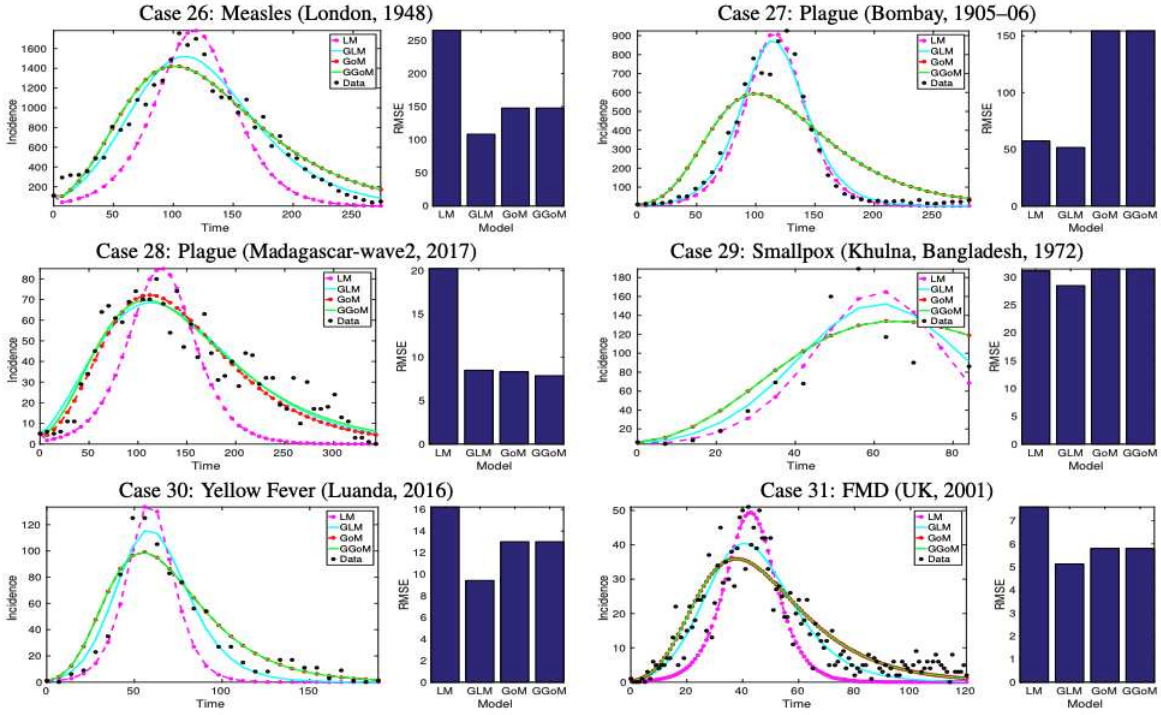
De similar maneira, a pesquisa proposta em (19) faz uso da função de crescimento de Gompertz para estimar o número máximo de óbitos nos estados brasileiros. Neste, os autores realizam a análise da capacidade descritiva do método, porém, não se aprofundam na discussão acerca da capacidade preditiva da função em questão. Ainda sobre a pesquisa, os autores utilizam utilizam o valor aproximado de 2,7 para a constante neperiana, assumindo a margem de erro causada pela supressão dos valores decimais presentes na mesma. Esta decisão pode ter impacto nas métricas obtidas, haja visto que, quando se trata de funções exponenciais e sigmoides, diferenças mínimas entre valores podem ser propagadas de modo a produzirem diferenças sensíveis nos resultados finais. Por fim, o referido artigo assume que existe um valor de pico para cada estado brasileiro estudado, o que, sob a ótica da presente proposta, se trata de situação hipotética e portanto não deveria ser considerada.

Dos trabalhos analisados, que mais se aproximou da presente proposta foi o (20). Neste, os autores tratam da analise de quatro funções de crescimento para avaliar pandemias ao longo da história: Logística, Gompertz e suas generalizadas. Eles estudaram diferentes surtos de doenças infecciosas (como HIV, Malaria e Febre amarela), traduzidas em séries temporais de incidência de casos, compreendendo diferentes contextos. Como principal contribuição, ficou comprovada a capacidade descritiva das funções estudadas para modelar diferentes pandemias, como demonstrado nos gráficos da figura 7.

Entretanto, a pesquisa feita em (20) não realiza uma verificação especifica para o caso da COVID-19. A pesquisa a ser apresentada nas próximas sessões, além tem como objeto de estudo a pandemia da COVID-19, inicia a analise pela função de Richards, tendo como casos particulares as funções de Gompertz e Logística. Essa diferenciação é um fator importante que marca a relevância da presente proposta. Além disto, a pesquisa tratada neste documento focou nos casos diários e acumulados, havendo observado a correlação entre os casos de infecção e os casos de óbitos, o que na prática determina uma melhor capacidade de análise, como será apresentado nas sessões posteriores. Outra diferença marcante encontra-se no fato da apresentação de dados preditivos, não observados na referencia tratada, e nesta proposta considerados. Além disso, todos os artigos analisados demonstraram esforços na direção da descrição de uma única onda da pandemia, de maneira isolada, o que também difere da presente

proposta, uma vez que esta também considera casos em que duas (ou mais) ondas se sobrepõem, e insere a necessidade do desenvolvimento de soluções que permitam identificar as interseções entre as ondas de modo a isolar e modelar dinâmica.

Figura 7 – Resultados de modelagem por meio das funções de Gompertz e Logística, para surtos epidêmicos.



Fonte: YISSEDT, 2019 (20)

2 MODELAGEM PARA ANÁLISE EPIDEMIOLOGICA

A pandemia mundial da COVID-19 obrigou uma reorganização social e econômica em todo o mundo. A falta de confirmação da eficácia das vacinas, atrelada a ausência de solução farmacológica, tem remetido à análise dos dados epidemiológicos que por sua vez, remetem a especulações sobre a evolução da doença o que por si só tem subsidiado a adoção de políticas públicas de contingenciamento como a exemplo da implantação do distanciamento social, isolamento e até mesmo *lockdown* de populações inteiras, além de ações de educação em saúde voltadas ao uso correto das máscaras e à limpeza/assepsia das mãos.

Modelar a evolução de uma pandemia, com intuito de proporcionar dados epidemiológicos que sejam válidos para a correta apuração e predição da evolução da doença, não é tarefa simples e requer utilização de diversas variáveis que contemplem a situação real a ser modelada, sobretudo quando os dados epidemiológicos existentes se mostram bastante imprevisíveis, com comportamentos assimétricos e dispersos em diversos ciclos de propagação, considerados pela literatura como ondas epidemiológicas (21).

Todos os métodos de análise de dados epidemiológicos são baseados em algum modelo diferencial, integral ou híbrido, que descrevem a interação entre o patógeno e o hospedeiro no meio social em que ocorre a epidemia (22). Um determinado modelo também pode fornecer uma solução analítica, o que significa que algumas funções matemáticas estariam disponíveis e poderiam ser utilizadas diretamente para calcular as variáveis epidemiológicas que seriam dependentes do tempo (número de novas infecções e óbitos) usando parâmetros que representam as informações-alvo, como a taxa de crescimento e o tempo de recuperação (14).

Nesse caso, o modelo seria classificado como tipo funcional porque é composto por uma ou mais funções. Assim, quando a solução analítica não estiver disponível, ou seja, quando não houver um modelo funcional disponível, é necessário utilizar, como solução, técnicas que podem ser de natureza estocástica (Método das Cadeias de Monte Carlo (23), Redes Bayesianas (24), entre outros) ou determinísticas, onde é utilizada uma variedade de métodos de integração numérica de sistemas de equações diferenciais.

Em geral, modelos matemáticos são pautados em hipóteses que quantificam os

principais aspectos biológicos da propagação de epidemias e procuram fornecer informações basicamente sobre dois parâmetros epidemiológicos: a força de infecção, entendido como as observações novos casos em função do tempo) e a razão de reprodutibilidade basal, representando o número de casos secundários que um determinado caso primário é capaz de produzir em uma população totalmente suscetível (25).

A despeito dos modelos existentes para modelagem de epidemias como o modelo comportamental SIR (Susceptíveis Infectados e Recuperados) (26), composto por equações diferenciais ou mesmo o modelo MIB (Modelo Baseado em Indivíduos) (27), que utiliza o SIR como base operacional, a análise de dados através da utilização de funções de crescimento para modelagem epidemiológica tem surgido como alternativa mais simples e precisa, sobretudo no caso do surto epidemiológico da COVID-19 (28).

A análise desses dados fornece dois objetivos diferentes, mas relacionados: descritivo e preditivo. A análise descritiva visa extrair dados específicos de fases distintas da epidemia para fornecer informações relevantes como a taxa de crescimento, o ponto de inflexão (aqui definido como o momento em que a taxa de crescimento começa a diminuir após um período de crescimento contínuo) e o momento de pico, após o que o número médio de casos diários começa a diminuir com o tempo. Por outro lado, a abordagem preditiva visa prever a quantidade de casos, no pico e também quando (data) o pico ocorre, bem como o número total de ocorrências (infecções e óbitos) na epidemia e seu dia de término.

Embora os modelos descritivos e preditivos baseados em funções de crescimento tenham sido amplamente utilizados para modelar a atual pandemia COVID-19 (como visto em (20), (17), e (29)), é fato que os modelos fenomenológicos ignoram completamente essas variações dos parâmetros durante o curso das epidemias, dando a impressão de que os parâmetros estimados são descritos ao longo de toda a onda epidêmica.

Outros estudos já discutiram e analisaram as questões relativas a representatividade dos parâmetros das funções de crescimento (30) assim como já comprovaram a capacidade descritiva destas funções para outras epidemias como as produzidas pelos vírus Zica, Chikungunya e H1N1, dentre outras (20).

Não obstante, existem diferentes tipos de doenças infecciosas. Algumas delas assumem proporções epidêmicas ou pandêmicas, como a exemplo da COVID-19 ou mesmo da Gripe

Suína (H1N1). Quando o progresso destas doenças se prologa no tempo, é registrada uma série temporal que forma padrões de picos e vales que mostram os incidentes de infecção no período analisado. Alguns autores acordaram em denominar este padrão como onda epidemiológica. Sem embargo, não existe uma definição fixa para uma onda de pandemia não havendo uma agência de saúde que defina universalmente uma onda de pandemia apesar dos esforços de alguns trabalhos científicos em cobrir esta lacuna.

2.1 ONDAS EPIDEMIOLÓGICAS

Durante o ano de 2020, muito se ouviu falar sobre as ondas da COVID-19, como intervalos de tempo que marcam uma elevação no número de casos diários, seguida por uma queda do mesmo. No entanto, o termo "onda" ainda não possui um conceito definido, no que diz respeito à área de estudo da epidemiologia. O conceito de onda advém da física, e passou a ser aplicado para contextualizar o comportamento de pandemias e epidemias. Do que se foi pesquisado durante a revisão bibliográfica do presente trabalho, o termo em questão não é definido por critérios científicos da epidemiologia, mas está relacionado principalmente a um aumento acelerado no número de casos. Segundo (21) o termo "onda", associado a surtos de doenças, começou a ser utilizado na epidemia da Influenza, de 1889, caracterizada por várias fases ao longo de 3 anos (31). No entanto, os autores propõem uma discussão sobre a definição do termo "onda" ao afirmarem:

> “Ondas” sugerem uma falta de circulação viral que provavelmente é uma ilusão. É possível que algumas das “ondas” ou fases secundárias tenham sido causadas ou favorecidas pela co-circulação de outros microrganismos. Ondas também são visíveis e na maior parte das vezes rítmicas. Não parece haver qualquer padrão ou ritmo para as epidemias resumidas na tabela e suas idas e vindas só são visíveis devido aos efeitos sobre o corpo humano e seu impacto na sociedade. (21)

Em (32) é possível verificar que, para algumas fontes, o crescimento do número de casos e mortes já é um indicativo de uma nova fase (o que significaria que o Brasil, entre Março de 2020 e Julho de 2021, poderia ter alcançado a quarta onda, por exemplo), ao passo que para outros, uma onda só pode ser considerada encerrada com uma baixa significativa dos índices, chegando a ter apenas casos esporádicos (definição que induz que o Brasil ainda estaria em sua primeira e única onda).

Em um esforço para garantir uma definição generalista, Stephen Zhang em (33), propôs oferecer uma definição funcional de ondas epidêmicas ao analisar a pandemia utilizando como base o fator R, onde R é o número de reprodução ou número médio de pessoas infectadas por uma pessoa infecciosa.

Já os autores de (34) definem as ondas epidemiológicas em função das questões sociais, políticas, e econômicas, vivenciadas por uma determinada população. Indo portanto de encontro com a orientação da Organização Mundial da Saúde (OMS) e de analise dos dados epidemiológicos em séries temporais.

Em contrapartida, diversos autores tem apostado na classificação de ondas epidemiológicas como sendo aqueles períodos em que são observados, após um determinado pico, a diminuição do número de casos diários em um percentual em relação ao valor de pico (H) passado e em função de um período de tempo prefixado.

Partindo desta premissa, observa-se que tal situação pode ser considerada se e somente se o período de tempo considerar os valores sobrepostos da onda do período. Assim sendo estaríamos considerando a diferença dos valores anotados na série temporal estudada, dando o correto sentido à avaliação das ondas.

Por outro lado, faz-se necessário a definição do período de tempo utilizado. Este deve ser definido de tal forma que permita a analise fidedigna dos dados estudados, entendendo que quanto maior o valor de tempo, menor o número de ondas possivelmente identificadas e vice–versa.

Segundo a WHO, para que uma onda de pandemia termine, "o vírus deve ser controlado e os casos devem cair substancialmente. Para que uma segunda onda comece, é necessário um aumento sustentado das infecções" (6).

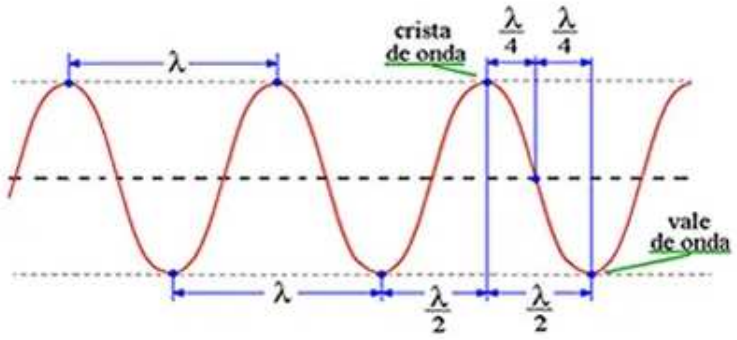
De qualquer maneira, tais definições tornam-se ambíguas ao adjetivarem a redução ou o aumento de novos casos (que definem uma onda) com termos como "significativas", "acentuadas", etc. Para fins de se estabelecer um critério de identificação de ondas, tais definições trazem uma subjetividade nociva à pesquisa, haja visto que o entendimento de queda significativa pode variar de pessoa pra pessoa, de região para região, etc. Ficando desta forma a comunidade científica livre para definir o quantitativo que traduzirá o termo "onda".

Dados os fatos, viu-se necessária uma definição que, ainda que de maneira direcio-

nada, pudesse nortear a presente pesquisa, com o objetivo único e exclusivo de eliminar o caráter ambíguo do que se entenderia por onda epidemiológica durante o projeto.

Baseando-se novamente no conceito físico de onda, entende-se como pico (ou crista) o ponto mais alto de uma onda, e como vale o ponto mais baixo da mesma (35). Desta forma, é possível afirmar que toda onda é marcada pela alternância entre picos e vales, como demonstrado na figura 8:

Figura 8 – Conceituação de "crista" e "vale" de uma onda sendo λ o cumprimento da onda, $\frac{\lambda}{2}$ os pontos críticos (picos e vales) e $\frac{\lambda}{4}$ os pontos de inflexão.



