Lucas Monteiro Bianchi

O impacto das principais variantes do Sars-Cov-2 na epidemia da COVID-19 no Brasil

Rio de Janeiro

2023

Lucas Monteiro Bianchi

O impacto das principais variantes do Sars-Cov-2 na epidemia da COVID-19 no Brasil

Tese apresentada ao Programa de Pós-graduação em Epidemiologia em Saúde Pública da Escola Nacional de Saúde Pública Sergio Arouca, na Fundação Oswaldo Cruz, como requisito parcial para obtenção do título de Doutor em Ciências. Área de Concentração: Métodos Quantitativos em Epidemiologia.

Orientador: Prof. Dr. Leonardo Soares Bastos.

Rio de Janeiro

2023

Título do trabalho em inglês: The impact of the main variants of Sars-Cov-2 on the COVID-19 epidemic in Brazil.

O presente trabalho foi realizado com apoio de Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) - Código de Financiamento 001.

B577i Bianchi, Lucas Monteiro.
O impacto das principais variantes do Sars-Cov-2 na epidemia da COVID-19 no Brasil / Lucas Monteiro Bianchi. -- 2023.
125 f. : il.color, mapas.

Orientador: Leonardo Soares Bastos.
Tese (Doutorado em Epidemiologia em Saúde Pública) - Fundação Oswaldo Cruz, Escola Nacional de Saúde Pública Sérgio Arouca, Rio de Janeiro, 2023.
Bibliografia: f. 118-125.

1. COVID-19. 2. Conglomerados Espaço-Temporais. 3. SARS-CoV-2. 4. Estatística. 5. Variantes. I. Título.

CDD 616.2

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da Rede de Bibliotecas da Fiocruz com os dados fornecidos pelo(a) autor(a).
Bibliotecário responsável pela elaboração da ficha catalográfica: Cláudia Menezes Freitas - CRB-7-5348
Biblioteca de Saúde Pública

Lucas Monteiro Bianchi

O impacto das principais variantes do Sars-Cov-2 na epidemia da COVID-19 no Brasil

Tese apresentada ao Programa de Pós-graduação em Epidemiologia em Saúde Pública da Escola Nacional de Saúde Pública Sergio Arouca, na Fundação Oswaldo Cruz, como requisito parcial para obtenção do título de Doutor em Ciências. Área de Concentração: Métodos Quantitativos em Epidemiologia.

Aprovada em: 30 de agosto de 2023.

Banca Examinadora

Prof.ª Dra. Jessica Quintanilha Kubrusly
Universidade Federal do Fluminense

Prof. Dr. Rafael Izbicki
Universidade Federal de São Carlos

Prof.ª Dra. Claudia Torres Codeço
Fundação Oswaldo Cruz – Programa de Computação Cientifica

Prof.ª Dra. Aline Araujo Nobre
Fundação Oswaldo Cruz – Programa de Computação Cientifica

Prof. Dr. Leonardo Soares Bastos (Orientador)
Fundação Oswaldo Cruz – Programa de Computação Cientifica

Rio de Janeiro

2023

Aos meus pais *(in memorian)*, minha esposa, minha irmã e amigos que comigo caminharam.

AGRADECIMENTOS

Essas palavras de agradecimento, por mais que tentem expressar meus sentimentos, ainda não alcançam o real significado da gratidão que tenho por todos.

Aos meus pais, Loris e Lucia, é em meio as lágrimas e memórias, que eternalizo esse meu agradecimento a vocês. Sei que vibram com as minhas conquistas e que mesmo sabendo dos desafios e dificuldades que eu poderia enfrentar, nunca me impediram de lutar pelo que acreditei. Eu sinto muitas saudades de vocês.

A minha companheira de vida, Iolanda, obrigado por estar comigo, mesmo nas horas mais difíceis, compartilhando, não apenas seu tempo, mas o seu amor. Obrigado pelo "tem que dar linha na pipa". Te amo mil milhões ou 7.

A minha irmã, Loriane, pela amizade, carinho e compreensão. Partilhamos de muitas dores e alegrias que me fizeram ver a mulher que tem se tornado, o pai e mãe, com certeza, estão orgulhosos.

Ao meu orientador, Leo, que sempre muito humilde, tranquilo e paciente, transmitiu o seu conhecimento, apoiou minhas ideias e me guiou durante toda essa jornada na Fiocruz.

Aos amigos que a Fiocruz me proporcionou, Fernanda Garrides, Leandro, Débora, Iasmim, Fernanda Oliveira, Deyvyd, Rodrigo e Eduardo. Nós estudamos, almoçamos, reclamamos, lutamos e vencemos juntos.

Aos professores que fizeram parte da minha formação. Vocês foram pilares que edificaram o meu conhecimento, me tornando um profissional qualificado para contribuir com a saúde pública.

Aos meus amigos e parentes, obrigado pelo carinho, apoio e por fazerem parte da minha caminhada.

A ENSP e à FIOCRUZ, instituições ímpares para o Brasil, seja no impacto na educação e na saúde, sinto-me orgulhoso por ter sido formado por elas e por poder retribuir à sociedade os frutos do meu estudo.

A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela concessão da bolsa de doutorado, resistindo aos difíceis tempos de pandemia e aos cortes na educação.

Quem escreverá a história do que poderia ter sido o irreparável do meu passado; este é o cadáver. Se a certa altura eu tivesse me voltado para a esquerda, ao invés que para direita; Se em certo momento eu tivesse dito não, ao invés que sim; Se em certas conversas eu tivesse dito as frases que só hoje elaboro; Seria outro hoje, e talvez o universo inteiro seria insensivelmente levado a ser outro também.

(PESSOA, 1942, p. 77)

RESUMO

Reconhecida como ameaça global pela Organização Mundial de Saúde (OMS), a COVID-19 (Coronavirus Disease 2019) se espalhou por mais de 200 países, somando mais de 767 milhões de casos e quase 7 milhões de mortes confirmadas. O Brasil foi palco para diversas variantes resultando em mais de 37.6 milhões casos e 702 mil óbitos. O objetivo dessa tese foi estudar a dinamicidade da COVID-19 no contexto espaço-temporal, fazendo uso desse cenário para propor um método para estimação do efeito causal em estudos ecológicos. Assim, essa tese foi construída na forma de três artigos. O primeiro artigo identifica, por meio do emprego da metodologia Estatística Scan de Kulldorff, clusters espaço-temporais de casos e óbitos por SRAG-Covid e estima os seus respectivos riscos relativos considerando as variantes em circulação, com o propósito evidenciar a importância da compreensão da movimentação das variantes, principalmente no que tange o auxílio de tomada de decisões baseada em dados. O segundo artigo propõe uma metodologia de estimação do efeito causal em delineamentos ecológicos de séries temporais, baseado no modelo de respostas potenciais proposto originalmente por Rubin (1974) para ensaios clínicos randomizados e não-randomizados. Essa metodologia utiliza a rede neural convolucional LSTM (Long Short Term Memory) para identificação de padrões, estima o cenário contrafactual. O terceiro artigo apresenta 4 aplicações do método proposto no artigo 2. Objetivou-se comparar as curvas geradas para o cenário contrafactual com a curva observada no mesmo período a fim de obter a carga atribuída de casos e óbitos de SRAG-Covid em idosos a partir dos 60 anos em decorrência da emergência da variante Gama. Como resultado, no primeiro artigo foram identificadas áreas com riscos elevados, variando de 1,43 a 8,25 vezes o risco de casos e de 1,41 a 11,33 óbitos por SRAG-Covid. O segundo artigo demonstrou que a metodologia proposta produz estimativas plausíveis, sendo um meio viável para a estimação de efeitos causais em estudos ecológicos que utilizam dados dispostos ao longo do tempo. Finalizando, o terceiro artigo estimou que ocorreram aproximadamente 50 mil internações e mais de 30 mil óbitos em todo o Brasil apenas em pessoas acima dos 60 anos, sendo esses números atribuídos à variante Gama.

Palavras-chave: COVID-19; satscan; variantes; clusters; LSTM.

ABSTRACT

Recognized as a global threat by the World Health Organization (WHO), COVID-19 (Coronavirus Disease 2019) has spread to more than 200 countries, totaling over 767 million cases and nearly 7 million confirmed deaths. Brazil has been a stage for various variants resulting in over 37.6 million cases and 702 thousand deaths. The objective of this thesis was to study the dynamism of COVID-19 in the spatio-temporal context, using this scenario to propose a method for estimating causal effects in ecological studies. Thus, this thesis was structured in the form of three articles. The first article identifies, through the use of Kulldorff's Spatial Scan Statistical methodology, spatio-temporal clusters of cases and deaths from COVID-related Severe Acute Respiratory Syndrome (SARS-Covid) and estimates their respective relative risks considering the circulating variants, with the purpose of highlighting the importance of understanding the movement of variants, particularly in supporting data-driven decision making. The second article proposes a methodology for estimating causal effects in ecological time series designs based on the potential outcomes model originally proposed by Rubin (1974) for randomized and non-randomized clinical trials. This methodology utilizes the *Long Short Term Memory* (*LSTM*) convolutional neural network for pattern identification and estimation of the counterfactual scenario. The third article presents 4 applications of the method proposed in article 2. The objective was to compare the curves generated for the counterfactual scenario with the observed curve in the same period in order to obtain the burden of COVID-related Severe Acute Respiratory Syndrome (SARS-Covid) cases and deaths in individuals aged 60 and older due to the emergence of the Gamma variant. As a result, the first article identified areas with elevated risks, ranging from 1.43 to 8.25 times the risk of cases and from 1.41 to 11.33 deaths from COVID-related Severe Acute Respiratory Syndrome (SARS-Covid). The second article demonstrated that the proposed methodology produces plausible estimates, providing a viable means for estimating causal effects in ecological studies that use temporally arranged data. In conclusion, the third article estimated that there were approximately 50 thousand hospitalizations and over 30 thousand deaths across Brazil, all attributed to the Gamma variant, among individuals aged 60 and above.

Keywords: COVID-19; satscan; variants; clusters; LSTM.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

ACE2	*Angiotensin-Converting Enzyme* 2
CAMG	Campos Aleatórios Markovianos Gaussiano
COVID-19	*Coronavirus Disease* 2019
ECLIA	Imunoensaio de Eletroquimioluminescência
ELISA	Ensaio de Imunoabsorção Enzimática
Ensp	Escola Nacional de Saúde Pública Sérgio Arouca
FAC	Função de Autocorrelação
FACP	Função de Autocorrelação Parcial
FACV	Função de Autocovariância
Fiocruz	Fundação Oswaldo Cruz
FMV	*Formally Monitored Variants*
GISAID	*Global Initiative on Sharing All Influenza Data*
IBGE	Instituto Brasileiro de Geografia e Estatística
LSTM	*Long Short Term Memory*
LUEM	*Local Uncertainty Estimation Model*
MERS	Síndrome Respiratória do Oriente Médio
OMS	Organização Mundial de Saúde
PIB	Produto Interno Bruto
RBD	*Receptor Binding Domain*
RNA	Redes Neurais Artificiais
RNR	Rede Neural Recorrente
SARS	*Severe Acute Respiratory Syndrome*
SG	Síndrome Gripal
SRAG	Síndrome Respiratória Aguda Grave
VOC	*Variant of Concern*
VOI	*Variant of Interest*
VUM	*Variant Under Monitoring*

SUMÁRIO

APRESENTAÇÃO

A principal motivação para o desenvolvimento dessa tese se deu diante a alguns fatores: o enfrentamento mundial contra o vírus Sars-Cov-2, ao meu interesse por técnicas de modelagem estatística e por buscar conectar à epidemiologia à aprendizagem de máquina. O COVID-19 se tornou tema da minha tese após um dashboard desenvolvido por mim para acompanhamento do número de casos e óbitos no Brasil por COVID-19 ganhou notoriedade pelo jornal da Fiocruz.

Esta tese segue o formato de coletânea proposto pela Escola Nacional de Saúde Pública Sergio Arouca. O capítulo 1 apresenta brevemente a contextualização da pandemia no Mundo e no Brasil. No capítulo 2 é apresentado a revisão de literatura, o qual aborda com mais detalhes e etiologia do vírus Sars-Cov-2, sua transmissibilidade, principais variantes que circularam até dezembro de 2022, definição de caso, sintomas e população de risco. Além disso, esse capítulo contextualiza o pensamento contrafactual, abrangendo sua história e a ligação com a estatística e epidemiologia.

O capítulo 3 apresenta a justificativa de tese. O capítulo 4 apresenta os objetivos gerais e específicos e as considerações éticas. O capítulo 5 apresenta as metodologias utilizadas nos três artigos que compõe essa tese. Os capítulos 6, 7 e 8 correspondem aos três artigos formatados para publicação. Por fim, o capítulo 9 apresenta as considerações finais.

1 **INTRODUÇÃO**

Em 31 de dezembro de 2019, o escritório da Organização Mundial da Saúde (OMS) estava sendo informado sobre casos de uma pneumonia de etiologia desconhecida ocorria em Wuhan, província de Hubei na China, relatando principalmente sintomas como febre, dor, dificuldade em respirar.

Conforme o boletim epidemiológico 1 emitido pelo Ministério da Saúde (2020a), o governo chinês já descartava como possíveis hipóteses a influenza, influenza aviária, adenovírus, pneumonia atípica infecciosa (SARS) e Síndrome Respiratória do Oriente Médio (MERS) como patógenos, sendo identificado no dia 07 de janeiro o agente responsável pelas enfermidades, um novo coronavírus. O primeiro caso de infecção pelo novo coronavírus detectado fora da China ocorreu na Tailândia, no dia 13 de janeiro. Atualmente, o vírus se espalhou por mais 200 de países, resultando em um número acumulado de mais de 251 milhões de casos e quase 7 milhões de mortes (WHO, 2023).

O terceiro relatório sobre o coronavírus apresentado pela Imai *et al.* (2020) se referia a transmissibilidade do vírus, o qual sustentavam a hipótese de transmissão pessoa a pessoa devido a alto número de casos no dado espaço de tempo. Imai *et al*. (2020) estimou que cada caso infectado poderia infectar, em média, outras 2,6 (1,5 − 3,5) pessoas. O conhecimento sobre a capacidade e modo de transmissão do vírus evoluiu com o tempo, descobrindo que até mesmo pessoas que não apresentam sintomas aparentes, denominados casos assintomáticos, ou mesmo quando o vírus se encontra fase de incubação, podem propagar a doença.

Quanto a etiologia do vírus, o Sars-Cov-2, vírus responsável pela COVID-19, pertence à ordem *Nidovirales* e à família *Coronaviridae* e possui elevada homologia ao vírus causador da SARS em 2003, o SARS-CoV (Uzunian, 2020; Lima, 2020). O Sars-Cov-2 é um vírus de ácido ribonucleico (RNA), envelopado com aspecto de coroa, característica visual sob microscopia eletrônica, cujo material genético é representado por uma única molécula de RNA positivo (RNA+) (Uzunian, 2020). Vírus como o Sars-Cov-2 evoluem continuamente à medida que ocorrem alterações no código genético (causadas por mutações genéticas ou recombinação viral) durante a replicação do genoma (*Centers for Disease Control and Prevention*, 2023). O Sars-Cov-2 sofreu mutações ao longo da pandemia, resultando em variantes diferentes do vírus Sars-Cov-2 original. Em um estudo sobre a origem e a evolução do Sars-Cov-2, considerando os dados genômicos até o dia 28 de fevereiro de 2020, Tang *et al*. (2020) ao realizar análises genéticas populacionais de 103 genomas Sars-Cov-2

observou que esses vírus tinham duas linhagens principais (designadas L e S), sendo a linhagem L mais prevalente do que a S nas amostras examinadas.

No Brasil, a vigilância epidemiológica dos vírus respiratórios relevantes para a saúde pública é realizada por meio de uma Rede de Vigilância Sentinela de síndrome gripal (SG) e da Vigilância de síndrome respiratória aguda grave (SRAG), conjuntamente articulada com Laboratórios de Saúde Pública (Ministério da Saúde, 2022). Desde a pandemia de Influenza A(H1N1)pdm09 que a notificação de casos hospitalizados por síndrome respiratória aguda grave (SRAG) e óbitos suspeitos de SRAG, independentemente da internação, se tornou compulsória (Ministério da Saúde, 2020b; 2022. Os serviços de saúde que fazem parte da rede têm como objetivo identificar casos de Síndrome Gripal (SG), Síndrome Respiratória Aguda Grave (SRAG) hospitalizada e/ou óbitos por SRAG, a fim de analisar o perfil epidemiológico dos casos e o conhecimento dos vírus em circulação, com o propósito de estabelecer medidas de prevenção e controle (Ministério da Saúde, 2022). Todos os casos notificados são inseridos no Sistema de Vigilância Epidemiológica da Gripe (SIVEP-Gripe) (Corrêa *et al*., 2020). Em 2020, a vigilância da COVID-19 foi integrada ao SIVEP-Gripe (Ministério da Saúde, 2020b).

