UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

JACIELE DE JESUS OLIVEIRA

MODELOS SIR E ALGORITMOS TIPO ENSEMBLE COM APLICAÇÕES A COVID-19

Recife

2022

JACIELE DE JESUS OLIVEIRA

MODELOS SIR E ALGORITMOS TIPO ENSEMBLE COM APLICAÇÕES A COVID-19

Dissertação apresentada ao Programa de Pós-Graduação em Estatística do Centro de Ciências Exatas e da Natureza da Universidade Federal de Pernambuco, como requisito parcial à obtenção do título de mestre em Estatística. Área de Concentração: Estatística Aplicada

Orientador: Prof. Dr. Raydonal Ospina Martínez

Co-Orientador: Prof. Dr. Cristiano Ferraz

Recife

2022

Catalogação na fonte
Bibliotecária Nataly Soares Leite Moro, CRB4-1722

O48m Oliveira, Jaciele de Jesus
Modelos SIR e algoritmos tipo ensemble com aplicações a COVID-19 / Jaciele de Jesus Oliveira. – 2022.
75 f.: il., fig., tab.

Orientador: Raydonal Ospina Martínez.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CCEN, Estatística, Recife, 2022.
Inclui referências e apêndice.

1. Estatística Aplicada. 2. Doenças infecciosas. 3. Subnotificação. 4. Modelos compartimentados. I. Ospina Martínez, Raydonal (orientador). II. Título.

310 CDD (23. ed.) UFPE- CCEN 2022 - 60

JACIELE DE JESUS OLIVEIRA

MODELOS SIR E ALGORITMOS TIPO ENSEMBLE COM APLICAÇÕES A COVID-19

Dissertação apresentada ao Programa de Pós-Graduação em Estatística do Centro de Ciências Exatas e da Natureza da Universidade Federal de Pernambuco, como requisito parcial à obtenção do título de mestre em Estatística. Área de Concentração: Estatística Aplicada

Aprovada em: 14 de Fevereiro de 2022

BANCA EXAMINADORA

Prof. Dr. Raydonal Ospina Martínez (Orientador)
Universidade Federal de Pernambuco - UFPE

Prof. Dr. Alex Dias Ramos
Universidade Federal de Pernambuco - UFPE

Prof. Dr. Hemílio Fernades Campos Coelho
Universidade Federal da Paraíba - UFPB

AGRADECIMENTOS

À Deus, pela vida, saúde, por ser a minha força e me permitir viver esse momento. Obrigada meu Senhor, sem Ti eu não teria conseguido. Toda honra e glória seja para o Senhor!

À meu amado noivo Felipe. Obrigada meu bem por tanto amor, cuidado, apoio e compreensão em todas vezes que não pude te dá a atenção que eu queria.

À minha mãe, minha vó e meus amados irmãos, por todo apoio, carinho, por confiarem e acreditarem em mim e nos meus sonhos. Obrigada por estarem comigo em todos os momentos da minha vida, amo vocês.

Aos meus amigos, especialmente a Rafaela Tavares e Weslen Carvalho. Obrigada por tanto apoio, pela amizade e por tanto carinho, vocês sabem o quanto foram importantes nessa caminhada.

À meu amigo Carlos Raphael pelos conselhos, preciosas discussões e pela amizade. Obrigada pelos "pitacos" e por me apoiar.

Aos meus colegas e amigos do mestrado: Jaime, Noemir, Pénelope e Suelem. Obrigada meus amigos, pelos grupos de estudos, pela partilha de conhecimentos, vocês tornaram essa caminhada mais simples e divertida.

Aos meus amigos da pós graduação que não são da minha turma, mas me auxiliaram na caminhada: Alexandro, Alisson, Luciene e Joás. Obrigada.

Ao meu orientador e co-orientador, professor doutor Raydonal Martínez e professor doutor Cristiano Ferraz, agradeço imensamente. Obrigada por todo suporte, paciência e pelos valiosos ensinamentos que tornaram possível a realização deste trabalho. Sem a orientação dos senhores não teria chegado até aqui.

Aos professores da banca examinadora pela disponibilidade, paciência e pelas valiosas sugestões para aperfeiçoamento deste trabalho.

Ao Programa de Pós-graduação em Estatística da UFPE e a todos os professores que fazem parte do corpo docente deste programa, agradeço profundamente por me proporcionar conhecimentos e ensino de excelência.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) pelo apoio financeiro que permitiu que eu pudesse me dedicar completamente a esta pesquisa.

RESUMO

Em janeiro de 2020 o mundo foi surpreendido com uma pandemia devido ao COVID-19, causada pelo vírus SARS-CoV-2. Os primeiros casos foram notificados na China e se espalhou rapidamente pelo mundo, de tal forma que no dia 11 de março de 2020 a Organização Mundial de Saúde (OMS) classificou a disseminação do vírus como uma pandemia. Por se tratar de um novo patógeno, até então, não havia conhecimento sobre sua taxa de infecção e sintomas que poderia causar, isso torna crucial o uso de modelos que permitissem descrever o curso da epidemia. Neste trabalho abordaremos alguns desses modelos, que podem ser utilizadas para descrever a propagação de doenças infecciosas. Utilizamos o modelo compartimentado SIR nos dados de COVID-19 do estado da Paraíba. Nosso objetivo é estimar as taxas de infecção e recuperação da doença e comparamos com resultados de prevalência estimados por uma pesquisa amostral sorológica probabilística realizada no estado. Os resultados obtidos pelo modelo SIR indicam subestimação, o que ocorre pelo fato do modelo ser ajustado com base em dados com subnotificação. Numa tentativa de aprimorar a análise dos dados, passamos trabalhar com as curvas acumuladas de óbitos, uma vez que essas curvas são mais estáveis e os números de óbitos não dependem do registro de casos confirmados. Para isso, utilizamos uma abordagem via modelo combinados (ensemble). Este tipo de abordagem usa modelos dinâmicos de crescimento integrando a predição de vários modelos através de uma combinação ponderada, o que permite diminuir o erro de previsão. Para a construção do modelo ensemble utilizamos os modelos de crescimento logístico, de Gompertz e de Richards. O modelo ensemble descreveu de forma satisfatória aos dados se mostrando uma metodologia promissora para predição dos dados da COVID-19.

Palavras-chave: doenças infecciosas; subnotificação; modelos compartimentados; modelagem de conjunto.

ABSTRACT

In January 2020, the world was surprised by a pandemic due to COVID-19, caused by the SARS-CoV-2 virus. The first cases were reported in China and spread rapidly around the world, so such that on March 11, 2020 the World Health Organization (WHO) classified the dissemination of the virus as a pandemic. As this is a new pathogen, until then, there was no knowledge about its rate of infection and symptoms it could cause, this makes the use of models that allow describing the course of the epidemic to be crucial. In this work we will approach some of the models, which can be used to describe the spread of infectious diseases. we use the SIR compartmented model in the COVID-19 data from the state of Paraíba. Our goal is estimate rates of infection and disease recovery and compared with prevalence results estimated by a probabilistic serological sample survey conducted in the state. The results obtained by the SIR model indicate underestimation, which occurs because the model is adjusted based on data with underreporting. In an attempt to improve data analysis, we started to work with the accumulated death curves, since these curves are more stable and the numbers of deaths do not depend on the registration of confirmed cases. For this we use an approach via combined model (ensemble). This type of approach uses dynamic growth models, integrating the prediction of several models through a weighted combination, which allows to reduce the forecast error. For the construction of the ensemble model we used the logistic growth models, by Gompertz and Richards. The ensemble model described the data satisfactorily, proving to be a promising methodology for predicting COVID-19 data.

Keywords: infectious diseases; underreporting; compartmentalized models; set modeling.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SÍMBOLOS

I	Infectados
S	Suscetíveis
R	Recuperados
E	Expostos
N	Tamanho da população
β	Taxa de infecção
R_0	Número básico de reprodução
K	Tamanho final da pandemia
γ	Taxa de crescimento
α	Parâmetro de forma
$\approx$	Aproximadamente

SUMÁRIO

1 INTRODUÇÃO

A humanidade sempre teve presente em sua história fatos ligados a epidemias. Uma das epidemias mais mortíferas da história foi a peste-negra que acometeu a Europa no período de 1348 a 1351, dizimando um terço da população deste continente (SARMENTO, 2020). No período de 1850 a 1960, a tuberculose, uma infecção causada pela bactéria Mycobacterium tuberculosis (MTB) matou cerca de um bilhão de pessoas (VENDRAMINI, 2001). Não podemos deixar de citar a pandemia da gripe espanhola, conhecida também como gripe de 1918, que afetou cerca de 500 milhões de pessoas em todo o mundo, essa gripe resultou na morte de 3% a 5% da população mundial (JILANI; JAMIL; SIDDIQUI, 2020).

Em janeiro de 2020 o mundo foi surpreendido com uma nova epidemia devido a COVID-19, causada pelo vírus SARS-CoV-2, mais conhecido como novo coronavírus. O surto desse vírus teve início na China, na cidade de Wuhan e se espalhou rapidamente pelos outros países no mundo, de tal forma que no dia 11 de março de 2020 a Organização Mundial de Saúde (OMS) classificou o alastramento do vírus como uma pandemia (UNA-SUS, 2020).

Cientistas chineses descobriram que o novo vírus teve origem zoonótica em morcegos (considerados excelentes reservatórios virais), porém houve um processo de mutação nesse vírus que até então atacava morcegos e passou a infectar humanos (FIOCRUZ, 2020). A sequência genética do novo coronavírus tem 96 % de similaridade com sequências genéticas de outros coronavírus que circulam em morcegos na China. Este vírus é mais letal e mais transmissível que outros vírus que causam infecções respiratórias. Sua transmissão é dada por secreções das vias aéreas (tosse, saliva, contato pessoal e contaminação de objetos), pode permanecer em superfícies por dias e é resistente a altas temperaturas (GALLASCH *et al.*, 2020).

O primeiro caso do novo coronavírus no Brasil foi registrado pelo Ministério da Saúde no dia 26 de fevereiro de 2020 no estado de São Paulo. Se tratava de um senhor de 61 anos de idade com histórico de viagem para Itália (Governo do Brasil, 2020). Quinze dias após a confirmação do primeiro caso de infecção no país, no dia 12 de março, ocorreu o primeiro óbito em virtude da doença, aqual tem como vítima, uma mulher de 57 anos que havia sido internada com sintomas do novo coronavírus um dia antes de falecer (SAÚDE, 2020). Desde então, o vírus se espalhou pelo país ocasionando o rápido aumento de vítimas que vinheram a óbito. Diante deste cenário surgiu grande preocupação por parte dos governantes de que grande parte da população fosse infectada durante o mesmo período, sobrecarregando desta forma o sistema público de saúde e causando um possível aumento na taxa de mortalidade pela

infecção. Para tentar frear a contaminação pelo SARS-CoV-2 foram aplicadas medidas restritivas de distanciamento social e até mesmo lockdown no Brasil e em vários países no mundo.

Por se tratar de um novo patógeno, não havia conhecimento sobre o comportamento do SARS-CoV-2, enquanto a pessoas no mundo estavam assustadas com o anúncio de uma pandemia. Portanto, era crucial o uso de ferramentas que possibilitassem descrever o curso da epidemia, avaliar o efeito das medidas restritivas, fazer possíveis previsões de cenários para a propagação do vírus e então auxiliar os governantes na implementação de medidas eficazes de enfrentamento contra a COVID-19.

Vários estudos utilizando modelos epidemiológicos vêm sendo realizados para tentar descrever a propagação do vírus. Estes tipos de modelos descrevem a transmissão de doenças infecciosas através dos indivíduos, estimando parâmetros epidemiológicos importantes como taxas de transmissão da doença e número básico de reprodução (detalhado no Capítulo 2). Os modelos epidemiológicos compartimentados são amplamente utilizados para investigar doenças infecciosas, eles subdividem a população em categorias conforme a fase de evolução da doença, de forma que a população vai percorrendo cada categoria. Outros modelos que são os modelos de crescimento, os quais estudam a curva de infectados acumulados em um determinado tempo, tem tido grande enfoque nesse tipo de pesquisa, eles estimam as taxas de crescimento.

Este trabalho tem por objetivo analisar as taxas de infecção da COVID-19, utilizando modelos epidemiológicos compartimentados e modelos de crescimento; comparando resultados obtidos pelos modelos com dados provenientes de uma pesquisa sorológica do estado da Paraíba, permitindo analisar o impacto causado pelos parâmetros estimados em ambas situações e melhorar o desempenho de previsão através dos modelos *ensemble*.

O trabalho está organizado assim: o Capítulo 2 discute os modelos matemáticos não lineares, descrevendo os modelos compartimentados e de crescimento, e definindo parâmetros importantes como a taxa de infecção, taxa de recuperação e taxa de reprodução, entre outros; o capítulo 3 aborda a estimação de parâmetros e algoritmos iterativos de otimização não linear e introduz os critérios de desempenho de modelos utilizados neste trabalho; o Capítulo 4 introduz os modelos de conjunto, conhecido também como modelos *ensemble*, e descreve a metodologia *ensemble* adotada aqui. O Capítulo 5 discute e mostra através de gráficos e tabelas os resultados obtidos por cada ajuste; finalmente o Capítulo 6 conclui este trabalho, mostrando as nossa considerações finais.

1.1 MODELOS MATEMÁTICOS

O uso de modelos matemáticos para modelar a disseminação de doenças contagiosas não é recente. No século VIII, o matemático Daniel Bernoulli recorreu a modelos baseados em equações diferenciais ordinárias para estudar a epidemia de varíola que ocorreu nessa mesma época na Europa (HETHCOTE, 2000).

Em 1906, com o objetivo de entender a recorrência da contaminação por sarampo, Hammer formulou o primeiro modelo em que se reconhecia que a taxa de incidência estava relacionada a fatores como o número de suscetíveis, o número de infectados e a taxa de contato entre suscetíveis e infectados. No ano de 1911, enquanto estudava a incidência da malária, Ronald Ross apontou a existência de um valor limite de densidade de mosquitos abaixo do qual a malária se extinguiria naturalmente. A hipótese de Ross pode ter sido o prenúncio para o teorema do limiar desenvolvido por Kermack e McKendrick em 1926, o qual indica uma densidade crítica de indivíduos abaixo do qual a entrada de indivíduos infecciosos não gera uma epidemia (BARROS, 2007). Em 1927, Kermack e McKendrick formularam o primeiro modelo epidemiológico compartimentado, que divide a população em categorias e utiliza de equações diferenciais, ele ficou conhecido como modelo SIR (*suscetíveis-infectados-removidos/recuperados*). A partir do modelo SIR surgem outros modelos compartimentados, a saber: SEIR (*suscetíveis-expostos-infectados-removidos*) e SIRD (*suscetíveis-infectados-recuperados-mortos (do inglês deceased)*).

1.2 MODELO SIR E SUAS VARIAÇÕES APLICADOS A COVID-19

Anastassopoulou *et al.* (2020) utilizou o modelo SIRD para estimar os principais parâmetros epidemiológicos (número de reprodução básico, as taxas de infecção e recuperação da doença por dia) da província de Hubei, China, no período de 11 de janeiro a 10 de fevereiro e realizou também uma predição da evolução da infecção três semanas depois. Neste trabalho, o número de reprodução básico foi estimado com base em dois cenários: no primeiro utilizando os dados oficiais e no segundo considera vinte vezes a quantidade de infectados e quarenta vezes o número de casos recuperados. Os autores concluíram que o modelo não consegue descrever de forma satisfatória a disseminação do vírus em ambos os cenários.

Através do modelo SIRD Silva *et al.* (2020) analisou o número de casos confirmados e o número de óbitos causados pela COVID-19 em cada estado da região Sul do Brasil para dados obtidos até o dia 6 de junho de 2020. Ele estimou o número de Reprodução Efetivo (NER) nesses estados. Neste trabalho o autor concluiu que o modelo SIRD reproduz com boa precisão

o número de casos confirmados e mortes causadas pela doença.

Em outro estudo, Silva (2020) utilizou o modelo SIR para dados de COVID-19 no estado do Paraná. Ele estimou o número de infectados no estado considerando quatro cenários diferentes: com base nos dados informados pela Secretaria de Estado da Saúde do referido estado, com menor isolamento social, com isolamento zero e um cenário com alto índice de isolamento. Os resultados mostram que o modelo SIR descreveu bem os dados oficiais, que no segundo e terceiro cenário anlisado o pico da do número de infectados ocorre antes da data que foi observado o pico nos dados originais e que a quantidade de infectados neste pico é maior que a observada nos dados, enquanto o quarto cenário retarda o momento do pico e estima um menor número de infecções em relação aos outros cenários.

COSTA *et al.* (2021), utilizaram o modelo SIR para estimar a taxa de reprodução de infectados pelo SARS-CoV-2 no estado do Maranhão e a evolução da doença no estado em 2020. Os resultados evidenciam que medidas restritivas adotadas pelo governo do Maranhão influenciaram na estabilidade e controle da infecção.

Com o intuito de melhorar a predição e adaptação do modelo SIR aos postulados da dinâmica de disseminação do novo coronavírus Gomes, Monteiro e Rocha (2020) propuseram três modificações no modelo: considerar um efeito dinâmico, intitulado zona de aderência, determinar a variação da taxa de crescimento em função do percentual de circulação (parcela da população que pode circular livremente pelo total de habitantes) e considerar que os infectados identificados não contribuem com a contaminação.

Larremore *et al.* (2020) utilizaram dados de pesquisas sorológicas para estimar a prevalência de SARS-CoV-2 e os parâmetros epidemiológicos. Para integrar a incerteza decorrente da especificidade e sensibilidade do teste sorológico, o autor utilizou um modelo bayesiano para encontrar a distribuição a posteriori da soroprevalência (prevalência de casos estimados pela pesquisa) e então integrou essa distribuição ao modelo SEIR considerando um modelo com 16 faixas de idade, com o intuito de estimar a data do pico do número de casos e o número básico de reprodução em dois cenários de amostragem sorológica: amostragem por conveniência de grupos etários específicos e amostras sorológicas estratificadas por idade.

Para estudar o comportamento da curva epidêmica da COVID-19 no estado de Sergipe, Sandes, Freitas *et al.* (2020) utilizaram o modelo SIR e sua variante, o modelo SIRE, o qual considera que nem todos os indivíduos suscetíveis têm a mesma probabilidade de contágio, nesse caso existem pessoas suscetíveis mais expostas ao vírus. Para comparar os dados obtidos por simulação com os dados oficiais, os autores assumiram que havia entre 3 e 9 casos reais para

cada caso registrado.

1.2.1 Modelos de Crescimento no Contexto de Doenças Infecciosas

No século VIII ocorreram significativas mudanças na população humana devido à revolução industrial, como o crescimento acelerado desta população e o processo de urbanização. Diante disso surgiu a teoria do Malthusianismo (BEZERRA *et al.*, 2016). Criada por Malthus em 1978, essa teoria apontava que a população crescia de forma mais acelerada do que a produção de alimentos, temendo assim a fome no mundo. Surgiu a partir daí primeiro modelo de crescimento. Modelos de crescimento ou modelos de dinâmica populacional de uma espécie estudam as taxas de variações na quantidade de indivíduos de uma determinada população no decorrer do tempo (FIGUEIREDO, 1997).

Os modelos de crescimento são utilizados em diversos trabalhos com o intuito de descrever por exemplo: a dinâmica de transmissão da dengue e tuberculose (CABELLA, 2012); o crescimento da doença mancha preta em citros (FATORETTO *et al.*, 2015) e o crescimento do tumor na próstata (CASTANHO, 2005); Há diversos trabalhos na literatura aplicando estes modelos à dados de COVID-19. (KAIO; BARROS, 2020) realizou uma análise preditiva do número de casos confirmados de COVID-19 no Brasil e em mais 8 países utilizando o modelo de crescimento de Gompertz. Em seu estudo, (DUTRA, 2021) utilizou também o modelo Gompertz para realizar predições sobre o número máximo de casos e morte por COVID-19. Silva, Melo e Leite (2021) utilizaram o modelo bi-logístico para descrever a tendência temporal de COVID-19 em indígenas do estado do Amapá e norte do Pará, o modelo obteve significância estatística e apontou os dias 12 de maio e 22 de julho, como os dias em que a doença desacelera na população. Em seus estudos Vasconcelos *et al.* (2020) utilizou o modelo de crescimento generalizado de Richards para analisar as curvas epidêmicas da COVID-19 para as cidades de Recife e Teresina, estado de Pernambuco.

A maioria dos estudos realizados sobre a COVID-19 utiliza dados oficiais do governo, isto é, provenientes do Ministério da Saúde. Mesmo sob tal acesso a essa informação muitas pessoas infectadas não entram nos dados oficiais por não apresentarem sintomas da doença ou até mesmo por escassez de testes, ou seja, muitas pessoas estão infectadas com o vírus e não realizam testes para detectá-lo. Devido a essa subnotificação de casos, os dados não são atualizados corretamente. Entretanto, são importantes estudos que possibilitem estimar a verdadeira quantidade de casos na população. Uma alternativa seria a realização de pesquisas

sorológicas.

A pesquisa sorológica é de extrema importância na investigação de prevalência de doenças infeciosas numa população, uma vez que permite estimar a quantidade de infecções causadas por um vírus sem testar toda a população. Uma pesquisa sorológica é realizada através da aplicação de testes sorológicos, os quais podem ser aplicados de acordo com um planejamento amostral ou testando o máximo de pessoas possíveis. Os testes mais utilizados são os testes de detecção direta (RT-PCR) e testes rápidos que detectam a presença dos anticorpos Imunoglobulina G (IgG) e Imunoglobulina monomérica (IgM). Nesse tipo de teste é possível identificar se o indivíduo está ou já esteve em algum momento em contato com o vírus. A partir dos resultados obtidos na pesquisa, é estimada então a soroprevalência da doença na população.