Fonte: SILVA, 2021 (35)

Tal conceito ganha uma maior complexidade quando aplicado a séries temporais de pandemias, haja visto que, o comportamento não determinístico das mesmas impede uma identificação previa do que seriam picos e vales. Desta forma, insere-se na pesquisa o conceito de picos e vales potenciais. Picos potenciais se definem como toda e qualquer medição cuja vizinhança (isto é, a dupla de medições anterior e posterior) apresentem valores menores. A definição de vales potenciais se dá de maneira análoga.

Para fins de definição, ficou-se compreendido que seria considerado como pico, todo pico potencial cujas n medições antecessoras e sucessoras (também chamado de *offset*) apresentem valor inferior. Restando portanto à conceituação de onda (no contexto do projeto), uma definição de qual valor seria considerado como *offset* ideal. Sobre este, o trabalho identificou que trata-se de um parâmetro sensível, haja visto que o seu valor pode determinar a identificação de mais, ou de menos picos em um gráfico, e consequentemente está diretamente relacionado à quantidade de ondas identificadas, que pode, apesar de ser melhor explicado nas sessões seguintes, pode ser objeto de uma segunda pesquisa no futuro.

2.2 PARÂMETROS QUE DEFINEM O CRESCIMENTO DE UMA EPIDEMIA

Estudos indicam que existem múltiplos fatores que podem influenciar na evolução da COVID-19 para um determinado país ou região. Em (2), fica evidenciada a capacidade de mutação dos vírus em geral, característica essa que protagoniza o surgimento de variantes de um mesmo vírus, que podem ser tão susceptíveis às medidas de proteção quanto seus antecessores, ou apresentarem resistência às mesmas, alterando diretamente a taxa de novos casos e óbitos decorrentes de uma determinada epidemia. Além disso, ainda em (2) fica evidenciada a influencia do clima e da sazonalidade no surgimento de novas variantes, haja visto que os autores propõem que climas com alta temperatura e umidade propiciam a proliferação de doenças.

Para além dos fatores climáticos e biológicos, fatores culturais também influenciam nos gráficos de uma doença para uma determinada região, como evidenciado em (36), um estudo que analisa como diferentes culturas enfrentam a pandemia. No trabalho, os autores deixam evidente que o forte reforço à coletividade das culturas orientais influenciou diretamente na velocidade com a qual países como China e Coreia do Sul trataram a doença provocada pelo novo coronavirus, quando comparados a países do ocidente, como Brasil e Estados Unidos.

No entanto, embora os artigos estudados apontem para uma grande diversidade de fatores que influenciam nos gráficos de uma pandemia, no que diz respeito à proposta da pesquisa neste documento tratada, estes se mostram irrelevantes. Tal irrelevância se deve ao fato de que o trabalho se focou nos casos de óbitos e novos casos diários, onde todas as variáveis supracitadas já exerceram a sua influência e estão traduzidas nos números analisados. Ainda nas predições, devido à metodologia adotada para a análise do potencial preditivo das funções de crescimento sigmoides possuir natureza retroativa, a ser explicada nas sessões seguintes, não foi necessário levar em consideração fatores como: clima, cultura, região, variantes, etc.

2.3 A FUNÇÃO DE RICHARDS

A função de Richards, originalmente desenvolvida em 1959 para modelagem de crescimento, é uma generalização do modelo de crescimento de Yon Bertalanffy (37). Na verdade, é uma variação da função sigmoide, que gera curvas em forma de S mais flexíveis. Sua

fórmula é definida por:

$$N(t) = \frac{k}{(1+\beta e^{-rt})^{\frac{1}{s}}} \tag{2.1}$$

Onde:

$$\beta = \frac{k}{N[0]^s} - 1 \tag{2.2}$$

Sendo:

$$N[0] = N(t=0) \tag{2.3}$$

A derivada da equação 2.1 com respeito ao tempo levará à equação diferencial não linear que governa os fenômenos descritos por esta classe de função (37):

$$\frac{dN}{dt} = \frac{rN}{s}[1-(\frac{N}{k})^s] \tag{2.4}$$

É possível portanto observar na equação 2.4 que o parâmetro s tem influência direta na taxa de crescimento da função, especialmente nas regiões de extremidade ou de adjascência, caracterizada pela presença das assíntotas, sendo esta variável influente no estabelecendo de simetria ou assimetria entre as mesmas. Portanto, entende-se existem dois casos particulares de grande importância: quando s tende a 1 e quando s tende a 0^+ (haja visto que $s \in R^+$), resultando, respectivamente, nas funções Logística e Gompertz.

2.4 A FUNÇÃO LOGÍSTICA

O modelo logístico foi originalmente formulado por Verhulst (1838) em resposta ao modelo de potencial crescimento populacional proposto por Thomas Robert Malthus (1798) (37), que defendia a ideia de que a população crescia de forma geométrica, enquanto a produção de recursos crescia de forma aritmética. Segundo Malthus, no seu livro intitulado "Ensaio sobre

o Princípio da População", esta relação, caso não observada, faria com que a população global ao longo do tempo crescesse de forma agressiva, superando a oferta de alimentos, o que resultaria em problemas como a fome e a miséria. Ao contrário da teoria Malthusiana, em que o número de indivíduos pode crescer indefinidamente, o modelo logístico estabelece um limite, ou valor assintótico, para o crescimento populacional em ambientes com recursos vitais, ou seja, onde haja capacidade máxima de convivência dos indivíduos. É expresso pela equação diferencial não linear:

$$\frac{dN}{dt} = rN(1 - \frac{N}{k}) \tag{2.5}$$

Onde $N = N(t)$ representa o número de indivíduos no instante t, r sendo a taxa de crescimento populacional e k o valor assintótico de $N(t)$.

Sua forma analítica se traduz na seguinte fórmula:

$$N(t) = \frac{k}{1 + \beta e^{-rt}} \tag{2.6}$$

Sendo $N[0] = N(t = 0)$ e $\beta = \frac{k}{N[0]} - 1$

É correto portanto afirmar que a função sigmoide Logística representa um caso específico da função de Richards (2.1), sendo esta o caso em que a taxa de simetria s vale 1, ou em outras palavras, os casos em que a função de Richards resulta em uma curva em S simétrica. Desta forma, a função Logística desenha uma curva em S, com ponto de inflexão localizado exatamente na metade do seu cumprimento, separando a curva em duas metades simétricas, como é possivel notar nas imagens 4, 5 e 6.

2.5 A FUNÇÃO DE GOMPERTZ

A equação diferencial associada à função de Gompertz pode ser obtida a partir da equação 2.1 (Richards), fazendo s $\rightarrow 0^+$ e aplicando a regra L'Hospital (29), sendo portanto:

$$\frac{dN}{dt} = rNln(\frac{k}{N}) \tag{2.7}$$

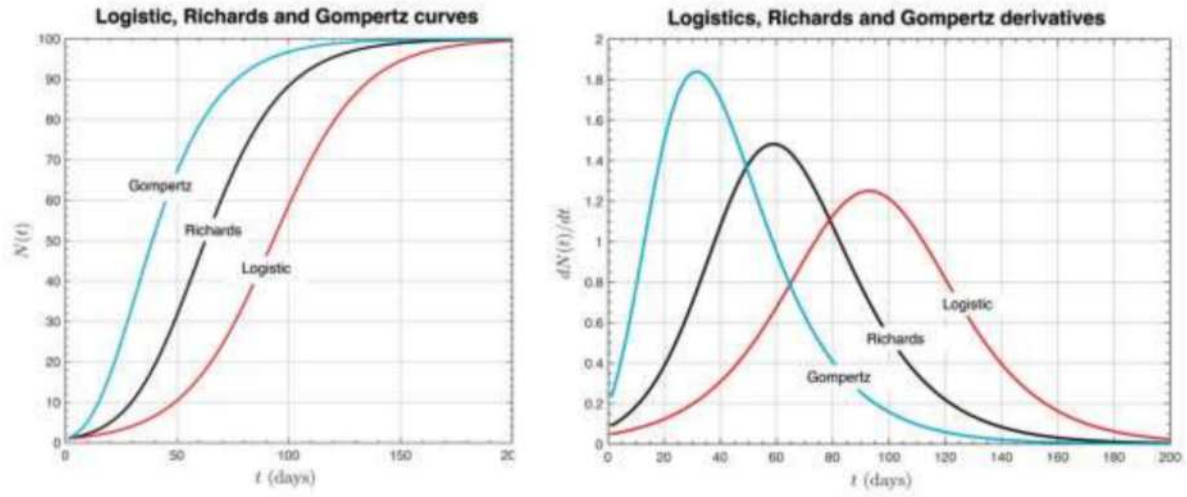
Respeitada a condição $N(t) = k$, a solução desta equação será a função de Gompertz, traduzida pela fórmula:

$$N(t) = ke^{-\beta e^{-rt}} \tag{2.8}$$

Sendo $\beta = ln\frac{k}{N[0]}$

Ao contrário da função Logística, a função de Gompertz apresenta uma assimetria, possuindo maior crescimento na assíntota inferior, ou seja, para os menores valores *t* do intervalo selecionado, conforme mostra a imagem 9. À esquerda, modelagem da curva de crescimento de casos acumulados de uma pandemia hipotética, utilizando as curvas de Richards, Logística e Gompertz, e assumindo como parâmetros: *N[0]=1, k=100, r=0.5, s=0.5 e t=200*, à direita, as suas respectivas derivadas, que, em termos matemáticos traduzem a modelagem dos casos diários, sendo portanto as funções as representações de uma onda epidêmica.

Figura 9 – Logistic, Richards and Gompertz curves and derivates



Fonte: FRIAS, 2020 (37)

2.6 SOBREPOSIÇÃO DE ONDAS

No que tange o universo dos artigos avaliados na presente pesquisa, nota-se que não há a ocorrência de algum trabalho que levem em consideração a possibilidade de duas ondas epidêmicas dividirem o mesmo espaço de tempo. A exemplos das pesquisas descritas

em (20), (37) e (18), as produções científicas trouxeram análises sob a ótica de uma única onda, para as quais, as curvas de crescimento apresentadas no presente documento possuem relevante capacidade descritiva. No entanto, em estágios mais avançados de uma pandemia, são registradas múltiplas ondas ao longo do tempo, cenário este não ocorre nos trabalhos previamente avaliados pelo aluno.

Do conteúdo analisado previamente, notou-se a prevalência de trabalhos que compreendiam ondas como fenômenos sequenciais, isto é, o surgimento de uma nova onda estava condicionado ao término da onda anterior. No entanto, baseando-se na ideia de que novas variantes do Coronavirus podem desencadear novas ondas epidêmicas (2), o surgimento de múltiplas variantes poderia ocasionar num quadro onde duas ou mais ondas possam coexistir em um mesmo intervalo de tempo.

Tendo em vista esta ideia, somado ao fato de que as funções sigmoides (Richards, Gompertz e Logistica) possuem poder descritivo sobre casos com uma única onda, foi necessária a adotação de estratégia para identificar ondas, e tratá-las de forma isolada. A estratégia adota consiste em 4 etapas:

- Normalização dos dados: Sabe-se que o fator humano pode influenciar negativamente na qualidade dos dados de uma série temporal. Por exemplo, para os dados da COVID-19 no Brasil, é possível notar dias com 0 casos registrados, com os registros anteriores e posteriores possuindo valores acima de 1.000 casos. Uma suposta explicação para esse fenômeno pode ser o esquecimento ou falha do sistema no dia em questão, e uma tentativa de compensação no dia seguinte.

- Identificação de picos: Conforme mencionado anteriormente, fontes como (21) e (32), definem uma onda epidemiológica, de forma geral, como um evento que registra um aumento no número de novos casos, seguido de uma queda acentuada do mesmo. Partindo desta linha de pensamento, é seguro afirmar que toda onda possui um pico, isto é, o maior valor registrado dentro do intervalo de tempo da onda. Portanto, um algorítimo capaz de identificar picos em séries temporais, neste contexto, poderá também ser usado para identificar a ocorrência de novas ondas.

- Identificação de limites: Após obtidos todos os picos de uma serie temporal, e portanto a

quantidade de ondas registradas, buscar-se-á identificar quais são os limites destas ondas, isto é, em quais pontos da série temporal as mesmas terminam. Desta forma, é possível construir um gráfico que contemple tanto casos sequencias, como eventuais casos de sobreposição de onda.

- Modelagem das ondas de forma isolada: Após identificados os inícios e términos de cada onda, é possível realizar a regressão não linear de cada uma, com base na função sigmoide de escolha (para a presente pesquisa, a função de Richards), e obter as ondas geradas pelo modelo.

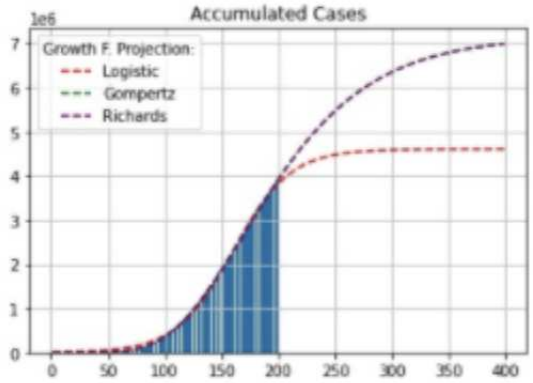
3 O PROJETO HERMES

O Projeto Hermes nasce da necessidade de fornecer à comunidade acadêmica uma ferramenta capaz de predizer o comportamento das curvas epidemiológicas da COVID-19 em diferentes cidades, estados, países e regiões do globo. O projeto em questão se destaca na literatura atual devido ao fato de incorporar uma metodologia inovadora que contempla a existência de múltiplas ondas, sobrepostas ou não, nos dados epidemiológicos.

Sua funcionalidade se baseia na utilização da técnica de regressão não linear da, já testada, função de Richards para, através da analise descritiva, disponibilizar a predição do cenário da pandemia a ser enfrentado. A metodologia escolhida para analisar a capacidade preditiva do modelo proposto pelo projeto não demanda diretamente a análise de fatores como clima, cultura, umidade, surgimentos de novas cepas, número total de vacinados, ou qualquer outro fator associado a disseminação de doenças. Isto se deve ao fato de que a análise será realizada com dados já existentes nos datasets, estando portanto todas as variáveis anteriormente citadas contempladas nos números analisados (estudo experimental). Ou seja, tanto os processos de análise descritiva ou preditiva foram realizados de forma retroativa.

Foram considerados os dados oriundos de *datasets* como (38), (13) e (7), separando em um primeiro momento, uma janela de t dias. Deste recorte, é realizada uma divisão entre dados para a processamento, e dados para comparação/confirmação. Os primeiros x dias (com $x < t \mid X >= t*0,3$) foram utilizados para adaptar as curvas de crescimento (Richards, Logistica e Gompertz), por meio de regressão não linear. Após esta etapa, é realizada a projeção dos t-x dias seguintes aos dias usados no processamento, e essas predições são comparadas com os dados reais, obtidos dos datasets. A figura 10, ilustra a predição o Brasil, em um processamento por meio das funções de Richards, Gompertz e Logística, dos casos acumulados de COVID-19 nos primeiros 200 dias de pandemia. É possível notar que para $t = 250$ (Novembro de 2020), as projeções das funções de Richards e de Gompertz apontam para o numero de casos acumulados de aproximadamente 5,6 milhões, ao passo que a função Logística aponta para um numero aproximado de 4,8 milhões. O que de fato ocorreu conforme figura 11.