Apesar dos sintomas da COVID-19 serem geralmente leves e inespecíficos, como mal-estar, febre e tosse seca, a evolução para o estágio mais severo da doença se manifesta como uma síndrome sistêmica extrapulmonar da hiperinflamação (Siddiqi; Mehra, 2020). Nas primeiras semanas, a taxa de fatalidade de caso em Hubei (China) foi estimada em 18% (intervalo de confiança de 95%: 11% − 81%) (Dorigatti *et al*., 2020). Porcheddu *et al.* (2020) comparou a taxa de fatalidade por COVID-19 observada na Itália com a da China e constatou, que apesar ainda da diferença na magnitude dos números de casos, 888 na Itália e 44.000 na China, os resultados foram similares, 2, 3%. Em ambas localidades, outra característica comum foi notada, mortalidade maior entre os mais idosos do que entre os mais jovens (Dorigatti *et al*., 2020; Porcheddu *et al*., 2020).

Ficou evidente que o risco de morrer por COVID-19 aumenta com a idade, consequentemente, tem-se que a predominância dos óbitos ocorrem em idosos, principalmente aqueles com doenças crônicas (hammerschmidt; Santana, 2020). Além disso, a imunossenescência aumenta a vulnerabilidade às doenças infectocontagiosas, dessa forma, tornando desfavoráveis os prognósticos para aqueles com doenças crônicas (Barbosa *et al*., 2020). Conforme o último boletim epidemiológico de 2021 publicado pelo Ministério da Saúde (2021a), aproximadamente 75% dos óbitos por síndrome respiratória aguda grave (SRAG) ocorreram em pessoas com 60 anos ou mais.

Perez-Guzman *et al.* (2020) realizou um estudo de coorte retrospectiva em pacientes hospitalizados por COVID-19 no Reino Unido e identificou que ser idoso, do sexo masculino e

apresentar hipóxia na admissão, trombocitopenia, insuficiência renal, hipoalbuminemia e aumento da bilirrubina estão associados a maiores chances de óbito. Corroborando com esses achados estão os resultados obtidos por Niquini *et al.* (2020), que ao fazer uma descrição do perfil sociodemográfico Brasileiro para os casos de SRAG-Covid, forma grave da COVID-19 e observou-se uma elevada proporção de homens, idosos ou com 40 a 59 anos, com comorbidades (diabetes mellitus, doença cardiovascular, doença renal crônica e pneumopatias crônicas) e também de gestantes/puérperas.

A aparição de novas variantes, resultado dos processos de mutações genéticas, preocupam os cientistas por três principais motivos, maior transmissibilidade, pode gerar casos mais graves e sobre a própria resposta imune, ou seja, indivíduos que já tiveram contatos com outras variantes ou ainda já foram vacinados, podem não estar imunes às novas formas do vírus. O Brasil devido a sua extensão e heterogeneidade territorial vivenciou diferentes níveis da epidemia acontecendo simultaneamente, bem como a co-circulação de diferentes variantes da Sars-Cov-2. Até novembro de 2020, as variantes B.1.1.28 e B.1.1.33 foram predominantes. Por meio de mutações no domínio de ligação da proteína Spike (S) originaram outras duas novas variantes, a Gama e a Zeta, que também se espalharam rapidamente (Faria *et al*., 2021b).

De fevereiro de 2020 a dezembro de 2022, a variante Gama, foi a que resultou no maior pico de casos e óbitos. Essa variante foi inicialmente identificada no Japão e se espalhou para diversos países, o que fez com que fosse classificada como variante de preocupação (VOC). No Brasil, essa variante foi primeiramente detectada em turistas japoneses no final de 2020 (Naveca *et al*., 2021a; Faria *et al*., 2021a). A emergência da Gama ocorreu em Manaus, capital do Amazonas, e se espalhou para os demais estados Brasileiros, se tornando a variante mais prevalente a partir de fevereiro (Fiocruz, 2023). De acordo com a Vigilância genômica do Sars-Cov-2 no Brasil, dentre as 21 mutações definidoras de variante Gama, alterações no domínio de ligação ao receptor RBD e da glicoproteína Spike resultaram em uma proteína diferente o bastante para reduzir a capacidade de neutralização da infecção por anticorpos adquiridos em infecções prévias ou por vacinação, o que pode ter contribuído para a ocorrência de reinfecções (Fiocruz, 2023). A Gama possui 12 mutações na proteína S, que incluem três mutações de preocupação em comum com a variante Beta (B.1.351), ou seja, K417N / T, E484K e N501Y. (Naveca *et al*., 2021b).

A partir de julho de 2021, com a diminuição da prevalência de casos causados pela Gama, outra variante começou a crescer, a Delta. Essa variante foi primeiramente identificada no estado de Maharashta, na Índia e se tornou uma dominante em diversos países, como Reino Unido, Israel, Rússia e Estados Unidos (Cherian *et al*., 2021; Chmielewska *et al*., 2021). Embora essa variante

tenha sido responsável pelo aumento do risco de casos e óbitos em algumas regiões do Brasil, principalmente no Sul e Sudeste, o seu período de prevalência foi de queda das taxas de hospitalização em adultos, ocasionado pela efetividade da vacinação na redução da transmissão e na gravidade dos casos de COVID-19 (Freitas *et al.*, 2022).

Ainda no mesmo ano, a variante Delta foi sucedida pela Ômicron. Segundo Freitas *et al.* (2022), a emergência da Ômicron coincidiu com o período de festas, férias e com o relaxamento de medidas de restrição à mobilidade, o que rapidamente resultou na sua predominância. A Ômicron foi registrada pela primeira vez na África do Sul e classificada como variante de preocupação (Kannan; Ali; Sheeza, 2021). As principais mutações dessa variante se davam na proteína Spike, aumentando a capacidade de infecção e transmissibilidade (Arora *et al.*, 2022). No Brasil, (Freitas *et al*., 2022) aponta que a transmissão comunitária da variante Ômicron aumentou o número de casos alcançando o seu maior pico em janeiro de 2022.

Nogueira (2008) disserta sobre o aspecto da saúde humana, dizendo que ela não resulta somente de aspectos biológicos e dos serviços médicos, mas de todo um conjunto de fatores sociais, econômicos e culturais que, em interação, constituem e fundamentam cada lugar. Nesse sentindo, a individualidade de cada região e a interação exercida entre os municípios, caracterizando uma rede de fluxo de bens e pessoas, pode contribuir para a interiorização dos casos de COVID-19, visto que o espalhamento da doença é acelerado pelas interações humanas presenciais. Assim, essas interações podem resultar no aumento do risco em municípios que se favorecem dessas relações socioeconômicas.

Na epidemiologia, é primordial a conceituação e operacionalização de questões metodológicas direcionadas à explicação de efeitos causais, a fim de garantir conhecimento sobre o fenômeno estudado e quando possível, intervir (Czeresnia; Albuquerque, 1995a). Deste modo, as premissas da inferência causal são baseadas em diretrizes lógicas, possibilitando o uso de técnicas estatísticas que não apenas evidenciam associações, mas estimam, via modelos, efeitos causais (Czeresnia; Albuquerque, 1995b). Entretanto, principalmente seguindo o modelo de respostas potenciais, as estimativas contrafactuais são geralmente obtidas em ensaios clínicos randomizados e não-randomizados. Assim, ao basear-se no uso de metodologias conhecidas, essa tese apresenta uma abordagem para aplicação em estudos ecológicos de séries temporais. Visa-se assim, não apenas inovar na questão metodológica, mas também possibilitar a estimação da carga de hospitalizações e óbitos por SRAG-Covid em pessoas a partir dos 60 anos atribuídas à variante Gama. Assim, a tese colabora com o conhecimento sob a dinâmica da COVID-19 no Brasil, bem como fortalecer e fornecer bases para melhores tomadas de decisões.

2 REVISÃO DE LITERATURA

A partir de dezembro de 2019, o Sars-Cov-2 se disseminou rapidamente pelo mundo. Globalmente, a emergência do novo coronavírus se encontra em diferentes estágios dado à combinação singular de características que cada país apresenta, podendo acelerar ou desacelerar a taxa de transmissão de novos casos (Lima-Costa, 2020). Um dos principais desafios para o controle da pandemia da COVID-19 está no reconhecimento de uma pessoa infectada e na interrupção da rota de transmissão do Sars-Cov-2, pois mesmos casos assintomáticos ou pessoas com sintomas leves podem transmitir o vírus (Moraes *et al.*, 2020).

Billah, Miah e Khan (2020) realizou uma revisão sistemática sobre a estimação do número reprodutivo do Sars-Cov-2, isso é, quantas pessoas em média um indivíduo infectado contamina e identificou uma alta heterogeneidade entre os valores documentados. Por exemplo, Zhou *et al.* (2020a) estimou que o número reprodutivo basal variava de 2,83 a 3,34, enquanto Zhou *et al.* (2020b) obteve um valor de 5,32. De acordo com Medeiros (2020), um indivíduo infectado pelo novo coronavírus transmite, em média, para outras 2 ou 3 pessoas dependendo das condições ambientais. Outros autores, dizem que o número reprodutivo varia entre 1,4 e 3,9, dobrando o número de casos a cada 6,4 a 7,4 dias (Wu; Leung; Leung, 2020; li *et al*., 2020). De toda forma, independentemente do valor exato, nota-se que o Sars-Cov-2 é altamente transmissível e essa transmissão se dá usualmente por gotículas e contato, principalmente em locais fechados, com pouca ventilação, baixa luminosidade e ambientes hospitalares facilitam a transmissão do vírus.

O comportamento característico da curva epidemiológica para o número de casos de COVID-19 referente às primeiras semanas é um aumento muito rápido, seguido de uma gradual redução na velocidade até atingir o pico (Villela, 2020). De acordo com Keeling e Rohani (2011), o número de novos casos pode ser avaliado como um produto entre a taxa de transmissão e ao número de indivíduos suscetíveis. O Brasil, por ser um país de dimensão continental, tem diferentes momentos da pandemia acontecendo simultaneamente. Além disso, parte da dinâmica da disseminação do vírus é potencializada pelas desigualdades sociais. Uma parcela significativa da população Brasileira reside em habitações precárias, sem acesso regular ao abastecimento de água, tratamento de esgoto e em situação de aglomeração, fatores que propiciam a disseminação do vírus (Werneck; Carvalho, 2020).

Outro fator característico da pandemia é a presença de grupos de riscos, em outras palavras, têm-se que algumas pessoas estão mais propensas a serem infectadas e evoluírem para o estágio grave da doença. São considerados grupo de risco para agravamento da COVID-19 pessoas portadores de

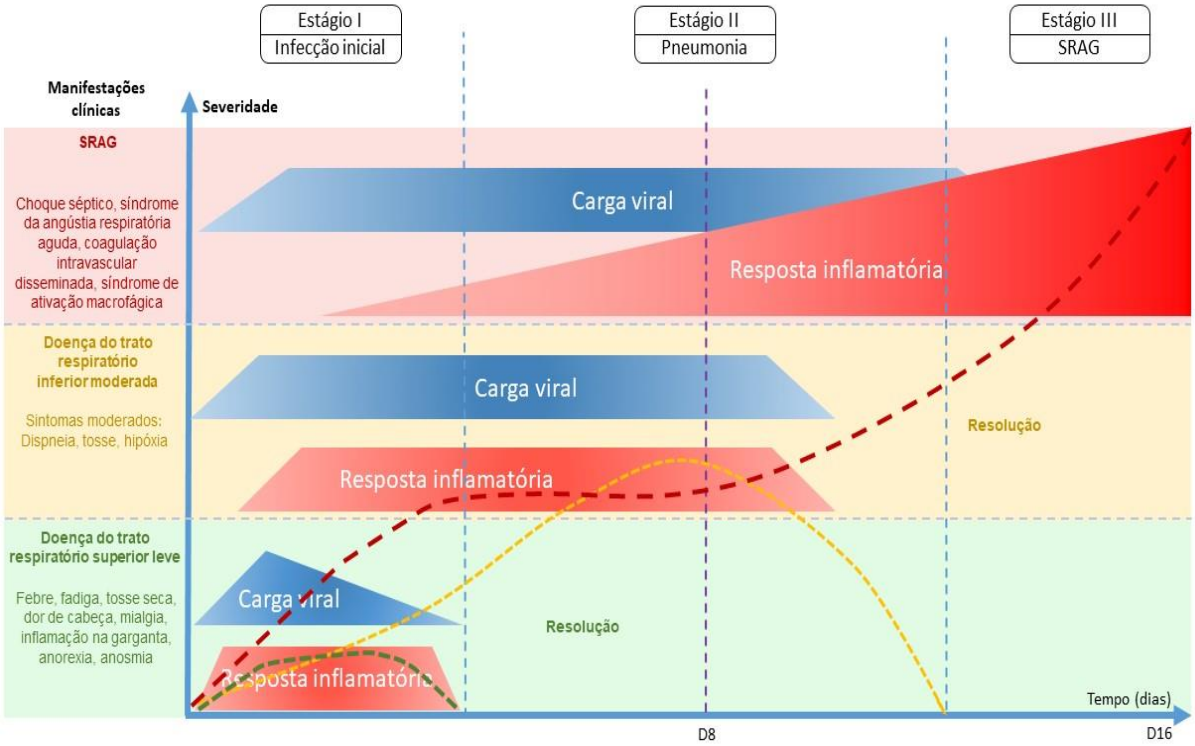
doenças crônicas, fumantes, gestantes, puérperas, crianças menores de 5 anos e pessoas acima de 60 anos (Ministério da Saúde, 2023). Esse último grupo, de maneira geral, costuma ser mais suscetível a diversas doenças infectocontagiosas, ao haver um enfraquecido do sistema imunológico ao longo da idade, o qual é denominado de imunossenescência. No Brasil, para o ano de 2020, quase um quarto dos óbitos por SRAG haviam ocorrido em pessoas com pelo menos 60 anos (Ministério da Saúde, 2021a).

Ao longo do decorrer da pandemia, uma das preocupações foi e ainda é o impacto das mutações e o que isso representaria no combate ao vírus. No início de março de 2020, em Wuhan, província chinesa de emergência do coronavírus, foi conduzido um estudo com 103 pacientes infectados, o qual indicou que o coronavírus havia se transformado em, pelo menos, duas novas cepas, uma mais agressiva e uma menos agressiva do que o coronavírus que estava se espalhando (Tang *et al*., 2020). Atualmente há diversas variantes circulando simultaneamente, responsáveis por reinfectar pessoas completamente imunizadas.

2.1 FASES DA COVID-19

Segundo Bordallo, Bellas e Cortez (2020), o vírus penetra no corpo por meio da inalação de partículas contaminantes, principalmente gotículas e aerossóis de hospedeiros infectados, alojando-se primeiro no trato respiratório superior e, em seguida, atingindo os pulmões. O Sars-Cov-2 usa ACE2 (Angiotensin-Converting Enzyme 2), uma proteína transmembrana presente em diversas células do corpo humano, para infectar células epiteliais da faringe, laringe, alvéolo (pneumócitos tipo II), macrófagos alveolares e células endoteliais, promovendo a vasodilatação e diminuição da pressão arterial (Hussain *et al*., 2020; Bordallo; Bellas; Cortez, 2020). A figura 1 ilustra os estágios da COVID-19, bem como sua manifestação por severidade.

Figura 1 – Espectro e severidade da COVID-19.



Fonte: Adaptado e traduzido de Bordallo; Bellas; Cortez (2020).

Siddiqi e Mehra (2020) foi um dos primeiros a descrever as fases patológicas da COVID-19, ao qual podem ser resumidas em:

- **Fase I ou de replicação viral:** os sintomas são leves e geralmente inespecíficos, como mal-estar, febre e tosse seca, podendo ser totalmente assintomático. Nos pacientes que conseguem manter o vírus limitado a este estágio, o prognóstico e a recuperação são excelentes.

- **Fase II ou inflamatória:** estão presentes na multiplicação viral e inflamação localizada no pulmão, os pacientes desenvolvem uma pneumonia viral, com tosse, febre. Inicialmente,na fase IIA os pacientes não apresentam hipóxia, apesar da inflamação pulmonar, passando a apresentá-la se evoluem à fase IIB.

- **Fase III ou de hiperinflamação sistêmica:** Uma minoria de pacientes com COVID-19fará a transição para o terceiro e mais grave estágio da doença e se manifesta como uma síndrome sistêmica extrapulmonar da hiperinflamação.

2.2 DEFINIÇÃO DE CASOS DE COVID E SRAG-COVID

Conforme as diretrizes do Ministério da Saúde (2021b), a definição de caso pode ser definida por síndrome gripal (SG) ou síndrome respiratória aguda grave (SRAG).