Inicialmente ajustamos o modelo determinístico SIR para dados da COVID no estado da Paraíba e comparamos aos resultados obtidos por uma pesquisa amostral realizada no estado. Em seguida ajustamos os modelos de crescimento logístico, Gompertz e Richards aos dados. Visando aprimorar o desempenho da previsão, ajustamos modelos de conjunto utilizando uma metodologia de bootstrap. Este tipo de abordagem consiste em combinar modelos individuais e integram a precisão preditiva entre estes modelos, possibilitando assim um controle dos erros de previsão.

2 MODELOS NÃO LINEARES

Modelos não lineares são amplamente utilizados na descrição e caracterização de uma dada população. Nesses modelos, o sistema de equações dependem dos parâmetros de interesse a serem estimados, o que impossibilita a resolução analítica do sistema. Portanto, é necessário a utilização de um processo de otimização não linear para obter as estimativas dos parâmetros nesses modelos (SARMENTO *et al.*, 2006).

2.1 MODELOS COMPARTIMENTADOS

Esses modelos admitem que a população é subdividida em categorias ou compartimentos, as quais interagem trocando elementos, de forma que os indivíduos percorrem cada compartimento. Os modelos compartimentados são muito utilizados na literatura epidemiológica para investigar a propagação de doenças infecciosas.

2.1.1 Modelo SIR

O modelo SIR foi criado por W. O. Kermack e A. G. McKendric em 1927 com o objetivo de obter mais informações sobre os efeitos dos vários fatores que regem a propagação de doenças contagiosas. Desde então, os modelos epidemiológicos compartimentados tem sido aprimorados. Este modelo considera uma população constante subdividida em três compartimentos: suscetíveis, $S(t)$; infectado, $I(t)$, e removido, $R(t)$ (HETHCOTE, 2000; MARTCHEVA, 2015; MAASSEN, 2020).

- $S(t)$ representa o número de indivíduos não infectados com a doença no momento t, ou aqueles suscetíveis à doença.
- $I(t)$ representa o número de indivíduos que tenham sido infectados com a doença no momento t e que podem transmitir a infecção para a população susceptível.
- $R(t)$ é usado para representar indivíduos que foram infectados e foram removidos da doença no momento t, devido a cura ou morte. Considera-se que os indivíduos nesta categoria não são infectados novamente.

O modelo SIR assume uma população constante N no tempo e em todo instante t, a soma da população em cada um dos compartimentos é igual a população total, ou seja, não há alteração no tamanho da população durante uma epidemia, assim $S(t)+I(t)+R(t)=N$. Como hipótese do modelo toda a população no inicio da propagação da doença é suscetível. É importante ressaltar

que a permanência de cada indivíduo em um dos estados (S, I ou R) é uma variável aleatória com distribuição exponencial.

O objetivo é descrever a transição dos indivíduos entre os compartimentos do modelo. Note que, tratando de população de pessoas e uma dada doença,temos que pessoas suscetíveis contraem a doença com uma certa taxa ao entrar em contato com indivíduos infeciosos, e estes, tornam-se infectados, os quais se recuperam e passam a integrar o compartimento recuperado. O modelo assume que os indivíduos recuperados adquirem imunidade e não voltam podem integrar a classe de suscetíveis novamente.

$$S \longrightarrow I \longrightarrow R$$

Portanto, o modelo SIR é descrito em 2.1 por um sistema de equações diferenciais que retratam como os indivíduos percorrem estes compartimentos.

$$\begin{aligned} \frac{dS}{dt} &= -\beta \frac{I}{N} S \\ \frac{dI}{dt} &= \beta \frac{I}{N} S - \gamma I \\ \frac{dR}{dt} &= \gamma I \end{aligned} \tag{2.1}$$

onde β e γ representam as taxas de infecção e recuperação do modelo, respectivamente. dS/dt, dI/dt e dR/dt representam as taxas de variação da classe de suscetíveis, de infectados e do compartimento recuperados, respectivamente.

O sinal da taxa de variação do compartimento é negativo por que o número de indivíduos nessa classe diminui no decorrer do tempo. O termo que está com sinal negativo na taxa de variação dos suscetíveis entra na taxa de variação com sinal positivo, pois os indivíduos saem de suscetíveis para infectados. O mesmo acontece com o termo negativo que integra a taxa de variação de infectados e entra na taxa de variação de recuperados com sinal positivo.

Seja k o número médio de contatos de um indivíduo infeccioso, então kS/N representa o número médio de contatos que um indivíduo infecioso tem com indivíduos suscetíveis. Suponha que ϕ representa a probabilidade de uma pessoa da classe suscetível se tornar infeciosa caso entre em contato com um indivíduo infeccioso. Assim $\phi kS/N$ é o número médio de infecções geradas pelo contato de um uma única pessoa infeciosa com pessoas suscetíveis durante todo seu período infeccioso. Portanto $\phi kSI/N$ é o numero de novas infecções geradas no tempo t, ou seja, $-dS/dt$ (2.1), dessa forma $\beta \equiv \phi k$ é a quantidade de contatos de um indivíduo

infeccioso que se tornam infectados, caso o indivíduo esteja no compartimento suscetível, ou seja, a taxa de transmissão da doença no modelo.

O parâmetro γ representa a taxa de recuperação, isto é, taxa na qual um indivíduo se move do compartimento infectado para o compartimento recuperado. Este parâmetro pode ser obtido do tempo médio $\bar{t}$ em que uma pessoa infectada permanece infecciosa.

Suponha que não há fluxo de entrada na classe infecciosa e foram colocados I_0 indivíduos na classe de infectados no tempo zero, $I(0) = I_0$. Portanto a equação diferencial que representa essa dinâmica é

$$\frac{dI(t)}{dt} = -\gamma I$$

Separando a equação diferencial

$$\frac{dI(t)}{I} = -\gamma dt$$

Integrando ambos os lados obtemos e resolvendo a equação para $I(t)$ obtemos,

$$I(t) = I_0 e^{-\gamma t}$$

ou

$$\frac{I(t)}{I_0} = e^{-\gamma t} \tag{2.2}$$

A equação 2.2 fornece a proporção de indivíduos infeciosos no tempo t, para $t \geq 0$. Assim podemos calular a fração de indivíduos que deixaram a classe infecciosa no instante t por

$$1 - e^{-\gamma t}$$

ou em termos de probabilidade,

$$F(T) = 1 - e^{-\gamma t} \qquad t \geq 0$$

Note que $F(T)$ é uma distribuição de probabilidade, portanto a densidade de probabilidade é dada pro $f(t) = dF/dt$, assim

$$f(t) = \gamma e^{-\gamma t} \tag{2.3}$$

Em seguida podemos calcular o tempo médio (tempo esperado) que um indivíduo permanece infeccioso,

$$\bar{t} = E(t) = \int_0^\infty t f(t) dt$$
$$= \int_0^\infty t\gamma e^{-\gamma t} dt \quad = \frac{1}{\gamma}$$

2.1.1.1 Número básico de reprodução

O número de reprodução básico R_0 é muito importante para prever se a doença infecciosa se espalhará na população ou se extinguirá. R_0 representa o número médio de infecções geradas pelo primeiro indivíduo contaminado numa população totalmente suscetível, portanto para o cálculo de R_0 é necessário que $S \approx N$. O R_0 pode ser obtido através dos parâmetros do modelo SIR apresentado em 2.4.

$$R_0 = \frac{\beta}{\gamma} \tag{2.4}$$

Para que uma doença infecciosa se espalhe na população é necessário que o $R_0 > 1$, pois caso $R_0 < 1$ a doença não se espalhará. Quanto maior o valor de R_0 maior é o potencial de propagação da epidemia. Ainda nesse sentido é possível afirmar que se a taxa de transmissão β for muito maior que as taxas de recuperação γ, então o número de infecção vai crescer muito rápido (MARTCHEVA, 2015).

Assumindo que o número de infectados inicial é $I(0)$, a solução para EDO (equação diferencial ordinária) dI/dt apresentada em 2.1 é:

$$I(t) = I(0)e^{(\beta \frac{S}{N} - \gamma)t} = I(0)e^{(\frac{\beta}{\gamma} \frac{S}{N} - 1)t} \tag{2.5}$$

Substituindo $R_0 = \beta/\gamma$ a equação 2.5 pode ser escrita como:

$$I(t) = I(0)e^{(R_0 \frac{S}{N} - 1)t} \tag{2.6}$$

Observe que $S \leq N$. Em 2.5 quando $\beta > \gamma$, o número de infectados cresce, porém quando $\beta < \gamma$ o número de infectados decresce para zero, o mesmo ocorre na equação 2.6, quando $R_0 \leq 1$, $I(t)$ decresce para zero, o que corrobora com o que foi dito anteriormente.

No modelo SIR a epidemia sempre morre. Quando $R_0 > 1$, $I(t)$ cresce e decai para zero quando I atinge um valor máximo I_{max}. Dividindo a taxa de variação de suscetíveis pela taxa de variação de infectados e resolvendo a equação diferencial separável obtida,

$$I + S - \frac{\gamma}{\beta} N ln(S) = 0$$

Utilizando as condições do modelo,

$$I(0)+S(0)-\frac{\gamma}{\beta}Nln(S(0))=I(t)+S(t)-\frac{\gamma}{\beta}Nln(S(t)) \tag{2.7}$$

A classe de infectados alcança o valor máximo quando sua taxa de variação é zero, ou seja, $dI/dt=0$, assim,

$$\frac{\beta}{N}I_{max}S-\gamma I_{max}=0\Longrightarrow S=\frac{\gamma}{\beta}N=\frac{N}{R_0}$$

Substituindo I_{max} e $S=\gamma N/\beta$ na equação 2.7 e resolvendo para I_{max},

$$I_{max}=I(0)+S(0)-\frac{\gamma}{\beta}N-\frac{\gamma}{\beta}Nln(S(0))-\frac{\gamma}{\beta}N+\frac{\gamma}{\beta}Nln(\frac{\gamma}{\beta}N)$$

Assumimos que no início de uma epidemia o número de pessoas recuperadas da infecção é zero, isto é, $R(0)=0$, então $I(0)+S(0)=N$. Se $S(0)\approx N$, então $ln(S(0))\approx ln(N)$. Por fim, substituindo $R_0=\beta/\gamma$, temos:

$$\begin{aligned}I_{max}&=N-\frac{N}{R_0}ln(N)-\frac{N}{R_0}+\frac{N}{R_0}(ln(N)-ln(R_0))\\&=N-N\frac{1+ln(R_0)}{R_0}\end{aligned}$$

Equivalentemente,

$$\frac{I_{max}}{N}=1-\frac{1+ln(R_0)}{R_0}$$

Este resultado é mais simples e interessante, pois depende apenas do valor de R_0.

Nas subseções 2.1.2 e 2.1.3 iremos descrever os modelos compartimentados SEIR e SIRD.

2.1.2 SEIR

No modelo SIR, uma pessoa infectada pode transmitir a infecção no momento em que contrai o vírus, porém muitas doenças infeciosas têm o que é chamado de período de incubação, nesse período a pessoa infectada não contagia outras imediatamente, ou seja, o indivíduo é exposto ao vírus, mas ainda não é infeccioso. Portanto se torna necessário criar um quarto compartimento que leve os expostos em consideração, nesse sentido surge o modelo SEIR, em que o compartimento E representa a classe de expostos (MAASSEN, 2020).

Assim como o modelo compartimentado SIR, o SEIR é descrito por equações diferenciais que retratam como os indivíduos percorrem os compartimentos. Como no modelo

SEIR existe um compartimento a mais, há também um parâmetro a mais quando comparado com o SIR, a taxa com que um indivíduo é exposto à infecção δ.

$$\begin{aligned} \frac{dS}{dt} &= -\beta \frac{I}{N} S \\ \frac{dE}{dt} &= \beta \frac{I}{N} S - \delta E \\ \frac{dI}{dt} &= \delta E - \gamma I \\ \frac{dR}{dt} &= \gamma I \end{aligned} \tag{2.8}$$

Note que a taxa de variação da classe de suscetíveis $\frac{dS}{dt}$ é negativa, pois o número de suscetíveis só diminui com o passar do tempo. A classe de expostos recebe os indivíduos que saíram da classe de suscetíveis e reduz aqueles que saíram do período de incubação com uma taxa δ. O compartimento de infectados recebe os indivíduos expostos que se tornaram infeciosos e reduz aqueles que foram removidos da doença e por último a classe R recebe os indivíduos que foram removidos da doença com uma taxa γ.

O número de reprodução básico é o mesmo apresentado para o modelo SIR, $R_0 = \frac{\beta}{\gamma}$.

2.1.3 SIRD

Diferente do SIR, o modelo SIRD inclui os dados de morte, ele divide indivíduos da categoria removidos em recuperados e mortos, adicionando assim taxa de recuperação e de mortalidade ao modelo. O modelo SIRD divide uma população constante N em 4 compartimentos: suscetíveis, infectados, recuperados e mortos. $D(t)$ representa o número de mortos (inglês deads) em decorrência da infecção (ANASTASSOPOULOU *et al.*, 2020). O modelo também é representado por um sistema de equações diferenciais dado por:

$$\begin{aligned} \frac{dS}{dt} &= -\beta \frac{I}{N} S \\ \frac{dI}{dt} &= \beta \frac{I}{N} S - \gamma I - \alpha I \\ \frac{dR}{dt} &= \gamma I \\ \frac{dD}{dt} &= \alpha I \end{aligned} \tag{2.9}$$

(2.10)

Aqui γ representa a taxa de recuperação e α representa a taxa de mortalidade (número de pessoas mortas devido a infecção no período t em relação as pessoas infectadas no mesmo período).

O número básico de reprodução do modelo é dado por 2.11.

$$R_0 = \frac{\alpha}{\beta + \gamma} \tag{2.11}$$

2.2 MODELOS DE CRESCIMENTO

Os modelos de crescimento não lineares são utilizados para estimar taxas de crescimento, é muito usado na economia, na nutrição animal, pode também ser aplicado no estudo de doenças infeciosas, entre outros. Diferente do modelo compartimentado SIR, os modelos de crescimento utilizam as curvas acumuladas em seu ajuste. Estes modelos são utilizados para modelar crescimentos populacionais, ou seja, para estudar comportamentos de curvas cumulativas em forma de S.

As técnicas de análise de modelos não lineares utilizam um processo iterativo para obter soluções de equações, pois nao é possível obter soluções analíticas de parâmetros como os modelos lineares. O processo iterativo precisa fornecer valores iniciais para os parâmetros e calcular a soma dos quadrados residuais (SQR) com base nesses valores. Os parâmetros são continuamente modificados até que o SQR seja minimizado (PAZ *et al.*, 2004).

2.2.1 Modelo Gompertz

Para descrever o crescimento de tumores sólidos, o matemático judeu Benjamin Gompertz desenvolveu uma equação em 1938, denominada equação de Gompertz (DOMINGUES, 2011). Gompertz apontou que em seu modelo a taxa de crescimento é mais alta nos estágios iniciais do processo e muda rapidamente para um crescimento mais lento. Este modelo é amplamente utilizado para o crescimento geral de células, como plantas, Bactérias, tumores, etc (BASSANEZI, 2011). A equação de Gompertz é a seguinte:

$$\frac{dC}{dt} = \gamma \ln\left(\frac{K}{C}\right) \tag{2.12}$$

onde C representa a quantidade de casos acumulado, K representa o número máximo de casos cumulativos ou o tamanho final da epidemia e γ é a taxa de crescimento per capita intrínseca da população infectada.

Para obter a solução dessa EDO é utilizado o método de separação de variáveis e integrando ambos os lados, temos:

$$\int \frac{dC}{c \ln\left(\frac{K}{C}\right)} = \int \gamma dt \tag{2.13}$$

resolvendo as integrais temos

$$-\ln\left(\ln\left(\frac{K}{C(t)}\right)\right) + \ln\left(\ln\left(\frac{K}{C(t)}\right)\right) = \gamma t \tag{2.14}$$

onde $C_0 = C(0)$ é a quantidade de casos acumulados inicial. A equação (2.14) é equivalente a 2.15

$$\ln\left(\frac{\ln\frac{K}{C_0}}{\ln\frac{K}{C(t)}}\right) = \gamma t \tag{2.15}$$

Agora aplicando a exponencial em ambos os lados para isolar $C(t)$, temos

$$\ln\left(\frac{K}{C(t)}\right) = e^{-\gamma t} \ln\frac{K}{C_0} \tag{2.16}$$

aplicando exponencial em ambos os lados novamente e isolando $C(t)$, finalmente temos

$$C(t) = Ke^{-e^{-\gamma t}\ln\frac{K}{C_0}} \tag{2.17}$$

onde $C(t)$ é a quantidade de casos acumulados no momento t. Neste modelo geralmente o crescimento é menor no começo e no fim do período da doença (BOYCE, 2000).

2.2.2 Modelo Exponencial

O modelo exponencial foi desenvolvido e apresentado por Thomas Robert Malthus em 1798. Neste modelo supõe-se que a taxa de variação de C no tempo t é proporcional a C. O modelo exponencial é dado por

$$\frac{dC}{dt} = \frac{\gamma}{N}C$$

Aqui γ é a taxa de crescimento. Quanto maior o valor de γ, maior será a quantidade C no tempo $t = 0$

Esta equação é uma EDO de primeira ordem que pode ser resolvida por integração:

$$\int \frac{dC}{C} = \int \frac{\gamma}{N} dt$$

Logo

$$C(t) = C_0 e^{\frac{\gamma}{N}t}$$

Onde γ é a taxa de crescimento exponencial, N é o tamanho da população, $C(t)$ é o número de acumulados no instante t e $C_0 = C(0)$ é a quantidade de casos inicial. Observe que γ/N é a taxa de crescimento relativa, ou seja, a taxa de crescimento do modelo dividida pelo tamanho da população, logo esta taxa de crescimento relativa é constante.

2.2.3 Modelo Logístico

O matemático belga Pierre F. Verhurst propôs um modelo em 1837, que presumia que a população poderia crescer até o limite máximo e então se estabilizar. O modelo proposto por Verhurst satisfaz a condição de que a taxa efetiva de crescimento da população muda ao longo do tempo. Este modelo é uma alternativa ao modelo de crescimento exponencial, onde a taxa de crescimento é constante e não há restrição ao crescimento populacional (BASSANEZI, 2011). O modelo logístico é dado por uma equação diferencial:

$$\frac{dC}{dt} = \gamma C(1 - \frac{C}{K}) \tag{2.18}$$

onde γ, C e K tem os mesmos significados apresntados para o modelo de Gompertz.

A equação logística é separável e pode ser resolvida:

$$\int \frac{dC}{C(1 - \frac{C}{K})} = \int \gamma dt$$

com o intuito de resolvermos a integral do lado esquerdo, multiplicamos e dividimos o mesmo por K

$$\int \frac{K}{C(K - C)} = \int \gamma dt$$

Utilizando frações parciais no lado esquerdo, temos:

$$\int \left(\frac{1}{C} - \frac{1}{K - C} \right) = \int \gamma dt$$

resolvendo as integrais temos:

$$\ln(C(t)) - \ln(K - C(t)) - (\ln(C_0) - \ln(K - C_0)) = \gamma t$$

usando propriedades logaritimicas podemos reescrever esta equação:

$$\ln\left(\frac{\frac{C(t)}{K-C(t)}}{\frac{C_0}{K-C_0}}\right) = \gamma t$$

aplicando exponencial em ambos os lados temos

$$\frac{\frac{C(t)}{K-C(t)}}{\frac{C_0}{K-C_0}} = e^{\gamma t}$$

Com um pouco mais de álgebra chegamos finalmente na equação 2.19

$$C(t) = \frac{K}{1 + \left(\frac{K}{C_0} - 1\right) e^{-\gamma t}} \tag{2.19}$$

2.2.4 Modelo de Richards

O famoso modelo de Richards é uma extensão do modelo de crescimento logístico simples, que depende de 3 parâmetros. Ele estende o modelo de crescimento logístico simples incluindo um parâmetro de forma α , que mede o desvio da curva de crescimento. Proposto por Richards em 1959, o modelo de Richards foi desenvolvido para modelar o crescimento de peixes. Este modelo é uma generalização do modelo de Bertalanffy. A equação de Richards é dada pela equação diferencial 2.20 (TSOULARIS; WALLACE, 2002).

$$\frac{dC}{dt} = \gamma C \left(1 - \frac{C}{K}\right)^{\frac{1}{\alpha}} \tag{2.20}$$

Resolvendo essa EDO obtemos então o modelo de Richards (equação 2.21)

$$C(t) = K\left(1 - e^{\alpha\gamma t}\left(1 - \left(\frac{C_0}{K}\right)^{-\alpha}\right)\right)^{\frac{-1}{\alpha}}, \tag{2.21}$$

onde K é o tamanho final da epidemia, γ é a taxa de crescimento, C_0 é o número de casos no início da pandemia e α é o parâmetro de forma que determina a curvatura. O parâmetro de forma nesse modelo permite uma maior flexibilidade a curva. O modelo de Richards admite que a curva de incidência diária é composta por um único pico de alta incidência, o que leva ao ponto de inflexão da curva epidêmica na forma de um único surto. Esses pontos de inflexão,

o momento em que a taxa de acumulação passa de crescente para decrescente ou vice-versa, podem ser encontrados ao observar o momento em que a trajetória da curva epidêmica começa a declinar (HSIEH, 2009). O ponto de inflexão N_{inf} nesse modelo é função de α e K e ocorre em

$$N_{inf} = \left(\frac{1}{1+\alpha}\right)^{\frac{1}{\alpha}} K$$

Para $\alpha = 1$ a equação (2.20) se reduz a equação de crescimento logístico de Verhulst (2.18). Essa quantidade é muito relevante na epidemiologia, apontando o início ou fim de uma fase, uma vez que encontra o momento de aceleração após a desaceleração ou o momento de desaceleração após a aceleração (HSIEH, 2009).