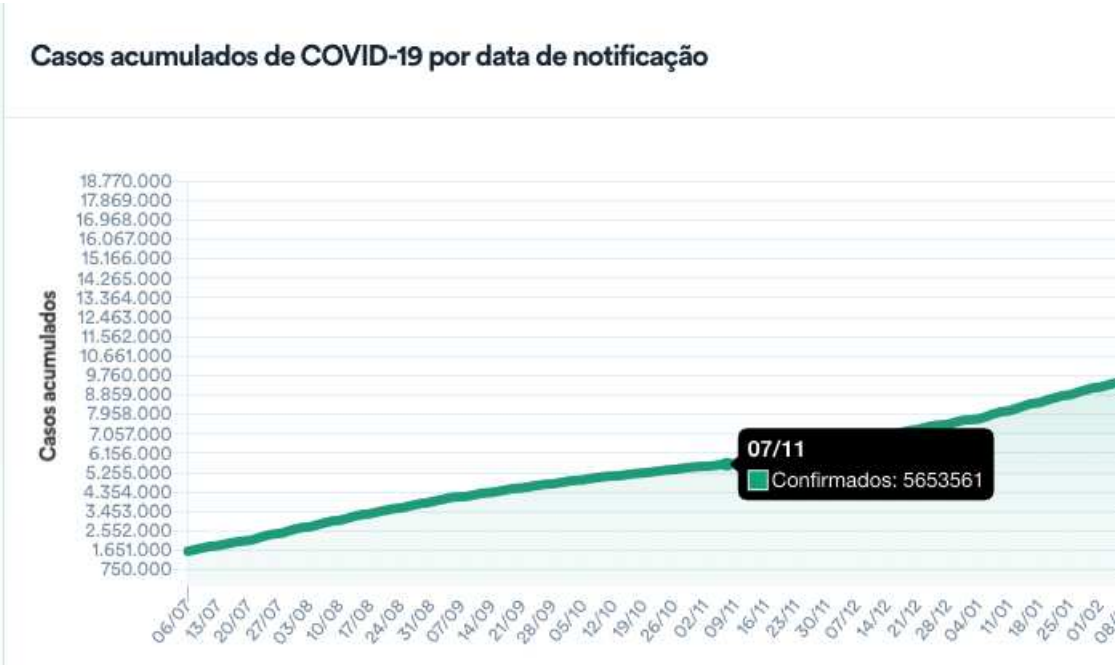
Figura 10 – Previsão de casos acumulados do Brasil e acordo com modelagem por funções sigmoides



.

Fonte: Autoral.

Figura 11 – Numero de casos acumulados no dia 07 de Novembro de 2020.



Fonte: METRICS, 2021. (16)

Desta forma é possível estimar que a projeção realizada pelo modelo para o intervalo de tempo selecionado foi acima de 90%. A proposta do trabalho é de se construir uma plataforma capaz de realizar estas projeções de forma automatizada e interativa, sendo portanto uma ferramenta que permite a análise da capacidade preditiva das curvas de crescimento referenciadas no trabalho, para variadas regiões do mundo, permitindo a extração de dados como: efetividade da campanha de vacinação, surgimento de novas ondas, estimativa da duração da pandemia, etc. Nesta perspectiva, diversas etapas foram necessárias para permitir a viabilidade do presente projeto, a serem descritas nos capítulos seguintes.

3.1 METODOLOGIA

A metodologia empregada para o desenvolvimento do trabalho está baseada na observação de dados experimentais, extraídos de base de dados confiáveis e disponibilizadas por entidades que se predispuseram a manter os dados atualizados e com acesso público. As bases de dados escolhidas foram:

1. Wesley Costa - Monitoramento do número de casos de COVID-19 no Brasil: Base de dados com números relacionados à COVID-19 no Brasil. Há dados de casos e óbitos por município, com informações oficiais do Ministério da Saúde, juntamente com os das Secretarias Estaduais de Saúde obtidos pelo Brasil.IO. Além disso, há dados de vacinados, bem como outros dados, como de recuperados e testes, correções e atualizações próprias (38).

2. Worldometers: Estatísticas do mundo em tempo real sobre população, governo, economia, sociedade, mídia, meio ambiente, alimentação, água, energia e saúde. Estatísticas interessantes como o relógio da população mundial, emissões de dióxido de carbono (CO2), fome, gastos públicos, valores de produção, dados de consumo (7), etc. O portal possui uma sessão específica para a pandemia da COVID-19, onde exibe, em tempo real, dados acerca de novos casos, mortes, recuperações e casos acumulados, com opções de filtro por país.

3. Johns Hopkins Coronavirus Resource Center (CRC): Base de dados da COVID-19 mundialmente consultada, atualizada constantemente. Além da diversidade de informações fornecidas gratuitamente, esta fonte foi considerada uma das 100 principais invenções de 2020, como citado no seguinte destaque:

> O Johns Hopkins Coronavirus Resource Center (CRC) é uma fonte continuamente atualizada de dados COVID-19 e orientação especializada. Coletamos e analisamos os melhores dados disponíveis sobre casos, mortes, testes, hospitalizações e vacinas para ajudar o público, legisladores e profissionais de saúde em todo o mundo a responder à pandemia. A TIME reconheceu o CRC como a "fonte de dados go-to" para COVID-19 e o nomeou entre as 100 principais invenções de 2020. Em 2021, a Research! America nomeou o CRC como recebedor do prêmio "Meeting the Moment for Public Health" (13).

De posse das informações disponibilizadas, os dados devem ser tratados de forma

isolada, por país, sendo então alimentados em um script Matlab® para realização de regressão linear através da função de Richards. Devem ser observados os parâmetros de correlação de Pearson (R) assim como o coeficiente de determinação R^2. Para verificação e validação dos resultados serão considerados os valores de RME e SSE para confirmar a qualidade do resultado do coeficiente de determinação.

Uma vez configurado o script para testes, deverá ser definido um valor mínimo aceitável para o coeficiente de determinação R^2 de 0,99 com RMS < 0.05 e um SSE inferior a 0.05. Estes parâmetros são os que são consagrados na literatura para atestar a eficiência do coeficiente de determinação.

Após verificação dos valores dos parâmetros estabelecidos deverão ser replicados os testes para *n* países que se enquadrem nas seguintes condições, entendidas como necessárias para viabilizar a analise descritiva e preditiva pretendida no estudo:

1. Tenham pelo menos uma onda epidemiológica completa.
2. Contemplem pelo menos um continente do globo.
3. Possuam população superior a 8 milhões de habitantes.
4. Possuam percentual de testagem (exames para diagnostico da Covid-19) superior a 60%

Estas condições irão garantir que os dados testados sejam representativos e permitam a execução da proposta considerando que:

1. Uma onda epidemiológica completa permite a analise descritiva e preditiva através da supressão de parte dos dados e comprovação posterior.
2. Regiões geográficas distintas permitem análises sobre diferentes regiões climáticas.
3. Um número mínimo de habitantes garante um volume de dados expressivo e por consequência significativo.
4. Países como elevada taxa de testagem permitem uma melhor fidelidade quanto aos registros de casos diários e por consequência de casos acumulados.

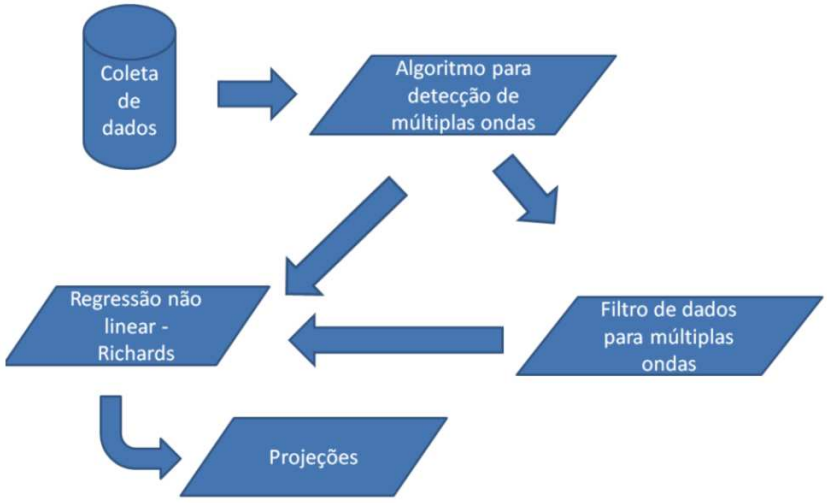
Uma vez comprovadas as capacidades descritivas e preditivas, pretende-se construir uma ferramenta capaz de isolar as ondas epidemiológicas de um objeto de estudo, permitindo assim a aplicação da metodologia de regressão não linear com a função de Richards em cada onda de forma a criar uma composição de regressões, capazes de descrever a evolução da pandemia de forma a englobar múltiplas ondas epidemiológicas. Neste caso a metodologia passa a vigorar

da seguinte forma: para cada conjunto de onda epidemiológica que não estiver sobreposta a outra, aplica-se a regressão não linear para modelagem descritiva. Para ondas que estejam sobrepostas, aplica-se o mesmo método porém desconsiderando os valores projetados (preditos) pela regressão não linear da onda predecessora. Assim sendo espera-se que este método seja capaz de modelar o comportamento não simétrico e não sigmoide da pandemia em seu curso direto.

3.2 ARQUITETURA DO SISTEMA

A arquitetura do sistema proposto pode ser simplificado e observado através da figura abaixo:

Figura 12 – Arquitetura do sistema de predição dinâmica.



Fonte: Autoral.

Inicialmente o sistema contará com um módulo para coleta dos dados diretamente das fontes mantenedoras das informações. Este módulo é responsável por capturar as informações e entrega-las para um algoritmo, que irá verificar se existem múltiplas ondas e se as mesmas estão sobrepostas. Caso não sejam identificadas múltiplas ondas o processo segue para o módulo de regressão não linear que tem a responsabilidade de executar a função de regressão de Richards sob os dados, identificando os parâmetros que melhor descrevem a curva estudada. Caso sejam identificadas múltiplas ondas, o sistema invocará um módulo específico que tratará de filtrar os dados que estão sobrepostos, reescrevendo-os e encaminhando para o módulo de regressão não linear. Finalmente o processo é finalizado no modulo de projeções que tem por finalidade

executar a projeção do número total de casos diários e acumulados do objeto de estudo.

Todo este processo dentro desta arquitetura gera um único arquivo em formato *JSON (Javascript Object Notation)* que será lido e apresentado em interface web para que sejam apresentados de forma gráfica aos interessados.

3.3 VALIDAÇÃO PARCIAL DO MODELO

Para validar a proposta, foram levados em consideração métricas comuns à analise de modelos matemáticos em geral, como R Quadrado (R^2) e o RMSE. No entanto, em um estudo prévio optou-se por adotar apenas o R^2 como métrica para analisar o poder descritivo das funções sigmoides analisadas. O entendimento de que a RMSE não se aplica ao modelo está fundamentado no fato de que a mesma é demasiadamente impactada por pontos fora da curva, bem como pela própria normalização de dados, variáveis estas que estão presentes em boa parte das abordagens para estudo de séries temporais. A análise baseada no RMSE demonstrou um elevado número falsos-positivos para boa parte dos casos. Desta forma, o processo de análise foi dividido em duas etapas:

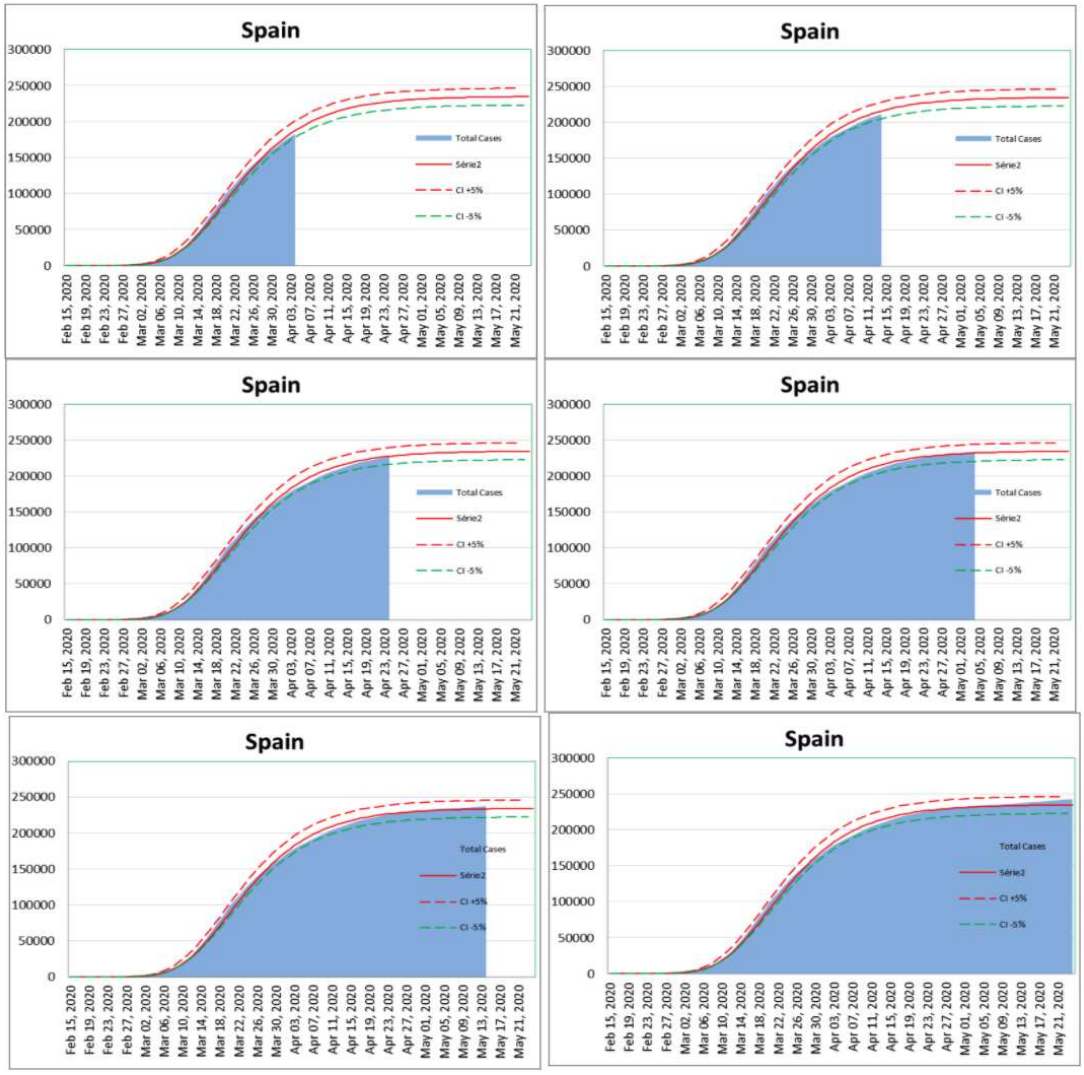
- Validação descritiva: Validação por capacidade descritiva do modelo proposto para um determinado país. Nesta etapa, é obtido o coeficiente R^2 comparando os dados reais da pandemia com os dados provenientes da modelagem, onde os valores de R^2 devem atender aos critérios determinados anteriormente (valores acima de 0.99).

- Validação Preditiva: Neste último, reforçamos a importância de utilizar países onde já se tenham a confirmação de pelo menos uma onda epidemiológica completa pois, a validação deverá utilizar uma parte dos dados da onda, por exemplo 60% dos dados para realizar a projeção de certo período à frente, que poderão ser contrastados com os 40% não utilizado e de fato ocorrido no objeto de estudo.

Estudos prévios utilizando ferramentas como Scilab, MatLab, permitiram a analise tanto da arquitetura, tanto da metologia e foi possível verificar que o projeto apresenta viabilidade com resultados preliminares que permitem a analise positiva da presente pesquisa. No estudo em questão, foram destacados os casos acumulados nos primeiros 50 dias da pandemia na Espanha (de 15/02/2020 a 03/04/2020). Sobre estes dados, foi realizada uma modelagem por meio de regressão não linear, utilizando a função de Richards. Com os melhores parâmetros obtidos,

foram encontrado os valores para os 50 dias analisados com base no modelo, bem como os valores dos próximos 48 dias (até o dia 21/05/2020), totalizando uma modelagem dos primeiros 98 dias de pandemia. Após isso, os 98 dias obtidos pela modelagem foram comparados aos valores reais dos 98 dias da COVID-19 na Espanha, o que nos permitiu elaborar a imagem 13, onde a linha vermelha contínua representa a projeção do modelo, e as linhas tracejadas os limites de erro (5% a mais ou a menos), e os dados reais são representados em azul. As imagens, analisadas da esquerda para a direita, de cima para baixo, representam uma série temporal de casos de COVID-19 (acumulados) na Espanha.