2.2.1 Síndrome gripal (SG)

Indivíduo com quadro respiratório agudo, caracterizado por, pelo menos, dois (2) dos seguintes sinais e sintomas: febre (mesmo que referida), calafrios, dor de garganta, dor de cabeça, tosse, coriza, distúrbios olfativos ou distúrbios gustativos.

Observações:

- Em crianças: além dos itens anteriores considera-se também obstrução nasal, na ausência de outro diagnóstico específico.
- Em idosos: deve-se considerar também critérios específicos de agravamento como sincope, confusão mental, sonolência excessiva, irritabilidade e inapetência.
- Na suspeita de COVID-19, a febre pode estar ausente e sintomas gastrointestinais (diarreia) podem estar presentes.

2.2.2 Síndrome respiratória aguda grave (SRAG):

Indivíduo com SG que apresente: dispneia/desconforto respiratório ou pressão persistente no tórax ou saturação de O2 menor que 95% em ar ambiente ou coloração azulada dos lábios ou rosto.

Observações:

- Em crianças: além dos itens anteriores, observar os batimentos de asa de nariz, cianose, tiragem intercostal, desidratação e inapetência;
- Para efeito de notificação no Sivep-Gripe, devem ser considerados os casos de SRAG hospitalizados ou os óbitos por SRAG independente de hospitalização.

O diagnóstico da COVID-19 pode ser realizado a partir de critérios como:

1. O diagnóstico clínico é realizado pelo médico atendente, que deve avaliar a possibilidade da doença, principalmente, em pacientes com a associação dos seguintes sinais e sintomas:

- Febre, que pode estar presente no momento do exame clínico ou referida pelo paciente (sensação febril) de ocorrência recente.
- Sintomas do trato respiratório (por exemplo, tosse, dispneia, coriza, dor de garganta)
- Outros sintomas consistentes incluindo, mialgias, distúrbios gastrointestinais (diarreia/náuseas/ vômitos), perda ou diminuição do olfato (anosmia), perda ou diminuição do paladar (ageusia).

2. O diagnóstico clínico-epidemiológico é realizado pelo médico atendente no qual se considera:

- Casos de paciente com a associação dos sinais e sintomas supracitados ou SRAG mais histórico de contato próximo, ou domiciliar, nos últimos 14 dias antes do aparecimento dos sintomas, com caso confirmado laboratorialmente para COVID-19 e para o qual não foi possível realizar a investigação laboratorial específica.

3. Diagnóstico clínico-imagem:

- Caso de sintomas respiratório mais febre ou SRAG ou óbito por SRAG que não foi possível confirmar ou descartar por critério laboratorial E que apresente alterações tomográficas.

4. Diagnóstico laboratorial - Caso o paciente apresente os sintomas respiratórios mais febre ou SRAG, o profissional de saúde poderá solicitar os seguintes exames laboratoriais:

- De biologia molecular, (RT-PCR em tempo real) que diagnostica tanto a COVID-19, a Influenza ou a presença de Vírus Sincicial Respiratório (VSR) normalmente até o oitavo dia de início de sintomas.
- Imunológico, que detecta, ou não, a presença de anticorpos em amostras coletadas a partir do oitavo dia de início dos sintomas. Sendo eles:
- Ensaio imunoenzimático (Enzyme-Linked Immunosorbent Assay - ELISA);
- Imunocromatografia (teste rápido) para detecção de anticorpos;
- Imunoensaio por Eletroquimioluminescência (ECLIA).

- Pesquisa de antígenos: resultado reagente para Sars-Cov-2 pelo método de Imuno-cromatografia para detecção de antígeno.

5. Diagnóstico laboratorial em indivíduo assintomático que realizou:

- Exame de Biologia Molecular com resultado DETECTÁVEL para Sars-Cov-2 realizado pelo método RT-PCR em tempo real.
- Exame de Imunológico com resultado REAGENTE para IgM e/ou IgA realizado pelos seguintes métodos: Ensaio imunoenzimático (ELISA) e Imunocromatografia (teste rápido) para detecção de anticorpos.

2.3 VARIANTES DO CORONAVÍRUS

O surgimento de novas variantes é um processo natural, resultado de modificações genéticas do vírus circulantes em uma população. Esse processo pode introduzir características que facilite a propagação do vírus, a reinfecção de pessoas que já tiveram contato com a doença ou ainda redução da sensibilidade de tratamentos e vacinas (Chen *et al*., 2021; Barcellos; Villela, 2021). É necessário fazer o rastreamento do processo evolutivo das novas variantes para ser possível desenvolver vacinas mais eficientes (Chen *et al*., 2021; Barcellos; Villela, 2021). Um exemplo bem-sucedido de rastreio de variantes é o Reino Unido, que durante o início de 2020, apesar de ter sido um dos países fortemente afetados pela COVID-19, se dedicou a obter uma amostragem genômica representativa do vírus circulante (Plessis *et al*., 2021). Como resultado dessa ação, foi constatado que antes do lockdown, devido ao elevado volume de viagens e poucas restrições às chegadas internacionais foram identificados mais de 1000 linhagens de transmissão, sendo esperado que situações semelhantes também fossem esperadas em outros países (Plessis *et al*., 2021). No Brasil, segundo Alcantara *et al.* (2022), devido aos esforços para a construção de um banco de dados genômico referente ao vírus Sars-Cov-2, ao final de fevereiro de 2022 haviam sido identificados 240 linhagens distintas.

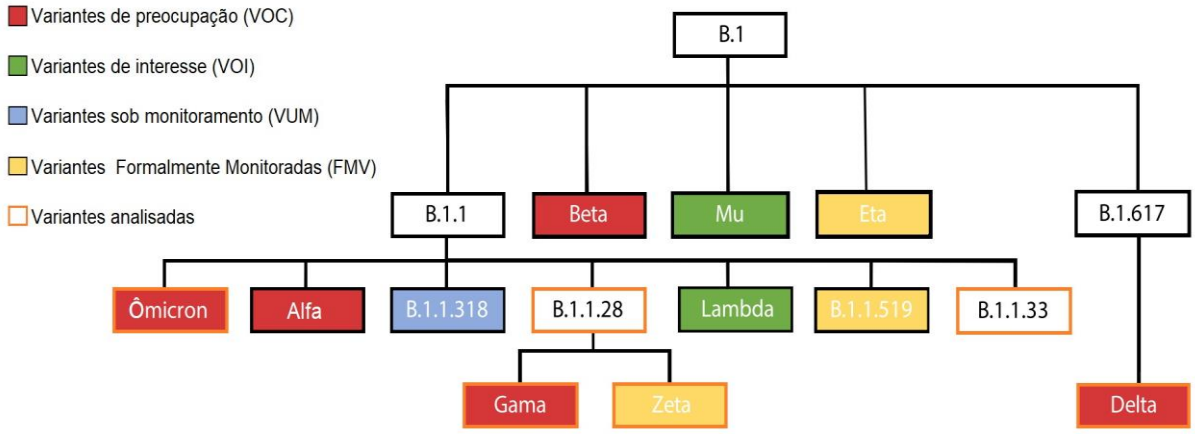
Com relação à introdução de novas variantes no Brasil, seis delas cocircularam, uma originária do Reino Unido (B.1.1.7), uma originária da África do Sul (B.1.351) e quatro que surgiram em diferentes regiões do país (Volz *et al.*, 2021; Tegally *et al*., 2021; Moreira *et al.*, 2021). Dessas, duas variantes se tornaram predominantes no Brasil, B.1.1.28 e B.1.1.33. Essas variantes apresentavam mutação E484K na ligação do RBD (do inglês *Receptor Binding Domain*) da proteína

S e foram associadas ao aumento da transmissibilidade, mudanças no perfil antigênico e reinfecção (Sabino *et al.,* 2021; Resende *et al.*, 2021b). Da variante B.1.1.28 originaram duas outras, denominadas como P.1 (Gama) e P.2 (Zeta), respectivamente, aliás da B.1.1.28.1 e da B.1.1.28.2 (Faria *et al*., 2021b). A P.1 (Gama) foi considerada variante de preocupação (*Variant of Concern - VOC)* devido à presença de múltiplas mutações na proteína S (incluindo K417T, E484K, N501Y), enquanto a P.2 (Zeta) foi considerada variante de interesse (Variant of Interest - VOI) por apresentar a mutação S:E484K (Resende *et al.*, 2021a).

Após a classificação da Ômicron pela OMS como variante de preocupação em 26 de novembro de 2021, ocorreu uma rápida disseminação a ponto de substituir a variante Delta que predominava até então (Zhou *et al.,* 2022). A Ômicron compartilha da mesma ancestralidade da B.1.1.28 e da B.1.1.33 identificada primeiramente na África do Sul em 14 de novembro de 2021 (Adamoski *et al*., 2022). Espenhain *et al.* (2021) constatou que a variante Ômicron causou uma elevada taxa de infecção em indivíduos já vacinados, mesmo naqueles que tomaram dose reforço e embora, não tenha levantado conclusões sobre a sua severidade comparada as demais variantes, ressalta a preocupação da variante levar a casos mais graves ou mortes devido a uma fuga da resposta imune. Corroborando com esses achados, Adamoski *et al.* (2022) e *Li et al.* (2023) apontam que o elevado nível de transmissibilidade se dá devido às numerosas mutações, incluindo o receptor RBD na proteína *Spike*, capazes até de subjugar a imunidade concedida pelas vacinas. Quanto a temporalidade da circulação das principais variantes identificadas no Brasil, têm-se que a B.1.1.28 circulou desde fevereiro de 2020, a P.2 (Zeta) foi detectada pela primeira vez no Rio de Janeiro e a P.1 (Gama) foi detectada pela primeira vez em viajantes japoneses retornando do Amazonas (Silva *et al*., 2021). Essa última variante, a P.1 (Gama), apareceu mais provavelmente em Manaus em meados de novembro (Silva *et al*., 2021). Entretanto, Silva *et al*. (2021) realizaram um estudo de sequenciamento do genoma completo o qual sugere que a origem da P.1 (Gama) ocorreu em meados de agosto de 2020. Com a atenuação da prevalência da Gama, outra variante começou a crescer, a Delta (B.1.617.2). Originalmente identificada na Índia em dezembro de 2020, foi apenas em maio de 2021 que as primeiras amostras foram identificadas no Brasil. A primeira amostra foi inicialmente identificada no Maranhão, seguido por Rio de Janeiro, Minas Gerais, Paraná e em Goiás (Menezes, 2021). Segundo Edara *et. al.* (2021), quando comparada à cepa original da Sars-Cov-2 (WA1/2020), a variante B.1.617 se demonstrou de 6 à 8 vezes mais resistentes aos soros de pessoas convalescentes e vacinadas com Pfizer e Moderna. Liu *et al.* (2021) destaca que os indivíduos infectados com as variantes B.1.351 (Beta) e P.1 (Gama) têm uma maior probabilidade de serem suscetíveis à

reinfecção pela variante Delta. A figura 2 apresenta o diagrama da evolução do vírus Sars-Cov-2 detectados no Brasil desde o início da epidemia até 19 de fevereiro de 2022 (Alcantara *et al*., 2022).

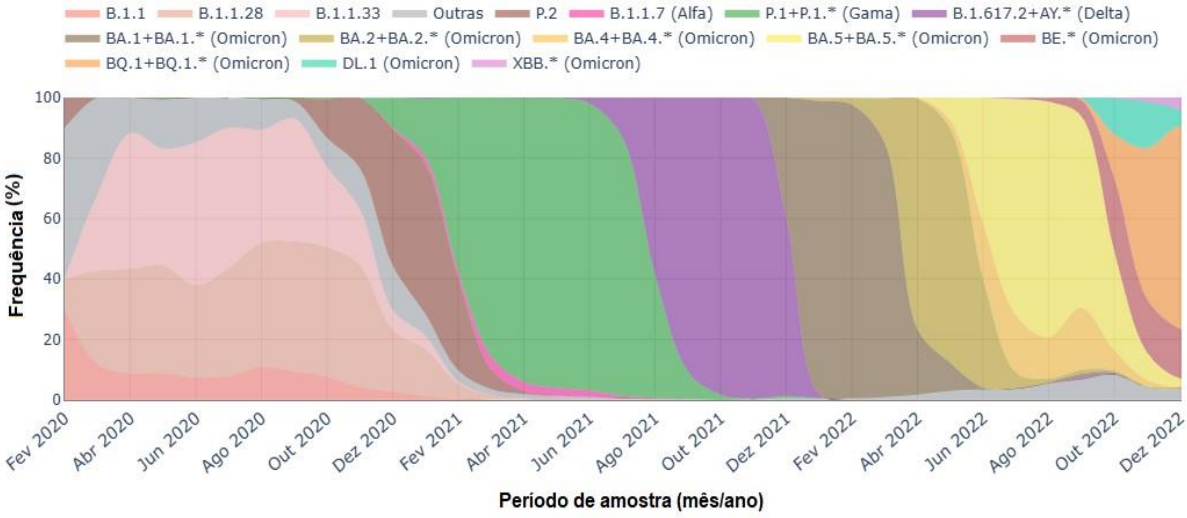
Figura 2 – Diagrama das variantes detectadas no Brasil



Fonte: Adaptado e traduzido de Alcantara *et al*., 2022.

A figura 3 apresenta as variantes relevantes no Brasil. É importante ressaltar que as frequências mostradas não são necessariamente representativas, podendo haver viés de seleção com a inclusão de investigação genômica de casos inusitados, rastreio de contactantes e seleção de amostras via protocolo de inferência de RT-PCR em tempo real para detecção de potenciais VOCs (Fiocruz, 2023).

Figura 3 – Variantes relevantes do vírus Sars-Cov-2.



Fonte: Fiocruz, 2023.

Por fim, COVID (2021) resume que as principais características das variantes mencionadas estão relacionadas ao aumento da capacidade de transmissão, resistência aumentada a anticorpos neutralizantes, sendo estas mais significativa nas variantes Beta (B.1.351), Gama (P.1) e Delta (B.1.617.2) e aumento do risco de reinfecção, principalmente para as variantes Gama (P.1) Zeta (P.2) e Delta (B.1.617.2). O aumento da capacidade de transmissão também foi observado na variante Ômicron, causando a “terceira onda" no Brasil (Alcantara *et al*., 2022).

2.4 COVID-19 E A POPULAÇÃO IDOSA

A pandemia causada pela emergência do vírus Sars-Cov-2 tem mostrado como os sistemas de saúde, mesmo em países desenvolvidos, tem enfrentado dificuldades em providenciar cuidados aos idosos (Mazumder; Hossain; Das, 2020). A taxa de óbito por COVID-19 em idosos é mais elevada quando comparado a indivíduos de outras faixas etárias. Na China, por exemplo, uma das primeiras nações severamente afetadas, a taxa de mortalidade chegou a 8% entre pessoas na faixa dos 70 a 79 anos, sendo ainda mais elevada para aqueles acima de 80 anos, 14,8% (Wu; Mcgoogan, 2020). Na Itália, padrões semelhantes foram observados, porém, em uma magnitude ainda maior, 32,4% para o grupo etário de 70 a 79 anos e 42,2% para o grupo de 80 a 89 anos (Remuzzi; Remuzzi, 2020). Porto *et al*. (2021) observou que durante as primeiras 6 semanas contadas a partir da confirmação do primeiro caso de COVID-19 no Brasil, cerca de 72% dos óbitos foram de pessoas com mais de 60 anos, embora 80% dos infectados não pertencessem a essa faixa etária. Porto *et al.*

(2021) aponta o aumento gradual da taxa de óbito a medida que se aumenta a idade, sendo, respectivamente, 0, 4%, 1, 3%, 3, 6% e 8% para as faixas etárias dos 40, 50, 60 e acima dos 70 anos. De acordo com o CDC (2020), isto ocorre devido a condições subjacentes, como hipertensão, diabetes, doenças cardiovasculares e respiratórias crônicas e câncer. No que se diz respeito a idosos que residem em instituições de longa permanência, em países da América do Norte, Ásia, Europa e Oceania, têm-se que entre 30% a 60% dos óbitos ocorreram nessa população (Moraes *et al.,* 2020). Cabrero (2020) faz uma reflexão sobre os outros fatores responsáveis pelo elevado número de óbitos em idosos na Espanha, apontando causas como negligência resultante da sobrecarga causada pela pandemia nos serviços de saúde, recursos humanos e materiais insuficientes, falta de controle e legislação específicas voltadas para os idosos que viviam em lares de idosos durante a pandemia, porém esses problemas não se limitam a esse país. Machado *et al.* (2020) estimou o impacto da COVID-19 na mortalidade de idosos institucionalizados no Brasil e estimou quase metade (44,7%) dos casos de COVID-19 ocorreriam entre idosos que vivem em lares de idosos. Ainda no Brasil, até 2 de janeiro de 2021, 73, 30% dos óbitos por SRAG haviam ocorrido entre pessoas com 60 anos ou mais (Ministério da Saúde, 2021a).