3 ESTIMAÇÃO DE PARÂMETROS

Informalmente falando, a estimação de parâmetros é o procedimento utilizado para encontrar parâmetros através de modelos que melhor ajustem os dados empíricos. A estimação pode ser classificada de duas formas: pontual ou intervalar. Na estimativa pontual, apenas um valor é obtido para estimar o parâmetro. Na estimação intervalar é construído um intervalo com uma probabilidade determinada de conter o parâmetro verdadeiro, essa probabilidade pré determinada é chamada de nível de confiança. Existem várias maneiras de se obter as estimativas dos parâmetros, os métodos mais utilizados na literatura são os métodos de estimativa por máxima verossimilhança e o método dos mínimos quadrados (utilizado neste trabalho).

3.1 MÉTODO DOS MÍNIMOS QUADRADOS

A soma dos quadrados dos desvios é uma medida de variação ou desvio da média, é usada para identificar dispersão nos dados e como os dados se ajustam a um determinado modelo. Considerando todos os modelos possíveis, um valor menor da soma de quadrados indica um melhor modelo, pois há menos variação nos dados. O nome soma de quadrados se dá pelo fato de ser calculada através da soma das diferenças ao quadrado (RIBEIRO, 2014).

A estimativa de mínimos quadrados do vetor de parâmetros θ, denotado por $\hat{\theta}$, minimiza a soma de quadrados dos erros (diferença entre o i-ésimo valor observado y_i e o respectivo valor ajustado).

$$SQ_R = \sum_{i=1}^{n} [y_i - f(x_i, \hat{\theta})]^2 \tag{3.1}$$

onde $f()$ é o modelo.

Em seguida para encontrar o estimador de mínimos quadrados é necessário diferenciar a equação 3.1 com relação a cada um dos parâmetros e igualar cada equação a zero

$$\frac{\delta SQ_R}{\delta \theta_r}|_{\hat{\theta}} = 0 \qquad r = 1, 2, \cdots, p$$

onde θ_r representa o r-ésimo parâmetro do modelo. O resultado disso é um sistema de equações não lineares. Para a maioria dos modelos lineares não é possível resolver os sistema de equações analiticamente, sendo necessário assim utilizar um método de otimização não linear.

3.2 OTIMIZAÇÃO NÃO LINEAR

Otimização é um problema matemático que visa encontrar o valor mínimo e máximo de uma função de várias variáveis que possuem valores em uma determinada área do espaço Multidimensional (Martinez e Santos, 1995). São eficientes na resolução de problemas matemáticos e estatísticos cujas soluções das funções de interesse não podem ser resolvidas de forma analítica, como por exemplo, as estimativas de máxima verossimilhança de um modelo estatístico, que em muitas vezes não possuem forma fechada. Dessa forma a função de verossimilhança deve ser maximizada, obtendo assim as estimativas (FRERY; CRIBARI-NETO, 2011). Aqui iremos utilizar dois métodos de otimização: métodos L-BFGS-B e Levenberg-Marquardt.

3.2.1 L-BFGS-B

O método L-BFGS-B (Byrd et al., 1995) é um algoritmo da família quasi Newton, que lida com um problema de otimização não linear com restrições simples. Este método usa muito menos memória do computador do que outros algoritmos para o mesmo problema, sendo assim muito conhecido como L-BFGS-B com limitação de memória. O método assume o cálculo da matriz Hessiana Impraticável ou muito custoso, e usa uma aproximação de memória limitada BFGS (Broyden-Fletcher-Goldfarb-Shanno) que usa informações de iterações passadas para atualizar a matriz aproximada em cada iteração. O problema pode ser escrito como:

$$minimizar \quad f(x)$$

sujeita a

$$l \leq x \leq u$$

Onde f é uma função não linear cujo gradiente g existe, l representa o limite inferior (do inglês lower) e u representa o limite superior (do inglês upper) dos parâmetros. Para realizar o algoritmo não é necessário derivadas secundárias ou conhecer a estrutura da função objetivo e pode, portanto, ser aplicada quando a obtenção da matriz Hessiana não é trivial.

3.2.2 Levenberg-Marquardt

Um método muito utilizado na literatura para resolver problemas não lineares são os método de Newton e Levenberg-Marquardt. Desenvolvido pelo estatístico americano Kenneth

Levenberg em 1944 e aperfeiçoado pelo também estatístico Donald Marquardt em 19 com o objetivo de corrigir uma falha no método padrão, o método de Levenberg-Marquardt é uma extensão do método de Gauss-Newton, pode ser usado para resolver equações simultâneas não lineares. Este método é baseado no gradiente da função e encontra o mínimo local da mesma. Este método realiza a regressão dos resíduos em relação a primeira derivada do modelo não linear de interesse com respeito aos seus parâmetros até que ocorra convergência (SANTOS, 2019) (LEVENBERG, 1944).

3.3 CRITÉRIOS DE DESEMPENHO DO MODELO

Em geral, um modelo não ajsuta com exatidão os dados empíricos, mas apresenta uma simplificação da realidade, ou seja, tenta reproduzir resultados os mais próximos possíveis dos dados observados. O desempenho de um determinado modelo pode ser verificado através de algumas medidas de desempenho, como: o coeficiente de determinação ajustado R^2, o EQM (Erro quadrático médio) e EMA (erro médio absoluto). Um ponto em comum entre esses critérios de desempenho é o fato de levarem em consideração os resíduos (distância entre os dados observados e valores ajustados) do modelo, isto é, estes critérios mostram o quão longe ou próximo os resultados ajustados estão dos dados.

3.3.1 Coeficiente de Determinação

O coeficiente de determinação R^2, também conhecido como quadrado do coeficiente de correlação de Pearson, é uma métrica de desempenho muito utilizada na literatura para avaliar a qualidade do ajuste do modelo aos dados. Este coeficiente assume valores entre 0 e 1, quanto mais próximo de 1 for o resultado, melhor o ajuste, ou seja, o modelo consegue explicar a maior parte das variáveis resposta. (PALA, 2019).

Para calcular o coeficiente de determinação ajustado, precisamos encontrar antes a soma de quadrados dos resíduos SQ_R apresentada na equação 3.1 e a soma de quadrados totais SQ_T dada por

$$SQ_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad ,$$

onde n é o número de observações, y_i representa o i-ésimo valor observado e $\bar{y}$ a média de todas as observações.

Então o coeficiente de determinação ajustado é dado por

$$R^2 = 1 - \frac{SQ_R}{SQ_T}$$

3.3.2 Erro Médio Absoluto e Erro Quadrático Médio

O erro médio absoluto (EMA) é a média aritmética da diferença entre o parâmetro e o valor estimado, enquanto o erro quadrático médio (EQM) é dado pela média do quadrado dessa diferença como nas equações 3.2 e 3.3

$$EMA = \frac{1}{n}\sum_{i=1}^{n} \mid y_i - \hat{y}_i \mid \tag{3.2}$$

$$EQM = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3.3}$$

3.3.3 Pontuação Média do Intervalo de Predição

O PMI (pontuação média do intervalo de predição) analisa a largura do intervalo considerando a incerteza da predição, diferente do R^2, EMA e EQM que consideram a distância entre o modelo e os dados (GNEITING; RAFTERY, 2007). Essa medida de desempenho é dada por

$$PMI = \frac{1}{h}\sum_{i=1}^{h}\left[(U_{ti} - L_{ti}) + \frac{2}{0,05}(L_{ti} - y_{ti})\, I\{y_{ti} < L_{ti}\} \frac{2}{0,05}(y_{ti} - U_{ti}) I\{y_{ti} > U_{ti}\}\right]$$

onde L_{ti} e U_{ti} são os respectivos limites inferior e superior do intervalo de predição no tempo t v 95% e $I\{\}$é uma função indicadora.

4 MODELOS DE CONJUNTO (ENSEMBLES)

Os modelos de conjunto, conhecidos também como modelos *ensemble* tem se destacado devido a sua robustez em processos de predição e previsão. A abordagem de conjunto combina as vantagens de vários modelos ao invés de escolher o melhor modelo segundo algum critério de seleção. Uma das vantagens de se utilizar este tipo de abordagem é a redução de erros de predição e previsão (KOTU; DESHPANDE, 2015). Chowell e Luo (2021) apresentaram uma modelagem *ensemble* via bootstrap que visa melhorar o desempenho da previsão integrando sistematicamente a precisão de predição de cada modelo. Esta metodologia é utilizada para prever a trajetória de um processo de crescimento dinâmico definido por um sistema de equações diferenciais não lineares, gerando soluções mais precisas.

4.1 BOOTSTRAP

Bootstrap, conhecido também como bootstraping, é um método de reamostragem proposto por Efron (1979). Geralmente é usado para aproximar, isto é, estimar o desvio ou variância de um conjunto de dados estatísticos e para estabelecer um intervalo de confiança. Este método tem um custo computacional baixo.

O método bootstrap é classificado em paramétrico e não paramétrico. O boostrtap paramétrico é utilizado quando existe alguma hipótese sobre a distribuição dos dados, sendo geradas B réplicas bootstrap utilizando os parâmetros estimados para os dados a partir desse modelo. No bootstrap paramétrico não é conhecida a distribuição dos dados, utilizando assim diretamente a amostra original dos dados na obtenção das réplicas bootstrap. A forma de se obter as réplicas bootstrap é a mesma nos dois casos, tanto no caso paramétrico quanto no não paramétrico (ALVES, 2013).

Para realizar o procedimento bootstrap, colhe-se uma amostra de tamanho n, denominada amostra mestre, ou ajusta-se um modelo para a população em estudo. Essa amostra ou modelo ajustado representa a população da qual a amostra foi retirada.

4.1.1 Réplicas Bootstrap

Suponha que $Y = (y_1, \ldots, y_n)$ é uma amostra aleatória de uma distribuição com fda (função de distribuição acumulada) F desconhecida e desejamos estimar um parâmetro de interesse θ com base em y. Para isso calcula-se uma estimativa $\hat{\theta} = s(y)$. A estatística de

interesse pode ser o erro padrão ou média por exemplo.

Seja $\hat{F}$ a distribuição empírica de $y_i's$, onde cada valor observado $y_i, \quad i = 1,\ldots,n$ possui probabilidade $1/n$. Uma amostra aleatória retirada de $\hat{F}$ é definida como uma amostra bootstrap.

$$\hat{F} \rightarrow (y_1^*, y_2^*, \ldots, y_n^*)$$

A notação $*$ indica que Y^* não são os dados originais Y, mas sim uma versão reamostrada de Y.

Várias réplicas bootstrap devem ser geradas através dessa amostra para que os parâmetros estimados sejam mais precisos. Em relação ao número de réplicas bootstrap, para obter uma boa estimativa de erro padrão são suficientes entre 25 e 200 replicações e mais de 500 réplicas para uma boa estimativa do intervalo de confiança (EFROM; TIBSHIRANI, 1983).

O parâmetro de interesse θ é estimado através das réplicas bootstrap

$$\hat{\theta}_b^* = s(x_b^*) \quad b = 1,2,\ldots,B \quad ,$$

onde B é o número de réplicas bootstrap geradas.

4.1.2 Erro Padrão

A estimativa do erro padrão bootstrap é dedfinido como o desvio padrão das réplicas bootstrap, sendo calculado por

$$\hat{D}_B = \frac{\sum_{b=1}^{B} [\theta^*(b) - \theta^*(.)]^2}{B-1},$$

em que $\theta^*(b)$ é igual ao valor da estatística de interesse para cada réplica B gerada. Além disso tem-se que

$$\hat{\theta}^*(.) = \frac{\sum_{b=1}^{B} \theta^*(b)}{B}$$

O erro padrão é uma medida de variabilidade, ou seja, mostra o quanto de variação ou dispersão existe em relação à média (ALVES, 2013). Em seguida apresentamos um algoritmo com uma melhor descrição para estimar o erro padrão de $\hat{\theta} = s(y)$ dos dados observados y.

1. selecione B amostras bootstrap independentes $y_1^*, y_2^*, \ldots, y_B^*$, cada amostra constituindo de n valores retiradas com reposição de $y = (y_1, y_2, .., y_n)$.

2. Calcule a réplica bootstrap correspondente para cada amostra bootstrap,

$$\hat{\theta}^*(B) = s(y^{*b}) \quad b = 1,2,\ldots,B.$$

3. Estime o erro padrão bootstrap utilizando as B estimações $\hat{\theta}^*(B)$

$$\hat{D}_B = \frac{\sum_{b=1}^{B}[\theta^*(b) - \theta^*(.)]^2}{B-1},$$

Em que, $\hat{\theta}^*(.) = \frac{\sum_{b=1}^{B}\theta^*(b)}{B}$

Este algoritmo está relacionado ao caso não paramétrico. No caso paramétrico o processo de estimação de erros é semelhante, diferindo apenas na forma como cada amostra bootstrap é obtida. A estimativa de bootstrap paramétrica do erro padrão é definida como

$$\hat{D}_{\hat{F}_{par}(\hat{\theta}^*)},$$

onde $\hat{F}_{par}$ é uma estimativa de F derivada de um modelo paramétrico.

4.1.3 Intervalo de Confiança

O processo de estimação intervalar muito utilizado na literatura, consiste na construção de intervalos que contenha o valor do parâmetro de interesse desconhecido θ, com determinado grau de confiança. Este tipo de estimação fornece uma série de estimativas possíveis, em que o coeficiente de confiança $(1-\alpha)$ com $0 < \alpha < 1$, determina o quanto essas estimativas são prováveis. O intervalo de confiança permite avaliar o erro gerado na estimação pontual e permite determinar o erro máximo cometido na estimação, com certa confiança.

O intervalo de confiança bootstrap padrão para θ com coeficiente de confiança $100(1-\alpha)\%$é dado por,

$$\left[\hat{\theta} - z_c\hat{D}_B, \hat{\theta} + z_c\hat{D}_B\right],$$

em que $\hat{D}$ é a estimativa bootstrap do erro padrão de $\hat{\theta}^*$

Tomando o o coeficiente de confiança $100(1-\alpha)\%$ com com $0 < \alpha < 1$, pode-se obter a quantidade z_c na tabela de distribuição normal padrão. Na tabela, z_c é o valor crítico de forma que

$$P\left(0 < \frac{\hat{\theta} - \theta}{\hat{D}_B} < z_c\right) = \frac{1-\alpha}{2}$$

4.2 MÉTODO ENSEMBLE

A abordagem *ensemble* combina a força de vários modelos através do cálculo de uma "média ponderada", isto é, esse tipo de modelo é basicamente uma combinação linear de modelos não lineares. Abordamos aqui um método *ensemble* baseado na combinação ponderada de modelos individuais, apresentado por Chowell e Luo (2021).

Considere I modelos paramétricos. A partir dos dados de treinamento estima-se o conjunto de parâmetros e a curva incidente média *ensemble* para o i-ésimo modelo, $i = 1, \ldots, I$. Baseado na qualidade do ajuste de cada modelo, medido através do erro quadrático médio (EQM) ou outros critérios como AIC, calcula-se o peso w_i para o i-ésimo modelo, $i = 1, 2, \ldots, I$, onde a soma de todos os pesos é igual a 1, $\sum_{i=1}^{I} w_i = 1$. Neste trabalho utilizamos o EQM para avaliar a qualidade do ajuste, então o peso para cada modelo é dado por

$$w_i = \frac{\frac{1}{EQM_i}}{\frac{1}{EQM_1} + \frac{1}{EQM_2} + \ldots + \frac{1}{EQM_I}}, \quad i = 1, \ldots, I$$

onde

$$EQM_i = \frac{1}{n} \sum_{j=1}^{n} (f_i(t_j, \hat{\theta}_i) - y_{t_j})^2$$

Em que, $f_i(t_j, \hat{\theta}_i)$ representa a curva ajustada pelo i-ésimo modelo.

Logo, a curva incidente média estimada do modelo *ensemble* é dada por

$$f_{ens}(t) = \sum_{i=1}^{I} w_i f_i(t, \hat{\theta}_i)$$

Assumindo que os dados observados têm uma estrutura de erros com distribuição Poisson com média $f_{ens}(t)$, pode-se construir um intervalo de confiança 95% ou intervalo de previsão para curva incidente no tempo t usando o método de bootstrap paramétrico. Para isso, suponhamos que o tamanho da amostra de treinamento seja n com pontos de tempo $t_1, \ldots, t_n$ (CHOWELL; LUO, 2021). Para gerar uma amostra Bootstrap, geramos uma variável aleatória y_i a partir da distribuição de Poisson com média $f_{ens}(t_j)$, ou seja, é gerada uma variável aleatória para cada ponto de tempo $t_j, \quad j = 1, \ldots, n$:

$$y_j \sim Poisson(f_{ens}(t_j)), \quad j = 1, \ldots, n.$$

Assim $\{y_1, \ldots, y_n\}$ é uma amostra bootstrap, utilizada para reajustar cada modelo I, calcular pesos para cada modelo reajustado, obter as estimativas dos parâmetros e gerar a previsão do modelo *ensemble*. Repetindo esse procedimento B vezes, pode-se construir o intervalo de confiança 95% ou predição baseado nos quantis 2,5 e 97,5%.

Como ilustração, suponha que temos três modelos individuais a partir dos quais iremos construir o modelo *ensemble*. Considere $n = 100$, então temos 100 pontos de tempo.

1. Ajuste cada um dos três modelos para a série original e estime os parâmetros.
2. Calcule o EQM de cada modelo e obtenha o peso w_i de cada modelo com base no EQM;
3. Obtenha a curva incidente média *ensemble*,

$$f_{ens}(t) = \sum_{i=1}^{3} w_i f_i(t, \hat{\theta}_i)$$

4. Assumindo que os dados observados tem uma estrutura de erros Poisson com média $f_{ens}(t)$, pode-se construir um intervalo de confiança 95% ou intervalo de previsão para curva incidente no tempo t usando o método de bootstrap paramétrico.
5. Gere uma variável aleatória y_j para a incidência em cada ponto de tempo t_j, $\quad j = 1, \ldots, 100$ com distribuição poisson com média $f_{ens}(t_j)$.

$$y_j \sim Poisson(f_{ens}(t_j)), \quad j = 1, \ldots, 100.$$

6. Repita o procedimento do item anterior B vezes, gerando assim B réplicas bootstrap e construa o intervalo de confiança;
7. Reajuste os I modelos de crescimento para cada réplica, calculando os respectivos EQMs e pesos dos modelos reajustados para cada uma dessas réplicas e construa intervalos de predição e previsão para cada modelo;
8. Obtenha B curvas de incidência média ensemble utilizando o item anterior, calculando EQM e construindo intervalos de confiança para cada uma dessas curvas médias;

5 ESTUDO DE CASO NO ESTADO DA PARAÍBA

O primeiro caso de COVID-19 no estado da Paraíba foi registrado no dia 18 de março de 2020. O paciente, um homem de 60 anos, morador do município de João pessoa, havia retornado de uma viagem à Europa no dia 29 de fevereiro. Desde o primeiro caso registrado , o vírus se espalhou por todo o estado. Com o objetivo de prevenir e combater a pandemia, o Governo do estado da Paraíba decretou situação de emergência no estado. Para auxiliar na tomada de decisão de enfrentamento à COVID-19, o Consórcio Nordeste anunciou a criação de um comitê científico que reúne os 9 governadores do nordeste (UOL, 2020).

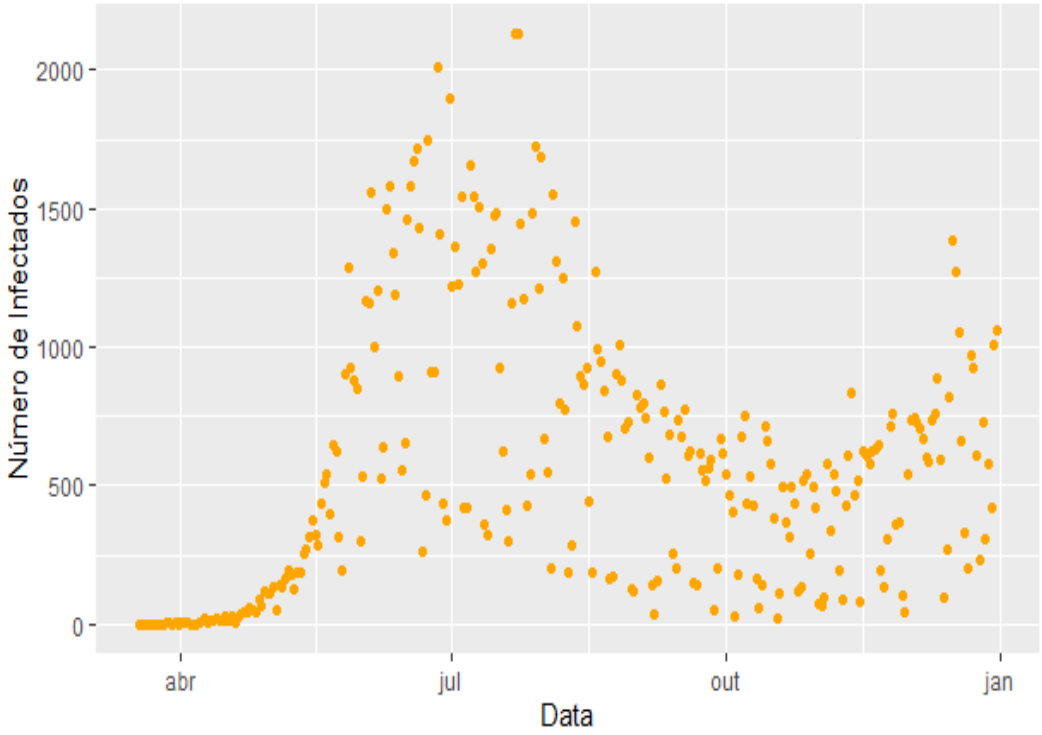
Estudamos as características da propagação do SARS-CoV-2 na Paraíba. Portanto, utilizamos o modelo SIR, três modelos de crescimento e modelagem de conjuntos. Os dados em estudo são referentes apenas ao ano de 2020, período de 18/03/20 a 31/12/20, quando ocorreu a primeira onda de COVID-19, pelo fato do modelo ser unimodal, ou seja, o modelo suporta apenas um pico (máximo), crescendo até o pico e decrescendo em seguida. Os resultados do modelo SIR serão comparados com resultados obtidos através de uma pesquisa amostral.