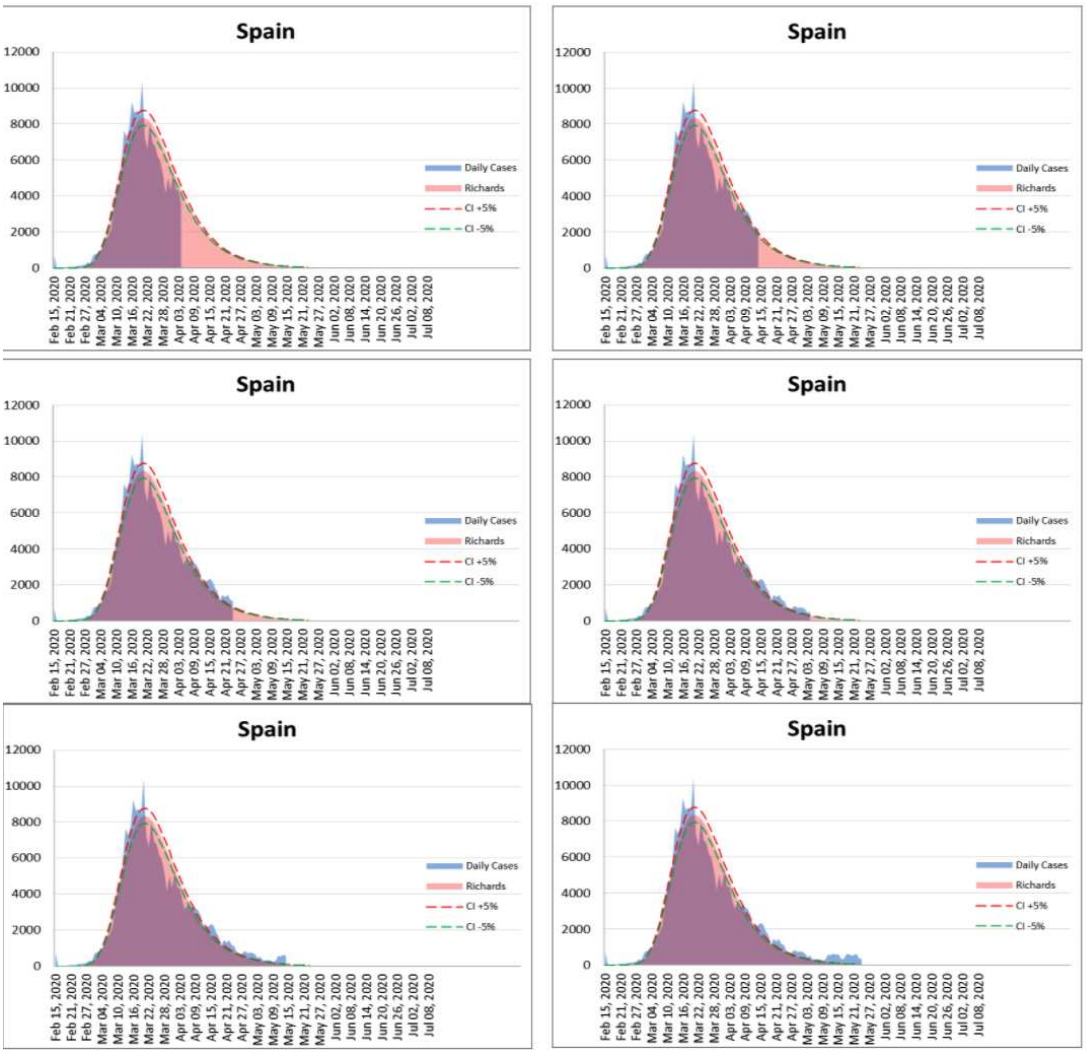
Figura 13 – Predição realizada para Espanha (Casos acumulados) em 03/04/2020. Da esquerda para direita, de cima para baixo, as imagens mostram a evolução dos casos reais (barras em azul) nas datas: 03/04, 15/04, 24/04, 05/05, 13/05 e 23/05 do ano de 2020.



Fonte: Autoral.

Como consequência, foi possível obter a representação da projeção no formato de onda, sobre os casos diários, por meio da derivação dos casos acumulados, como demonstra a figura 14:

Figura 14 – Predição realizada para Espanha (Casos diários) em 03/04/2020. Da esquerda para direita, de cima para baixo, as imagens mostram a evolução dos casos reais (barras em azul) nas datas: 03/04, 15/04, 24/04, 05/05, 13/05 e 23/05 do ano de 2020.



Fonte: Autoral.

Para este caso específico, foi possível concluir que o comportamento da pandemia nos 48 dias subsequentes ao recorte inicial respeitou a projeção realizada pelo modelo, com margem de erro inferior a 5%.

Além da Espanha, foram analisados países como França, Suécia, Itália, Grécia, Finlândia, Dinamarca, Nova Zelândia, China, Alemanha e Reino Unido, utilizando a mesma

metodologia. Os resultados apontaram para uma capacidade preditiva de até 20 dias em média (37).

4 RESULTADOS EXPERIMENTAIS

A execução dos experimentos práticos, com a finalidade de proporcionar a análise de resultados produzidos no âmbito do projeto Hermes demandou a observação de questões importantes como:

- **Qualidade dos dados:** a qualidade dos dados epidemiológicos utilizados para análise, entendidos como a série temporal dos valores de novos casos diários ou de novos casos de óbitos diários,bem como as séries destes dados de forma cumulativa ao longo do tempo.

- **Adequação do método:** a adequação do método de identificação de ondas epidemiológicas permitindo o conceito de sobreposição de ondas.

- **Capacidade preditiva:** a capacidade preditiva do modelo de regressão não linear baseado da função decrescimento de Richards.

Neste sentido as considerações necessárias para viabilizar o estudo experimental são também descritas nesta seção conforme seguem:

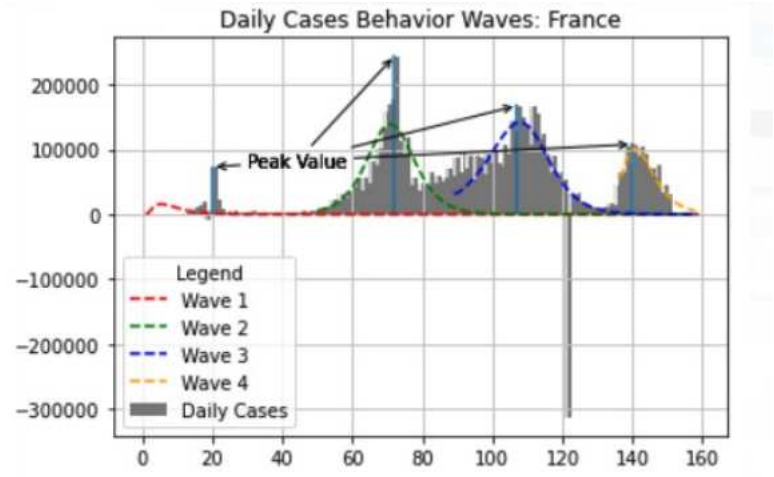
4.1 NORMALIZAÇÃO DE DADOS

Considerando que os dados epidemiológicos são fornecidos diariamente através da observação, por seres humanos, dos eventos que lhes são comunicados, a possibilidade de equívocos nos dados é quase inevitável. Nesta perspectiva, as bases de dados disponíveis possuem ruídos que podem, se não tratados, interferir nos resultados da análise. Dentre os ruídos mais comuns destacam-se:

- **Dados acumulados invertidos:** Em alguns casos, era possível notar evidências de erro humano no registro de casos acumulados pela relação entre os valores na série temporal, como por exemplo os registros da França, que em determinado intervalo da série temporal apresentava valores para o dia seguinte menores do que o do dia anterior. Esta é uma evidência clara de erro, haja visto que, para casos acumulados, valores subsequentes serão sempre maiores ou iguais aos valores anteriores, isto é, não há como no dia 10

registrarem-se *x* casos acumulados, e no dia 11 obter-se *x-10*, haja visto que a contagem do dia 11 corresponde à contagem de casos do dia 10, acrescido do registro dos casos diários no dia 11. Desta forma, o fato do dia 11 assumir um valor inferior ao dia 10 implica dizer que foi registrado um número negativo de casos diários, o que por definição não se encaixa no contexto da pandemia. Esta inconsistência se apresentou graficamente, na forma de valores negativos no gráfico de casos diários, como mostra a figura 15:

Figura 15 – França, Agosto 2021: Casos diários, não normalizados



Fonte: Autoral.

- **Sub-notificação e Super-notificação:** Outra inconsistência, de mais difícil detecção, é a presença de dias em que são registrados *x* casos quando em verdade ocorreram *y* casos no dia em questão, com $y > x$ ou $y < x$. Uma possível interpretação para o referido fato é de falha no processamento dos dados no dia em questão, gerando por consequência, o registro de *z* casos, deixando-se de registrar ou registrando-se a mais $|z\text{-}x|$ casos na data, de tal maneira que se $z\text{-}x < 0$ seria uma sub-notificação, e se $z\text{-}x > 0$, seria uma super-notificação. Em um rito natural, a existência de sub-notificações, em geral, culminam numa super-notificação no dia seguinte ou subsequentes, produzindo um risco para analistas no momento do estudo destes dados.

Optou-se como tratativa para o primeiro caso, a ordenação crescente dos registros de casos acumulados, uma vez que, dado o dataset, esta abordagem é a que mais se aproxima do cenário real, pois elimina os casos em que o dia seguinte apresenta valores menores que o dia anterior, eliminando portanto o registro de casos diários negativos.

Para os casos de sub-notificação e super-notificação, foi adotada como estratégia a

análise dos valores em intervalos de tempos maiores que um dia. A medida se baseia no fato de que a analise de intervalos de dias ao invés de dias faz com que os eventos de sub-notificação e super-notificação estejam abstraídos dentro do intervalo. Ao número de dias considerados em cada intervalo deu-se o nome de *Chunk Size*. Desta forma, entende-se como o gráfico de casos semanais, o registro dos casos diários agrupados de 7 em 7, isto é, com *Chunk Size = 7*.

Após adotadas as medidas de normalização, foi possível eliminar os ruídos que inviabilizavam uma modelagem assertiva.

4.2 IDENTIFICAÇÃO DE ONDAS

Após a normalização dos dados, foi necessária construção de um algorítimo para identificação dos picos, uma vez que este está diretamente relacionado à quantidade de ondas numa série temporal epidêmica. A solução consiste em encontrar dias cujo número de casos diários registrado permaneça o maior por um número pré-determinado de dias seguidos. Este número de dias em questão foi denominado *Wave Offset*. Portanto, para que um dia seja considerado o pico de uma onda pandêmica, este deve permanecer possuindo o maior registro quando comparado aos valores dos *Wave Offset* dias seguintes. Desta forma, por definição, a relação entre o número de picos encontrados e o *Wave Offset* é inversamente proporcional.

Além dos parâmetros *Chunk Size* para normalização de dados, e *Wave Offset* para identificação de picos, foi necessária também a definição de uma estratégia para identificar quando uma onda termina e a outra começa, do contrário, a solução para os picos estaria levando em consideração que uma onda termina exatamente *Wave Offset* dias após o seu pico, o que não é uma verdade. O entendimento adotado na pesquisa para o término de uma onda é o retorno do aumento dos casos diários (o que, no âmbito de trabalhos sobre séries temporais em geral, recebe o nome de "reversão"), uma vez que após um pico, os valores seguintes da onda tendem a serem inferiores. Tendo em vista os fatos, viu-se a necessidade da adição de uma solução para identificar a reversão do gráfico. A solução adotada foi a análise da Média Móvel Aritmética (MMA) (39), que consiste em atribuir a cada ponto um valor, correspondente à média aritmética dos *n* valores anteriores. Este processo desenha no gráfico uma curva suave e menos sensível a pontos fora da curva. Desta forma, o ponto de reversão é marcado pela inversão da curva formada pela média móvel, isto é, o início de um movimento de subida quando precedido de um histórico de descida, e vice-versa. Na figura 16, podemos analisar a aplicação da média móvel

para a identificação de reversão em uma série temporal, onde a linha contínua azul representa a MMA, enquanto as velas (candlesticks) vermelhas e verdes representam as medições da série temporal, e as setas em preto traduzem os pontos em que é confirmada uma reversão.

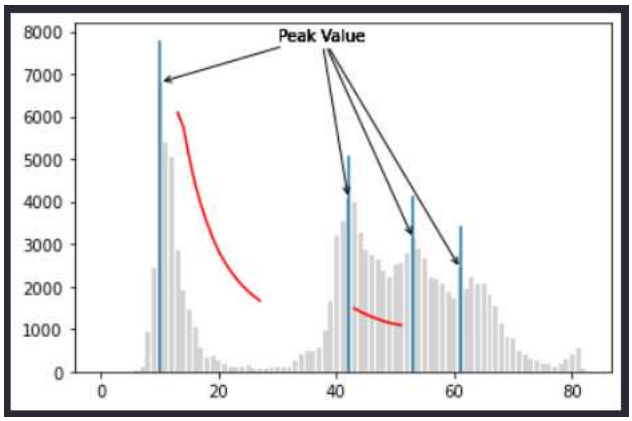
Figura 16 – Uso de média móvel para identificação de pontos de reversão em uma série temporal

Fonte: JIANG, 2020. (40)

Desta forma, inseriu-se à pesquisa, além dos parâmetro *Chunk Size* e *Wave Offset*, um terceiro parâmetro, nomeado *Moving Average Index*, que define qual será o tamanho do intervalo analisado para a obtenção do valor da MMA para um ponto, isto é, dado um ponto numa séria temporal, quantos pontos anteriores serão levados em consideração no cálculo da média aritmética.

Com base nestes três parâmetros, foi possível construir um algorítmo capaz de, da uma série temporal, identificar todos os picos na mesma, e, quando aplicado a séries temporais epidêmicas, obter também o início e um fim de uma onda, como demonstrado na figura 17, onde as linhas verticais azuis representam os picos, as linhas verticais cinzas representam os valores para cada dia (ou *chunk*) da série temporal da França, e a linha em vermelho representa o desenho da média móvel até o momento onde é identificada a reversão.

Figura 17 – França, Agosto 2021: Identificação de picos (*Chunk Size=7, Wave Offset = 6, Moving Average Index = 15*)



Fonte: Autoral.

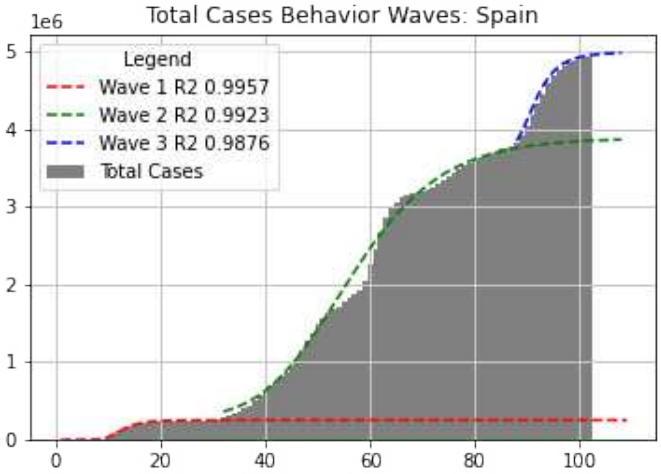
4.3 MÉTODO PARA DETECÇÃO DE ONDAS EPIDEMIOLÓGICAS

A solução para identificação de picos de uma série temporal permite obter, além de outros dados, os pontos na série temporal que marcam o início e o término de cada onda. Em posse destes dados, é possível dividir os registros de novos casos por onda. Neste ponto, é considerada a possibilidade de haverem múltiplas ondas coexistindo em um intervalo de tempo, de modo que é necessário realizar o isolamento dos dados pertencentes a cada onda.