2.5 O PENSAMENTO CONTRAFACTUAL

Como sabemos que a ocorrência de um evento causou ou está associada a um determinado resultado? Esse tipo de questionamento não é novo. Remetendo-se primeiramente ao século a.C, onde Aristóteles baseia o conceito de causalidade em quatro pilares: a causa formal (o que o ser é, a essência); a causa material (do que o ser é feito, a aparência); a causa final (para onde o ser se dirige, finalidade); e a causa eficiente (a passagem da potência ao ato, isto é, aquilo que determina a expressão da essência na aparência) (Barata, 1997). Esse conceito de causalidade foi ganhando diferentes interpretações ao longo da história, porém foi no século XVII que se teve uma abordagem aplicada a um contexto epidemiológico. John Graunt estabeleceu relações de causas e consequências para detectar, ao longo do tempo, diferenças no número de óbitos entre sexo, cidade e campo (Barata, 1997). Hume (2013), um dos principais filósofos empiristas, contrapondo os pensamentos até então predominado pela corrente racionalista, definiu causalidade como:

> (...) um objeto seguido por outro, tal que todos os objetos similares ao primeiro são seguidos por objetos similares ao segundo. Em outras palavras, se o primeiro não tivesse ocorrido, o segundo nunca teria existido.

Respaldado pela última sentença, o filósofo Kim, Korman e Sosa (2011) propôs o pensamento contrafactual da causalidade, sintetizando a afirmação de Hume (2013) de modo que se a causa não tivesse acontecido, o evento também não aconteceria. O pensamento contrafactual pode ser entendido como uma condicional, atuando no isolamento de possíveis eventos candidatos para a ocorrência de um determinado evento de interesse. Apesar de Hume (2013) relacionar a noção de causa com probabilidade, outro filósofo, Mackie (1965), foi mais eficaz em demonstrar que a noção popular de causa tem, na verdade, um pressuposto probabilístico (Araújo *et al*., 2013). Mackie (1965) apresentava uma nova proposta ao qual se baseava na ocorrência de um conjunto de fatores para um específico desfecho acontecer, porém, também reconhecendo que o mesmo desfecho poderia ser causado por outro conjunto de fatores. Assim, Mackie (1965) apresenta que a chamada causa é uma parte insuficiente, mas necessária, de uma condição que por si mesma não é necessária, mas suficiente para o resultado. Em outras palavras, tem-se que X é uma causa de Y se X for uma condição insuficiente, embora seja parte não redundante de uma condição desnecessária, mas suficiente de Y. Williamson (2007) resume, em duas principais regras, como a abordagem contrafactual pode ser vista como condicional por meio da redução das relações de causas e efeitos à condicionais subjuntivas em que X é uma causa direta de Y se e só:

- Se X ocorre, Y ocorre (ou a sua probabilidade de ocorrência aumenta significativamente);
- Se X não ocorrer, Y não ocorre (ou a sua probabilidade de ocorrência diminui significativamente).

Apesar das diferentes abordagens contrafactuais, Durlauf e Blume (2016) destaca três principais problemas que dificultam a interpretação dos efeitos causais a partir da observação de dados empíricos, são eles: os problemas da simultaneidade, da equivalência observacional e da identificação. O primeiro problema, a simultaneidade, se refere ao fato de que ao menos uma variável explicativa no modelo de regressão linear múltipla é determinada em conjunto com a variável dependente, embora esta não exclua necessariamente a ordenação causal, ela complica a inferência (Durlauf; Blume, 2016; Wooldridge, 2006; Fonseca; Sánchez-Rivero, 2020). O segundo problema, a equivalência observacional, reflete a dificuldade de identificar a direção do efeito causal, visto que por mais ainda que as modelos possam ter uma direção causal natural, os dados em si não revelam qual é a direção causal correta (Fonseca; Sanchez-Rivero, 2020). Assim, um dos problemas enfrentados pelas teorias probabilísticas é o de justificar como a causa ocorre antes do efeito, já que essa definição geral descreve uma relação simétrica entre ambos os elementos. O terceiro problema

é estimar relações causais a partir de um número menor de observações empíricas (Fonseca; Sánchez-Rivero, 2020).

Em tempos mais recentes, outra abordagem contrafactual ganhou espaço, a abordagem da causalidade pela manipulabilidade, a qual as causas são expedientes que permitem manipular efeitos: X é uma causa de Y se e só se for possível manipular X de modo a induzir alterações em Y (Neuberg, 2003). Entretanto, a principal crítica à causalidade por meio da manipulação é que o ato de manipular (ou intervir) também é ser uma causa, criando assim um loop.

Finalmente, saindo das reflexões sobre as diferentes compreensões de causalidade contrafactual, podemos interpretar, em uma visão mais voltada para a estatística, que a inferência dos efeitos causais pode ser aplicadas a uma unidade em um contexto real desde que consideradas as suas complexidades, ou seja, estruturas de dependências (Rubin, 1974). Dessa forma, é importante definir as limitações estatísticas a cerca deste tema, ao qual Rubin (1974) expressa:

> Em qualquer situação relativamente complicada do mundo real, envolvendo plantas, animais, pessoas, aviões ou reatores nucleares, eu acredito que é geralmente impossível examinar uma resposta observada e realisticamente encontrar a causa dela. A única esperança para tal atribuição é descrever a situação muito cuidadosamente para limitar o tratamento (causa) sugerido que tenha ocorrido e indicar a) quais outros fatores (outros potenciais causas) estão sendo assumidos fixados em seus valores observados; e b) quais tratamentos (causas) alternativos contrafactuais estão sendo considerados terem acontecidos preferivelmente ao tratamento observado, que é a causa postulada.

Outro fator importante a ser considerado é que cada unidade observacional deve estar potencialmente exposta à ação de qualquer uma das causas cujo efeito poderia ser medido, pois assim, atributos pessoais imutáveis, tais como sexo ou raça, não podem ser vistos como causas, pois não se pode observar seu efeito sob a condição alternativa àquela que a unidade possui (Struchiner, 2002).

Holland (1986) propôs um modelo que considera a ideia de que um fator se torna uma causa se o resultado não tivesse ocorrido na ausência desse mesmo fator, assegurando que tudo tivesse permanecido constante, incluindo espaço e tempo, o qual denominou de modelo de respostas potenciais. Luiz e Struchiner (2002) descrevem a estrutura do modelo:

- Uma população de unidades, U;
- Um conjunto, K, de agentes causais bem definidos (também chamados tratamentos ou causas) para os quais cada unidade $u \in U$ possa ser exposta. Para efeito de simplificação, serão considerados apenas dois agentes causais, $K = \{t; c\}$, onde t representa tratamento ou exposição e c controle ou não exposição;

- Uma resposta, *Y*, a variável dependente, que pode ser registrada para cada unidade após a exposição aos agentes causais em *K*. *Y* será considerada dicotômica, enquanto em epidemiologia o principal interesse é a ocorrência ou não de uma determinada doença.

Para esse modelo proposto, a presença de temporalidade é necessária, pois a exposição da unidade observacional a uma causa deve ocorrer num intervalo de tempo específico (Struchiner, 2002). O papel de *Y* é medir o efeito de uma causa que, consequentemente, foi observado após a exposição, podendo o seu valor estar afetado pela causa *t* ou *c* ao qual foi exposto. Assim, tem-se Y_k em vez de *Y*, sendo *k* o indicador referente a exposição ao agente causal *t* ou *c*, associando-se o vetor (Y_t, Y_c) para cada unidade observacional $u \in U$, onde $Y_k(u)$ é a resposta obtida para a unidade *u* quando exposta à causa $k \in K$ (Struchiner, 2002). Dessa forma, tem-se que o modo mais básico para expressar a diferença do efeito causal de t e c sobre u quando medido por *Y* é dado por:

$$Y_t(u) - Y_c(u)$$

Holland (1986) ressalta que a impossibilidade de se observar simultaneamente $Y_t(u)$ e $Y_c(u)$ não é sinônimo de falta de informação.

3 JUSTIFICATIVA

Segundo a ONU (2020), a pandemia de COVID-19 é o maior desafio enfrentado no Mundo após a Segunda Guerra Mundial. Com mais de dois anos da emergência do vírus Sars-Cov-2, o agente causador da COVID-19, contabilizamos mais de 767 milhões de casos e quase 7 milhões de mortes e mais de 200 países atingidos (WHO, 2023). O coronavírus se dissemina muito rápido e apresenta alta taxa de hospitalização, causando grandes impactos econômicos e sociais. Sua transmissão pode ocorrer mesmo em pacientes assintomáticos, ou ainda quando o vírus está em fase de incubação, já em casos sintomáticos, a transmissibilidade é, em média, sete dias após o início dos sintomas (Adami; Imig; Ribas, 2020).

Embora a pandemia do coronavírus tenha sido um tema de estudo muito explorado, esta tese, além de considerar as principais variantes do Sars-Cov-2 por período e área de abrangência, o primeiro artigo intitulado "Movimentação espaço-temporal das variantes da COVID-19 nos números de casos e óbitos por SRAG-Covid no Brasil em 2020 a 2022", aplica uma metodologia para identificação de *clusters* espaço-temporais de municípios com alto risco de casos e óbitos por SRAG-Covid. Esse tipo de conhecimento se faz necessário, pois entender a dinamicidade da espacialização e temporalidade dos agravos de saúde são subsídios importantes para o processo de tomadas de decisões.

Taubes (1995) reflete sob os epidemiologistas se demonstrarem muitas vezes incapazes de construir modelos de explicação causais convincentes, gerando uma situação incômoda tanto na comunidade científica quanto também na política. Logo, o segundo artigo dessa tese, ao propor uma nova abordagem para estimação de efeitos causais por meio da análise de dados de séries temporais, faz isso, baseando-se em conceitos já validados, sendo um desses, o modelo de respostas potenciais de Rubin (1974). Essa nova metodologia visa preencher uma lacuna existente na lógica contrafactual, a aplicabilidade nos estudos ecológicos. Grande parte das políticas públicas são projetadas em nível de população, logo, se faz necessário desenvolver meios para que a estimação de efeitos causais possa ser realizada com plausibilidade, respeitando o histórico do evento de interesse. Além disso, a metodologia proposta ao utilizar a rede neural recorrente *LSTM*, amplia o leque de técnicas que podem ser empregadas no contexto de saúde pública, principalmente no que tange o emprego de métodos de aprendizagem de máquina.

Por fim, essa tesa encerra apresentando o último artigo, o qual reúne 8 aplicações dessa nova metodologia para estimação do efeito causal. Objetiva-se compreender o efeito da emergência de uma nova variante, especialmente nos idosos, um dos grupos de risco da COVID-19. Dessa forma, o terceiro artigo proposto "Modelo *LSTM* para estimação contrafactual da carga de hospitalizações e

óbitos por SRAG-Covid atribuível à variante Gama em pessoas a partir de 60 anos" sugere o que aconteceria "se a Gama fosse mais parecida com a outras variantes que a antecederam?" apresentando padrões alternativos e respondendo à pergunta, "quantas hospitalizações e óbitos teríamos a menos?". Dessa forma, esse estudo transita pelos campos das possibilidades e inova ao ser o primeiro artigo a aplicar uma metodologia de aprendizado de máquina, o modelo *LSTM*, ao pensamento contrafactual em um estudo de delineamento ecológico de séries temporais no contexto da pandemia de COVID-19. Assim, este estudo corrobora com o conceito de "causa" trazido por Rothman, Greenland e Lash (2016), que ao ressaltar a importância deste mecanismo para a discussão sobre causalidade, apresenta a ideia de efeito causal como algo relativo que deve ser compreendido dentre as alternativas concebíveis. Assim, sob a ótica da natureza probabilística que se encontra a epidemiologia moderna, uma das áreas de aplicação dos métodos quantitativos é a compreensão da dinâmica da distribuição dos indicadores de saúde. Desta forma, uma justificativa adicional quanto à relevância desse estudo, se faz por buscar compreender as tendências da transmissão da COVID-19 no Brasil. A COVID-19 continua sendo um problema de saúde pública para o mundo todo, inclusive para o Brasil. Sendo assim, essa tese se faz importante por reconhecer e explorar aspectos inerentes a esta pandemia, como a facilidade de transmissão do vírus, idosos como grupo de risco e como essas características se interagem com a emergência de uma nova variante.

4 OBJETIVOS

4.1 OBJETIVO GERAL

Investigar e quantificar medidas de impacto referentes a introdução de novas variantes da COVID-19 durante o período de fevereiro de 2020 a dezembro de 2022 e propor um método de estimação do efeito causal em estudos ecológicos de séries temporais.

4.2 OBJETIVOS ESPECÍFICOS E HIPÓTESES

Artigo 1: Movimentação espaço-temporal das variantes da COVID-19: Números de casos e óbitos por SRAG-Covid no Brasil em 2020 a 2022.

- Identificar a variante prevalente do vírus Sars-Cov-2 segundo o período e localidade da ocorrência dos *clusters*;
- Detectar a ocorrência *clusters* espaço-temporais de casos e óbitos por SRAG-Covid no Brasil;
- Estimar o risco relativo de se pertencer aos *clusters* de casos e óbitos de SRAG-Covid.

Hipótese: A co-circulação de diferentes variantes facilitaram a transmissão do vírus Sars-Cov-2, causando aumento nos riscos de casos e óbitos por SRAG-Covid em diferentes momentos e localidades.

Artigo 2: Cenários contrafactuais em estudos ecológicos de séries temporais utilizando a rede neural *LSTM*.

- Propor uma metodologia para a estimação do efeito causal em estudos ecológicos de séries temporais utilizando a rede neural *LSTM*
- Explicar a utilização dessa nova metodologia;
- Apresentar exemplo usando dados reais.

Hipótese: A rede *LSTM* se demonstra uma alternativa viável para a construção de cenários contrafactuais em estudos ecológicos de séries temporais.

Artigo 3: Modelo *LSTM* para estimação contrafactual da carga de hospitalizações e óbitos por SRAG-Covid atribuível à variante Gama em pessoas a partir de 60 anos.

- Aplicar a rede *LSTM* para estimação de cenários contrafactuais no qual a introdução variante Gama tivesse impacto similar às variantes já conhecidas;

- Mensurar o número de hospitalizações e óbitos no Brasil e nos estados de São Paulo, Rio de Janeiro e Amazonas devido à emergência da variante Gama.

Hipótese: Se a variante Gama tivesse um comportamento similar as variantes que a antecederam, o número esperado de hospitalizações e óbitos em pessoas acima dos 60 anos por SRAG-Covid seria menor.

4.3 CONSIDERAÇÕES ÉTICAS

A presente tese não necessitou de aprovação no Comitê de Ética em Pesquisa (CEP) da Escola Nacional de Saúde Pública Sergio Arouca (ENSP), devido aos dados utilizados estarem abertamente disponíveis na plataforma de dados abertos do Ministério da Saúde, *OpenDataSUS*.

5 MATERIAIS E MÉTODOS

5.1 DADOS

Visando identificar o perfil epidemiológico, compreender o processo de circulação do vírus e estabelecer medidas de prevenção e controle, a vigilância epidemiológica no Brasil é realizada por meio de uma Rede de Vigilância Sentinela de síndrome gripal (SG) e da Vigilância de síndrome respiratória aguda grave (SRAG) que atuam em conjunto com os Laboratórios de Saúde Pública. A notificação de casos hospitalizados por síndrome respiratória aguda grave (SRAG) e óbitos suspeitos de SRAG, independentemente da internação, se tornou compulsória desde a pandemia de Influenza em 2009 (Ministério da Saúde, 2020b; 2022). Em 2020, a vigilância da COVID-19 foi integrada ao SIVEP-Gripe, onde todos os casos notificados são inseridos (Corrêa *et al.*, 2020). Nessa tese foram utilizados dados referentes ao número diário de casos hospitalizados e óbitos em 2020 e 2022 por SRAG-COVID disponibilizados na plataforma *OpenDataSUS*, sendo cada um dos artigos utilizou um recorte temporal ou arranjo distinto. Para o primeiro artigo os dados são casos hospitalizados e óbitos diários por município para o período de 26/02/2020 a 31/12/2022. No segundo artigo os dados são número de casos hospitalizados diários para o período de 26/02/2020 a 31/07/2021 no Brasil. O terceiro artigo utilizou ao número de casos hospitalizados e óbitos de pessoas acima dos 60 anos para o período de 01/03/2020 a 31/07/2021 e tem como regiões de análise: Brasil e os estados de São Paulo, Rio de Janeiro e Amazonas. Para os três artigos foram utilizadas taxas de incidência por 100 mil pessoas. As informações referentes às populações estimadas para os anos de 2020 a 2022 foram obtidas pelo IBGE e as informações referentes às variantes predominantes pela Rede Genômica Fiocruz. O delineamento dos artigos apresentados é de caráter ecológico. Todas as análises serão feitas utilizando a linguagem R.4.2.2.