A figura 1 mostra a curva de incidência diária de casos da COVID-19 na Paraíba do dia 18 de março a 31 de dezembro do ano 2020. Estes dados foram fornecidos pelo Ministério da saúde. O pico de casos no estado ocorreu no dia 23 de julho com um registro de 2132 novos casos de coronavírus e 1618 mortes pela doença. Neste momento da pandemia a maior parte dos municípios estavam classificados na bandeira amarela.

No dia 20 de março de 2020 foi aberto o hospital de campanha montado no estacionamento do Hospital Metropolitano de Santa Rita, na Grande João Pessoa. O hospital, com 130 leitos, foi aberto para receber pacientes com sintomas moderados e graves de COVID-19 (PGE-PB, 2020).

Como medida de combate ao espalhamento do vírus, em maio de 2020 foi decretado que barreiras sanitárias fossem instaladas em rodovias estaduais da Paraíba para restringirem deslocamento entre municípios. Por outro lado, é importante frisar que em maio de 2020, o estado da Paraíba adotou o plano "novo normal Paraíba"que gradualmente planejava a retomada das atividades comerciais. Esse plano foi desenvolvido pela Secretaria da Saúde e pela Auditoria Geral do Estado sendo baseado em indicadores como o percentual de isolamento social, o percentual de casos novos, a taxa de ocupação da rede hospitalar da região e a taxa de letalidade. Os municípios paraibanos passaram a ser avaliados por quatro bandeiras de classificação diferenciadas pelas cores vermelha, laranja, amarela e verde. A cor vermelha indica nível de mobilidade

Figura 1 – Número de infectados por dia na Paraíba no período de 18/03/2020 a 31/12/2020.



Fonte: A autoria (2022)

impedida do município, a cor laranja representa nível de mobilidade restrita, o amarelo indica nível de mobilidade reduzida e o verde representa nível de mobilidade normal (PARAÍBA, 2020).

5.1 MODELO SIR

No início da pandemia pouco se sabia sobre o comportamento da disseminação do vírus SARS-CoV-19, logo, havia extrema importância e urgência estudos para se obter tais informações para que os gestores públicos tivessem um instrumento científico que auxiliasse na implementação de medidas de combate ao novo coronavírus e avaliar também o desempenho dessas medidas.

O modelo SIR foi aplicado a dados oficiais de COVID-19 do estado da Paraíba fornecidos pelo Ministério da Saúde (MS). Foram considerados dados do ano de 2020 para estudo, do dia 18 de março a 31 de dezembro, totalizando 289 dias.

Nesta análise, utilizamos o *software RStudio*. Especificamente, usamos o algoritmo L-BFGS-B através da função *optim* implementada nesse *software* para realizar a otimização não linear pelo método dos mínimos quadrados e utilizou-se também o pacote *ggplot2* para a construção dos gráficos apresentados.

As taxas de infecção β e de recuperação γ foram obtidas pelo algoritmo de otimização de forma que a soma de quadrados da diferença entre o valor real e o valor ajustado fosse a menor possível e assim o número de infectados simulados seriam o mais próximo dos valores fornecidos pelo MS.

Com as taxas de recuperação e infecção estimadas foi possível estimar a taxa de reprodução $R_0 = \beta/\gamma$ (como descrevemos na seção (2.1.1.1)), nos permitindo inferir sobre a propagação da epidemia no período analisado.

As taxas de infecção e recuperação não são únicas durante todo o período de uma epidemia, dado que sofrem muitas alterações no decorrer do tempo. Nosso período em análise foi dividido em nove sub períodos, o ajuste do SIR foi feito para cada um destes. Esses períodos foram separados segundo a regra de Sturges, desenvolvida pelo matemático alemão Herbert Sturges em 1926, que permite criar classes ou intervalos de frequência. Os nove períodos ficaram assim: oito períodos com 32 dias e um período (o último) com 34 dias.

A tabela 1 mostra os resultados obtidos pelo ajuste do modelo SIR, a saber: o número de suscetíveis S e infectados I_0 no início de cada período, a quantidade de infectados acumulados durante todo período especificado CAP, a taxa de infecção estimada β, a taxa de recuperação estimada α, os erros padrão (E.P) das estimativas de β e α e o número básico de reprodução R_0. Nota-se que o R_0 estimado para o primeiro período (18 de março a 18 de abril) foi $1,6351$, ou seja, nesse período 100 indivíduos infectados contaminou em média 164 pessoas durante seu período infecioso. Os erros padrão para estimativas de β e α nesse primeiro período foram bem altos (6,6791), entretanto para os outros períodos foram pequenos (abaixo de $0,0040$). Os resultados mostraram que a taxa de contágio e consequentemente o R_0 estimados para o segundo período (19 de março a 20 de maio) apresentaram maior valor em comparação com os outros períodos. Os valores estimados para β e R_0 começaram a decrescer 4 períodos seguintes (21 de maio a 25 de setembro), simulando um possível fim da pandemia, mas voltaram a crescer a partir do dia 26 de setembro do mesmo ano.

A redução das taxas de infecção nos períodos entre 21 de maio e 25 de setembro pode ter sido uma consequência das medidas restritivas adotadas no estado, bem como o aumento dessas taxas a partir do dia 26 de novembro teve relação com a flexibilização das restrição referentes ao combate ao novo coronavírus. Segundo Leung *et al.* (2020), embora essas medidas auxiliem na redução do número de reprodução básico, essa quantidade pode voltar a crescer rapidamente devido à ausência de imunidade coletiva da população, tornando possível a ocorrência de uma segunda onda da pandemia, como aconteceu na Paraíba, nos demais estados

do Brasil e vários países no mundo. Como observamos nos resultados mostrados na tabela 1, o número de reprodução básico no estado da Paraíba havia caído para valores menores que 1, mas voltou a crescer e se tornar maior que 1 novamente. É importante ressaltar que neste momento da pandemia, ainda não havia vacinação para combater o vírus.

Tabela 1 – Estimativas das taxas de infecção β, taxa de recuperação α e número de reprodução básico obtidas através do ajuste do modelo SIR para cada um dos nove períodos subdivididos do conjunto de dados.

Período	S	I_0	CAP	β	E.P. de β	α	E.P. de α	R_0
18/03/2020 a 18/04/2020	4051319	1	236	0,2704	6,6791	0,1654	6,6791	1,6351
19/04/2020 a 20/05/2020	4051075	9	5602	0,2743	0,0313	0,1395	0,0313	1,9654
21/05/2020 a 21/06/2020	4045082	400	30951	0,2606	0,0019	0,2169	0,0019	1,2017
22/06/2020 a 23/07/2020	4014062	469	36315	0,2584	0,0021	0,2175	0,0021	1,1882
24/07/2020 a 24/08/2020	3976770	1146	28199	0,2215	0,0037	0,2489	0,0036	0,8901
25/08/2020 a 25/09/2020	3949112	905	17831	0,1989	0,0040	0,2244	0,0039	0,8863
26/09/2020 a 27/10/2020	3931589	597	12301	0,2005	0,0095	0,2198	0,0092	0,9120
28/10/2020 a 28/11/2020	3919341	544	13675	0,2147	0,0062	0,2198	0,0060	0,9120
28/11/2020 a 31/12/2020	3906108	102	20316	0,2867	0,0040	0,2038	0,0038	1,4067

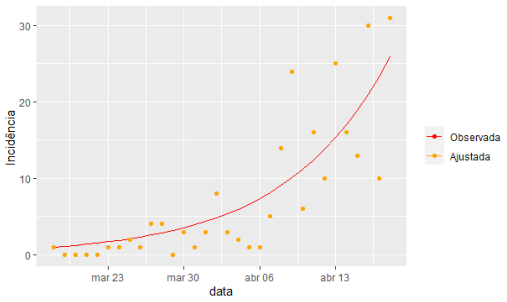
Fonte: A autoria (2022)

A figura 2 mostra a comparação entre as curvas acumuladas de casos diários e curvas simuladas para cada período de tempo. As simulações foram realizadas de acordo com os parâmetros iniciais e parâmetros estimados pelo modelo SIR apresentados na tabela 1.

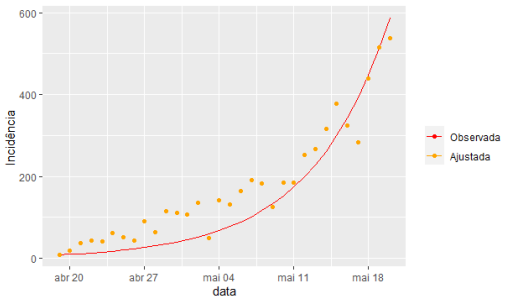
Os gráficos na figura 2 corroboram com o que foi dito anteriormente. No gráfico da figura 2b notamos um crescimento mais acelerado na curva de infectados, apesar de no mês de março terem se dado início as medidas restritivas e de distanciamento social, esse resultado se dá pelo fato de que há uma defasagem entre o momento em que ocorre a infecção e o momento que se inicia os sintomas, esse período é chamado incubação. O tempo de incubação de COVID-19 varia entre 2 e 14 dias (LAUER *et al.*, 2020). As figuras 2e, 2f, 2g e 2h mostram que a curva diária do número de infectados foram decrescentes nos períodos: 24 de julho a 24 de agosto, 25 de agosto a 25 de setembro, 26 de setembro a 27 de outubro e 28 de outubro a 28 de novembro do ano de 2020. O que indica que o número de reprodução básico foi menor que 1 nesses momentos da pandemia, como mostrado na tabela 1.

A figura 3 mostra os gráficos das curvas acumuladas dos dados oficiais versus as respectivas curvas acumuladas simuladas para cada período especificado. De forma geral, as simulações chegaram próximo aos valores fornecidos pelo ministério da saúde, apresentando uma maior discrepância no ajuste do último período (29 de novembro a 31 de dezembro).

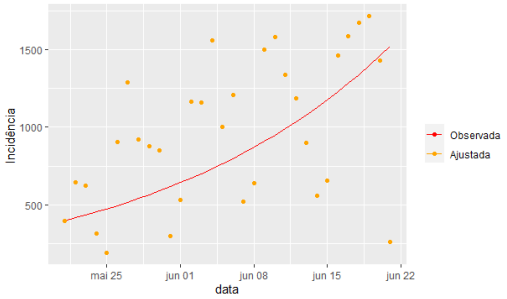
Figura 2 – Incidência diária observada versus ajustada pelo modelo SIR.



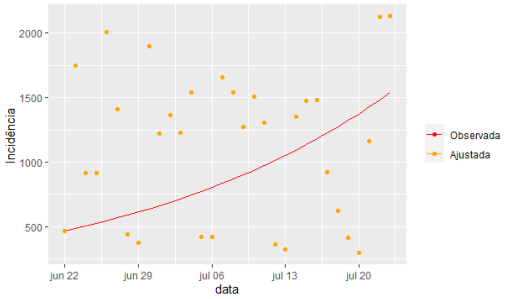
(a) 18/03/2020 a 18/04/2020



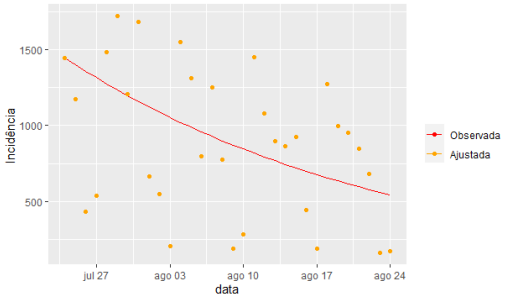
(b) 19/04/2020 a 20/05/2020



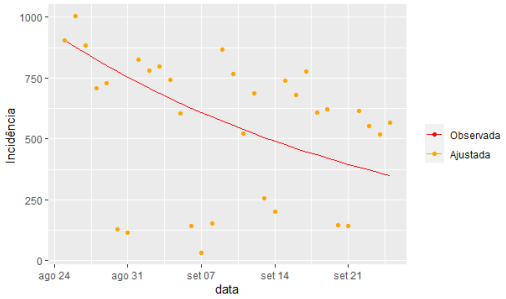
(c) 21/05/2020 a 21/06/2020



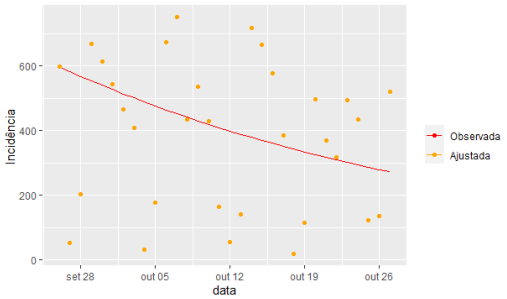
(d) 22/06/2020 a 23/07/2020



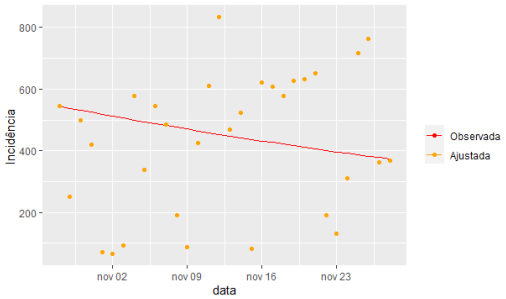
(e) 24/07/2020 a 24/08/2020



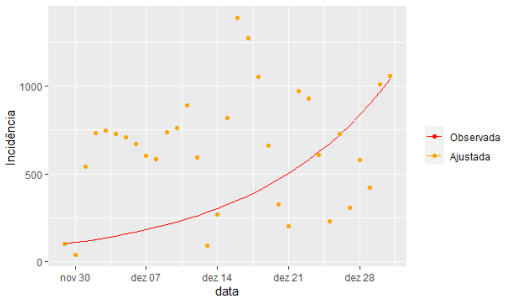
(f) 25/08/2020 a 25/09/2020



(g) 26/09/2020 a 27/10/2020



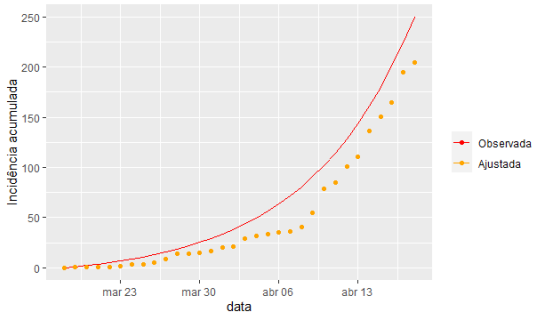
(h) 28/10/2020 a 28/11/2020



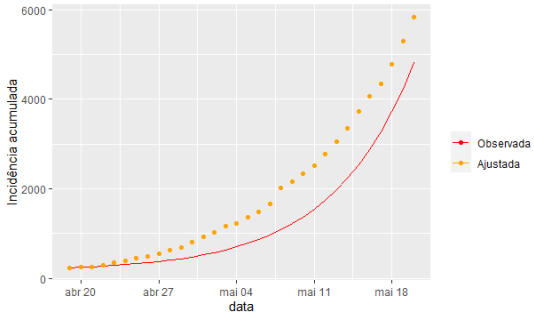
(i) 29/11/2020 a 31/12/2020

Fonte: A autoria (2022)

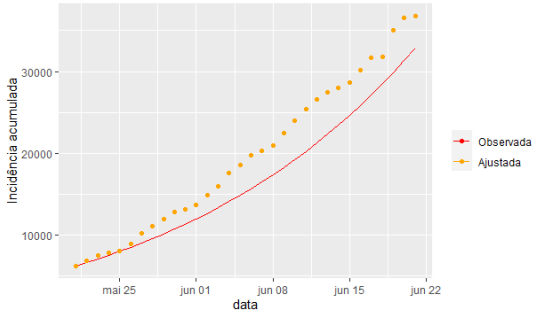
Figura 3 – Incidência acumulada observada versus obtida pelo modelo SIR.



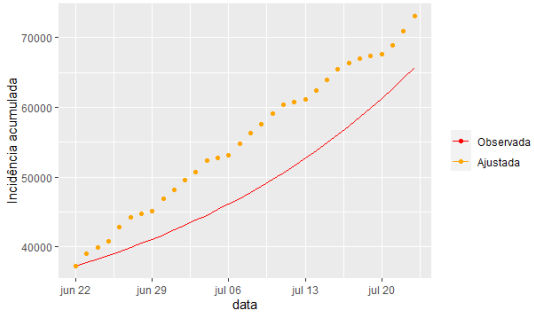
(a) 18/03/2020 a 18/04/2020



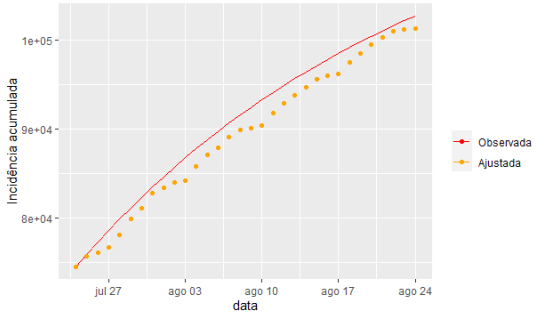
(b) 19/04/2020 a 20/05/2020



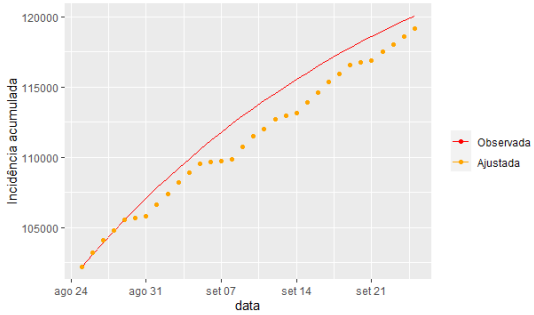
(c) 21/05/2020 a 21/06/2020



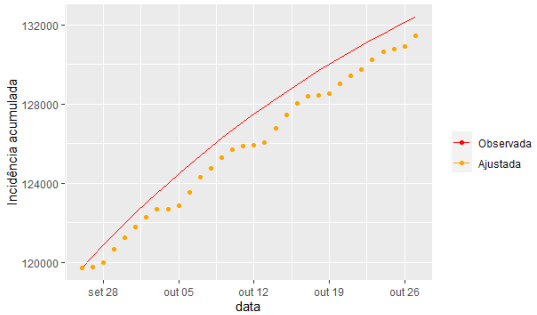
(d) 22/06/2020 a 23/07/2020



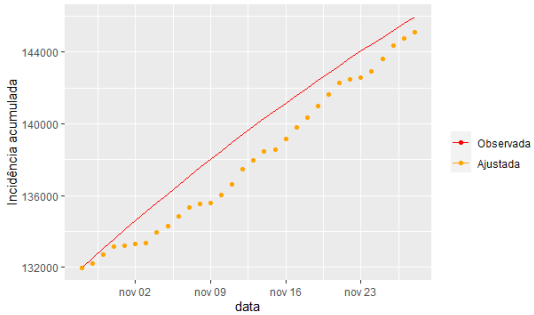
(e) 24/07/2020 a 24/08/2020



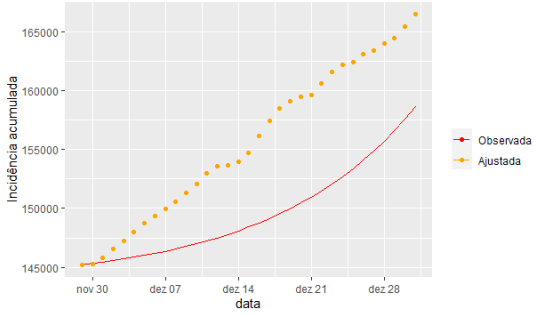
(f) 25/08/2020 a 25/09/2020



(g) 26/09/2020 a 27/10/2020



(h) 28/10/2020 a 28/11/2020



(i) 29/11/2020 a 31/12/2020

Fonte: A autoria (2022)

Um dos maiores desafios que surgem nos estudos que descrevem a transmissão da COVID-19, é a subnotificação de casos da doença. Estima-se que o número de casos sejam muito maiores que aqueles apresentados nos dados oficiais do governo. Isso ocorre pelo fato de que nem todas as pessoas que contraem o vírus SARS-CoV-2 realizam o exame de diagnóstico por não apresentarem sintomas ou devido à escassez de testes (nesse caso, somente pessoas com sintomas mais graves realizam o teste). Visando estimar a verdadeira quantidade de infectados no estado, o Programa Continuar Cuidando da Paraíba, em parceria com o Observatório de Síndromes Respiratórias da Universidade Federal da Paraíba, realizou uma pesquisa sorológica amostral no estado. Nesta pesquisa foram aplicados testes rápidos (IgM e IgG) e testes RT-PCR em uma amostra da população. Esta, se trata de uma amostra estratificada e conglomerada de domicílios, em que os estratos dividem o estado da Paraíba em quatro macrorregiões. Na amostra de domicilios selecionada foi levantada informações sobre todos os moradores. A amostra da pesquisa foi ponderada por um processo tradicional de cálculo de pesos para amostras probabilísticas.