Para se obter os valores de casos acumulados correspondentes única e exclusivamente a uma determinada onda, é preciso garantir que os dados a serem analisados de cada onda não se sobreponham. A técnica proposta, analisa os dados de cada onda epidemiológica de forma independente de tal maneira que, ondas que se sobrepõem tem seus valores de sobreposição, subtraídos dos valores projetados de ondas anteriores, evitando assim a contabilização equivocada de dados já considerados no modelo. Isto garante que os dados processados para a onda em questão não levem em consideração os registros da onda anterior, tornando portanto o cenário propício para a aplicação da modelagem por meio de regressão não-linear baseada na função de Richards sobre o intervalo, por tratar dados isolados sem influência de ondas anteriores, ou seja, em condições similares às encontradas em (20) e (37). Na figura 18, é possível observar a modelagem completa do registro de casos acumulados da COVID-19 na Espanha.

Além da modelagem completa dos casos acumulados, realizou-se também uma adaptação do gráfico, agrupando todas as ondas identificadas e modeladas pela aplicação em

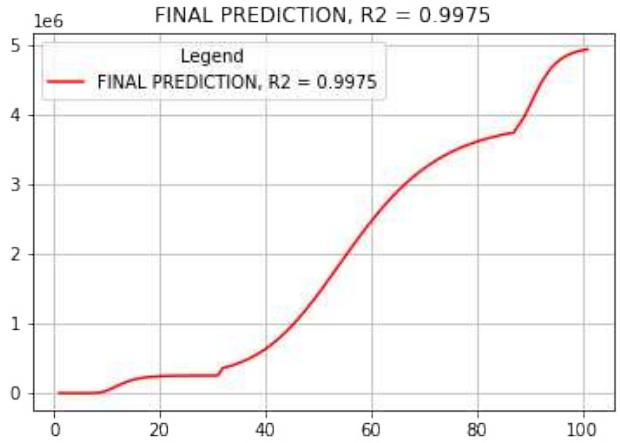
Figura 18 – Espanha, Setembro 2021: Modelagem de múltiplas ondas para caso acumulado, Richards. (*Chunk Size=6, Wave Offset = 15, Moving Average Index = 20*)



Fonte: Autoral.

uma única onda, nomeada de onda integrada, que define, de forma geral e com alto grau de assertividade (para o caso da Espanha, com um R^2=0.9975), o formato de toda a série temporal formada pela pandemia no país conforme mostrado na figura 19:

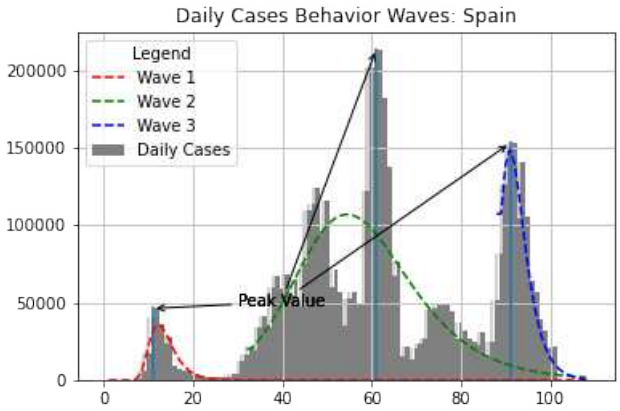
Figura 19 – Espanha, Setembro 2021: Modelagem de onda integrada para caso acumulado, Richards. (*Chunk Size=6, Wave Offset = 15, Moving Average Index = 20*)



Fonte: Autoral.

Assim como demonstrado na fase de verificação parcial do modelo (prova de conceito), foi possível derivar a modelagem dos casos acumulados, obtendo os casos diários e suas respectivas ondas sobrepostas.

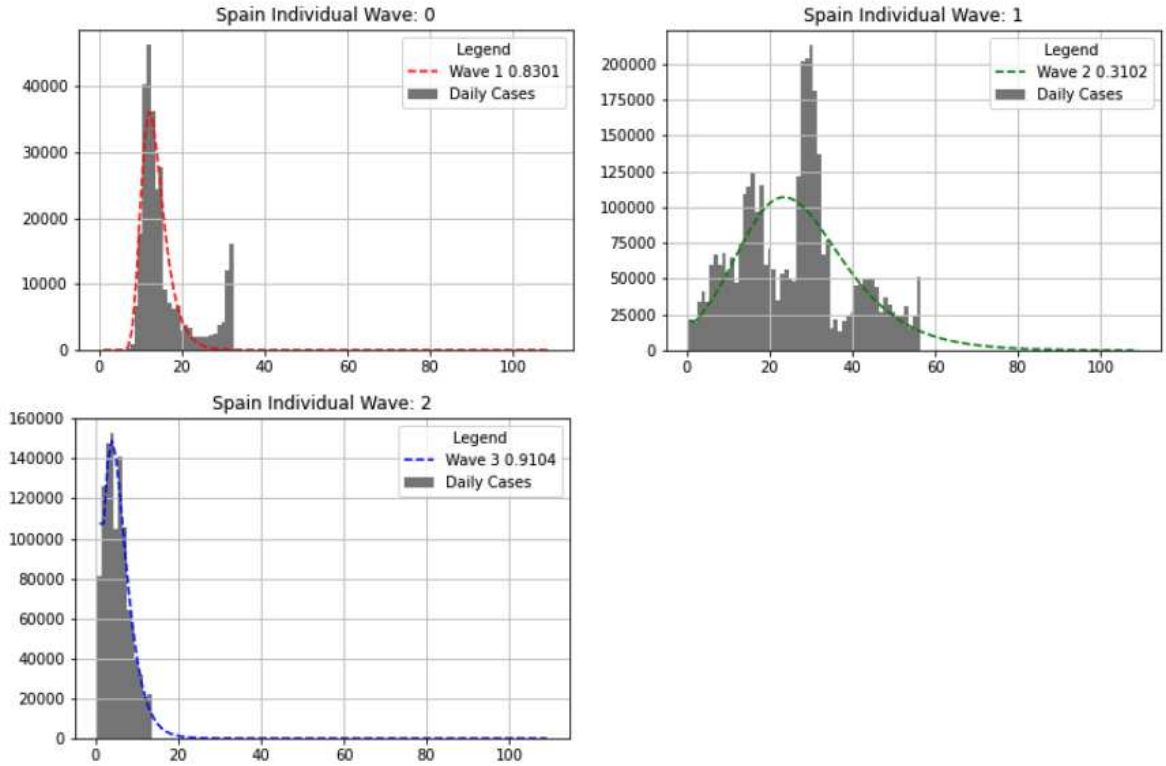
Figura 20 – Espanha, Setembro 2021: Modelagem de múltiplas ondas para caso diário, Richards. (*Chunk Size=6, Wave Offset = 15, Moving Average Index = 20*)



Fonte: Autoral.

Uma vez em posse dos pontos de início e fim de cada onda, foi possível também realizar a análise do R^2 de cada onda de forma isolada.

Figura 21 – Espanha, Setembro 2021: Modelagem de múltiplas ondas para caso diário, Richards. (*Chunk Size=6, Wave Offset = 15, Moving Average Index = 20*)



Fonte: Autoral.

4.4 SINTONIZAÇÃO AUTOMÁTICA DOS PARÂMETROS DO MODELO

Todo o processo de modelagem e predição é realiado com base em 3 variáveis: *Chunk Size*, *Wave Offset* e *Moving Average Index*. Desta forma, a precisão da modelagem está condicionada à escolha ideal destes 3 valores. Portanto, fez-se necessário o desenvolvimento de uma abordagem que permitisse, para cada país, obter-se os melhores valores dos parâmetros supracitados.

Entende-se como melhores valores aqueles que resultando numa modelagem com menor taxa de erro, isto é, um valor de R^2 próximo a 1. Desta forma, optou-se por se realizar um processamento iterativo sobre os dados de cada um dos 193 países, testando todas as possíveis combinações entre os parâmetros em questão, dentro dos seguintes limites:

- **Chunk Size**: Variando de 3 a 7
- **Wave Offset**: Variando de 4 a 15
- **Moving Average Index**: Variando de 5 a 20

Para cada combinação de parâmetros foi realizada a modelagem e obtido o valor do R^2 da onda integrada. Os dados foram organizados em arquivos no formato *csv* (campos separados por vírgula), onde pra cada país se gerou um arquivo com os resultados de cada uma das combinações testadas.

O teste de todas as possíveis combinações para *Chunk Size*, *Wave Offset* e *Moving Average Index* (dentro dos intervalos previamente definidos) se mostrou um processo com duração média de 35 minutos por país. Desta forma, a previsão para o término do processamento dos 193 país seria de 4 dias e 14 horas. No entanto, esta estimativa foi otimizada com o uso de processamento paralelo. Ao invés de definir um processo que iteraria sobre os 193 países, foram definidos múltiplos processos que iteraram sobre sub-listas (também compreendidas como páginas), contendo 15 países cada, da lista de países. Estes processos aconteceram de forma paralela, de modo que a cada ciclo 12 países eram processados simultaneamente. Com o uso desta estratégia, o processamento de todos os países foi finalizado em aproximadamente 12 horas, representando uma otimização de aproximadamente 90% do tempo previsto.

Posteriormente, os arquivos gerados foram processados para se extrair os melhores parâmetros para cada país, conforme os seguintes critérios de desempate caso dois registros apresentassem o mesmo valor para o R^2:

- **Menor Chunk Size**: Entendeu-se que, caso duas ou mais combinações resultassem no mesmo valor de R^2, deveria-se optar por aquela que apresentasse menor *Chunk Size*, partindo da premissa de que valores baixos de *Chunk Size* aumentam a sensibiliade do modelo à detecção de novas ondas, o que é visto como algo muito positivo.

- **Maior Wave Offset**: Caso duas combinações já possuam o menor valor definido para *Chunk Size* (*Chunk Size= 3*), deveria-se optar pela combinação que apresentasse maior valor de *Wave Offset*, baseado na ideia de que um intervalo maior para confirmação de um pico resulta na obtenção de menos picos e portanto ondas com mais dados a serem processados, o que aumenta faz com que o modelo tenda a ser mais assertivo, e evita o surgimento de novas ondas com quantidade de dados insuficientes para se realizar a modelagem.

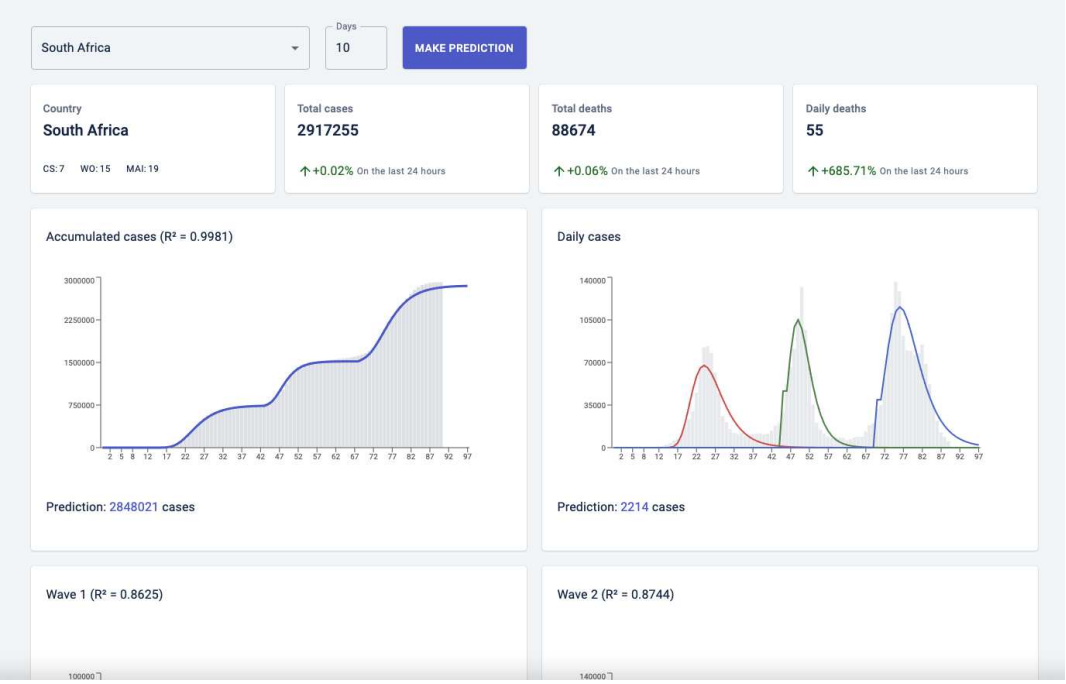
- **Maior Moving Average Index**: Por fim, caso duas ou mais combinações possuam os mesmos valores de *Chunk Size* e *Wave Offset*, deve-se optar pela combinação que possua maior valor de *Moving Average Index*, partindo do fato de que este fará com que o modelo releve mais dados na identificação de uma reversão, e portanto, seja mais assertivo no que tange a definição dos limites das ondas.

Indo de acordo com os resultados obtidos na etapa de validação parcial, após filtragem, todos os parâmetros obtidos apresentados resultados com $R^2 > 0,99$.

4.5 A PLATAFORMA WEB

Após obtidos os melhores parâmetros por país, iniciou-se o processo de publicação e democratização dos resultados da pesquisa, na forma de uma aplicação acessível por meio dos browsers Google Chrome, Internet Explorer, Microsoft Edge e Mozilla Firefox, onde os usuários pudessem avaliar de forma gráfica a capacidade descritiva e preditiva do modelo desenvolvido na pesquisa.

Figura 22 – África do Sul, Outubro 2021: Predição do modelo Hermes para os casos de COVID-19 (*Chunk Size = 7, Wave Offset = 15, Moving Average Index = 19*)



Fonte: Autoral.

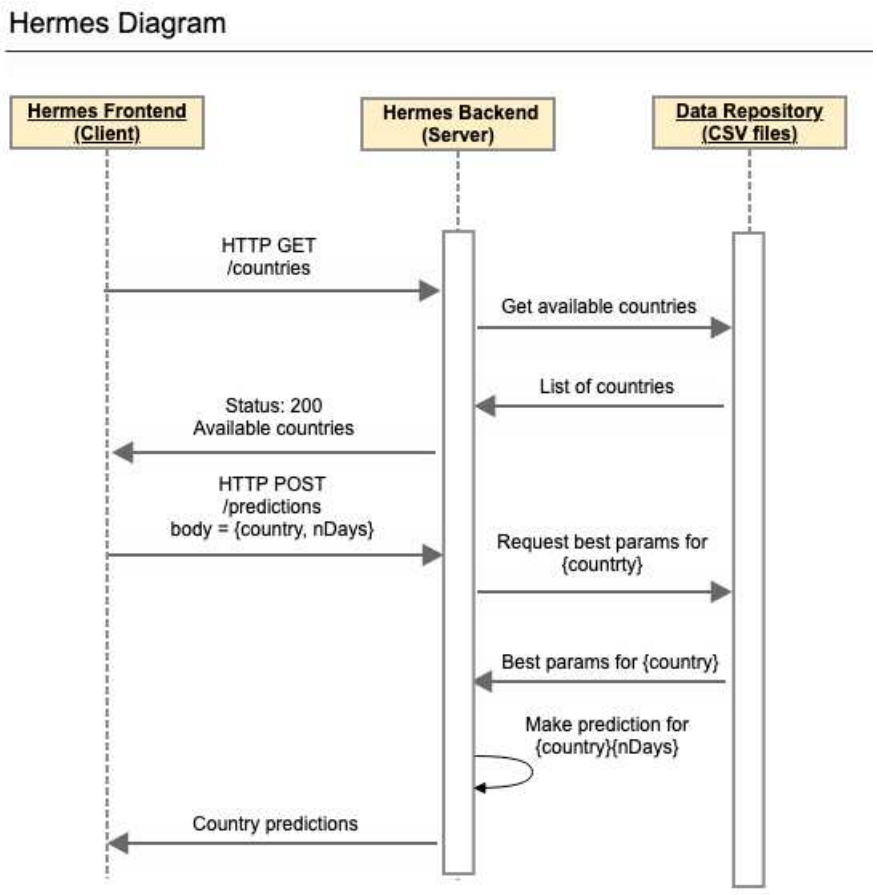
A plataforma, intitulada como UNEB Hermes, encontra-se na sua primeira versão, possuindo apenas os requisitos básicos para se obter as predições realizadas pelo modelo. A aplicação web, consiste em 2 campos e um botão. No primeiro campo, encontra-se um seletor, onde o usuário pode definir qual país (dos processados na pesquisa) será usado na predição, bem como um segundo campo onde o usuário pode definir quantos dias (ou chunks) deseja que o modelo preveja, conforme ilustrado na figura 23.