5.2 SÉRIES TEMPORAIS

Segundo Morettin e Toloi (2006), uma série temporal é qualquer conjunto de observações ordenadas no tempo. As séries podem ser discretas ou contínuas e a análise da série temporal pode ser feita no domínio do tempo ou no domínio das frequências, sendo os modelos propostos, paramétricos e não-paramétricos. Em ambos os domínios, esses modelos são controlados por leis probabilísticas, chamado de processo estocástico.

O objetivo da análise de séries temporais é investigar mecanismos geradores da série, fazer previsões, descrever o comportamento e procurar periodicidade relevante (Morettin; Toloi, 2006).

Um modelo clássico para séries temporais supõe que $\{Z_1, \ldots, Z_n\}$ pode ser escrita como:

$$Z_t = T_t + S_t + a_t = 1, 2, \ldots, n$$

em que Z_t representa a série temporal, T_t um componente de tendência, S_t um componente de sazonalidade e a_t é um componente aleatório. Uma das suposições mais frequentes que se faz a respeito de uma série temporal é que ela é estacionária, em outras palavras, ela se desenvolve no tempo aleatoriamente ao redor de uma média constante (Morettin; Toloi, 2006).

Para obter a estacionariedade, geralmente, é tomado as diferenças da série original, tendo a primeira diferença definida como:

$$\Delta Z_t = Z_t - Z_t - 1 \qquad t = 1, 2, \ldots, n.$$

Generalizando, tem-se que

$$\Delta^n Z_t = \Delta\{\Delta^{n-1} - Z_t\} \qquad t = 1, 2, \ldots, n.$$

Em geral, muitas vezes uma ou duas diferenças são suficientes para que as séries se tornem. Muitas séries temporais exibem um comportamento que tende a se repetir a cada intervalo de tempo, ou seja, apresentam a componente de sazonalidade, assim faz-se necessário estimar S_t e subtrair a série estimada de Z_t. As sazonalidades podem ser classificadas em dois tipos:

- Aditiva: A série apresenta flutuações sazonais mais ou menos constantes não importando o nível global da série.
- Multiplicativa: O tamanho das flutuações sazonais varia dependendo do nível global da série.

Conforme Morettin e Toloi (2006), empiricamente considera-se como sazonais os fenômenos que ocorrem regularmente no período máximo de 12 meses. Toda periodicidade acima do período de 12 meses é considerada ciclo. Retirando-se o efeito do ciclo, a série perde muitas observações, reduzindo seu tamanho, o que prejudica e dificulta o ajuste do modelo. Pode-se afirmar que as séries sazonais apresentam alta correlação em lags sazonais. Os procedimentos mais comuns para estimar a sazonalidade são o método de regressão (sazonalidade determinística) e o método de médias móveis (sazonalidade estocástica).

5.2.1 Função de autocovariância e autocorrelação

Conforme Morettin e Toloi (2006), a função de autocovariância (FACV) é dada por $\Gamma_\tau = E\{Z_t Z_{t+\tau}\}$, em que Z_t é um processo estacionário real discreto de média zero e τ representa a defasagem no tempo. O estimador da função de autocovariância Γ é dado por

$$c_\tau = \frac{1}{N}\sum_{t=1}^{N-\tau}(Z_t - \bar{Z})(Z_{t+\tau} - \bar{Z})$$

em que $\tau = 1, 2, ..., N-1, N$ é o número de observações da série e Z é a média amostral $\bar{z} = \frac{1}{N}\sum_{i=1}^{N} Z_i$. A função da autocorrelação (FAC) é dada por:

$$\rho_\tau = \frac{\Gamma_\tau}{\Gamma_0}$$

em que $\tau \in Z$ e Γ_0 a função de autocovariância sem defasagem. O estimador da função de autocorrelação ρ_τ é dada por:

$$r_\tau = \frac{c_\tau}{c_0}$$

sendo c_τ a função de autocovariância na defasagem τ e c_0 a variância. A função de autocorrelação é importante para conhecer a relação entre as observações atuais e as anteriores. A autocorrelação entre as séries z_t e z_{t-1} (autocorrelação com lag 1) indicará como os valores da série estão relacionados com seus valores imediatamente precedentes, enquanto a autocorrelação entre z_t e z_{t-2} (autocorrelação com lag 2) fornecerá uma relação dos valores da série z_t com aqueles atrasados em dois intervalos de tempo. Desta forma, a generalização da autocorrelação com τ atrasos para uma série temporal com n elementos é dada pela expressão abaixo:

$$r_\tau = \frac{\sum_{t=1}^{n-\tau}(Z_t - \bar{Z})(Z_{t+\tau} - \bar{Z})}{\sum_{t=1}^{n}(Z_t - \bar{Z})^2}$$

5.2.2 Função de autocorrelação parcial

A ideia de autocorrelação pode ser estendida de modo que se medirmos a correlação entre duas observações seriais, z_t e $z_{t+\tau}$, eliminando a dependência dos termos intermediários, $z_{t+1}, z_{t+2}, \ldots z_{t+\tau-1}$, temos o que se denomina autocorrelação parcial, representada por:

$$Cov(z_t, z_{t+\tau} | z_{t-1}, \ldots, z_{t-(\tau+1)})$$

O coeficiente de correlação parcial é utilizado para medir o grau de associação entre as observações z_t e $z_{t+\tau}$, quando os efeitos das defasagens até $\tau - 1$ são fixadas e são geralmente apresentados em um correlograma. O correlograma é um gráfico com os τ primeiros coeficientes de autocorrelação como função de τ e é um instrumento importante para identificar características da série temporal. Pode-se pensar, de forma meramente ilustrativa, que a autocorrelação parcial (FACP) pode ser definida como a contribuição da correlação em uma determinada defasagem dada a ausência dos coeficientes das demais defasagens. A autocorrelação parcial de um processo é definida como a sequência dos $\phi'_{\tau\tau}s$ obtidos pela resolução sucessiva das equações de Yule-Walker para τ = 1, 2, 3, ..., expressas na forma matricial por:

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{-2} \\ \vdots & & & & \vdots \\ \rho_{\tau-1} & \rho_{\tau-2} & \rho_{\tau-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \phi_{\tau 1} \\ \phi_{\tau 2} \\ \cdots \\ \phi_{\tau\tau} \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \tau \\ \rho_\tau \end{bmatrix}$$

em que $\phi_{\tau j}$ é o *j*-ésimo coeficiente do modelo AR(τ) e $\phi_{\tau\tau}$ é o último coeficiente.

5.3 ESTATÍSTICA ESPACIAL E ESPAÇO-TEMPORAL

A Estatística espacial pode ser definida como uma coleção de técnicas que visam descrever padrões existentes nos quais os dados estão espacialmente distribuídos, considerando o arranjo espacial na análise e nas interpretações dos resultados (Bailey; Gatrell, 1995). De acordo com Blangiardo *et al.* (2013), os dados podem ser descritos como um processo estocástico indexado no espaço:

$$Y(s) = \{y(s), s \in \mathcal{D}\}$$

onde $\mathcal{D}$ é um subconjunto fixo de $\mathbb{R}^d$ para d igual a 2. Os dados podem ser representados como uma coleção de observações $y = \{y(s_1), \dots, y(s_n)\}$ onde o conjunto $(s_1, \dots, s_n)$ corresponde a localização espacial da variável observada. Dependendo de $\mathbb{R}^d$, distribuição espacial da variável resposta pode ser dada de modo contínuo ou discreto.

O conceito de processo espacial pode ser expandido para o espaço-temporal caso seja incluída a dimensão tempo. Segundo Blangiardo *et al.* (2013), os dados podem ser descritos como um processo:

$$Y(s,t) = \{y(s,t), s \in \mathcal{D} \in \mathbb{R}^2 \times \mathbb{R}\}$$

e são observados em *S* áreas (municípios) e em *T* tempos. Quando se interessa identificar o padrão espacial e temporal de doenças, os modelos espaço-temporais são amplamente utilizados (Abellan; Richardson; Best, 2008; Lawson, 2013).

A dependência espacial é considerada em termos de uma estrutura de vizinhança. Simplificando a notação anteriormente apresentada, ao invés de escrevermos $(s_1, \dots, s_n)$, passaremos a escrever $(1, \dots, n)$, assim temos um município i, seus vizinhos $\mathcal{N}_{(i)}$ são os que fazem fronteira com o município *i* (vizinhos de 1ª ordem) ou os que fazem fronteiras com os vizinhos de primeira ordem (vizinhos de 2ª ordem).

5.4 ESTATÍSTICA SCAN DE KULLDORFF

A detecção dos *clusters* espaços-temporais é feita através da estatística *Scan* de *Kulldorff* por meio do *software SaTScan*^{TM} v10.1. Essa metodologia permite a detecção de *clusters* espaço-temporais utilizando a distribuição de probabilidade Poisson; testa a significância estatística e a corrige para múltiplos testes; examina a dinâmica da doença em um intervalo de tempo contínuo; e estima o risco relativo (RR) para cada *cluster* considerando o tamanho populacional daquele *cluster* (Kulldorff, 1999). A estatística Scan espaço-temporal utiliza-se de "cilindros" que se movimentam ao longo do tempo e espaço para a identificação de *clusters* com altos riscos através da comparação do número de eventos observados com o esperado dentro do cilindro. A base do cilindro representa a área geográfica enquanto a altura representa o tempo e são independentes um do outra (Kulldorff, 1999). Assim, o número de eventos esperados contidos nessas janelas cilíndricas assume a distribuição de Poisson:

$$\left(\frac{C}{E[c]}\right)^{c}\left(\frac{C-c}{C-E[c]}\right)^{C-c}$$

onde *C* é o total de eventos, *c* é o número de eventos dentro da janela cilíndrica e *E[c]* é o número eventos esperados dentro da janela cilíndrica. No modelo de Poisson, o número esperado de eventos em cada município é proporcional à sua população. Além disso, a probabilidade de um evento ocorrer em um município m assumindo uma data *d* é a mesma para os demais dias (Kulldorff, 1999).

Foi considerado como áreas vizinhas, os municípios nos quais os seus centroides pertençam à base do cilindro. A hipótese nula é que o risco dentro do cilindro é igual ao risco fora do cilindro. Assim, para cada cilindro, o número de casos *(E[c])* é igual ao total de eventos no município *(C)* dividido pela população *(P)*, vezes a população dentro do cilindro p (Kulldorff, 1999):

$$E[c] = \frac{C}{P} \times p$$

Os *clusters* detectados são ordenados pelo log da razão da verossimilhança (LLR) expresso pela fórmula abaixo (Kulldorff, 2022)

$$LLR = \left(\frac{C}{E[c]}\right)^{c} \left(\frac{C-c}{C-E[c]}\right)^{C-c} I()$$

onde *c* é o número de casos dentro do *cluster* e *I()* é uma função indicadora que é igual a 1 quando o cilindro tem mais casos do que o esperado e 0, caso contrário. Assim os *clusters* com maiores valores para a (LLR) são os mais verosímeis. O risco relativo (RR) para cada *cluster* é calculado pela seguinte expressão (Kulldorff, 2022):

$$RR = \frac{c/E[c]}{(C-c)/(C-E[c])}$$

Os valores-p dos *clusters* detectados foram calculados usando simulação de Monte Carlo com 999 replicações, combinando três diferentes métodos, são eles:

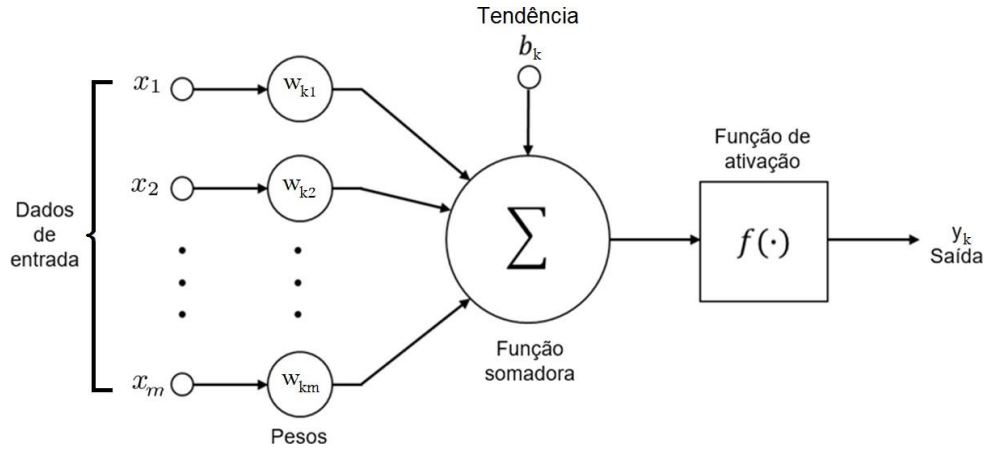
- Monte Carlo (Padrão): Proposto por Dwass (1957), a estatística de teste é calculada para cada simulação, bem como para o conjunto de dados reais, se o resultado estiver entre os 5% mais altos, por exemplo, o teste é significativo no nível 0,05.
- Monte Carlo (Sequencial): Os cálculos terminam assim que o número de simulações tiver uma razão de verossimilhança maior do que a razão de verossimilhança do conjunto de dados reais. Se essa condição nunca for atingida, os cálculos continuarão até que o número máximo de simulações (999). De toda forma, não há perda de poder ao nível de 5% de significância ao comparar o teste sequencial com o teste padrão de Monte Carlo (Kulldorff, 2022).
- Aproximação de Gumbel: Nessa abordagem não há limite inferior nos valores-p ($valor - p > 0$). Os métodos funcionam gerando simulações dos dados sob a hipótese nula. A razão de verossimilhança máxima de cada simulação é usada para ajustar uma distribuição de Gumbel pelo método de estimativa de momentos. Uma vez obtida a distribuição de Gumbel que melhor se ajusta aos dados, calcula-se o valor-p como a probabilidade dessa distribuição gerar um valor maior que a razão de verossimilhança máxima observada para o *cluster* mais provável do conjunto de dados reais.

O nível de significância adotado para a detecção dos *clusters* foi de 5% ($valor - p < 0{,}05$)., porém foram considerados apenas *clusters* significativos à 1% ($valor - p < 0{,}01$). Os municípios não podem pertencer à diferentes *clusters* dentro do mesmo período de análise, ou seja, não há intersecção espacial. Devido à magnitude da área de estudo (Brasil), o tamanho máximo permitido para população de cada *clusters* foi igual a 5% da população total, isso são 10.665.882 indivíduos, fator esse que limita também o tamanho do raio da janela espacial (círculo). O período de duração de cada *cluster* pode variar entre 7 à 14 dias e deve conter pelo menos 120 casos (óbitos).

5.5 REDES NEURAIS ARTIFICIAIS (RNA)

O primeiro modelo de rede neural artificial (RNA) foi proposto por Rosenblatt (1958), o *perceptron*. Essa rede consistia somente em uma camada e usava o modelo McCulloch-Pitts.

Figura 4 – Modelo de neurônio artificial de McCulloch e Pitts



Fonte: Fernandes, 1998.

O neurônio Figura 4 matematicamente expresso por:

$$u_k = \sum_{j=1}^{m} w_{kj} x_j,$$

e

$$y_k = \rho(u_k + b)$$

em que $x_1, x_2, \ldots x_m$ são as entradas, $w_1, w_2, \ldots w_m$ os pesos sinápticos correspondentes do neurônio *k*, u_k é a combinação linear das entradas com os pesos sinápticos, b_k é a tendência, $\rho(.)$ é função de ativação e e y_k é a saída do neurônio. A função de ativação visa limitar o espaço de variação da saída do modelo. Algumas principais etapas da implementação de uma rede neural são:

- Pré-processamento: Há diferentes formas de se trabalhar com redes neurais, embora funcionem melhor com dados normalizados ou padronizados, por ajudarem a prevenir os problemas de gradientes (Vallbo, 2019).
- Treino e validação: Para evitar *overfitting*, o conjunto de treinamento é dividido em duas partes, conjunto de treino e validação, sendo o desempenho obtido no conjunto de validação utilizada

para encerrar a etapa de treinamento quando o erro para de cair (Frau *et. al.*, 2021). As proporções mais utilizadas para a divisão dos dados em conjunto de treino e validação são 60:40, 70:30 e 80:20 (Santos *et al*., 2019).

- Teste: Os dados que fazem parte do conjunto de testes são informações não vistas durante a etapa de treinamento/validação do modelo. Dessa forma, espera-se que o modelo apresente boa capacidade de generalização para novas observações (Santos *et al*., 2019).