O período da pesquisa está compreendido entre 3 de novembro e 22 de dezembro de 2020 (oito semanas). A análise foi feita a cada duas semanas, considerando as datas: 03/11/2020 a 12/11/2020, 03/11/2020 a 27/11/2020, 03/11/2020 a 10/12/2020 e 03/11/2020 a 22/12/2020.

A coleta de dados da pesquisa foi realizada pela SCIENCE (Sociedade para o Desenvolvimento da Pesquisa Científica) e a aplicação dos testes (rápido e RT-PCR) foi feita por profissionais de saúde capacitados.

Para avaliar o impacto da subnotificação de casos, foi ajustado o modelo SIR utilizando exatamente as datas da pesquisa e comparamos os números de infectados estimados pelos modelos com as quantidades estimadas utilizando os resultados da pesquisa.

A tabela 2 apresenta os parâmetros iniciais e os resultados obtidos pelo ajuste do modelo SIR para cada período da pesquisa: período considerado no ajuste, quantidade de suscetíveis S, casos acumulados no período analisado CAP, quantidade de casos acumulados do início da pandemia até o último dia do período dado $CAUP$, taxa de infecção estimada β, erro padrão de β, taxa de recuperação estimada α, erro padrão de α e o número básico de reprodução R_0. O número de casos inicial de cada modelo I_0 é 93, quantidade registrada no dia 03 de novembro de 2020. A taxa de infecção estimada no período de 03 de novembro de 2020 a 12 de novembro de 2020 é $0,3982$ e o número de reprodução básico é $2,6466$, ou seja, estima-se que nesse período um único indivíduo infectado contamine em média quase 3 pessoas. Observa-se que conforme é acrescentada o número de semanas epidemiológicas no modelo, a

taxa de infecção diminui, reduzindo assim o parâmetro R_0.

Tabela 2 – Estimativas das taxas de infecção β, taxa de recuperação α e número de reprodução básico para os períodos da pesquisa amostral, obtidas através do ajuste do modelo SIR.

Período	S	*CAP*	*CAUP*	β	E.P. de β	α	E.P. de α	R_0
03/11/2020 a 12/11/2020	3917941	4188	137474	0,3982	0,1227	0,1504	0,1156	2,6466
03/11/2020 a 27/11/2020	3917941	11485	144741	0,2610	0,0092	0,1675	0,0089	1,5578
03/11/2020 a 10/12/2020	3917941	18787	152073	0,2553	0,0052	0,1862	0,0051	1,3713
03/11/2020 a 22/12/2020	3917941	27325	160611	0,2455	0,0025	0,1886	0,0024	1,3000

Fonte: A autoria (2022)

A tabela 3 mostra o número de casos acumulados estimados pelo modelo SIR em cada período analisado *CAP*, quantidade estimada de casos acumulados do início da pandemia até o último dia do período dado *CAUP* e quantidade de casos estimados pela pesquisa amostral na Paraíba (quantidades de testes com resultados positivos para IgG e positivos para IgM). O resultado do teste reagente para IgG significa que o indivíduo esteve em contato com o vírus em algum momento e reagente para IgM significa que o indivíduo está ou esteve em contato com o vírus recentemente. Para estimar a prevalência de COVID-19 na população, foi utilizada a proporção de casos obtidos na pesquisa. Percebe-se que há uma diferença muito grande entre os resultados do modelo e da pesquisa, sendo as quantidades obtidas pela pesquisa mais que o dobro das quantidades ajustadas. Ao final da pesquisa, estimou-se que 10% da população (405132 pessoas) da Paraíba possuíam anticorpos para o SARS-CoV-2, apontando que já haviam sido infectadas com o vírus, enquanto a quantidade estimada pelo modelo nesse período foi apenas 152750. Essa subestimação do número de infectados ajustados com relação aos estimados pela pesquisa ocorre pelo fato de que o modelo foi ajustado com base em dados subnotificados. Essas subnotificações, como foi discutido acontece devido à limitação de testes de COVID-19 na população, pois geralmente somente pessoas com sintomas hospitalares são testadas, resultando assim na subnotificação de casos de COVID-19.

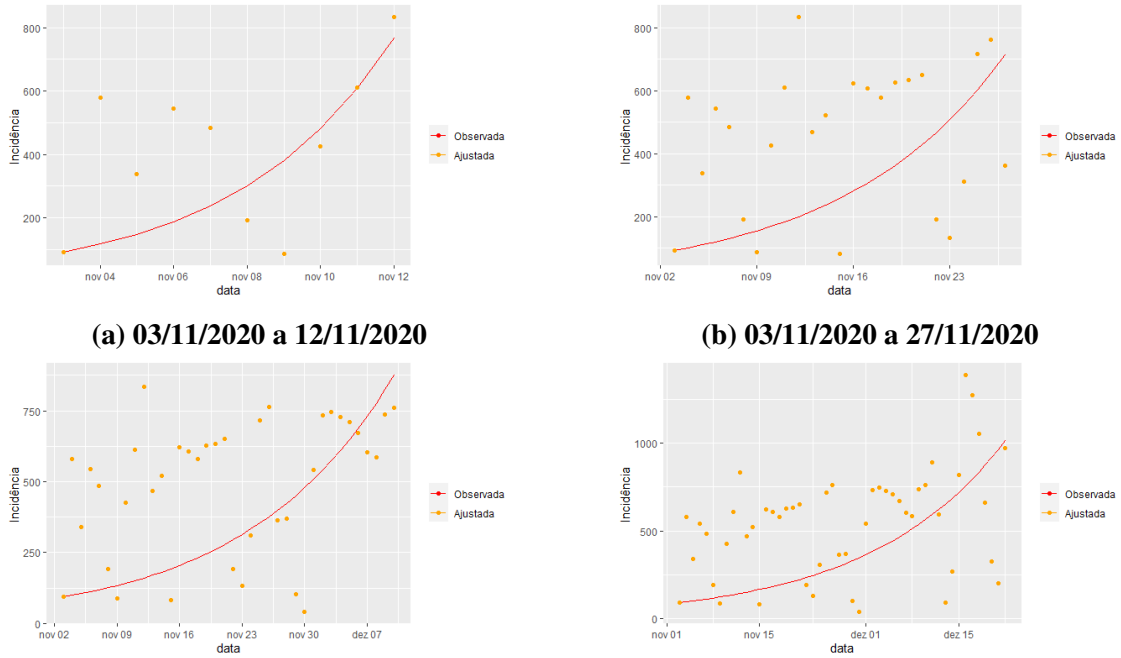
Tabela 3 – Prevalência de COVID-19 estimada pelo modelo SIR e pela pesquisa amostral.

Período	*CAP*	*CAUP*	IgG	IgM
03/11/2020 a 12/11/2020	3322	136608	397029	80665
03/11/2020 a 27/11/2020	7707	140993	376773	100348
03/11/2020 a 10/12/2020	13429	146715	384875	107657
03/11/2020 a 22/12/2020	19465	152750	405132	135445

Fonte: A autoria (2022)

A figura 4 mostra os gráficos de casos diários em cada período com a linha de tendência do modelo ajustado, nota-se que o modelo capta de forma satisfatória a tendência dos dados. A figura 5 mostra a curva acumulada observada em cada período versus a curva acumulada ajustada pelo modelo. Pode-se observar que a curva de infectados acumulada estimada pelo modelo SIR fica abaixo da curva fornecida pelos dados observados.

Figura 4 – Incidência diária observada versus ajustada pelo modelo SIR, considerando as datas da pesquisa.



(a) 03/11/2020 a 12/11/2020

(b) 03/11/2020 a 27/11/2020

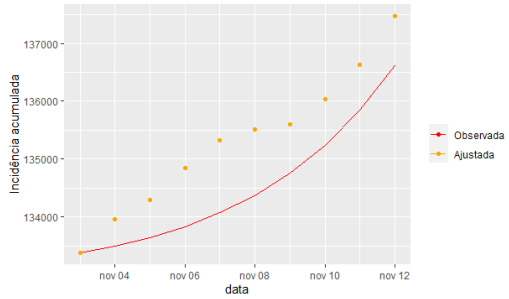
(c) 03/11/2020 a 10/12/2020

(d) 03/11/2020 a 22/12/2020

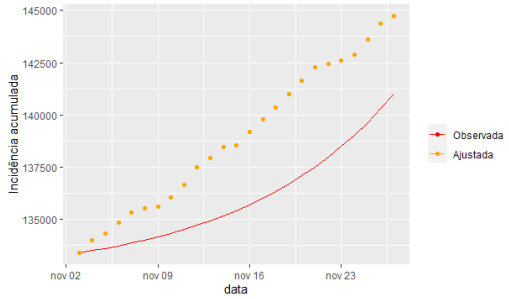
Fonte: A autoria (2022)

A curva foi ajustada considerando os dados a partir do dia 3 de novembro, neste dia o número de infectados relatado foi 93, muito abaixo dos números observados nos dias seguintes. Dessa forma para avaliar o impacto desse fato no modelo ajustou-se um modelo considerando os dados a partir do dia 4 de novembro, neste dia o número de casos relatado é 579. Os resultados obtidos pelo modelo estão apresentados na tabela 4 e figura 6. Na tabela 4 observa-se que as taxas obtidas diferiram das estimadas considerando o dia 3 de novembro como o primeiro dia do período. As taxas de infecção e número básico de reprodução obtidos nestes modelos foram bem menores em relação aos modelos ajustados considerando o dia 03, como o primeiro dia. Observamos nos gráficos da figura 6 que a curva estimada pelo modelo que considera o dia 04 como data inicial, ficou acima da curva de infectados fornecida pelo ministério da saúde. Estes resultados mostram que o modelo foi sensível às condições inicias e também às subnotificações de casos de coronavírus.

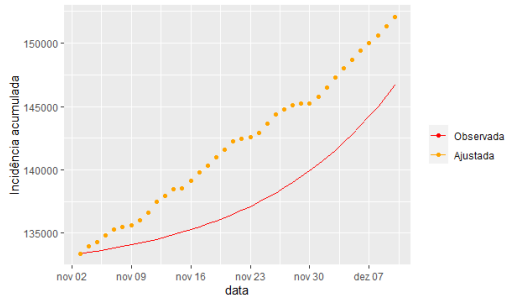
Figura 5 – Incidência acumulada observada versus ajustada pelo modelo SIR, considerando as datas da pesquisa.



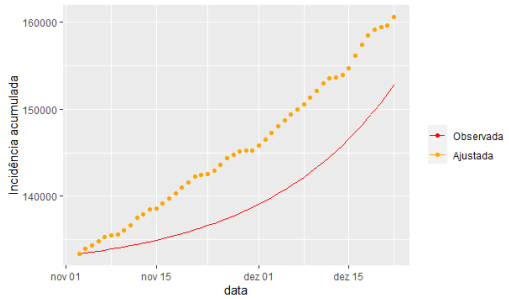
(a) 03/11/2020 a 12/11/2020



(b) 03/11/2020 a 27/11/2020



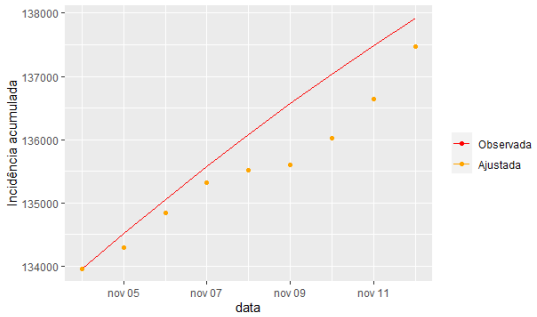
(c) 03/11/2020 a 10/12/2020



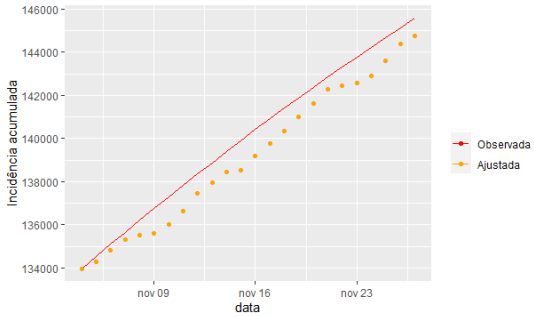
(d) 03/11/2020 a 22/12/2020

Fonte: A autoria (2022)

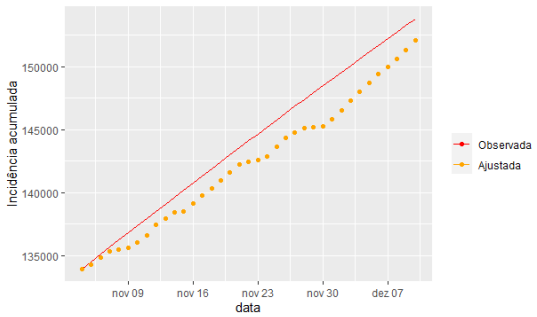
Figura 6 – Incidência acumulada observada versus ajustada pelo modelo SIR ajustado à cada período da pesquisa com início no dia 04 de novembro.



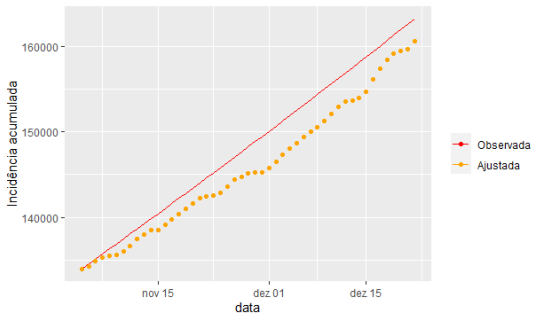
(a) 04/11/2020 a 12/11/2020



(b) 04/11/2020 a 27/11/2020



(c) 04/11/2020 a 10/12/2020



(d) 04/11/2020 a 22/12/2020

Fonte: A autoria (2022)

Tabela 4 – Estimativas do modelo SIR, considerando as datas da pesquisa com início no dia 04 de novembro de 2020.

Período	S	CAP	β	E.P. de β	α	E.P. de α	R_0
04/11/2020 a 12/11/2020	3917362	4095	0,1942	0,0257	0,2230	0,0249	0,8707
04/11/2020 a 27/11/2020	3917362	11362	0,1257	0,0112	0,1331	0,0109	0,9443
04/11/2020 a 10/12/2020	3917362	18694	0,1519	0,0037	0,1496	0,0035	1,0152
04/11/2020 a 22/12/2020	3917362	27232	0,1629	0,0022	0,1554	0,0021	1,0482

Fonte: A autoria (2022)

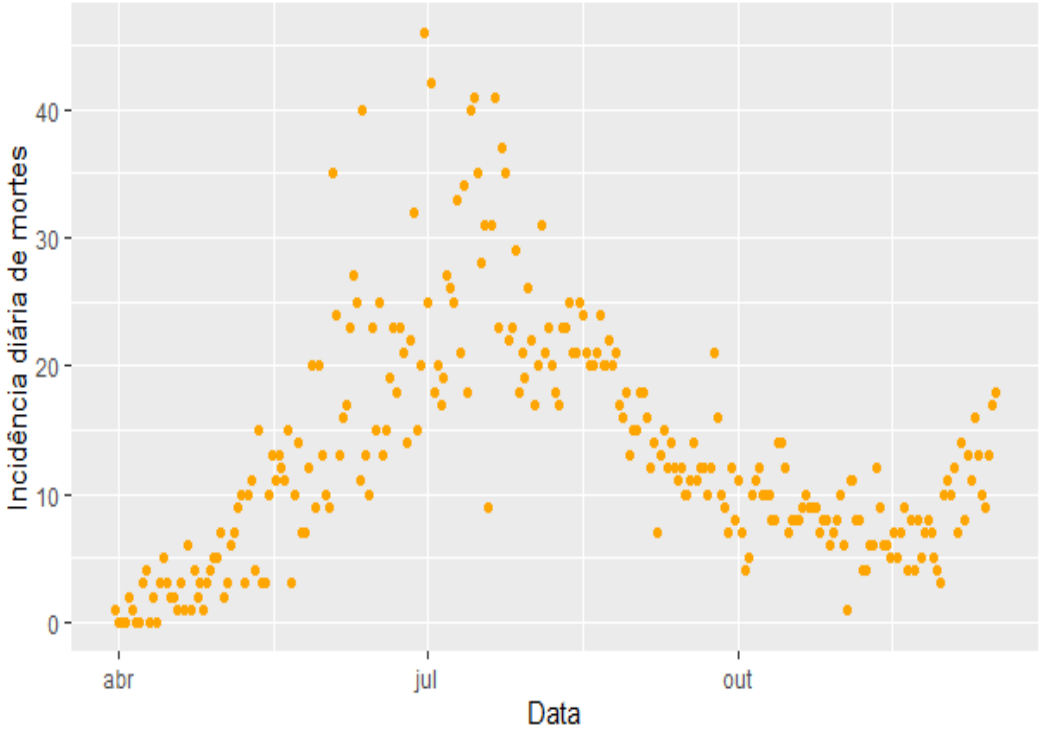
O ajuste do modelo SIR é realizado através dos casos diários da doença, os quais são muito voláteis, variado muito entre um dia e outro. Este fato pode influenciar também no ajuste. Como alternativa para o modelo compartimentado podem ser utilizados modelos de crescimento, que ao contrário do modelo SIR utiliza as curvas acumuladas que possui menos variabilidade que as curvas diárias. Dessa forma, foram ajustados modelos de crescimento, os quais serão utilizados na abordagem de conjunto.

5.2 MODELOS DE CRESCIMENTO E MODELO *ENSEMBLE*

Considerando que a taxa de subnotificação na contagem do número de mortes por COVID-19 é menor do que a taxa de subnotificação do número de casos, uma vez que esta contagem não depende da realização de testes, utilizamos aqui as curvas acumuladas de óbitos decorrentes da doença no estado da Paraíba. O registro do primeiro óbito por coronavírus na Paraíba ocorreu no dia 31 de março no ano de 2020, exatamente 14 dias após o primeiro caso de infecção no estado. A vítima foi um homem diabético de 36 anos que residia no sertão do estado, na cidade de Patos. O homem apresentou os primeiros sintomas no dia 25 de março, apenas 6 dias antes do óbito.

A figura 7 mostra o número diário de óbitos por COVID-19 no estado da Paraíba, após o primeiro caso de morte registrado até o dia 31 de dezembro do ano de 2020. Os picos dos números de óbitos no estado aconteceram nos dias 25 de maio e 5 de junho de 2020, quando foram registrados a ocorrência de 41 óbitos em cada um desses dias.

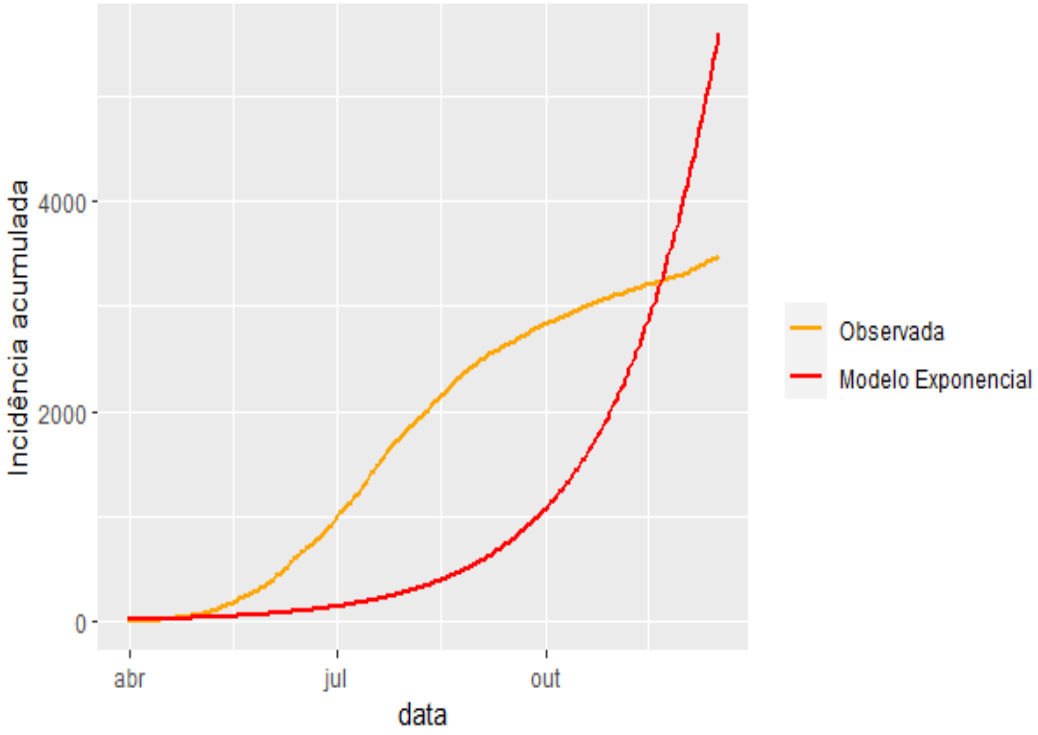
Figura 7 – Número de mortes por dia na Paraíba no período de 31/03/2020 a 31/12/2020.