Figura 23 – Projeto Hermes: Formulário para predição

Fonte: Autoral.

Após definidos os dois parâmetros, ao clicar no botão de ação (que possui como rótulo a frase "MAKE PREDICTION"), o sistema então envia uma requisição HTTP POST, para uma máquina servidora, que realiza o processamento dos dados enviados no corpo da requisição (país e número de dias a serem preditos), busca quais são os melhores valores de *Chunk Size*, *Wave Offset* e *Moving Average Index* para o país selecionado, realiza a modelagem pro regressão não linear com base na função de Richards, e por fim envia os resultados do processamento de volta para a aplicação web, para que esta possa exibir os resultados da modelagem em forma de gráfico. O diagrama representado na figura 24 traduz de forma visual todo o fluxo supracitado:

Figura 24 – Projeto Hermes: Diagrama de Fluxo



Fonte: Autoral.

Após os resultados da predição serem retornados da máquina servidora para a máquina cliente (no caso, a máquina do usuário), são gerados os gráficos para os casos acumulados e diários, e a análise de cada onda de forma isolada, com seus respectivos valores para R^2.

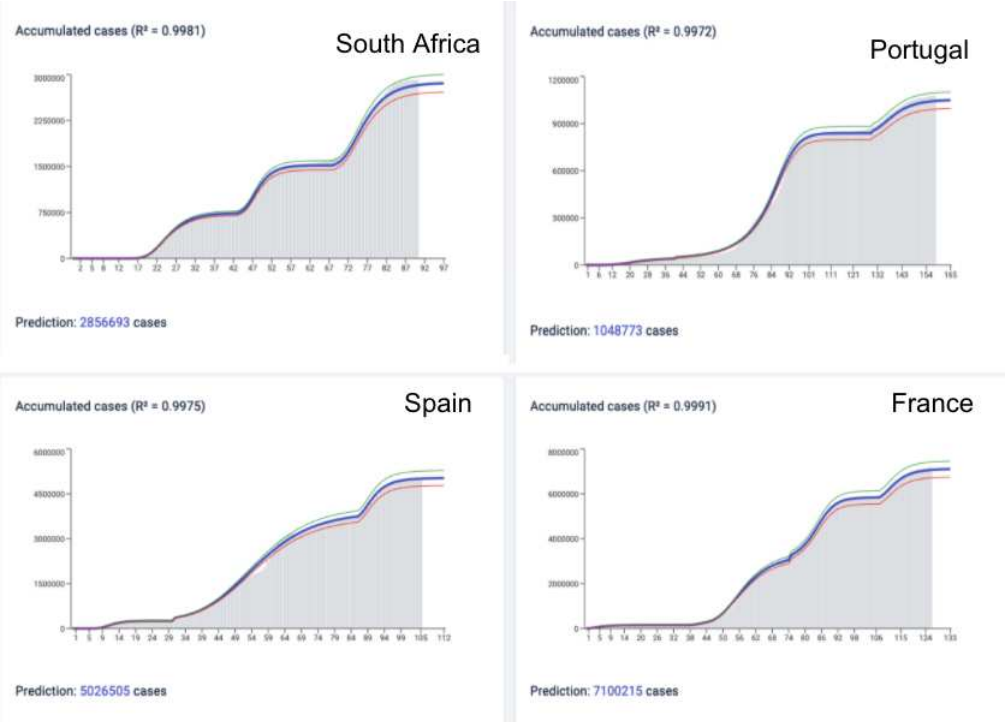
4.6 ANÁLISE DE CAPACIDADE DESCRITIVA DO MODELO

Os resultados experimentais demonstraram que, a utilização do método proposto, considerando as particularidades das múltiplas ondas epidemiológicas, alcançou a marca de *99,95%*, em média, de fator de determinação entre os países analisados (193). Este fato demonstra que a função de Richards pode ser usada tanto para modelar surtos epidêmicos de uma única onda (ou de ondas isoladas), como demonstrados em (37) e (20), bem como, agora comprovado, para casos de surtos com múltiplas ondas epidemiológicas, sobrepostas ou não.

Considerando um intervalo de confiança em *5%* para mais ou para menos, os resultados experimentais demonstraram que o modelo proposto foi capaz de descrever com *100%* de fidelidade, o desenvolvimento da doença em todos os países estudados. Assim sendo, comprovou-se que a proposta de analise descritiva através da utilização do método proposto, integrado com a função de Richards para regressão não linear é eficiente para casos da COVID-19.

A figura 25 ilustra os resultados para quatro países estudados. A linha azul representa a modelagem realizada, em verde e vermelho estão respectivamente os limites superiores e inferiores de confiança, e em cinza os valores das medições reais.

Figura 25 – Projeto Hermes: Ondas integradas para África do Sul, Portugal, Espanha e França

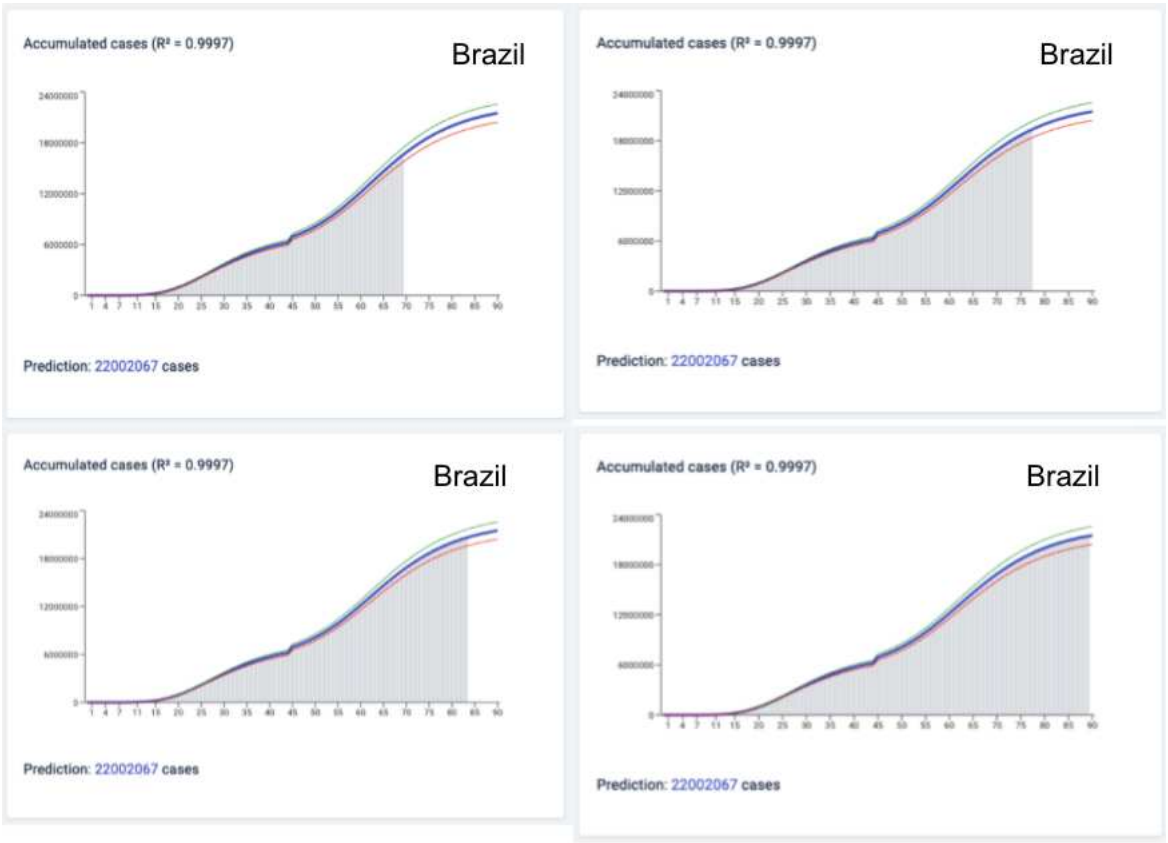


Fonte: Autoral.

4.7 ANÁLISE DE CAPACIDADE PREDITIVA DO MODELO

No que diz respeito à capacidade preditiva do modelo, os testes realizados confirmaram a hipótese levantada no processo de validação, tendo o modelo apresentado capacidade preditiva com alta assertividade (baseando-se nos mesmos critérios definidos para avaliar a capacidade descritiva), para as primeiras 20 chunks (que podem compreender 3 ou mais dias por chunck), como demonstram as simulações representadas na figura 26:

Figura 26 – Projeto Hermes: Teste de capacidade preditiva para 20 chunks (Brasil)



Fonte: Autoral.

De forma semelhante, foram realizados testes (obtidos resultados similares) sobre os dados de casos acumulados em diversos outros países, tendo sido o critério de diversidade definido pela localização geográfica, isto é, países que de preferência não compartilhassem o mesmo continente, de modo a avaliar se a assertividade do modelo poderia estar de alguma forma relacionada a características de um continente, hipótese esta que não se confirmou. Denota-se portanto que o modelo pode ser usado para modelar os dados da COVID-19 para qualquer região no mundo, bem como para qualquer surto epidêmico (como SARS, HIV, Ebola, etc). Entendeu-se que, o motivo para este resultado se deve ao fato de o modelo ser baseado em fundamentos estatísticos, estando portanto todas as particularidades de cada região ou doença abstraídas nos próprios valores registrados diariamente, conforme sugerido na prova de conceito do mesmo.

É válido ressaltar que, a assertividade das predições está condicionada à existência de uma quantidade mínima de registros, exigidas tanto para a realização da regressão não linear dos dados, quanto pela natureza da modelagem com base na função de Richards. Nos testes

realizados, foi identificado que as projeções para uma onda tendem a ser mais assertivas caso esta se encontre próximo ao seu ponto de inflexão, ou tenha o ultrapassado. Em outro termos, as projeções do modelo se mostraram mais confiáveis para ondas que já registraram o seu pico e se mostram em tendência de descida. Para casos de ondas em crescente subida, o modelo não apresentou resultados promissores ao tentar projetar em qual momento ocorreria a inflexão da curva.

5 TRABALHOS FUTUROS

A análise dos resultados obtidos na pesquisa relatada no presente documento, bem como a ferramenta gerada no processo, apresenta à comunidade científica uma extensa gama de pesquisas derivadas. Traduzem-se como exemplos de possíveis discussões os tópicos a seguir.

5.1 PERÍODO DE VALIDADE DOS MELHORES PARÂMETROS

Conforme mencionado em capítulos anteriores, os melhores parâmetros de cada país foram obtidos com base em um processamento iterativo dos tados, comparando todas as possíveis combinações para os valores de *Wave Offset*, *Chunk Size* e *Moving Average Index* dentro dos seus respectivos limites.

Dado que o processamento foi realizado em cima de informações que são atualizados diariamente, há de se pensar que, passado um determinado intervalo de tempo, os parâmetros gerados precisem ser atualizados mediante novo processamento dos países. Foi notado durante a fase de testes que, após 15 dias, a modelagem de alguns países apresentou divergências que se aproximam e tendem a ultrapassar a tolerância a falha, definida em 5% na etapa de validação da proposta. Desta forma, a necessidade de atualização dos parâmetros deixa um espaço aberto para que a comunidade científica, em posse dos resultados neste documento apresentados, realize pesquisas com o intuito de desenvolver estratégias para otimizar e automatizar o processo de definição da melhor combinação de valores para as variáveis supracitadas, bem como um estudo sobre a periodicidade ideal para realização do processamento em questão.

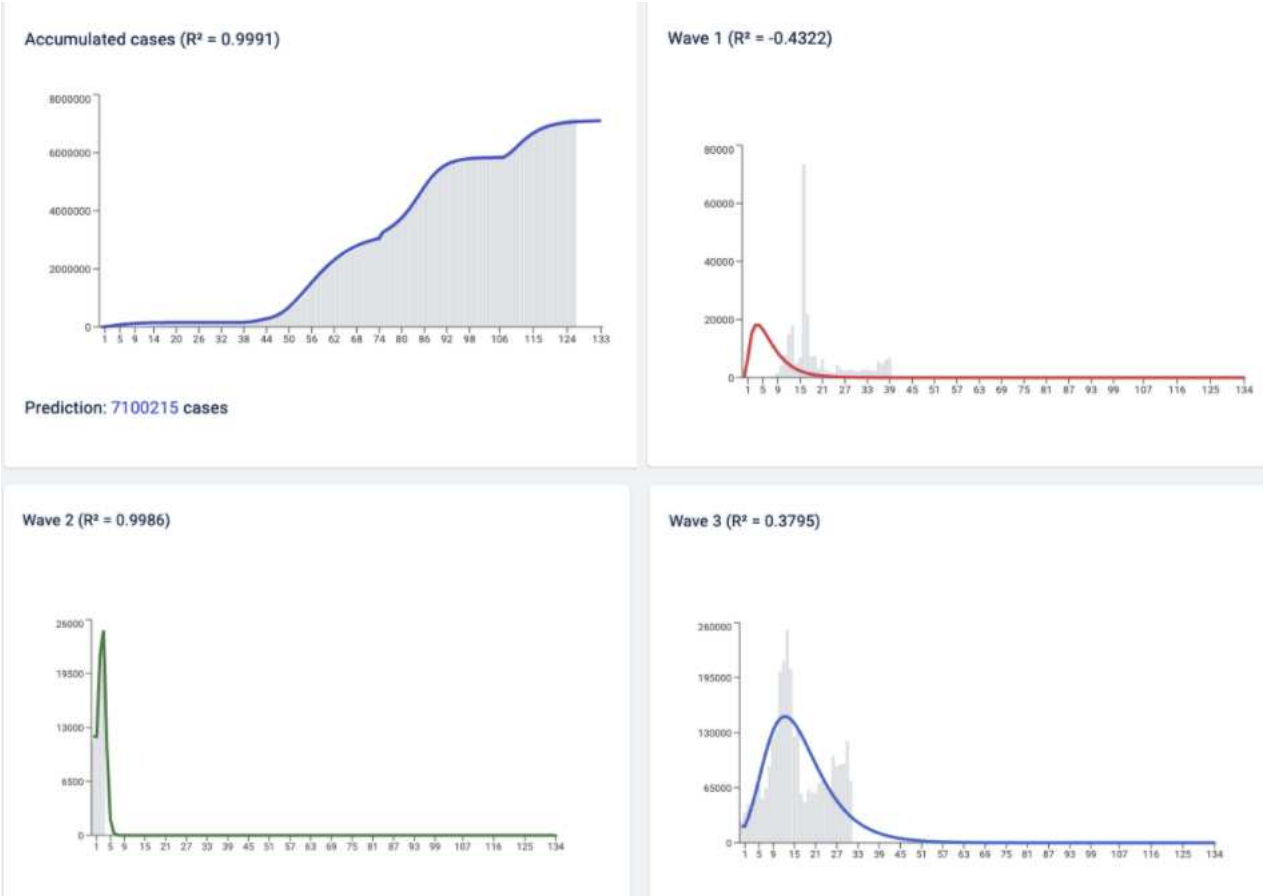
5.2 PROCESSAMENTO BASEADO NA MÉDIA DA ASSERTIVIDADE DAS ONDAS

O critério para seleção da melhor combinação de *Wave Offset*, *Chunk Size* e *Moving Average Index* está baseado na análise do valor de R^2, obtido pela comparação entre os valores gerados para a onda integrada, e os valores reais presentes na base de dados.