Quanto aos tipos de aprendizagem de máquina, os principais são:

- Supervisionado: esse método se dá pela comparação da saída atual por meio da rede neural com a saída desejada, os pesos têm os seus valores iniciais aleatorizados que são ajustados através do algoritmo de aprendizagem na próxima iteração. Esse ajuste depende do valor esperado e do sinal atual de saída (Anderson; Mcneill, 1992). Esse processo reduz os erros e modifica os pesos sinápticos até alcançar uma determinada precisão, sendo necessário verificar o desempenho da rede, observando se as saídas produzidas se devem de fato ao processo de generalização. (Anderson; Mcneill, 1992).
- Não-supervisionado: esse procedimento se caracteriza por não oferecer a rede neural a saída desejada, sendo o ajuste dos pesos sinápticos dado pelo reconhecimento de padrões de entrada que vão se adaptando conforme a necessidade da rede mediante um próprio monitoramento interno (Anderson; Mcneill,1992).
- Semi-supervisionado: essa metodologia, como o nome sugere, incorpora características tanto do aprendizado supervisionado como do não-supervisionado. Na aprendizagem-supervisionada os dados são parcialmente rotulados, de modo que grande parte seja não-rotulado (Géron, 2019).
- Por reforço: nesse caso o sistema de aprendizado é denominado por agente. O agente pode observar o ambiente, selecionar e executar ações e obter recompensas ou penalidades, aprendendo por si própria a melhor estratégia, visando maximizar o número de recompensas ao longo do tempo (Géron, 2019).

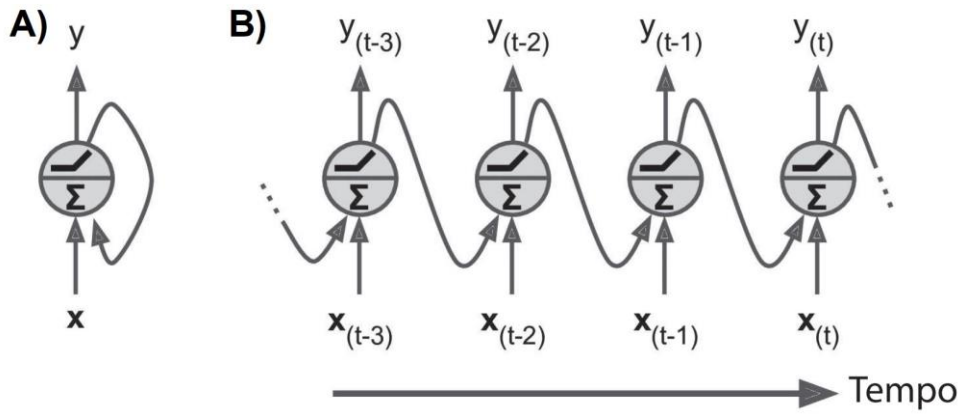
Além disso, outra característica importante é a taxa de aprendizagem, por exercer forte influência no treinamento da rede neural. Uma taxa de aprendizado muito baixa consumirá muito mais tempo, será menos sensível na presença de novos dados, porém pode produzir saídas mais adequadas, enquanto uma taxa de aprendizagem muito alta, a rede se adaptará rapidamente aos novos dados, embora tenderá a esquecer os antigos (Anderson; Mcneill, 1992; Géron, 2019).

5.5.1 Redes neurais recorrentes (RNR)

O neurônio apresentado na figura 5 é do tipo *feedforward*, ou seja, fluem em apenas uma direção, iniciando da camada de entrada dos dados até a saída da rede. As redes neurais recorrentes (RNR) se diferenciam por conter conexões com neurônios de trás.

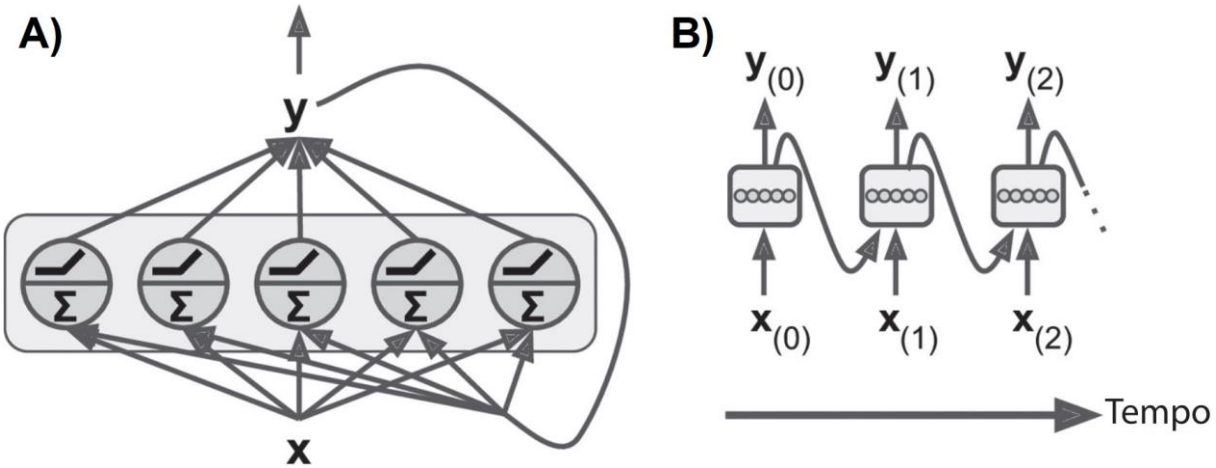
A figura 5 ilustra a conexão entre os neurônios onde a cada intervalo de tempo *t*, o neurônio recebe as entradas $\mathbf{x}_t$ e a própria saída no tempo anterior $y_{(t-1)}$ enquanto a figura 6 apresenta o mesmo neurônio desenrolado no tempo (Géron, 2019).

Figura 5 – Neurônio recorrente (A), desenrolado através do tempo (B)



Fonte: Géron, 2019.

Figura 6 – Camada de neurônios recorrentes (A) desenrolados através do tempo (B)



Fonte: Géron,2019.

Cada neurônio recorrente possui pesos específicos para as informações vindas da camada de entrada e para as camadas de saída do tempo anterior $y_{(t-1)}$, respectivamente, $\mathbf{w}_{(x)}$ *e* $\mathbf{w}_{(y)}$. Assim, podemos escrever a equação de uma camada recorrente como:

$$y_t = \phi\left(W_x^T \cdot x_t + W_y^T \cdot y_{(t-1)} + \mathrm{b}\right)$$

em que $\mathbf{W}_{(x)}$ *e* $\mathbf{W}_{(y)}$ são matrizes de peso correspondentes a formadas pelos vetores $\mathbf{w}_{(x)}$ *e* $\mathbf{w}_{(y)}$, *b* é o viés e ϕ a função de ativação.

Conforme Géron (2019), a equação 6.9 descreve o cálculo da saída de batch ("minilote" de dados) uma camada recorrente considerando todas as entradas no intervalo de tempo em uma matriz de entrada $\mathbf{X}_{(t)}$

$$y_t = \phi\left(\mathrm{X}_{(t)} \cdot \mathbf{W}_{(x)} + Y_{(t-1)} \cdot \mathbf{W}_y + \mathrm{b}\right) = \phi\left(\left[\mathbf{W}_{(x)} + Y_{(t-1)}\right] \cdot \mathbf{W}_y + \mathrm{b}\right) \text{ com } W = \left|\frac{\mathbf{W}_{(x)}}{\mathbf{W}_{(y)}}\right.$$

sendo

- $\mathrm{X}_{(t)}$ é uma matriz $m \times n_{entradas}$ que contém as entradas para todas as instâncias (sendo $n_{entradas}$é o número de características de entrada);
- $\mathrm{Y}_{(t)}$é uma matriz $m \times n_{neurônios}$ que contém as saídas de um intervalo de tempo *t* para cada instância no *batch* (sendo *m* o número de instâncias no batch e $n_{neurônios}$ é o número de neurônios);
- $\mathbf{W}_{(x)}$ é uma matriz $n_{entradas} \times n_{neurônios}$ contendo os pesos das conexões para as entradas no tempo atual.
- $\mathbf{W}_{(y)}$ é uma matriz $n_{neurônios} \times n_{neurônios}$ contendo os pesos das conexões para as saídas no tempo anterior.
- *b* é um vetor com tamanho n neurônio e contém os vieses do neurônio
- As matrizes de peso $\mathbf{W}_{(x)}$ e $\mathbf{W}_{(y)}$ são geralmente concatenadas verticalmente.

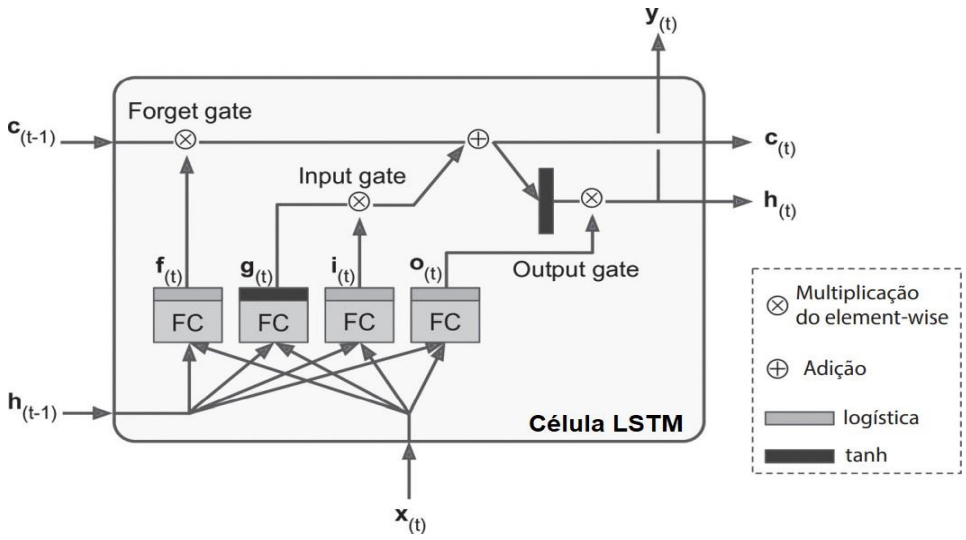
A notação $\left[\mathrm{X}_{(t)}, Y_{(t-1)}\right]$representa a concatenação horizontal das matrizes $\mathbf{W}_{(x)}$ e $\mathbf{W}_{(y)}$

Para t = 0 não há saídas anteriores, logo como $Y_{(t)}$ é uma função de $X_{(t)}$ e $Y_{(t-1)}$que assume zero. Além disso, devido à saída de um neurônio recorrente ser função de todas as saídas anteriores, tem-se a característica de memória (Géron, 2019).

5.5.2 Rede *LSTM*

Proposto por Hochreiter e Schmidhuber (1997), a rede *LSTM* é um tipo de RNR capaz de aprender dependências de curto e longo prazo através da retenção de informações passadas. A figura 7 abaixo ilustra a representação da *LSTM*.

Figura 7 – Célula *LSTM*



Fonte: GÉRON, 2019.

Quanto ao funcionamento da rede *LSTM*, Géron (2019) descreve que o vetor de entrada atual x_t e a memória de curto prazo anterior $h_{(t-1)}$ são fornecidas para quatro diferentes camadas que estão conectadas, porém, servem a um propósito diferente:

5.5.3 Estimação de incertezas para a rede *LSTM*

A rede *LSTM* produz estimação pontual, algo menos útil que quando comparado às estimações associadas às incertezas, pois estas assumem um maior escopo de possibilidades. Além disso, a estimação de incertezas ajuda mais na interpretação da saída do modelo do que apenas apresentar uma estatística pontual (Vallbo, 2019). As metodologias tradicionais geralmente permitem estimar incertezas e consequentemente a construção de intervalos de confiança. Entretanto, para redes neurais a obtenção desses intervalos são, na maioria das vezes, obtidos via algum método de estimação, sendo o mais comum o *dropout* (Frau *et al*., 2021). Outros métodos também são utilizados como abordagens bayesianas, *bagging, bootstrap, Local Uncertainty Estimation Model* (LUEM) (Gal; Ghahramani, 2016; Khosravi *et al*., 2011; Zhang *et al*., 2019). Usualmente, podemos definir um intervalo de predição como:

$$\hat{y}_{T+h|T} \pm c\hat{\sigma}_h$$

onde $\hat{y}$ é o valor predito, T o tempo até a disponibilidade das informações, h é o intervalo de tempo para o qual a previsão é efetuada, c é o valor que define a probabilidade de cobertura desejada, $\hat{\sigma}_h$ é o desvio padrão da distribuição da previsão no tempo $T + h$. Em geral, os intervalos de predição aumentam à medida que mais longe for o futuro da previsão. De acordo com Gal e Ghahramani (2016), há diversas formas de estimar $\hat{\sigma}_h$, sendo em alguns casos, ao assumir que os erros têm distribuição normal e são independentes, pode-se calcular o desvio-padrão diretamente. É importante ressaltar que enquanto os intervalos de confiança assumem apenas as incertezas inerentes ao modelo, os intervalos de predição levam consideram também a variância do ruído dos dados (Khosravi *et al*., 2011). Consequentemente, podemos escrever a variância do valor predito como:

$$\hat{\sigma}_y^2 = \hat{\sigma}_{modelo}^2 + \hat{\sigma}_{ruído}^2$$

Segundo Pearce *et al.* (2018), a incerteza do modelo pode ser atribuída a diversos fatores como a própria especificação incorreta na etapa de construção do modelo, incerteza associada aos parâmetros ou ainda devido à falta de representatividade do conjunto de treinamento. Assim, para criar um intervalo de predição que seja adequado, essas incertezas precisam ser consideradas, sendo isto um desafio quando se trata de redes neurais, pois não há uma forma fechada para predizer esses intervalos (Pearce *et al*., 2018).

Gal e Ghahramani (2016) propuseram a estimação do intervalo de predição utilizando *dropout* sem comprometer a complexidade computacional e a acurácia do modelo. O uso de *dropout* foi inicialmente proposta visando evitar *overfitting* e melhorar a generalização dos erros (Srivastava *et al*., 2014; Frau *et al*., 2021). Uma das principais vantagens de se utilizar *dropout* é por ser conceitualmente simples, rápido de implementar e computacionalmente barato (Frau *et al*., 2021). *Dropout* é o ato da rede desligar aleatoriamente alguns neurônios durante o processo de treinamento, sendo esse desligamento dado pelo hiperparâmetro *dropout rate*. Esse hiperparâmetro expressa a probabilidade p de um neurônio ser mantido, sendo 1 a ausência de *dropout* e valores mais baixos, maior é a quantidade de neurônios desligados (Srivastava *et al*., 2014). Baixo valor para p torna necessário um n grande, cujo qual deixa mais lento a etapa de treinamento e causa falta de ajuste (*underfitting*), enquanto um p grande pode não gerar *dropouts* suficientes para evitar o *overfitting* (Srivastava *et al*., 2014).

No processo de estimação da incerteza, o *dropout* é ativado durante o período de inferência, resultando em múltiplas predições para um único tempo devido à ativação de diferentes neurônios. Como resultado, isso permite obter a distribuição do valor predito e consequentemente a estimação

da incerteza associada a predição (Frau *et al*., 2021). Segundo Gal e Ghahramani (2016), uma rede com *dropout* aplicada em cada camada é matematicamente equivalente a uma aproximação de um processo gaussiano profundo.

Para estimação da incerteza utilizando *dropout*, a distribuição posterior é estimada usando processo estocástico *forward*, sendo o passe para frente uma previsão onde o *dropout* é aplicado. Quando passes estocásticos para frente são usados durante o treinamento e teste, são chamados de Monte Carlo (Vallbo, 2019). Para B iterações (passos à frente) são obtidos $\{\hat{y}_1^*, \dots, \hat{y}_B^*\}$. O valor da estimação pontual é calculado como a média dos B passos à frente, como demonstrado nas equações abaixo:

$$\hat{y}^* = \frac{1}{B}\sum_{b=1}^{B} \hat{y}_B^*$$

$$\hat{\sigma}^2 = \frac{1}{V}\sum_{v=1}^{V} (\hat{y}_v' - f(\hat{y}_v'))^2$$

Zhu e Laptev (2017) considerou tanto a incerteza inerente aos dados de treinamento quanto ao ruído para produzir os intervalos de predição, sendo a incerteza atribuída aos ruídos estimados pelo conjunto de validação. O conjunto de validação V consiste em amostras de $X' = \{x_1', \dots, x_V'\}$ e $Y' = \{y_1', \dots, y_V'\}$. Para cada amostra no conjunto de validação é calculado o ruído conforme a equação abaixo:

$$var\big(f(x)\big) = \frac{1}{B}\sum_{b=1}^{B} (\hat{y}_B^* - \overline{\hat{y}^*})$$

Por fim, Zhu e Laptev (2017) estimam a incerteza inerente aos dados de treinamento utilizando *autoencoder*, que deve aproximar a distância entre a entrada e os dados na fase de teste.