Fonte: A autoria (2022)

Foram ajustados quatro modelos de crescimento e construído o modelo *ensemble* com o objetivo de estudar o comportamento da curva de óbitos na Paraíba e as respectivas taxas de crescimento. Os modelos foram ajustados considerando o período de 31 de março a 16 de dezembro de 2020 (tempo de 261 dias), deixando os últimos 15 dias do ano de fora do ajuste, para realizar a previsão de 15 dias a frente. Os modelos utilizados foram o modelo de crescimento exponencial, logístico, Gompertz e de Richards. A figura 8 mostra a comparação entre a curva acumulada observada nos dados (laranja) e a curva ajustada pelo modelo exponencial (vermelha). Observamos que o modelo não se ajustou bem aos dados, a curva simulada apresentou um comportamento muito diferente da curva real. Isso ocorreu provavelmente porque é impossível um crescimento exponencial infinito aqui, considerando que a taxa de crescimento exponencial acelera cada vez mais com o passar do tempo e que a curva inevitavelmente teria que desacelerar porque a população é finita.

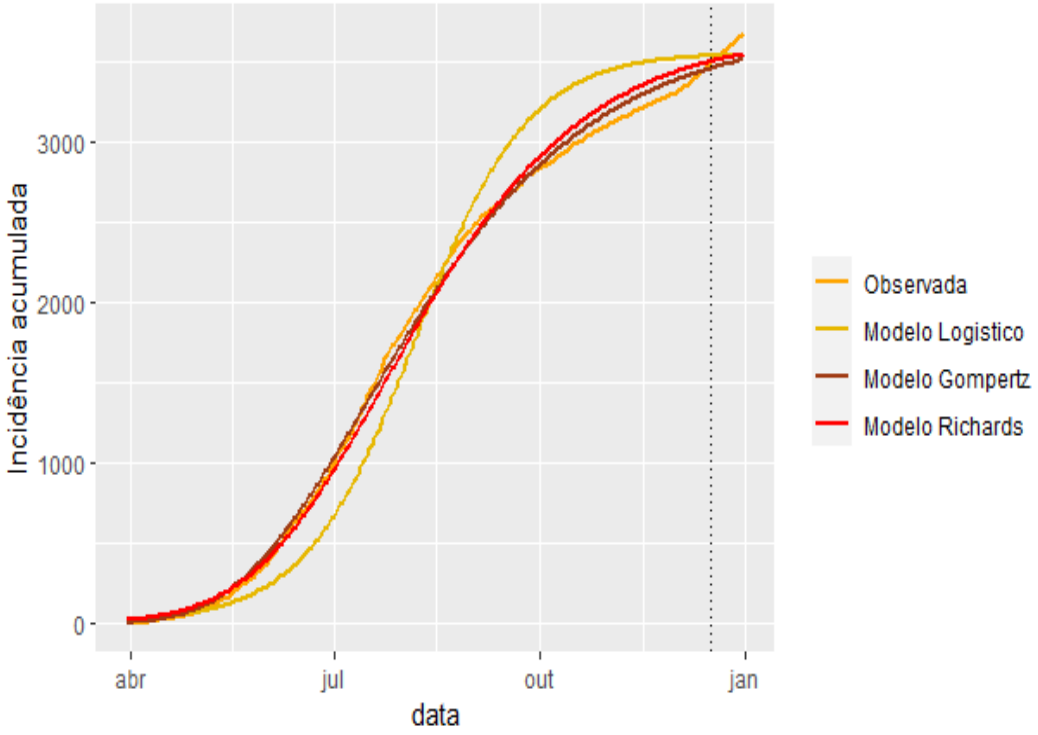
Figura 8 – Incidência acumulada diária do número de óbitos versus curva simulada pelo modelo exponencial



Fonte: A autoria (2022)

A figura 9 mostra a curva obtida pelos modelos logístico, Gompertz e de Richards e o número acumulado de óbitos fornecido pelo MS. Além da predição, foi realizada a previsão de 15 dias a frente para cada um destes modelos. A linha tracejada vertical indica o momento a partir do qual é gerada a previsão (após a linha), referente ao dia 17 de dezembro de 2020. Pelo comportamento das curvas é possível ver que os modelos de gompertz e de Richards se ajustam melhor aos dados, quando comparado com o modelo logístico. No que se refere a previsão, os três modelos subestimaram a curva de óbitos observada. O gráfico mostra que um pouco antes do dia 17 de dezembro, o comportamento da curva de dados muda de forma brusca, o que dificulta um melhor desempenho de previsão dos modelos.

Figura 9 – Incidência acumulada diária do número de óbitos versus curvas simuladas pelos modelos de crescimento



Fonte: A autoria (2022)

A partir dos modelos de crescimento, exceto o modelo exponencial que apresenta um comportamento bem diferente dos dados, foi construído o modelo *ensemble*. Para a construção deste modelo foi calculada a curva média ponderada dos modelos de crescimento com base no erro quadrático médio. A partir do modelo *ensemble* foram geradas 1000 réplicas de bootstrap, adotando uma estrutura de erros com distribuição Poisson considerando como parâmetro essa média ponderada. Em seguida foram construídos intervalos com 95% de confiança; os modelos logístico, Gompertz e Richards foram reajustados para cada uma das 1000 réplicas; calculamos o

EQM e peso de cada ajuste (das mil réplicas); construímos o modelo ensemble para cada réplica e foi finalmente obtida a nova curva média *ensemble* através dos modelos reajustados às réplicas.

A tabela 5 mostra os parâmetros estimados por cada modelo de crescimento e pelo modelo *ensemble*, a saber: o tamanho final da pandemia K, a taxa de crescimento γ, o parâmetro de forma α e os respectivos erros padrão das estimativas. A tabela mostra também o peso calculado para cada modelo, com base no erro quadrático médio, como detalhado na seção 4. O modelo Gompertz apresentou maior peso devido ao menor EQM (tabela 6) quando comparado aos modelos logístico e de Richards, por outro lado, o peso calculado para o modelo logístico foi bem pequeno. No que diz respeito as taxas de crescimento, o modelo logístico estimou que o número cumulativo de óbitos crescia com uma taxa de 3,99%, o modelo Gompertz estimou que a taxa de crescimento foi de 1,73%, já o modelo Richards estimou uma taxa de crescimento de 8%, uma taxa alta em relação às estimativas dos outros modelos, e o modelo *ensemble* estimou uma taxa de crescimento de 3,54%.

Tabela 5 – Parâmetros estimados de cada modelo de crescimento e do modelo ensemble.

Modelo	K	E.P(K)	γ	E.P(γ)	α	E.P(α)	peso
logístico	3550	0,1056	0,0399	$4,3107x10^{-06}$	–	–	0,0329
Gompertz	3700	0,2182	0,0173	$2,3139x10^{-06}$	–	–	0,6902
Richards	3672	0,2937	0,0800	$5,7843x10^{-04}$	0,2585	0,0006	0,2769
ensemble	3687	0,2354	0,0354	$1,6191x10^{-04}$	–	–	1

Fonte: A autoria (2022)

Para fins comparativos, geramos também 1000 réplicas dos modelos Gompertz, logístico e de Richards para construir o respectivo intervalo de confiança e calcular o PMI (pontuação média do intervalo).

A tabela 6 apresenta as métricas de desempenho de predição para cada modelo: o coeficiente de determinação R^2, erro médio absoluto (EMA), erro quadrático médio (EQM) e pontuação média do intervalo (PMI) de confiança 95% para o número acumulado de óbitos. Os resultados mostram que o modelo logístico obteve menor coeficiente de determinação, maior EMA e maior EQM quando comparado aos outros modelos, o que corrobora com os resultados observados na figura 9 e com o peso atribuído a este modelo (tabela 5). Os modelos de Gompertz, de Richards e *ensemble* obtiveram ótimos coeficientes de determinação (maior que 0,99), indicando um bom ajuste dos modelos aos dados. Ainda nesse sentido as curvas de Gompertz se destacam também por obterem menor MAE e EQM, entretanto o modelo Gompertz

conseguiu ainda ter melhor desempenho que o modelo *ensemble*. Observando o PMI, nota-se que os intervalos de confiança construídos considerando o modelo Gompertz e o modelo *ensemble* apresentaram menor PMI, apontando melhor desempenho destes intervalos.

Tabela 6 – Coeficiente de determinação, erro médio absoluto, erro quadrático médio e pontuação média do intervalo para avaliar o desempenho da predição.

Modelo	R^2	EMA	EQM	PMI
logístico	0,9614	205,8528	59393,68	5640
Gompertz	0,99825	46,1195	2831,737	205,9703
Richards	0,9955	71,0488	7058,065	608,7089
ensemble	0,9963	52,0208	3914.654	234,8977

Fonte: A autoria (2022)

A tabela 7 mostra as medidas de desempenho e PMI da previsão de cada modelo. Os resultados de EMA e EQM nesta tabela indicam que o modelo logístico e de Richards apresentaram melhor adequação, seguidos do modelo *ensemble*. Não obstante, o modelo logístico e Richards também obtiveram os menores PMI, apontando melhor desempenho da estimação intervalar feita através destes modelos.

Tabela 7 – Erro médio absoluto, erro quadrático médio e pontuação média do intervalo para avaliar o desempenho da previsão.

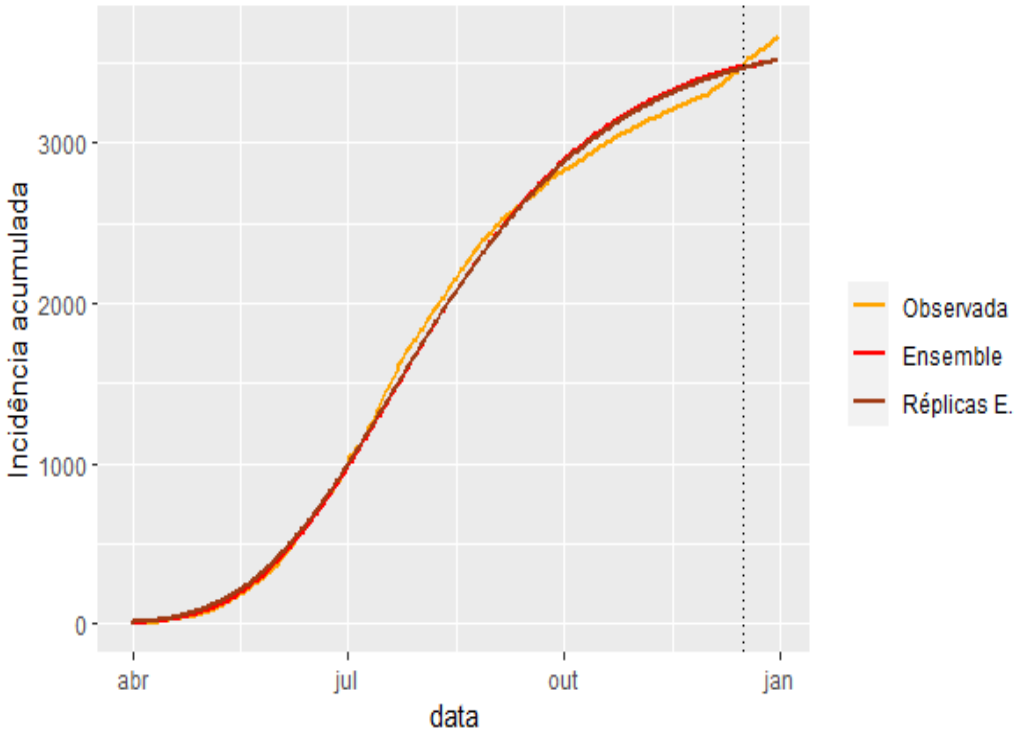
Modelo	EMA	EQM	PMI
logístico	55,6799	4860,5070	278,7717
Gompertz	106,8016	12591,6100	642,5833
Richards	69,2534	6162,4490	270,2633
ensemble	73,2892	6629,3380	293,6600

Fonte: A autoria (2022)

A figura 10 mostra a comparação da curva ajustada pelo modelo *ensemble* utilizando os dados originais diretamente e a média das curvas dos modelos *ensembles* reajustados para cada uma das 1000 réplicas *ensemble* com os dados. Observamos que a curva *ensemble* e a curva média do modelo *ensemble* para cada réplica estão bem próxima e se ajustam muito bem aos dados, mostrando uma discrepância maior nas predições após o mês de outubro e também na previsão. Mais uma vez, essa "dificuldade"na previsão de óbitos pode estar relacionada a repentina inclinação da curva no final do ano de 2020.

Na figura 11 podemos ver o intervalo com confiança de 95% construído para o número acumulado de óbitos considerando modelo *ensemble*. A linha vertical indica o momento

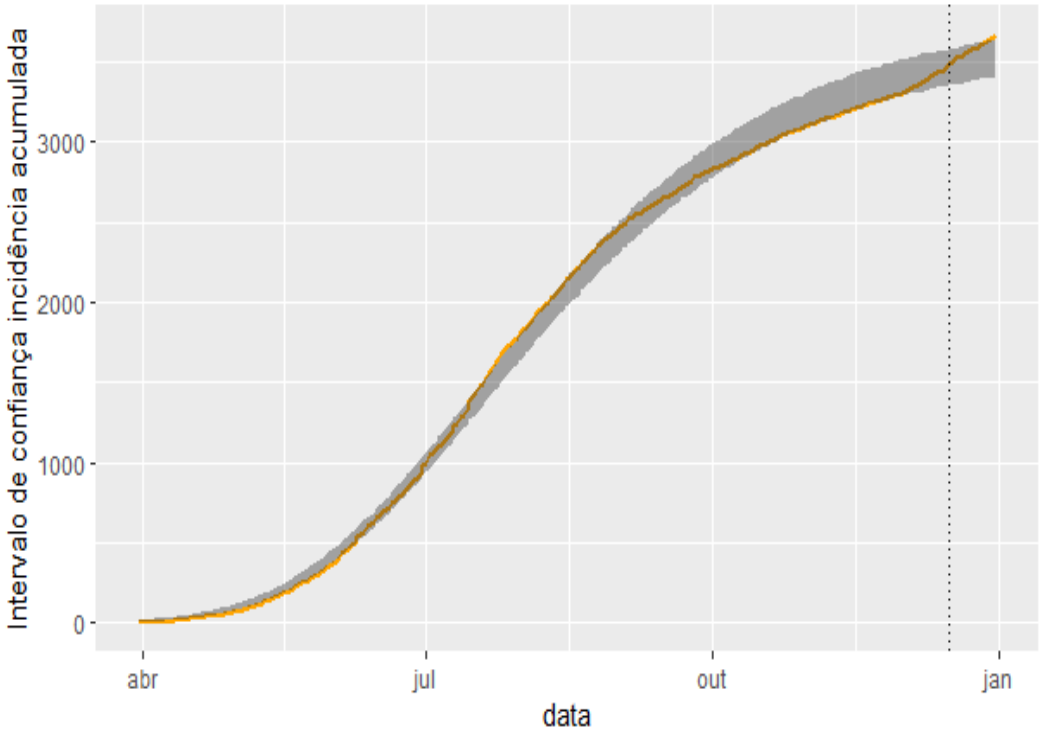
Figura 10 – Incidência acumulada ,curva ensemble e média das réplicas ensemble.



Fonte: A autoria (2022)

em que inicia a previsão. Notamos que a amplitude do intervalo é pequena, indicando boa precisão do intervalo, e que alguns pontos dos dados estão fora das margens do intervalo. Apesar disso, a distância entre os pontos que estão fora das margens e o intervalo não se mostra grande e o intervalo parece cobrir de forma satisfatória a curva de mortes observada.

Figura 11 – Intervalo de confiança pelo método ensemble



Fonte: A autoria (2022)

Após gerar as réplicas do modelo *ensemble*, os modelos de crescimento foram reajustados a cada uma das 1000 réplicas, obtendo assim 1000 curvas ajustadas de cada um dos três modelos. Em seguida foram calculados os EQMs, MAEs e pesos dos modelos ajustados às respectivas réplicas, e então construímos uma curva *ensemble* para cada réplica. Dessa forma, a tabela 8 mostra a média dos resultados obtidos nesses ajustes: estimativa do tamanho final da pandemia K, estimativa da taxa de crescimento γ, parâmetro de forma α, respectivos erros padrões das estimativas e o peso obtido de cada modelo. Notamos que as estimativas para o tamanho final da pandemia e taxa de crescimento obtidas para o modelo *ensemble* foram um pouco maiores do que as apresentadas na tabela 5, o que está relacionado aos pesos dos modelos de crescimento que também mudaram. O peso do modelo Gompertz diminuiu de 0,6902 (tabela 5) para 0,5649 e o peso modelo de Richards aumentou de 0,2769 para 0,4123. A estimativa da taxa de crescimento média foi de 4,35%, um pouco maior que a taxa obtida pelo modelo ajustado diretamente aos dados.

Tabela 8 – Réplicas: Parâmetros estimados de cada modelo de crescimento e do modelo ensemble.

Modelo	K	E.P(K)	γ	E.P(γ)	α	E.P(α)	peso
logístico	3550	0,1078	0,0274	$2,7990x10^{-06}$	–	–	0,0228
Gompertz	3730	0.2188	0,0175	$2,4002x10^{-06}$	–	–	0,5649
Richards	3672	0,2935	0,08	$1,5098x10^{-04}$	0,2586	$5,7834x10^{-04}$	0,4123
ensemble	3702	0,2471	0,0435	$6.3664x10^{-05}$	–	–	1

Fonte: A autoria (2022)

Na tabela 9 mostramos a média das métricas de desempenho obtidas através dos modelos reajustados: coeficiente de determinação R^2, erro médio absoluto e erro quadrático médio. Para o cálculo do EQM do modelo *ensemble*, o peso de cada modelo de crescimento foi multiplicado pelo respectivo EQM obtidos na b-ésima réplica, $b = 1, 2, \ldots, 1000$, e então foi calculada a média dos EQMs gerados. O mesmo foi feito para o coeficiente de determinação e o erro médio absoluto. Os coeficientes de determinação mostram que os modelos se ajustaram bem aos dados. Observa-se que o EQM do modelo logístico foi muito alto, por isso ele apresentou um peso muito baixo em relação aos modelos Gompertz e Richards. Percebemos que os modelos reajustados às réplicas obtiveram em média menor EQM do que os modelos ajustados diretamente aos dados, inclusive o modelo *ensemble*.

A tabela 10 mostra as médias do erro médio absoluto e erro quadrático médio obtidos na previsão gerada através dos ajustes dos modelos às 1000 réplicas. Diferente do que

foi observado para os resultados da predição, o modelo Gompertz apresentou maiores valores de EMA e EQM.

Tabela 9 – Resultados das réplicas ensemble: erro médio absoluto e erro quadrático médio para avaliar o desempenho da predição.

Modelo	R^2	EMA	EQM
logístico	0,9707	206,7985	45501,1777
Gompertz	0,9988	48,2287	1846,1060
Richards	0,9984	71,0308	2528,5940
ensemble	0,9979	61,2420	3123,2370

Fonte: A autoria (2022)

Tabela 10 – Resultados das réplicas ensemble: erro médio absoluto e erro quadrático médio para avaliar o desempenho da previsão.

Modelo	EMA	EQM
logístico	56,2020	4950,2090
Gompertz	74,5657	6830,1650
Richards	65,9720	5740,0720
ensemble	70,6751	6349,1710

Fonte: A autoria (2022)

Observa-se que as previsões geradas pelos modelos de crescimento e *ensemble* não apresentaram bons desempenhos. Apesar da segunda onda de COVID-19 ter acontecido no ano de 2021, é possível perceber pela figura 7 que o número de mortes voltou a crescer entre novembro e dezembro de 2020. Portanto, o modelo *ensemble* foi ajustado para os dados de óbitos registrados até o dia 30 de setembro de 2020 sendo realizadas previsões para 15 e 30 dias à frente.

A tabela 11, mostra os resultados obtidos por cada modelo de crescimento e pelo modelo *ensemble*. Dentre os 3 modelos de crescimento, o modelo de Richards obteve o maior peso para construção do modelo de conjunto. O modelo *ensemble* estimou que a taxa de crescimento do número de óbitos na Paraíba é aproximadamente 8,3% e que no final da pandemia haveria um registro de 3281 mortes (observando os dados oficiais, o modelo estima que o final da pandemia seria por volta do dia 27 de novembro de 2020).

As tabelas 12 e 13, mostram as métricas de desempenho da predição e previsão 15 dias à frente. O coeficiente de determinação mostra que os 4 modelos se ajustaram os dados de forma satisfatória. Os modelos de Richards e *ensemble* obtiveram melhor desempenho de

Tabela 11 – Parâmetros estimados de cada modelo de crescimento e do modelo ensemble para o número de óbitos por COVID-19 na Paraíba registrados até o dia 30 de setembro.

Modelo	K	E.P(K)	γ	E.P(γ)	α	E.P(α)	peso
logístico	2775	0,2019	0,0458	$8,1582x10^{-06}$	–	–	0,0380
Gompertz	3444	0,5655	0,0203	$5,1910x10^{-06}$	–	–	0,2765
Richards	3244	0,8318	0,1102	$3,9789x10^{-04}$	0,2239	$9,7707x10^{-04}$	0,6855
ensemble	3281	0,7342	0,0829	$2,7450x10^{-04}$	–	–	1

Fonte: A autoria (2022)

predição e previsão no quesito EMA e EQM, o que concorda com o peso obtido pelo modelo Richards. Observamos que o modelo *ensemble* apresentou melhor desempenho de previsão em comparação com os outros modelos.