Esta abordagem, embora a um primeiro momento tenha demonstrado resultados promissores, deixa à comunidade científica o questionamento sobre esta ser de fato a metodologia

ideal. A dúvida tem como principal origem o fato de que, para uma parcela significativa da base de dados analisada, a onda integrada apresentou resultados dentro dos limites definidos na etapa de validação, no entanto a análise do R² de cada onda de forma isolada não refletiu a mesma realidade, como pode ser verificado na figura 27:

Figura 27 – França, Outubro 2021: Ondas integradas (casos acumulados) x Ondas isoladas (casos diários)



Fonte: Autoral.

Nota-se que há uma discrepância entre a assertividade apontada pela onda integrada, e a assertividade apontada por cada onda individual. Tal divergência deixa uma lacuna a ser preenchida, que se define na necessidade de um estudo sobre uma comparação entre obter os melhores parâmetros com base no R² da onda integrada, ou com base na média dos R²s das ondas individuais.

5.3 MELHORIAS INCREMENTAIS AO SISTEMA HERMES

Conforme dito, a pesquisa realizada no presente documento teve como foco o estudo da aplicação da função de Richards sobre os casos da COVID-19, sob uma perspectiva de múltiplas ondas, bem como a publicação dos resultados em forma de uma aplicação web, hospedada em domínio público. No entanto, no que diz respeito ao website, este se mostra na sua primeira versão, havendo portanto espaço para implementações de melhorias, tais quais:

- **Possibilidade de escolher a função modeladora**: Embora já comprovado em (20) e (37) que as funções Gompertz e Logística podem ser interpretadas como casos específicos da função de Richards, haveria valor científico em demonstrar tal fato de forma gráfica, havendo assim um novo campo no formulário da aplicação, no qual o usuário poderia escolher qual das três funções deveria ser usada na modelagem. Ou ainda, gerar um gráfico comparativo com a modelagem de cada uma das três funções para um determinado país.

- **Predição de mortes, ou de recuperados**: Na primeira versão da plataforma Hermes, foram desenvolvidas apenas as features que diziam respeito à predição de novos casos da doença no mundo. No entanto, seria de grande valor científico estender o uso da ferramenta, implementando a funcionalidade de realizar predições sobre número de mortos, bem como número de recuperados. Desta forma, a ferramenta poderia ser utilizada por entidades governamentais como mais uma forma de acompanhar o ritmo das campanhas de vacinação, ou mesmo cruzar as predições para novos casos e as predições para mortes de modo a se obter um parâmetro quanto a letalidade da variante que protagoniza a onda epidêmica em curso.

- **Análise de casos onda há falha no processamento**: Da base de países disponíveis, foi possível notar que as projeções para alguns países, como por exemplo Estados Unidos da América, não apresentaram resultados gráficos coerentes, como pode ser observado na imagem tal. Desta forma, faz-se necessário uma investigação acerca do motivo de alguns países (5/193) não apresentarem dados que viabilizem uma modelagem com alto grau de assertividade.

- **Possibilidade de trocar a base de dados usada na análise**: Sabe-se que existem diversas

bases de dados onde se pode acompanhar a situação da pandemia do Coronavirus no mundo. Podem ser citados como alguns exemplos a base de dados Jhon Hopkins (13), Wesley Costa (38), Worldometers (7), etc. Sabe-se também que, podem existir discrepâncias entre os dados registrados por cada uma destas bases. Desta maneira, entende-se que haveria valor científico em permitir ao usuário escolher, dentre as bases de dados disponíveis, qual usar para a realização da modelagem e predição.

5.4 APLICAÇÃO DE PARÂMETROS OTIMIZADOS PARA CENÁRIOS ALÉM DA ANÁLISE DE NOVOS CASOS

Durante o desenvolvimento a pesquisa, o levantamento dos melhores valores para os parâmetros que nortearão a modelagem foi feito sobre o número de novos casos acumulados. É denotado em (16) que o número de novos casos possui correlação com o número de mortes, bem como o número de recuperados. No entanto, no que tange a revisão bibliográfica realizada sobre os artigos científicos já publicados sobre o tema, não foram identificados trabalhos que, de forma explícita, metrificassem e provassem qual a força das correlações em questão.

Uma boa forma de endereçar isto seria o elencar os melhores parâmetros com base nos casos acumulados, e usá-los para modelar a contagem de mortes, ou de recuperados, no intuito de identificar se, mesmo tendo sido gerados sobre os números de novos casos, os parâmetros ainda resultariam uma curva integrada com alto grau de assertividade quando aplicados aos dados de pessoas recuperadas, por exemplo. Entende-se que, em caso de resultados positivos no que tange a aplicabilidade dos parâmetros otimizados sobre as diferentes métricas da pandemia, isto denotaria uma forte correlação entre as mesmas.

6 CONSIDERAÇÕES FINAIS

A motivação para o desenvolvimento da presente pesquisa foi oriunda do fato de que, na presente data (Outubro, 2021), a pandemia protagonizada pelo novo Coronavírus (Sars-Cov-2019) ainda se mostrar preocupante, num contexto global. O anseio pelo retorno das atividades tal qual eram antes do surgimento da doença, somado à falta de consenso sobre qual seria o jeito ideal de conciliar os avanços econômicos com a saúde pública, bem como a própria natureza viral da doença em questão deixa todo o globo em estado de alerta, haja visto que não existe ainda uma certeza de que não surgirão mais ondas epidêmicas.

Desta maneira, viu-se a necessidade de criação de ferramentas capazes de fornecer informações acerca da gravidade do quadro da COVID-19 em uma determinada região, bem como realizar uma projeção do cenários para os dias seguintes, de modo a auxiliar instituições médicas ou governamentais em processos de tomada de decisão, como por exemplo, enrijecer ou flexibilizar os protocolos de controle de contaminação.

Neste contexto, está inserido o projeto Hermes, uma pesquisa que, baseando-se na já comprovada capacidade descritiva da função de Richards no que diz respeito à modelagem de dados de uma onda epidêmica, visa definir uma metodologia capaz de capturar as múltiplas ondas presentes nas séries temporais formadas pelo registro novos casos de COVID-19 em uma região, identificar interseções e realizar a modelagem de cada onda isoladamente, bem como, avaliar se (ou o quanto) é eficaz a função de Richards no que diz respeito à projeção de novos casos.

O modelo Hermes se mostrou capaz de identificar as ocorrências de novas ondas, modelar a série temporal completa com alto grau de assertividade - segundo critérios definidos na etapa de validação, que correspondiam ao valor do R^2 calculado comparando os dados da modelagem com os dados reais estar entre 0,99 e 1 para os casos ideais, com uma tolerância a erro de até 5% (pra mais ou pra menos) - para casos acumulados de uma determinada epidemia, não estando este limitado à COVID-19, embora esta tenha sido o objeto de estudo da pesquisa. Além disso, o modelo também se mostrou capaz de realizar projeções com relevante nível de confiança para intervalos de até 20 dias no futuro, conforme esperado pela etapa de validação,

estando os resultados mencionado no corpo do documento.

Dentre as limitações do modelo Hermes, foi identificado que a assertividade está condicionada a uma quantidade mínima de registros para uma onda, bem como ao registro prévio do pico da onda analisada, uma vez que o modelo não é eficaz em prever quando ocorrerá a reversão da tendência da onda (término da fase de subida e início da fase de descida).

Por fim, o projeto da plataforma web foi capaz de demonstrar a relevância científica do modelo desenvolvido, tornando as modelagens acessíveis a todo e qualquer indivíduo em posse de um navegador e de acesso à internet. Espera-se com isso que mais pesquisas se derivem deste trabalho, conforme as possíveis discussões levantadas no documento.

REFERÊNCIAS

1 LIU REI-LIN KUO, S.-R. S. Y.-C. Covid-19: The first documented coronavirus pandemic in history. **Biomedical Journal**, n. 2, p. 8–9, mai 2020.

2 SANJUáN, P. D.-C. R. Mechanisms of viral mutation. **Cellular and Molecular Life Sciences**, n. 3, p. 5 – 11, jul 2016.

3 WALSH, N. P. **China to test thousands of Wuhan blood samples in Covid-19 probe**. 2021. Disponível em: <https://edition.cnn.com/2021/10/12/asia/china-wuhan-blood-samples-covid-19-probe-intl-cmd/index.html>.

4 M, T. D. F. Cold wars: The fight against the common cold. **Oxford University Press.**, p. 96, 2002.

5 ASHOUR, H. M. Insights into the recent 2019 novel coronavirus (sars-cov-2) in light of past human coronavirus outbreaks. **Pathogens (Bazel, Switzerland)**, n. 4, mar 2020.

6 WHO, W. H. O. **Coronavirus disease (COVID-19) pandemic**. 2020. Disponível em: <https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/novel-coronavirus-2019-ncov>.

7 WORLDOMETERS. **COVID-19 CORONAVIRUS PANDEMIC**. 2021. Disponível em: <https://www.worldometers.info/coronavirus/>.

8 HO CORNELIA YI CHEE, R. C. H. C. S. Mental health strategies to combat the psychological impact of covid-19 beyond paranoia and panic. **Ann Acad Med Singap**, n. 6, mar 2020.

9 GULATI, B. D. K. G. Domestic violence against women and the covid-19 pandemic: What is the role of psychiatry? **Int J Law Psychiatry**, n. 5, jun 2020.

10 BRENNER, D. B. M. H. Acceleration of anxiety, depression, and suicide: Secondary effects of economic disruption related to covid-19. **Front Psychiatry**, n. 8, 2020.

11 YELOWITZ, C. C. J. G. A. L. J. P. A. Strong social distancing measures in the united states reduced the covid-19 growth rate. **Health Affairs**, 2020. Disponível em: <https://www.healthaffairs.org/doi/full/10.1377/hlthaff.2020.00608>.

12 ROSER, H. R. E. M. L. R.-G. C. A. C. G. E. O.-O. J. H. B. M. D. B. M. Coronavirus pandemic (covid-19). **Our World in Data**, 2020. Https://ourworldindata.org/coronavirus.

13 MEDICINE, J. H. U. of. **Coronavirus Resource Center (CRC) - About Us**. 2020. Disponível em: <https://coronavirus.jhu.edu/about>.

14 BAHRI MOETEZ KDAYEM, N. Z. S. Deep learning for covid-19 prediction. **International Conference on Advanced Systems and Emergent Technologies**, v. 4, 2020.

15 YUDISTIRA, N. Covid-19 growth prediction using multivariate long short term memory. **JOURNAL OF LATEX CLASS FILES**, v. 14, n. 8, 2015.

16 METRICS, I. of H.; IHME, E. **COVID-19 Projections**. 2021. Disponível em: <https://covid19.healthdata.org/global?view=cumulative-deaths&tab=trend>.

17 ISLAM, M. K. M. S. M. M. S. Coronavirus outbreak and the mathematical growth map of covid-19. **School of Mathematical and Computational Sciences, University of Prince Edward Island,Charlottetown, PE, Canada**, 2020. ISSN 2347-565X.

18 MELO, C. M. D. F. M. F. C. A. R. de. New approach of non-linear fitting to estimate the temporal trajectory of the covid-19 cases. **Brazilian Journal of Health**, 2020.

19 MELO, C. M. D. F. M. F. M. G. M. C. A. R. de. Estimated number of deaths, confirmed cases and duration of the covid-19 pandemic in brazil. **Brazilian Journal of Health**, n. 13, 2020.

20 YISSEDT, R. B. G. C. L. **Comparative analysis of phenomenological growth models applied to epidemic outbreaks**. 2019.

21 JEFFERSON, C. H. T. **COVID 19 – "ONDAS" EPIDÊMICAS**. 2020. Disponível em: <https://oxfordbrazilebm.com/index.php/covid-19-ondas-epidemicas/>.

22 ALVARENGA, L. de R. Modelagem de epidemias atravÉs de modelos baseados em indivÍduos. **Universidade Federal de Minas Gerais**, 2008.

23 FIGUEIREDO, D. R. **Algoritmos de Monte Carlo e Cadeias de Markov**. 2021. Disponível em: <https://www.cos.ufrj.br/~daniel/mcmc/>.

24 ARA-SOUZA, A. L. Redes bayesianas: Uma introduÇÃo aplicada a credit scoring. **19º Simpósio Nacional de Probabilidade e Estatística (SINAPE)**, 2010.

25 HETHCOTE, H. W. The mathematics of infectious diseases. **SIAM - Society for Industrial and Applied Mathematics**, 2000. Disponível em: <https://epubs.siam.org/doi/10.1137/S0036144500371907>.

26 NETO Élis Gardel da Costa Mesquita; Janeisi de Lima Meira; José de R. L. D. AplicaÇÃo do modelo sir À covid-19: distanciamento social e (des)evolução da pandemia no tocantins. **Revista Observatório**, 2020.

27 KERMACK W. ; MCKENDRICK, A. A contribution to the mathematical theory of epidemics. **Proceedings of the Royal Society of London Series A Mathematical and Physical Sciences**, 1927.

28 NEPOMUCENO, E. G. Dinâmica, modelagem e controle de epidemias. **City, University of London**, 2005.

29 TAYLOR, A. E. **The American Mathematical Monthly**. [S.l.: s.n.]. v. 59. 20 – 24 p.

30 CHOWELL, G. Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts. 2017.

31 KEMPIńSKA-MIROSłAWSKA, A. W.-K. B. The influenza epidemic of 1889–90 in selected european cities – a picture based on the reports of two poznań daily newspapers from the second half of the nineteenth century. **Medical Science Monitor**, 2013.

32 ROCHA, L. **O que são ondas da Covid-19 e por que o Brasil pode estar diante da terceira**. 2021. Disponível em: <https://www.cnnbrasil.com.br/saude/2021/05/30/o-que-sao-ondas-da-covid-19-e-por-que-o-brasil-pode-estar-diante-da-3>.

33 ZHANG FRANCISCO ARROYO MARIOLI, R. G. S. A second wave? what do people mean by covid waves? – a working definition of epidemic waves. **Medrxvi**, 2021.

34 FISAYO, S. T. T. Three waves of the covid-19 pandemic. **Postgraduate Medical Journal**, v. 97, p. 332 — 332, 2021.

35 SILVA, D. C. M. da. **Velocidade e comprimento de onda**. Disponível em: <https://mundoeducacao.uol.com.br/fisica/velocidade-comprimento-onda.htm>.

36 GROSSMANN, M. **Coronavírus: como diferentes culturas enfrentam a pandemia**. Disponível em: <https://jornal.usp.br/radio-usp/coronavirus-como-diferentes-culturas-enfrentam-a-pandemia/>.

37 FRIAS, V. F. E. M. P. R. L. C. D. A study of the application of growth functions for the direct prediction of the total number of deaths in the covid-19 pandemic. n. 12, 2020.

38 COSTA, W. **Monitoramento do número de casos de COVID-19 no Brasil**. Disponível em: <https://covid19br.wcota.me/>.

39 FARIAS, R. G. da Silva; Alef Berg de Oliveira;Igor Cruz da Silva;Thulio de O. Application of a demand forecasting model in a rental company of billiard tables. p. 15–58, 2018.

40 MITCHELL, C. **How to Use a Moving Average to Buy Stocks**. 2020. Disponível em: <https://www.investopedia.com/articles/active-trading/052014/how-use-moving-average-buy-stocks.asp>.

41 MATOS, D. **Por que Cientistas de Dados escolhem Python?** 2020. Disponível em: <https://www.cienciaedados.com/por-que-cientistas-de-dados-escolhem-python/>.

42 ATLASSIAN. **O que é Git**. Disponível em: <https://www.atlassian.com/br/git/tutorials/what-is-git>.

43 LONGEN, A. **O Que é GitHub e Para Que é Usado?** 2021. Disponível em: <https://www.hostinger.com.br/tutoriais/o-que-github>.

APÊNDICES

APÊNDICE A – Tecnologias utilizadas

A.1 LINGUAGEM DE PROGRAMAÇÃO

Das linguagens de programação a serem utilizadas no projeto, foram definidas as seguintes:

- Javascript: Encontrando-se na versão 12 (*EcmaScript 12*) durante o desenvolvimento do projeto, esta linguagem se mostra de grande valor por possuir suporte à maioria dos padrões de design e arquitetura ja validados científicamente, sintaxe amigável, extensa comunidade, diversidade de frameworks ja validados no mercado, performance acima do desejado para o projeto, bem como ja ser de domínio do aluno, por fazer parte da sua atividade laboral há anos. A linguagem será utilizada na construção da interface gráfica, bem como na execução de eventuais scripts para automatização.