6 ARTIGO 1 - MOVIMENTAÇÃO ESPAÇO-TEMPORAL DAS VARIANTES DA COVID-19: NÚMEROS DE CASOS E ÓBITOS POR SRAG-COVID NO BRASIL EM 2020 A 2022

RESUMO

Background. A epidemia de COVID-19 no Brasil foi caracterizada pela co-circulação de múltiplas variantes, consequência da introdução de variantes independentes do vírus e das mutações no material genético que facilitaram sua transmissão. Esse artigo analisa a distribuição espaço-temporal *clusters* de casos e óbitos por SRAG-Covid causadas por 6 variantes da Sars-Cov-2, são elas: B.1.1.28, B.1.1.33, Zeta, Gama, Delta e Ômicron.

Métodos. Foram analisados dados diários de casos e óbitos por SRAG-Covid período de 26/02/2020 a 31/12/2022 ao nível de município para todo o Brasil. As análises foram divididas conforme os períodos de prevalência de cada variante. Para a identificação dos *clusters,* estimação da população sob risco e estimação do risco relativo foi utilizado a Estatística *Scan* de *Kulldorff.*

Resultados. Foram identificados 95 *clusters* de alto risco para casos e 82 para óbitos por SRAG-Covid. As áreas identificadas com risco elevado apresentaram de 1,43 a 8,25 vezes o risco de casos por SRAG-Covid. Para óbitos, esses valores foram de 1,41 a 11,33 vezes.

Conclusões. Foi possível observar o espalhamento espaço-temporal das variantes, resultando em áreas de elevados riscos relativos de casos e óbitos em diversas regiões do Brasil. A emergência de novas variantes deve ser vista como um fator de preocupação, pois as mutações podem permitir que os vírus se tornem capazes de evadir a resposta dos anticorpos neutralizantes, resultando em novas ondas de casos e óbitos.

Palavras-chave: *SRAG-Covid*, SaTScan, variantes, *clusters*.

1 Introdução

A COVID-19, doença reconhecida como Emergência de Saúde Pública de Importância Internacional (ESPII) durante o período de 30/1/2020 até 5/5/2023 pela Organização Mundial de Saúde (OMS), alcançou mais de 200 países, ultrapassando 743 milhões de casos e mais de 6.6 milhões de mortes. Entre esses países, o Brasil foi um dos mais afetados, terminando o ano de 2022 superando os 36 milhões de casos e aproximadamente 695 mil óbitos registrados (WHO, 2023; Burki, 2023). Com o avanço da pandemia, novas variantes do vírus Sars-Cov-2 emergiram. Muitas dessas novas variantes apresentaram mutações que as tornaram mais transmissíveis e como consequência, constituíram outras variantes. Durante o primeiro ano da pandemia no Brasil, as variantes B.1.1.28 e B.1.1.33 foram predominantes até novembro e partir delas, mediante às mutações no domínio de ligação da proteína Spike (S) deram origem a outras duas novas variantes, a Gama e a Zeta, que também se espalharam rapidamente (Faria et al., 2021). Dessas, a Gama, foi a variante que resultou no maior pico no número de casos e óbitos. Com a redução da prevalência de casos causados pela Gama, a Delta ganhou espaço a partir de julho de 2021 e ainda no mesmo ano, foi sucedida pela Ômicron, variante que ainda se mantém prevalente.

Assim, este artigo visa, por meio do uso da estatística *Scan* de *Kulldorff*, identificar *clusters* de alto riscos e estimar os riscos relativos para o número de casos e óbitos por SRAG-Covid considerando as variantes circulantes para o período de 26/02/2020 a 31/12/2022. Finalmente, este artigo tem como propósito evidenciar a importância da compreensão da movimentação das variantes, principalmente no que tange o auxílio de tomada de decisões baseada em dados.

2 Materiais e métodos

2.1 Dados

O presente estudo utilizou dados individuais com sintomas que caracterizam a Síndrome Respiratória Aguda Grave (SRAG) notificados no Sistema de Informação da Vigilância Epidemiológica da Gripe (SIVEP-Gripe) com confirmação laboratorial da COVID-19 (SRAG-Covid) para todos os municípios Brasileiros para o período de 26/02/2020 a 31/12/2022. Os dados são de livre acesso e disponíveis para download na plataforma *OpenDataSUS* (url: https://opendatasus.saude.gov.br/). É importante ressaltar que a notificação de casos suspeitos ou confirmados de COVID-19 é compulsória, isto é, sua ocorrência implica na obrigação de se notificar o caso (Ministério da Saúde, 2020). Ainda, segundo Ministério da Saúde (2021b), considera-se como hospitalização por COVID-19 qualquer registro hospitalar com a classificação final de SRAG por

COVID-19. Essa classificação pode ser baseada em critérios laboratoriais, clínico-epidemiológicos, clínicos ou clínico-imagem para encerramento do registro. Segundo as diretrizes do Ministério da Saúde (2021b), os casos de Síndrome Respiratória Aguda Grave (SRAG) são definidos pela presença simultânea de quatro critérios: (i.) febre (mesmo que autorreferida), (ii.) dispneia/desconforto respiratório ou pressão persistente no tórax ou saturação de O2 menor que 95%, (iii.) dor de garganta ou tosse e (iv.) hospitalização ou evolução para óbito, independentemente de hospitalização prévia.

Foram também utilizados dados de sequenciamentos genéticos obtidos através da vigilância genômica do Sars-Cov-2 pela Rede Genômica Fiocruz (2021b). Foi considerado como variante mais prevalente aquela com maior percentual relativo ao número de amostras coletadas naquele mês. Em casos de empates no percentual, foram adotadas as seguintes regras:

- Se uma das variantes que empatou for a mais prevalente no mês seguinte, ela também será considerada a mais prevalente no mês do empate.
- Se a condição 1) não for atendida, ou seja, a variante prevalente do próximo mês for diferente, então é verificado se uma das variantes que empatou foi a mais prevalente no mês anterior, considerando-a como a mais prevalente no mês do empate.

Por último, os dados foram agregados por município de notificação, considerando o dia de primeiros sintomas para as hospitalizações e a data do óbito quando o caso evoluiu para óbito. Foram calculadas as taxas de incidência considerando o tamanho da população consoante as estimativas do IBGE para o ano de 2021 (IBGE, 2021).

2.2 Áreas e períodos de estudo

O Brasil é um país localizado na América do Sul, com uma extensão territorial de 8.515 milhões km^2, dividido em 26 estados. Possui uma população estimada em cerca de 213,3 milhões de habitantes, com uma densidade demográfica média de 25 habitantes por km^2 (IBGE, 2021). As análises foram divididas em 5 fases, correspondendo à atuação das variantes. A determinação do momento que se inicia e termina cada fase foi feito considerando que as variantes devam ter pelo menos 5% de prevalência a cada mês. Assim, temos as seguintes fases de análises:

Fase 1: 26/02/2020 à 31/12/2020 - B.1.1.28, B.1.1.33, Zeta e outras.

Fase 2: 01/10/2020 à 28/02/2021 - B.1.1.28, Zeta e Gama.

Fase 3: 01/01/2021 à 31/08/2021 - Zeta e Gama.

Fase 4: 01/07/2021 à 31/12/2021 - Gama e Delta.

Fase 5: 01/12/2021 à 31/12/2022 - Ômicron.

Foi considerado a sobreposição de tempo para as fases das análises visando atender o critério adotado de prevalência mínima das variantes. Além disso, como as variantes emergem e se findam em momentos diferentes, muitas vezes, compartilhando da mesma temporalidade, é necessário que essa característica seja incorporada nas análises. Por fim, essa sobreposição do tempo também é importante, pois possibilita obter novas estimativas para locais onde o risco relativo já havia sido estimado.

2.3 Análises estatísticas

A detecção dos *clusters* espaços-temporais dos números de casos (óbitos) de SRAG-Covid se deu através do uso da estatística *Scan* de *Kulldorff* por meio do *software SaTScanTM* v10.1. Essa metodologia permite a detecção de *clusters* espaço-temporais utilizando a distribuição de probabilidade Poisson; testa a significância estatística e a corrige para múltiplos testes; examina a dinâmica da doença em um intervalo de tempo contínuo; e estima o risco relativo (RR) para cada *cluster* considerando o tamanho populacional daquele *cluster* (Kulldorff, 1999). A estatística *Scan* espaço-temporal utiliza-se de "cilindros" que se movimentam ao longo do tempo e espaço para a identificação de *clusters* com altos riscos através da comparação do número de casos (óbitos) observados com o esperado dentro do cilindro. A base do cilindro representa a área geográfica enquanto a altura representa o intervalo de tempo, sendo o tamanho da base do cilindro, como também a sua altura, independentes uma da outra (Kulldorff, 1999). Dessa forma, o número de casos (óbitos) esperados contidos nessas janelas cilíndricas assumem a distribuição de Poisson:

$$\left(\frac{C}{E[c]}\right)^{c}\left(\frac{C-c}{C-E[c]}\right)^{C-c}$$

onde C é o total de casos (óbitos), c é o número de casos (óbitos) dentro da janela cilíndrica é o número esperado de casos (óbitos) dentro da janela cilíndrica. No modelo de Poisson, o número esperado de casos (óbitos) de SRAG-Covid em cada município é proporcional à sua população. Além disso, a probabilidade de um caso (óbitos) estar em um município m assumindo que a data do primeiro sintoma foi dia d é a mesma para todos os dias d (Kulldorff, 1999). Nesse estudo foi

considerado como áreas vizinhas, municípios em que os seus centroides estejam dentro da base do cilindro. A hipótese nula é que o risco dentro do cilindro é igual ao risco fora do cilindro. Assim, para cada cilindro, o número de casos esperados ($E[c]$) é igual ao total de casos no município (C) dividido pela população (P), vezes a população dentro do cilindro p (Kulldorff, 1999):

$$E[c] = \frac{C}{P} \times p$$

Os *clusters* detectados são ordenados segundo o log da razão da verossimilhança (LLR) expresso pela fórmula abaixo (Kulldorff, 2022):

$$LLR = \left(\frac{C}{E[c]}\right)^{c} \left(\frac{C-c}{C-E[c]}\right)^{C-c} I()$$

onde c é o número de casos dentro do *cluster* e *I()* é uma função indicadora que é igual a 1 quando o cilindro tem mais casos do que o esperado e 0, caso contrário. Assim, os *clusters* com maiores valores para a LLR são os mais verosímeis. O risco relativo (RR) para cada *cluster* é calculado pela seguinte expressão (Kulldorff, 2022):

$$RR = \frac{c/E[c]}{(C-c)/(C-E[c])}$$

Os valores-p dos *clusters* detectados foram calculados usando simulação de Monte Carlo com 999 replicações. Essas simulações são comparadas aos dados reais, assim se a razão de verossimilhança for maior nos dados reais do que nos dados simulados, rejeita-se a hipótese nula, ou seja, há evidências suficientes para dizer que o *cluster* detectado apresenta risco relativo estatisticamente maior do que 1. Foi considerado o método padrão do *SaTScan*^{TM} para o cálculo do valor-p, ou seja, são combinados três diferentes métodos, são eles:

- Monte Carlo (Padrão): Proposto por Dwass (1957), a estatística de teste é calculada para cada simulação, bem como para o conjunto de dados reais, se o resultado estiver entre os 5% mais altos, por exemplo, o teste é significativo no nível 0,05.
- Monte Carlo (Sequencial): Conforme se avança nos números de simulações, maior se torna o poder do teste, porém também se aumenta o tempo de processamento. Quando o valor-p é pequeno, é interessante continuar utilizando a simulação de Monte Carlo padrão, entretanto, para valores-p elevados, o método de Monte Carlo sequencial permite acelerar o término da análise sem comprometer o nível de significância das análises. Os cálculos terminam assim que o número de simulações tiver uma razão de verossimilhança maior do que a razão de

verossimilhança do conjunto de dados reais. Se essa condição nunca for atingida, os cálculos continuarão até que o número máximo de simulações (999) (Kulldorff, 2022).

- Aproximação de Gumbel: Nessa abordagem não há limite inferior nos valores-p ($valor - p > 0$). Os métodos funcionam gerando simulações dos dados sob a hipótese nula. A razão de verossimilhança máxima de cada simulação é usada para ajustar uma distribuição de Gumbel pelo método de estimativa de momentos. Uma vez obtida a distribuição de Gumbel que melhor se ajusta aos dados, calcula-se o valor-p como a probabilidade dessa distribuição gerar um valor maior que a razão de verossimilhança máxima observada para o cluster mais provável do conjunto de dados reais.

Inicialmente, todas as simulações utilizam o método de Monte Carlo padrão, porém combinação desses outros métodos ocorre da seguinte forma:

- Em casos de valores-p altos, o método de Monte Carlo sequencial pode ser aplicado;
- Em casos em que os valores-p são muito pequenos, utiliza-se a aproximação de Gumbel.

O nível de significância adotado para a detecção dos *clusters* foi de 5% ($valor - p < 0{,}05$), porém foram considerados apenas *clusters* significativos à 1% ($valor - p < 0{,}01$). Os municípios não podem pertencer à diferentes *clusters* dentro do mesmo período de análise, ou seja, não há intersecção espacial. Devido à magnitude da área de estudo (Brasil), o tamanho máximo permitido para população de cada *clusters* foi igual a 10% da população total, isso são 21.331.764 indivíduos, fator esse que limita também o tamanho do raio da janela espacial (círculo). O período de duração de cada *cluster* pode variar entre 7 e 21 dias e deve conter pelo menos 100 casos (óbitos).

3 Resultados

O Brasil é um país de contrastes, o que naturalmente fez com que múltiplas epidemias da COVID-19 iniciassem e terminassem em diferentes momentos e lugares, muitas vezes causadas por diferentes variantes de vírus. Considerando essa característica, a figura 1 apresenta uma linha temporal enfatizando o percentual de municípios sob a presença de cada variante.

Nota-se que durante os primeiros meses, a variante B.1.1.28 foi a mais prevalente, embora disputasse espaço com a B.1.1.33 e outras variantes menores. Essa interação perdeu força em dezembro de 2020, com a expansão da Zeta iniciada em setembro. Não muito diferente, a Gama (P.1), tornou-se a mais prevalente em fevereiro de 2021, perdendo “protagonismo” para a Delta em agosto do mesmo ano. Ainda em 2021, ocorreu a emergência da Ômicron, variante que até então é prevalente.

A tabela 1 apresenta as principais medidas descritivas para as características dos *clusters* de casos e óbitos por SRAG-Covid. A duração e o raio dos *clusters* foi, em média, de 18,3 dias e 231 km, respectivamente. Ainda, o raio mínimo foi de 0 km para *clusters* de casos com apenas com 1 município, chegando até 1.888 km. A média de municípios pertencentes aos *clusters* de casos é de 226 com mínimo e máximo, respectivamente de 1 a 894 municípios e risco relativo médio de 3,03 variando entre 1,43 (Gama) a 8,25 (B.1.1.33). Para os *clusters* de óbitos a duração temporal média foi praticamente a mesma observada nos *clusters* de casos, 18,9 dias, porém a média dos raios apresentaram alcances superiores, 332 km. O risco relativo médio foi de 3,65 variando entre 1,41 (Gama) a 11,33 (Ômicron).

Figura 1 – Percentual de municípios sob predominância da COVID-19 (A), série temporal dos casos e óbitos por SRAG-Covid (B).