Tabela 12 – Métricas de desempenho da predição para cada modelo ajustado para o número de óbitos registrados até 30 de setembro de 2020: coeficiente de determinação, erro médio absoluto e erro quadrático médio.

Modelo	R^2	EMA	EQM
logístico	0,9970	43,7131	3003
Gompertz	0,9996	17,1538	412
Richards	0,9998	9,7566	166
ensemble	0,9997	10,4711	183

Fonte: A autoria (2022)

Tabela 13 – Métricas de desempenho da previsão para cada modelo ajustado para o número de óbitos registrados até 30 de setembro de 2020: coeficiente de determinação, erro médio absoluto e erro quadrático médio.

Modelo	EMA	EQM
logístico	181,4285	33874
Gompertz	25,9068	698
Richards	9,7615	160
ensemble	7,0135	75

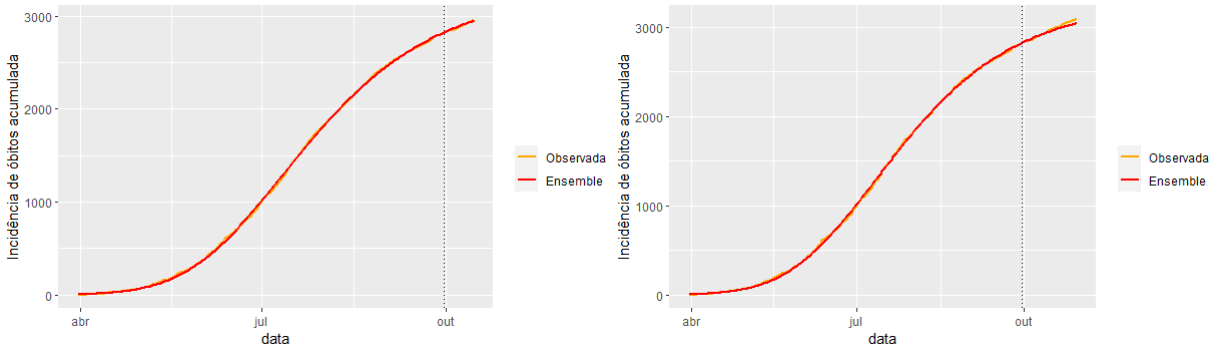
Fonte: A autoria (2022)

A figura 12 mostra a curva de óbitos simulada pelo modelo *ensemble* e a previsão para 15 e 30 dias à frente, respectivamente. O modelo se ajustou perfeitamente aos dados e, além disso, gerou previsões excelentes tanto para duas semanas quanto para um mês à frente. Dessa forma, observamos que ao utilizar os dados antes que a curva real voltasse a crescer, o modelo

apresentou melhor performance na predição e previsão dos dados.

Mostramos na figura 13, o intervalo de confiança gerado pelo modelo *ensemble* para o número de óbitos registrados até 30 de setembro e para a previsão de 30 dias à frente. O intervalo teve um ótimo desempenho e cobriu muito bem a curva de óbitos fornecida pelo ministério da saúde.

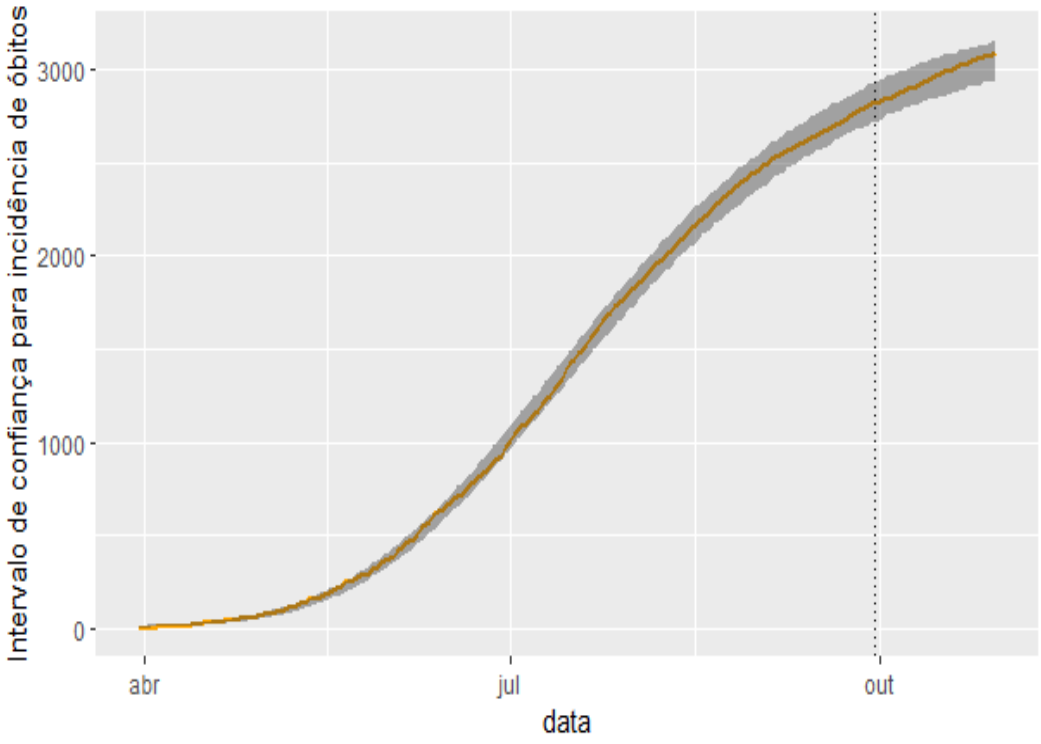
Figura 12 – Incidência de óbitos acumulada até 30 de setembro de 2020 versus curva ensemble.



(a) previsão de 15 dias à frente **(b) previsão de 30 dias à frente**

Fonte: A autoria (2022)

Figura 13 – Intervalo de confiança para o número de óbitos registrado até 30 de setembro de 2020 e para a previsão de 30 dias à frente.



Fonte: A autoria (2022)

O modelo *ensemble* foi ajustado também para o número acumulado de casos de COVID-19 na Paraíba. Os resultados se encontram no apêndice A. A tabela 14 mostra os parâmetros estimados pelo modelo *ensemble* e cada modelo de crescimento, seus respectivos erros padrão e o peso obtido de cada modelo. O modelo *ensemble* estimou que a taxa média de crescimento do número de casos de COVID-19 foi de 3,38%. As tabelas 15 e 16, mostram as métricas de desempenho de predição e previsão respectivamente dos modelos ajustados. Assim como no ajuste para o número de óbitos, o modelo Gompertz foi o modelo com menor erro quadrático médio e consequentemente com maior peso. Os resultados mostram bom desempenho do modelo Gompertz e *ensemble*, ambos com maior dificuldade de estimar bem o final da curva, provavelmente uma consequência da mudança repentina de inclinação da curva dos dados. O coeficiente de determinação R^2 mostra que os modelos Gompertz, Richards e *ensemble* se ajustaram muito bem aos dados. Os modelos Gompertz e ensemble tiveram melhor desempenho considerando o erro quadrático médio, erro médio absoluto e pontuação média do intervalo de confiança.

A figura 14 no apêndice A mostra as curvas simuladas pelos modelos de crescimento e a curva do número de infectados. A linha vertical tracejada indica o momento que começa a previsão (17 de dezembro de 2020). Os gráficos apontam que o modelo Gompertz e Richards obtiveram melhor desempenho de predição. O gráfico da figura 15 compara a curva simulada do modelo ensemble e curva de dados. Observamos que os modelos tem maior dificuldade para previsão do número de casos. Embora o modelo Gompertz reduza o erro de previsão, nesse caso é mais complicado, uma vez que em dezembro de 2020 já estava no início da segunda onda.

Observamos através do intervalo de confiança na figura 16 que apesar dos valores reais estarem próximos dos limites inferior e superior do intervalo, muitas observações estão fora do intervalo. Diferente dos números de óbitos por COVID-19, o número cumulativo de infecções por SARS-CoV-2 são muito maiores e a variação destes também é maior, uma vez que o registro de casos depende da manifestação de sintomas e realização de testes.

Assim como na curva de óbitos, a curva de casos começa a crescer em novembro de 2020, então os modelos de crescimento e *ensemble* foram ajustados para o número de casos registrados até o dia 15 de setembro para analisar o desempenho dos modelos nessas condições sendo gerada previsão de 15 dias à frente. As tabelas e gráficos estão disponíveis no apêndice A. Observamos que o desempenho de predição e previsão foram muitos melhores.

6 CONSIDERAÇÕES FINAIS

No início da pandemia, havia pouco conhecimento sobre a propagação do vírus SARS-CoV-2. Pensando nisso alguns estudos científicos tem sido realizados. Com a finalidade de estimar as taxas médias de infecção e de recuperação de COVID-19 na Paraíba no ano de 2020, foi ajustado o modelo compartimentado SIR para os dados registrados nesse período, porém o ajuste não obteve bom desempenho, visto que as taxas de infecção e recuperação de COVID-19 variam muito no decorrer do tempo.

O ano de 2020 foi dividido em 9 períodos, de acordo com método de Sturges e foi ajustado o modelo SIR para os números de infectados registrados em cada período. O modelo SIR conseguiu descrever bem a tendência da curva diária de novos casos de infectados. O número de reprodução básico foi maior nos dois primeiros períodos, diminuindo nos períodos seguintes e voltando a crescer próximo do final do ano.

Os dados de COVID-19 são subnotificados devido aos indivíduos assintomáticos ou a falta de testes de diagnóstico, o que impede o registro do real número de infectados. Por este motivo, foi realizada a pesquisa sorológica amostral no estado, com o objetivo de estimar o real número de infectados. A pesquisa foi realizada em 8 semanas, sendo que as análises eram feitas a cada duas semanas. Portanto, ajustamos novamente o modelo SIR para esses quatro períodos específicos entre novembro e dezembro de 2020, segundo as datas da pesquisa amostral. Observamos que os números estimados pela pesquisa são quase 5 vezes maiores que os registrados pelo MS e que os resultados estimados pelo modelo SIR também, dado que são ajustados com base em dados subnotificados. Percebemos também que o número de infectados inicial influência nos resultados estimados.

A subnotificação de casos tem sido um dos maiores desafios em estudos de COVID-19. Numa tentativa de aprimorar a análise dos dados, passamos a trabalhar com a curva acumulada dos números de óbitos, uma vez que as curvas acumuladas são mais estáveis que as diárias e que os números de óbitos não dependem do registro de infectados. Para isso utilizamos os modelos logístico, de Richards e Gompertz para ajustar os dados de março a dezembro de 2021, os quais foram utilizados na construção de um modelo de conjunto. O modelo logístico não ajustou tão bem os dados. Já os modelos de Richards e Gompertz tiveram melhor desempenho de predição Os modelos de Gompertz, Richards e de conjunto se ajustaram bem aos dados, mas não realizaram previsões muito precisas, pois em novembro de 2020 a curva apresenta uma inclinação que os modelos não conseguem captar.

Com isso construímos os modelo ensemble aos dados, o qual apresentou ótimo desempenho de predição, porém para previsão o resultado foi parecido com os obtidos pelos modelos de crescimento. Apesar da segunda onda de COVID-19 ter ocorrido em 2021, percebemos que no fim do ano de 2020 já havia um crescimento no número de casos. Entretanto, ajustamos os modelos para os dados registrados até 31 de setembro e realizamos previsões para 15 e 30 dias à frente. Os resultados mostraram que a abordagem de conjunto obteve um ótimo desempenho na predição dos dados e previsão de 15 e até 30 dias à frente. Portanto, o método de conjunto utilizado aqui se mostrou eficaz na predição e previsão de dados de COVID-19, considerando uma onda.

REFERÊNCIAS

ALVES, E. J. **Métodos de bootstrap e aplicações em problemas biológicos**. Dissertação de Mestrado — Instituto de Geociências e Ciências Exatas da Universidade Estadual Paulista, 2013.

ANASTASSOPOULOU, C.; RUSSO, L.; TSAKRIS, A.; SIETTOS, C. Data-based analysis, modelling and forecasting of the COVID-19 outbreak. **PLOS ONE**, v. 2, n. 2, 2020.

BARROS, A. M. R. de. Modelos matemáticos de equações diferenciais ordinárias aplicados à epidemiologia. **Revista de Ciências Exatas e Tecnologia**, v. 2, n. 2, p. 62–67, 2007.

BASSANEZI, R. C. **Ensino - aprendizagem com Modelagem matemática**. 3ª edição. ed. [S.l.]: Editora Contexto, 2011. ISBN 9788572442077.

BEZERRA, A. D. M.; FAVORETO, G. d. S.; QUEIROZ, K. F.; SILVA, M. d.; MORAES, W. d. Equaçoes diferenciais aplicadas ao modelo de malthus na dinâmica de crescimento da populaçao de bataguassu–ms. **Revista Conexao Eletrônica**, v. 13, n. 1, 2016.

BOYCE, W. E. **Elementary Differential Equations and Boundary Value Problems**. 7ª edição. ed. [S.l.]: John Wiley & Sons, 2000. ISBN 978-0471319993.

CABELLA, B. C. T. **Modelos aplicados ao crescimento e tratamento de tumores e à disseminação da dengue e tuberculose**. Tese de Doutorado — Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, 2012.

CASTANHO, M. J. de P. **Construção e avaliação de um modelo matemático para predizer a evolução do câncer de próstata e descrever seu crescimento utilizando a teoria dos conjuntos**. Tese de Doutorado — UNICAMP, 2005.

CHOWELL, G.; LUO, R. Ensemble bootstrap methodology for forecasting dynamic growth processes using differential equations: application to epidemic outbreaks. **BMC Med Res Methodol**, v. 21, n. 34, 2021.

COSTA, F.; SOUSA, Í. J. de; SANTOS, J. A.; SILVA, L. Modelo SIR aplicado na dinâmica da COVID-19 no Estado do Maranhão, Brasil. **Revista de Matemática**, v. 1, n. 1, p. 18–34, 2021.

DOMINGUES, J. S. Análise do modelo de gompertz no crescimento de tumores sólidos e inserção de um fator de tratamento. **Biomatemática**, v. 21, p. 103–12, 2011.

DUTRA, C. M. Previsões de máximo de casos confirmados e óbitos de COVID-19 no brasil. **Vigilância Sanitária em Debate: Sociedade, Ciência & Tecnologia (Health Surveillance under Debate: Society, Science & Technology)–Visa em Debate**, v. 9, n. 1, p. 12–17, 2021.

EFROM, B.; TIBSHIRANI, R. J. **An Introduction to the bootstrap**. 1ª edição. ed. [S.l.]: Chapman and Hall, New York, 1983. ISBN 978-0-412-04231-7.

FATORETTO, M. B.; SANTOS, A.; SAVIAN, T. V.; SPÓSITO, M. B. Modelos não lineares aplicados ao crescimento da doença mancha preta em citros. **60ª Rbras e 16º Seagro**, 2015.

FIGUEIREDO, D. G. de. **Equações Diferenciais Aplicadas**. [S.l.]: Instituto de Matemática Pura e Aplicada, 1997.

FIOCRUZ. **Qual a origem desse novo coronavírus?** 2020. Disponível em: <https://portal.fiocruz.br/pergunta/qual-origem-desse-novo-coronavirus>. Acesso em: 28 outubro 2021.

FRERY, A. C.; CRIBARI-NETO, F. **Elementos de Estatística Computacional Usando Plataformas de Software Livre/Gratuito**. [S.l.]: Copyright, 2011.

GALLASCH, C. H.; CUNHA, M. L. da; PEREIRA, L. A. de S.; SILVA-JUNIOR, J. S. Prevenção relacionada à exposição ocupacional do profissional de saúde no cenário de COVID-19 [Prevention related to the occupational exposure of health professionals workers in the COVID-19 scenario][Prevención relacionada cone la exposición ocupacional de profesionales de la salud en el escenario COVID-19]. **Revista Enfermagem UERJ**, v. 28, p. 49596, 2020.

GNEITING, T.; RAFTERY, A. E. Strictly proper scoring rules, prediction, and estimation. **Journal of the American Statistical Association**, Taylor & Francis, v. 102, n. 477, p. 359–378, 2007.

GOMES, S. C. P.; MONTEIRO, I. O.; ROCHA, C. R. Modelagem dinâmica da COVID-19 com aplicação a algumas cidades brasileiras. **Revista Thema**, v. 18, p. 1–25, 2020.

Governo do Brasil. **Brasil confirma primeiro caso do novo coronavírus**. 2020. Disponível em: <http://www.theregister.co.uk/2011/02/01/arm_holdings_q4_2010_numbers/>. Acesso em: 10 julho 2021.

HETHCOTE, H. W. The mathematics of infectious diseases. **SIAM Review**, v. 42, n. 4, p. 599–653, 2000.

HSIEH, Y.-H. Richards model: A simple procedure for real-time prediction of outbreak severity. **Modeling and Dynamics of Infectious Diseases Series in Contemporary Applied Mathematics (CAM)**, v. 11, 04 2009.

JILANI, T. N.; JAMIL, R. T.; SIDDIQUI, A. H. H1N1 Influenza. **Escola de Enfermagem de Ribeirão Preto**, 2020.

KAIO, M.; BARROS, N. Análise preditiva de casos confirmados de COVID-19 no brasil e em oito países baseada no modelo não linear de gompertz. **Scielo Peru**, v. 202, 2020.

KOTU, V.; DESHPANDE, B. **Predictive Analytics and Data Mining**. 1ª edição. ed. [S.l.]: Morgan Kaufmann, 2015. ISBN 9780128016503.

LARREMORE, D. B.; FOSDICK, B. K.; BUBAR, K. M.; ZHANG, S.; KISSLER, S. M.; METCALF, C. J. E.; BCKEE, C. O.; GRAD, Y. H. Estimating SARS-CoV-2 seroprevalence and epidemiological parameters with uncertainty from serological surveys. 2020.

LAUER, S. A.; GRANTZ, K. H.; BI, Q.; JONES, F. K.; ZHENG, Q.; MEREDITH, H. R.; AZMAN, A. S.; REICH LESSLER, J. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Esti-mation and application. **Ann Intern Med**, v. 82, 2020.

LEUNG, K.; WU, J. T.; LIU, D.; LEUNG, G. M. First-wave COVID-19 transmissibility and severity in china outside hubei after control measures, and second-wave scenario planning: a modelling impact assessment. **Lancet**, v. 395, p. 1382–1393, 2020.

LEVENBERG, K. A method for the solution of certain non-linear problems in least squares. **Quarterly of Applied Mathematics**, p. 164–168, 1944.

MAASSEN, J. P. The SIR and SEIR epidemiological models revisited. **Preprints**, 2020.

MARTCHEVA, M. **An Introduction to Mathematical Epidemiology**. [S.l.]: Springer, 2015. v. 61.

PALA, L. O. d. O. Revisitando a estimação de coeficiente de determinação. Universidade Federal de Alfenas, 2019.

PARAÍBA, G. do Estado da. **Plano Novo Normal PB**. 2020. Disponível em: <https://paraiba.pb.gov.br/diretas/saude/coronavirus/novonormalpb>. Acesso em: 10 outubro 2021.

PAZ, C. C. Paro de; PACKER, I. U.; FREITAS, A. R. d.; TAMBASCO-TALHARI, D.; REGITANO, L. C. d. A.; ALENCAR, M. M. d.; CRUZ, G. M. d. Ajuste de modelos não-lineares em estudos de associação entre polimorfismos genéticos e crescimento em bovinos de corte. **Revista Brasileira de Zootecnia**, SciELO Brasil, v. 33, p. 1416–1425, 2004.

PGE-PB. **Governo da Paraíba presta contas das medidas tomadas até agora durante a pandemia**. 2020. Disponível em: <https://pge.pb.gov.br/noticias/governo-da-paraiba-presta-contas-das-medidas-tomadas-ate-agora-durante-a-pandemia-1>. Acesso em: 10 junho 2021.

RIBEIRO, M. J. B. **Curva de crescimento de codornas ajustadas por modelos não lineares**. Dissertacç ao de Mestrado — Universidade Federal de Sergipe, 2014.

SANDES, S.; FREITAS, A. dos S. *et al.* Modelo sir com taxa de exposição para estudo da projeção de casos de COVID-19 no estado de sergipe. SciELO Preprints, 2020.

SANTOS, A. L. P. d. Métodos geradores de modelos de crescimento e decrescimento aplicados ás ciências agrárias. **Tese de Doutorado-Universidade Federal Rural de Pernambuco**, 2019.

SARMENTO, C. M. Paradoxos de uma pandemial paradoxes of a pandemic. **Political Observer| Revista Portuguesa de Ciência Política**, n. 14, 2020.

SARMENTO, J. L. R.; REGAZZI, A. J.; SOUSA, W. H. d.; TORRES, R. d. A.; BREDA, F. C.; MENEZES, G. R. d. O. Estudo da curva de crescimento de ovinos Santa Inês. **Revista Brasileira de Zootecnia**, SciELO Brasil, v. 35, p. 435–442, 2006.

SAÚDE, B. M. da Saúde. Secretaria de Vigilância em. **Boletim Epidemiológico Especial**. 2020. Disponível em: <https://antigo.saude.gov.br/images/pdf/2020/July/15/>. Acesso em: 10 agosto 2021.

SILVA, E. V. da; MELO, J. da S.; LEITE, M. A. Modelo bi-logístico aplicado aos primeiros 1015 casos de COVID-19 em indígenas do estado do amapá e norte do pará. **Science and Knowledge in Focus**, v. 3, n. 2, 2021.