- Python: Linguagem utilizada pela maioria dos cientistas de dados. Se mostra pertinente ao projeto por possuir uma grande comunidade, crescente numero de bibliotecas de analise de dados, possuir um ambiente de desenvolvimento de fácil montagem e que permite ao programador testar diferentes partes do código de forma assíncrona (*Jupyter Notebook*), fornecer boas soluções para visualização de gráficos e ser mais rapido que outros pacotes para se trabalhar com Ciência de Dados (como *Matlab, R e Stata*) (41). Apesar do fato do aluno não possuir tanta experiência com a linguagem, esta foi projetada para ser de simples e rapido aprendizado, sendo portanto perfeitamente aplicável ao projeto. Será utilizada para leitura e processamento dos dados das bases já referenciadas, bem como para a construção de relatórios e possivelmente a aplicação computacional principal do projeto.

A.2 CONTROLE DE VERSÃO

O controle de versão dos códigos será realizado por meio das seguintes ferramentas:

- Git: Projeto *open source* (de código aberto) para controle de versão, inicialmente voltado

para o controle de versões de códigos de aplicações computacionais, porém hoje utilizado também para versionamento de arquivos em geral. É de amplo uso no mercado, com alto grau de confiabilidade, como afirma (42):

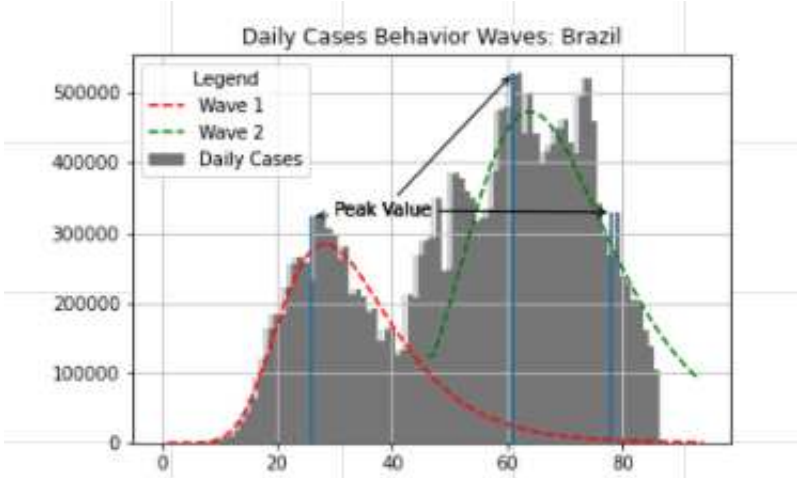
> O Git é um projeto de código aberto maduro e com manutenção ativa desenvolvido em 2005 por Linus Torvalds, o famoso criador do *kernel* do sistema operacional *Linux*. Um número impressionante de projetos de software depende do Git para controle de versão, incluindo projetos comerciais e de código-fonte aberto. [...] Tendo uma arquitetura distribuída, o Git é um exemplo de DVCS (portanto, Sistema de Controle de Versão Distribuído). Em vez de ter apenas um único local para o histórico completo da versão do software, como é comum em sistemas de controle de versão outrora populares como CVS ou Subversion (também conhecido como SVN), no Git, a cópia de trabalho de todo desenvolvedor do código também é um repositório que pode conter o histórico completo de todas as alterações. Além de ser distribuído, o Git foi projetado com desempenho, segurança e flexibilidade em mente (42).

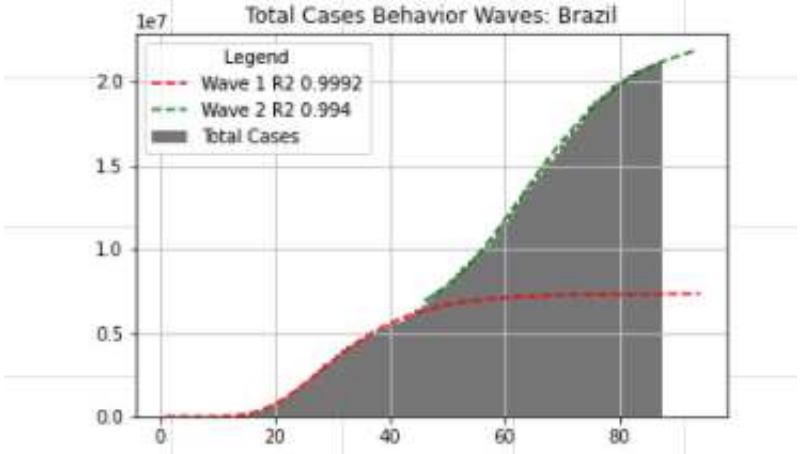
- GitHub: Serviço baseado em nuvem, que possui o sistema de controle de versão Git hospedado. O GitHub permite que os desenvolvedores colaborem e façam mudanças em projetos compartilhados enquanto mantêm um registro detalhado do seu progresso (43).

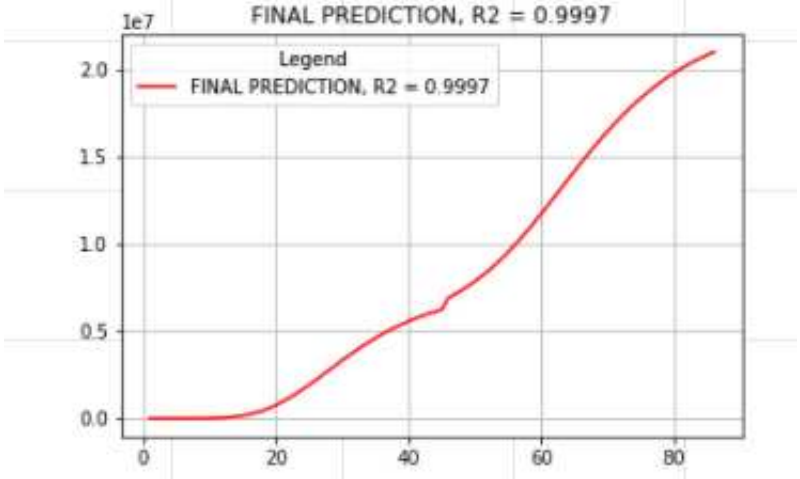
APÊNDICE B – Tabela com melhores parâmetros encontrados por país (Página 1)

country_name	chunck_size	wave_offset	moving_average_index	r_squared
Africa	3	8	20	0.9995
Albania	3	14	8	0.9995
Algeria	4	7	9	0.9993
Andorra	6	15	20	0.9987
Angola	5	15	19	0.9993
Argentina	6	15	14	0.9996
Armenia	4	15	4	0.9991
Asia	5	15	20	0.9992
Australia	6	3	5	0.9999
Austria	4	11	19	0.9987
Azerbaijan	3	9	11	0.9991
Bahamas	4	15	16	0.9986
Bahrain	3	15	20	0.9993
Bangladesh	5	15	16	0.9992
Barbados	3	9	6	0.9984
Belarus	7	15	20	0.9995
Belgium	4	6	10	0.9962
Belize	3	13	4	0.9976
Benin	4	8	5	0.9927
Bhutan	4	14	9	0.9728
Bolivia	7	5	6	0.9986
Bosnia and Herzegovina	4	15	20	0.9995
Brazil	7	15	19	0.9997
Brunei	4	13	11	0.9936
Bulgaria	5	13	17	0.9994
Burundi	6	15	19	0.9965
Cambodia	3	15	8	0.9982
Cameroon	3	6	4	0.9987
Canada	7	11	10	0.9945
Cape Verde	5	15	20	0.9984
Central African Republic	3	9	6	0.9941
Chad	3	3	17	0.9988

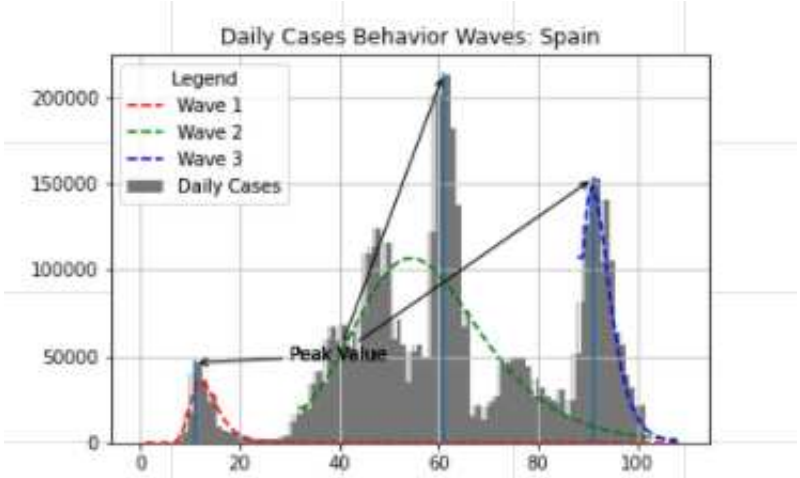
APÊNDICE C – Brazil: Modelagem utilizando parâmetros otimizados (Chunk Size = 7, Wave Offset = 15, Moving Average Index = 19)

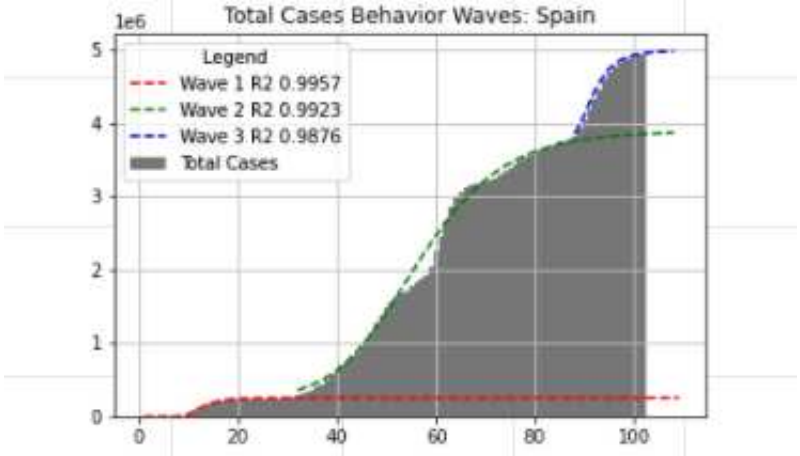


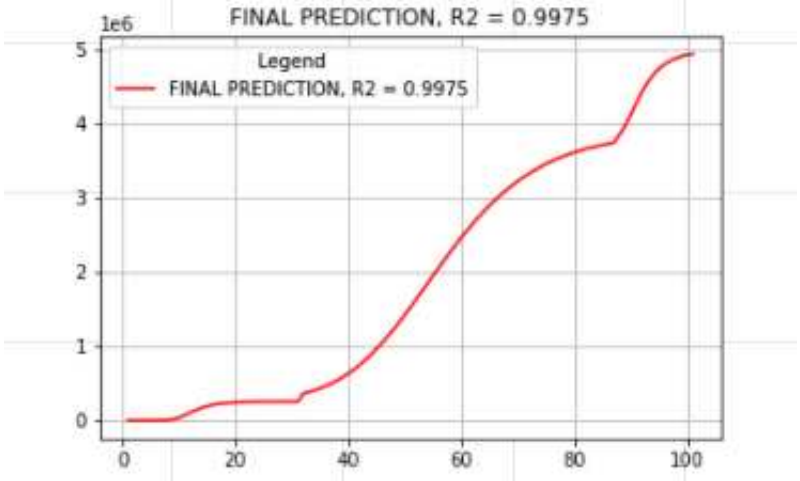




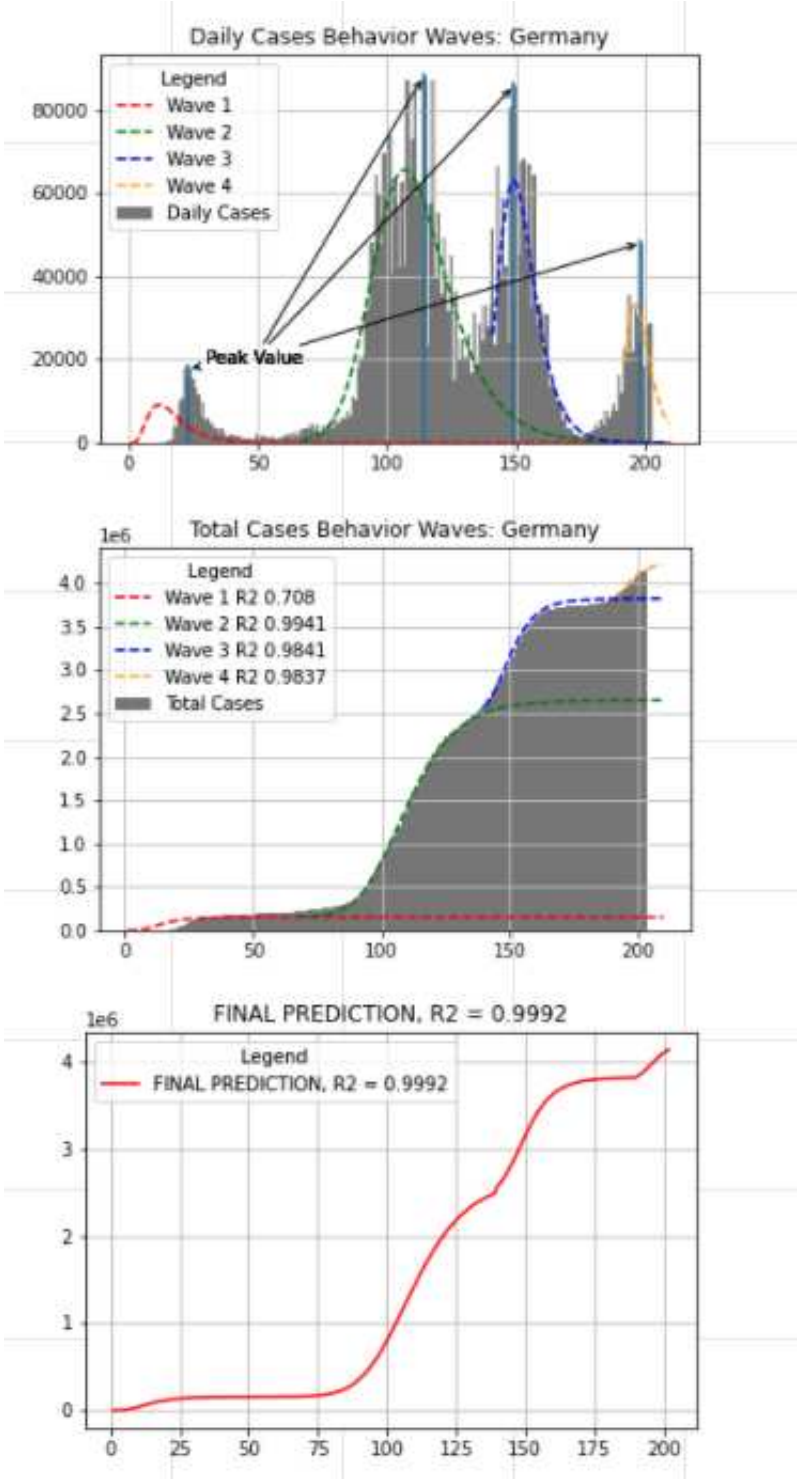
APÊNDICE D – Espanha: Modelagem utilizando parâmetros otimizados (Chunk Size = 6, Wave Offset = 15, Moving Average Index = 20)







APÊNDICE E – Alemanha: Modelagem utilizando parâmetros otimizados (Chunk Size = 3, Wave Offset = 15, Moving Average Index = 20)



APÊNDICE F – Itália: Modelagem utilizando parâmetros otimizados (Chunk Size = 3, Wave Offset = 15, Moving Average Index = 20)