SILVA, R. F. da. **Estudos por meio do modelo Epidemiológico SIR para os Casos de Infecção pelo COVID-19 no Paraná**. 2020. Disponível em: <https://www.researchgate.net/publication/344380496>. Acesso em: 10 junho 2021.

SILVA, R. M. da *et al.* Using the SIRD model to characterize the COVID-19 spreading in the states of Paraná, Rio Grande do Sul, and Santa Catarina. SciELO Preprints, 2020.

TSOULARIS, A.; WALLACE, J. Analysis of logistic growth models. **Mathematical biosciences**, v. 179, p. 21–55, 07 2002.

UNA-SUS. **Organização Mundial da Saúde declara pandemia do novo coronavírus**. 2020. Disponível em: <https://www.unasus.gov.br/noticia/organizacao-mundial-de-saude-declara-pandemia-de-coronavirus>. Acesso em: 20 agosto 2021.

UOL, P. **Com Nicolelis, Consórcio Nordeste cria Comitê Científico contra COVID-19**. 2020. Disponível em: <https://noticias.uol.com.br/colunas/chico-alves/2020/03/30/governadores-do-nordeste-criam-comite-cientifico-contra-o-coronavirus.htm>. Acesso em: 18 fevereiro 2022.

VASCONCELOS, G. L.; DUARTE-FILHO, G. C.; BRUM, A. A.; OSPINA, R.; ALMEIDA, F. A.; MACÊDO, A. M. *et al.* Análise de curvas epidêmicas da COVID-19 via modelos generalizados de crescimento: Estudo de caso para as cidades de recife e teresina. SciELO Preprints, 2020.

VENDRAMINI, S. H. F. **O tratamento supervisionado no controle da tuberculose em Ribeirão Preto sob a percepção do doente**. Dissertação de Mestrado — Escola de Enfermagem de Ribeirão Preto, 2001.

GLOSSÁRIO

C

Curvas cumulativas: curvas de crescimento de uma determinada população no decorrer do tempo.

D

Daniel Bernoulli: importante matemático, físico e professor suíço (1700-1782).

E

Ensembles: são combinações de diferentes modelos, criando um modelo único.

Epidemia: manifestação coletiva de uma doença que rapidamente se espalha em diversas regiões, estados ou cidades.

H

Hospital de Campanha: centros de assistência médica construídos durante emergências de saúde pública.

K

Kermack e McKendrik: importantes matemáticos que introduziram o primeiro modelo compartimentado.

N

Número de Reprodução Básico: número médio de infecções geradas por um único individuo infectado numa população totalmente suscetível .

Número de Reprodução Efetivo (NER): número médio de infecções geradas por um único individuo infectado em um determinado momento.

P

Paraíba: uma das 27 unidades federativas do Brasil, localizada na região nordeste.

Período de Incubação: intervalo entre a data de contato com o vírus até o início dos sintomas.

Pesquisa Sorológica: pesquisa de soro prevalência, na qual é necessário a realização de testes baseado na sorologia.

R

Ronald Ross: Bacteriologista britânico nascido em 1857, que descobriu o parasita da malária.

T

Taxa de Transmissão: serve como uma estimativa de como a doença se espalha entre a população.

Tumores Sólidos: formado pelo crescimento anormal de células de um tecido.

Z

Zoonótica: doenças que são transmitidas de animais para humanos, ou de humanos para os animais.

APÊNDICE A

Tabela 14 – Parâmetros estimados de cada modelo de crescimento e do modelo ensemble para o número de casos de COVID na Paraíba.

Modelo	K	E.P(K)	γ	E.P(γ)	α	E.P(α)	peso
logístico	157000	0,0807	0,0638	$1,7410x10^{-07}$	–	–	0,0177
Gompertz	157000	0,1920	0,0187	$5,7978x10^{-08}$	–	–	0,8507
Richards	156000	0,1981	0,1272	$7,1226x10^{-06}$	0,2000	$1,2581x10^{-05}$	0,1316
ensemble	156868	0,1908	0,0338	$9,8974x10^{-07}$	–	–	1

Fonte: A autoria (2022)

Tabela 15 – Métricas de desempenho da predição para cada modelo ajustado para o número de casos: coeficiente de determinação, erro médio absoluto, erro quadrático médio e pontuação média do intervalo

Modelo	R^2	EMA	EQM	PMI
logístico	0,8672	16439,8400	415689947	641000,90
Gompertz	0,9965	2846,6950	11092662	96847,57
Richards	0,9919	5909,6810	55978676	219579,10
ensemble	0,9928	3377,0000	16077781	117934,40

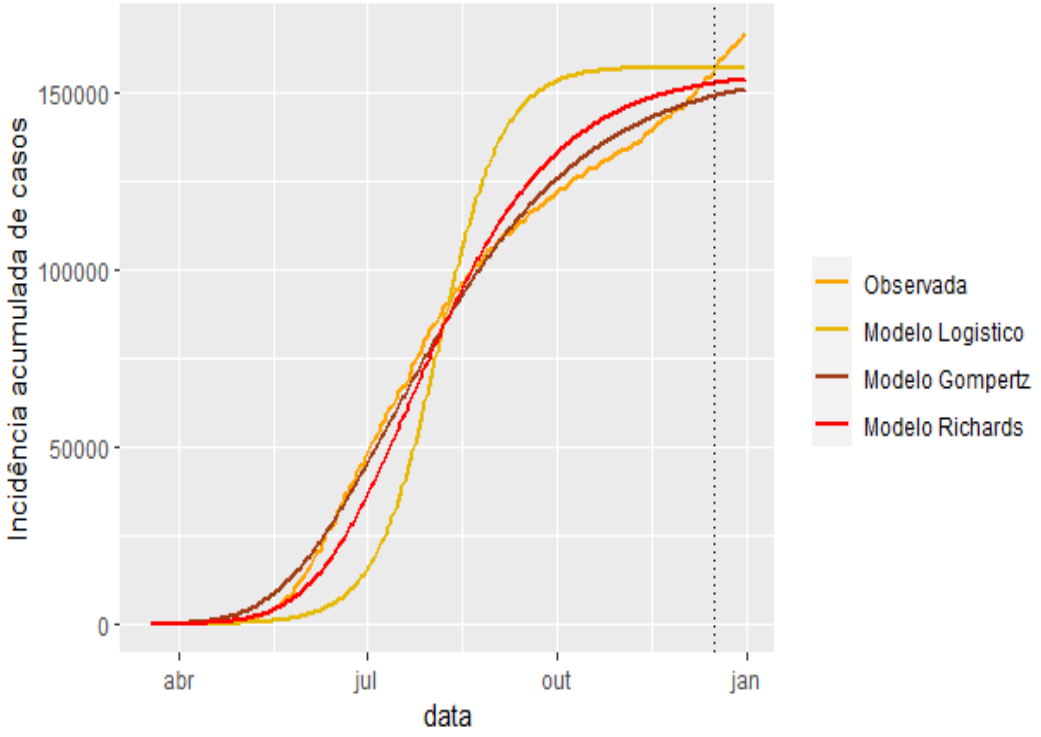
Fonte: A autoria (2022)

Tabela 16 – Métricas de desempenho da previsão para cada modelo ajustado para o número de casos: erro médio absoluto, erro quadrático médio e pontuação média do intervalo.

Modelo	EMA	EQM	PMI
logístico	4856,1240	30294665	165875,60
Gompertz	12337,7500	156308045	450719,90
Richards	8814,1330	82866898	323308,00
ensemble	11316,0000	132431578	423900,00

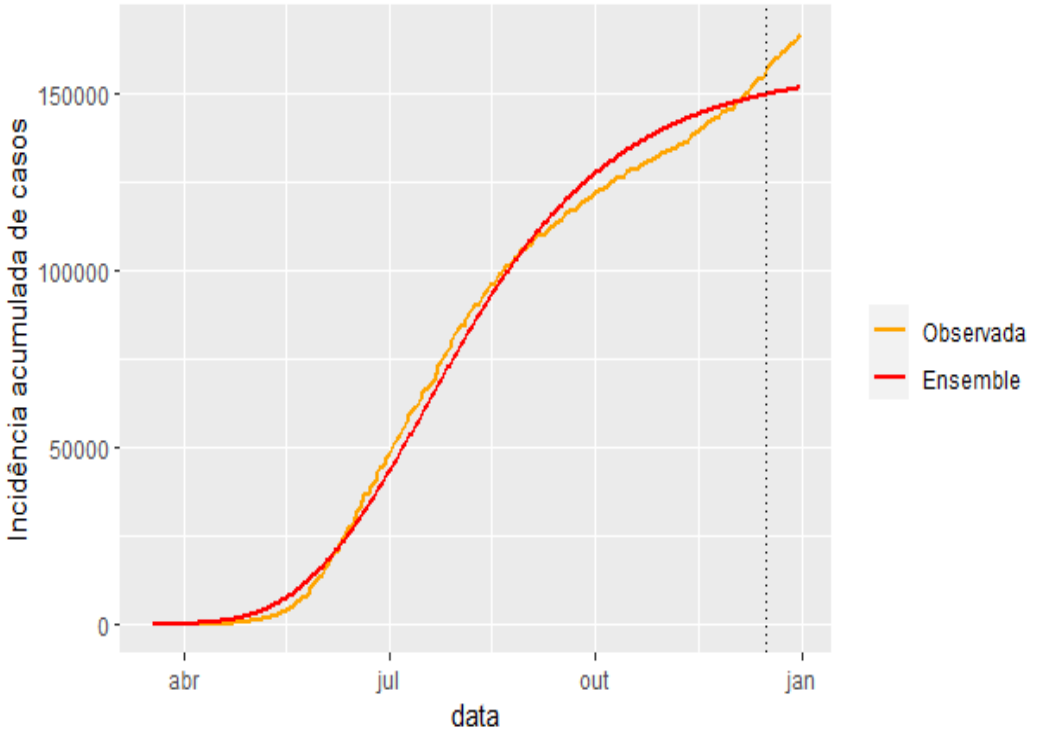
Fonte: A autoria (2022)

Figura 14 – Incidência acumulada observada do número de casos versus as respectivas curvas do modelo logístico, Gompertz e Richards.



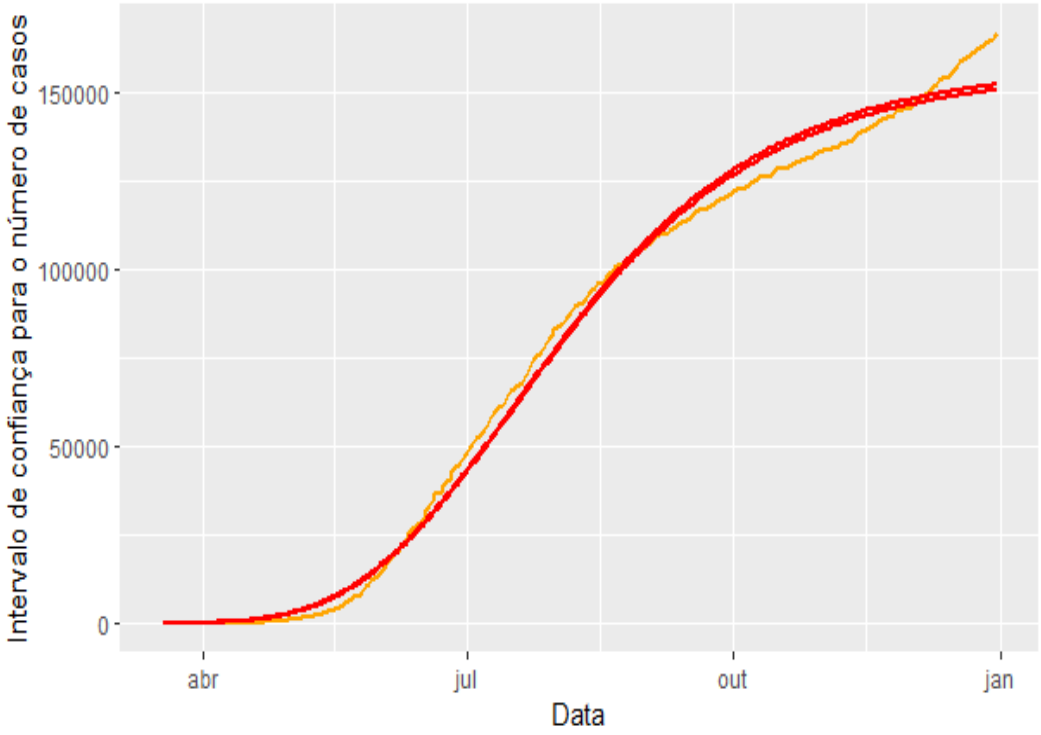
Fonte: A autoria (2022)

Figura 15 – Incidência acumulada observada do número de casos versus curva ensemble.



Fonte: A autoria (2022)

Figura 16 – Intervalo de confiança para a incidência acumulada de casos.



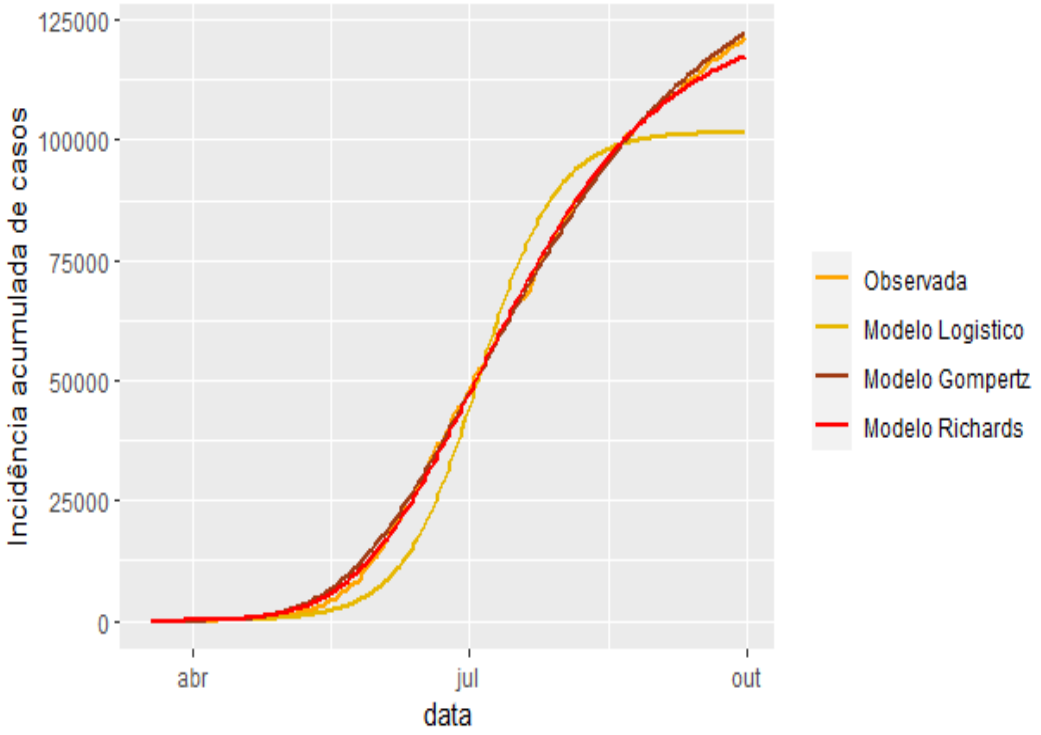
Fonte: A autoria (2022)

Tabela 17 – Parâmetros estimados de cada modelo de crescimento e do modelo ensemble para o número de casos de COVID na Paraíba registrados até o dia 15 de setembro.

Modelo	K	E.P(K)	γ	E.P(γ)	α	E.P(α)	peso
logístico	101664	0,0134	0,0780	$3,9937x10^{-07}$	–	–	0,0159
Gompertz	140398	0,0617	0,0225	$1,5158x10^{-07}$	–	–	0,3799
Richards	125936	0,7127	0,1495	$1,5833x10^{-05}$	0,2000	$2,4838x10^{-05}$	0,6041
ensemble	131043	0,4542	0,1001	$9,6151x10^{-04}$	–	–	1

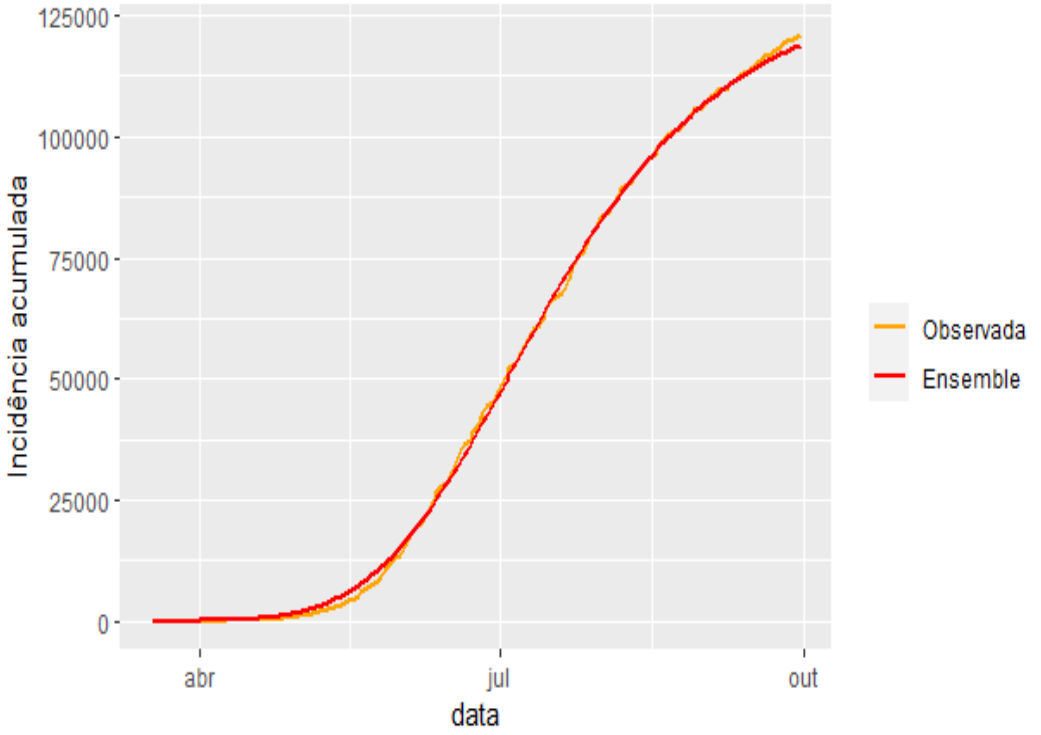
Fonte: A autoria (2022)

Figura 17 – Curva simulada pelo modelo ensemble para o número acumulado de óbitos registrado até o dia 15 de setembro.



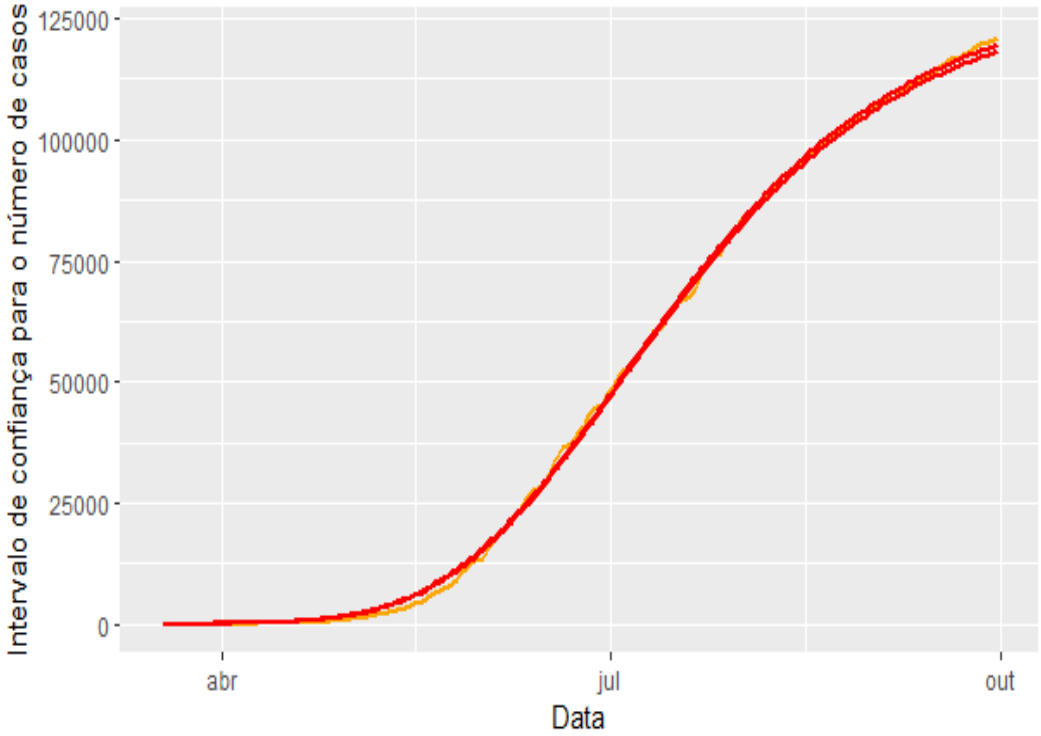
Fonte: A autoria (2022)

Figura 18 – Curva simulada pelo modelo ensemble para o número acumulado de óbitos registrado até o dia 15 de setembro.



Fonte: A autoria (2022)

Figura 19 – Intervalo de confiança para o número acumulado de óbitos registrado até o dia 15 de setembro e previsão de 15 dias á frente.



Fonte: A autoria (2022)