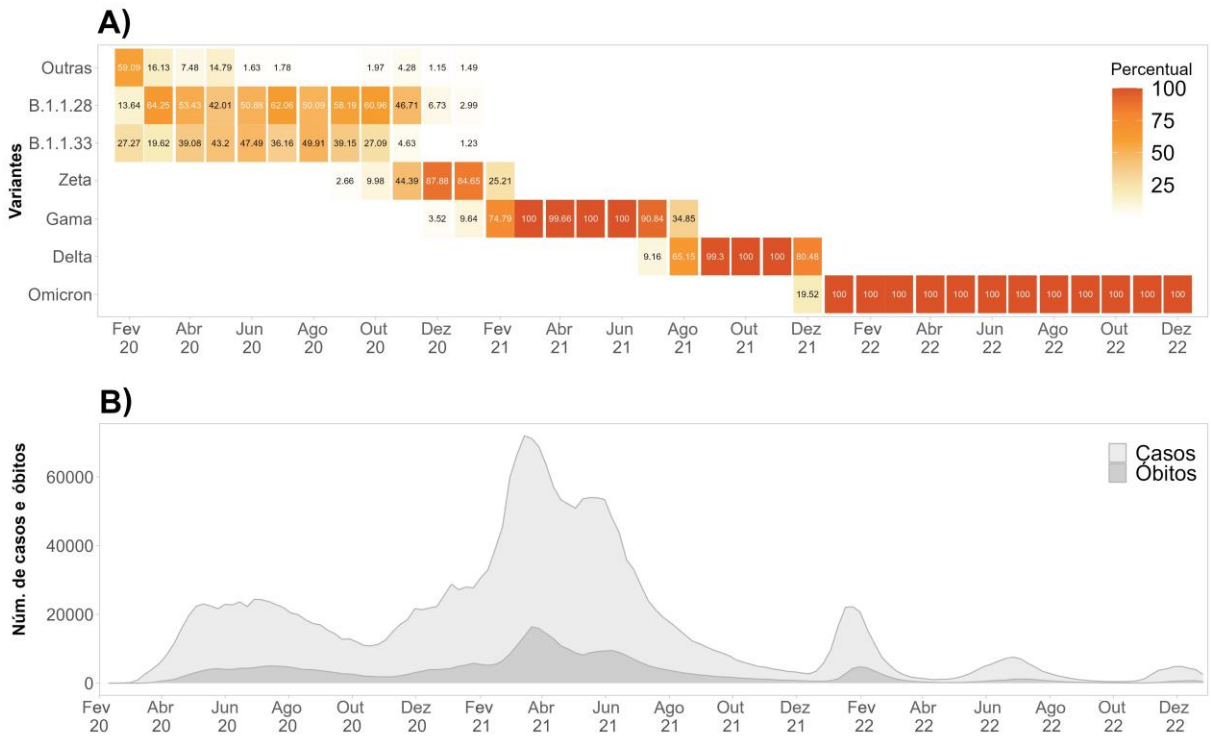


Tabela 1 – Resumo descritivo dos *clusters* de casos e óbitos por SRAG-Covid.

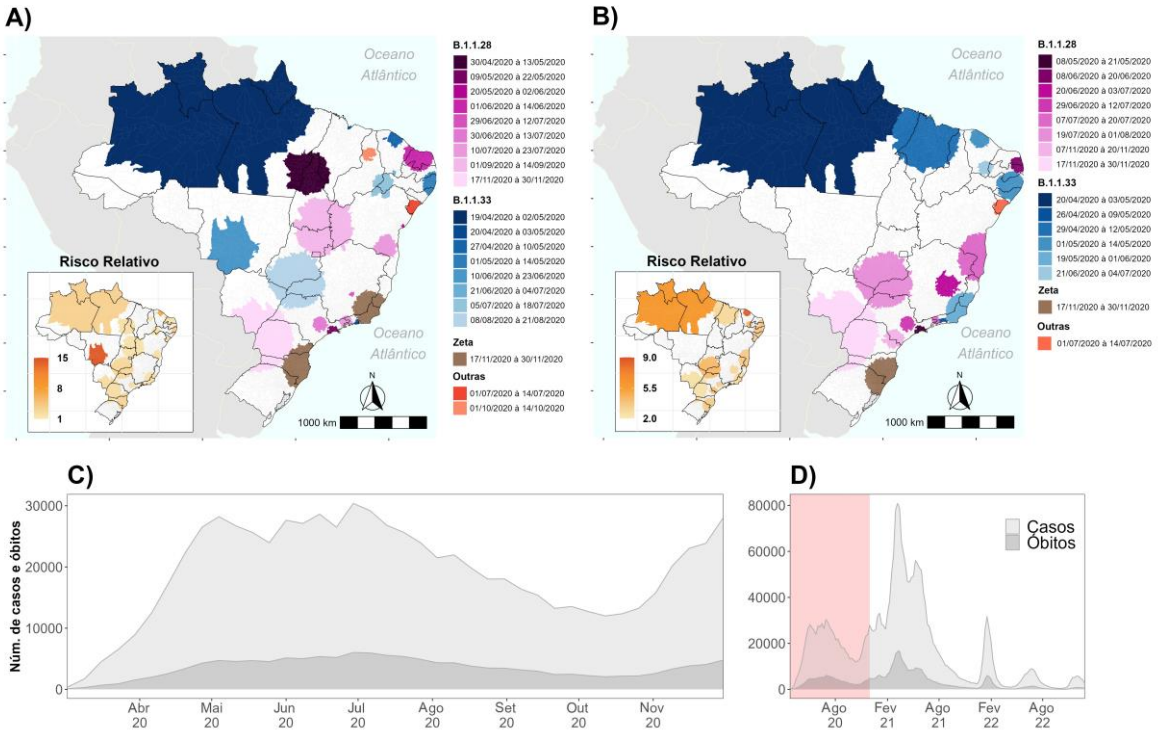
	Média (DP)	Mín	Mediana	Máx
Casos				
Duração (em dias)	18,8 (3)	6	20	20
População	9.831.105 (9.150.073)	58,02	3.829.810	21.320.828
Núm. de municípios	226,48 (290,45)	1	41	894
Raio	231 (317,61)	0	114,9	1888,38

LLR	2454,31 (2971,68)	28,08	669,91	11995,4
Observado	1962,61 (2641,82)	27,01	823,38	9921,55
Esperado	5428,06 (6402,29)	101	3026	25312
ODE	2,99 (1,46)	1,43	2,47	8,21
Risco Relativo	3,03 (1,5)	1,43	2,5	8,25
Óbitos				
Duração (em dias)	18,9 (2,7)	8	20	20
População	11.777.384 (8.378.600)	425,318	11.112.747	21.330.636
Núm. de municípios	339,98 (306,48)	1	249,5	1031
Raio	332,66 (333,43)	0	232,93	1658,97
LLR	1055,06 (1353,31)	19,47	656,54	5991,93
Observado	540,91 (589,98)	31,17	388,49	2320,57
Esperado	1799,42 (2019,15)	100	1276,5	9278
ODE	3,59 (1,94)	1,41	3,11	10,74
Risco Relativo	3,65 (2,02)	1,41	3,14	11,33

3.1 Fase 1: 26/02/2020 à 31/12/2020

As variantes predominantes nesse período foram a B.1.1.28 e B.1.1.33. Como apresentado na figura 2, até meados de maio de 2020, a curva de casos e óbitos por SRAG-Covid apenas aumentava e apesar de uma pequena queda, voltou a subir atingindo o seu pico em menos de 60 dias (SE 27). Apesar dos primeiros casos terem sido confirmados no estado de São Paulo, o primeiro *cluster* foi identificado predominantemente na região Norte, sendo a sua causa atribuída à variante B.1.1.33, abrangeu 145 municípios distribuídos por 4 estados (AM, RR, PA, AP e MA) ao longo de 20 dias.

Figura 2 – Distribuição espaço-temporal dos *clusters*, seus respectivos riscos relativos para casos (A) e óbitos (B), recorte da série temporal (C) e série temporal completa de casos e óbitos (D) por SRAG-Covid durante o período 26/02/2020 à 31/12/2020.



O primeiro *cluster* causado pela variante B.1.1.28 ocorreu na região Nordeste, nos estados do Alagoas, Pernambuco, Paraíba, Rio Grande do Norte e Ceará. Ao total, foram identificados 13 *clusters* (6 referentes à B.1.1.28, 5 à B.1.1.33, 1 à Zeta e 1 a outra variante) para casos e 8 para óbitos (4 referentes à B.1.1.28, 1 à B.1.1.33, 2 à Zeta e 1 a outra variante).

Tabela 2 – *Clusters* de casos e óbitos de SRAG-Covid para o período 26/02/2020 à 31/12/2020.

SRAG	Início	Fim	Variante	Núm. de municípios	População	Raio	LLR	Observado	Esperado	ODE	Risco Relativo
Casos											
	25/04/2020	15/05/2020	B.1.1.33	39	14.341.209	102,75	6112,54	10238	2715,5	3,77	3,82
	25/04/2020	15/05/2020	B.1.1.33	271	16.168.351	1007,3	2963,75	8201	3061,47	2,68	2,7
	30/04/2020	20/05/2020	B.1.1.28	470	18.403.603	688,61	5389,05	11181	3484,71	3,21	3,25
	01/05/2020	21/05/2020	B.1.1.28	24	21.209.040	37,75	6279,3	12936	4015,92	3,22	3,27
	21/05/2020	10/06/2020	B.1.1.28	4	3.329.925	16,93	451,31	1523	630,52	2,42	2,42
	12/06/2020	02/07/2020	B.1.1.33	80	2.413.610	383,85	4611,22	3751	457,02	8,21	8,25
	12/06/2020	02/07/2020	B.1.1.33	35	1.664.119	99,47	610,95	1116	315,1	3,54	3,55
	20/06/2020	10/07/2020	Outras	61	2.042.443	83,14	437,02	1100	386,74	2,84	2,85
	28/06/2020	18/07/2020	B.1.1.28	12	4.843.869	25,54	819,93	2392	917,18	2,61	2,61
	29/06/2020	16/07/2020	B.1.1.33	14	698.812	55,04	83,79	277	113,42	2,44	2,44
	17/07/2020	06/08/2020	B.1.1.28	9	458.692	30,16	150,16	293	86,85	3,37	3,37
	31/07/2020	20/08/2020	B.1.1.28	532	18.603.633	422,59	2162,28	8073	3522,58	2,29	2,31
	20/11/2020	10/12/2020	Zeta	870	21.071.020	367,41	1924,35	8485	3989,78	2,13	2,14
Óbitos											

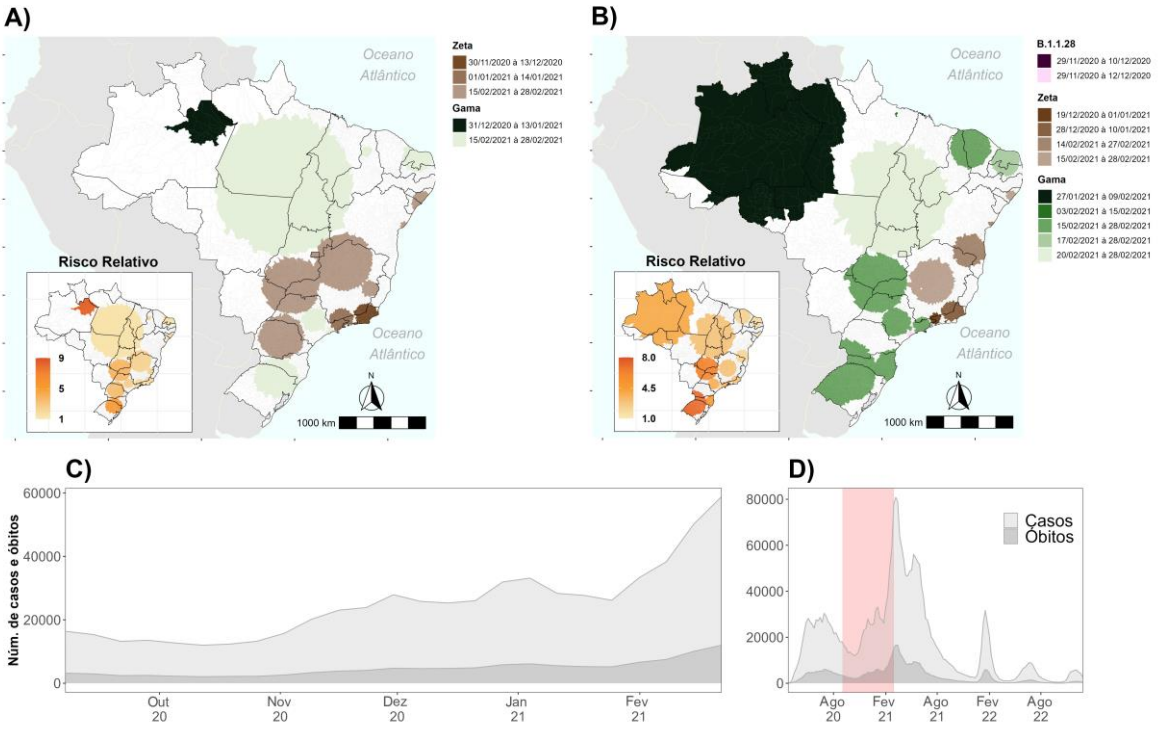
02/05/2020	22/05/2020	B.1.1.33	216	14.433.261	1659	2400,75	2942	589	4,99	5,09
10/05/2020	30/05/2020	Outras	78	3.104.334	152,05	802,79	800	126,68	6,31	6,35
16/05/2020	05/06/2020	B.1.1.28	235	11.404.286	180,21	876,28	1624	465,39	3,49	3,52
08/07/2020	28/07/2020	B.1.1.28	322	12.637.998	331,06	254,11	1105	515,74	2,14	2,15
10/07/2020	30/07/2020	B.1.1.28	264	19.625.760	161,94	811,31	2180	800,89	2,72	2,75
12/07/2020	01/08/2020	B.1.1.28	662	21.300.492	660,79	1231,13	2690	869,24	3,09	3,14
10/12/2020	30/12/2020	Zeta	613	21.316.968	413,51	722,77	2205	869,91	2,53	2,56
11/12/2020	31/12/2020	Zeta	458	21.069.506	230,23	600,98	2055	859,81	2,39	2,41

A variante B.1.1.33 foi a responsável por apresentar os riscos relativos mais elevados tanto para casos quanto para óbitos por SRAG-Covid, conforme apontado na tabela 2. Assim, para os *clusters* de casos, os menores e maiores riscos relativos foram, respectivamente, 2,14 (Zeta) e 8,25 (B.1.1.33) enquanto para os óbitos foram, 2,15 (B.1.1.28) e 6,35 (variante não especificada).

3.2 Fase 2: 01/10/2020 à 28/02/2021

As variantes predominantes nesse período foram a Zeta e Gama. A figura 3 apresenta um crescimento contínuo para o número de casos e óbitos por SRAG-Covid, atingindo os valores mais elevados em fevereiro de 2021, momento em que uma nova variante, a Gama, havia sido estabelecida. O primeiro *cluster* da Gama foi identificado ocupando a divisa do Amazonas com Roraima e Pará. Posteriormente, a Gama se espalhou para as 5 regiões do Brasil. Os *clusters* da Zeta foram detectados primeiramente na região Sudeste e então se espalharam a ponto de serem detectadas formações no Centro-Oeste e no Nordeste.

Figura 3 – Distribuição espaço-temporal dos *clusters*, seus respectivos riscos relativos para casos e óbitos (B), recorte da série temporal (C) e série temporal completa de casos e óbitos (D) por SRAG-Covid durante o período 01/10/2020 à 28/02/2021.



Os dois primeiros *clusters* de casos identificados para esse período são atribuídos a Zeta e a B.1.1.28 e ocorreram, respectivamente, nos estados do Rio de Janeiro e no Pará. Nesse período foram identificados 15 *clusters* para casos (1 referente B.1.1.28, 5 à Zeta e 9 à Gama) e 12 para óbitos (8 referentes à Zeta e 4 à Gama). Se destacam Belém e Recife, que sozinhos foram classificados respectivamente como *clusters* de casos e óbitos, apresentando riscos relativos de 2,03 e 1,79.

Tabela 3 – *Clusters* de casos e óbitos de SRAG-Covid para o período 01/10/2020 à 28/02/2021.

SRAG	Início	Fim	Variante	Núm. de municípios	População	Raio	LLR	Observado	Esperado	ODE	Risco Relativo
Casos											
	30/11/2020	20/12/2020	Zeta	97	17.089.171	150	1648,57	8910	4553,07	1,96	1,98
	01/01/2021	21/01/2021	B.1.1.28	7	353.883	106	31,34	181	94,29	1,92	1,92
	08/01/2021	28/01/2021	Gama	20	2.908.639	254,22	5611,14	5259	774,95	6,79	6,86
	08/02/2021	28/02/2021	Gama	834	21.287.140	326,15	8860,1	18176	5671,54	3,2	3,31
	08/02/2021	28/02/2021	Gama	3	2.989.412	15,15	1815,7	3026	796,47	3,8	3,82
	08/02/2021	28/02/2021	Gama	1	825.796	0	391,37	751	220,02	3,41	3,42
	08/02/2021	28/02/2021	Zeta	777	21.282.802	518,21	5623,34	15251	5670,38	2,69	2,76
	08/02/2021	28/02/2021	Zeta	9	3.829.810	29,86	658,34	2380	1020,38	2,33	2,34
	10/02/2021	28/02/2021	Gama	25	20.050.907	43,14	2859,81	10920	4833,39	2,26	2,29
	10/02/2021	28/02/2021	Gama	57	1.668.838	129,39	35,63	583	402,28	1,45	1,45
	12/02/2021	28/02/2021	Gama	58	1.044.979	198,03	76,47	435	225,38	1,93	1,93

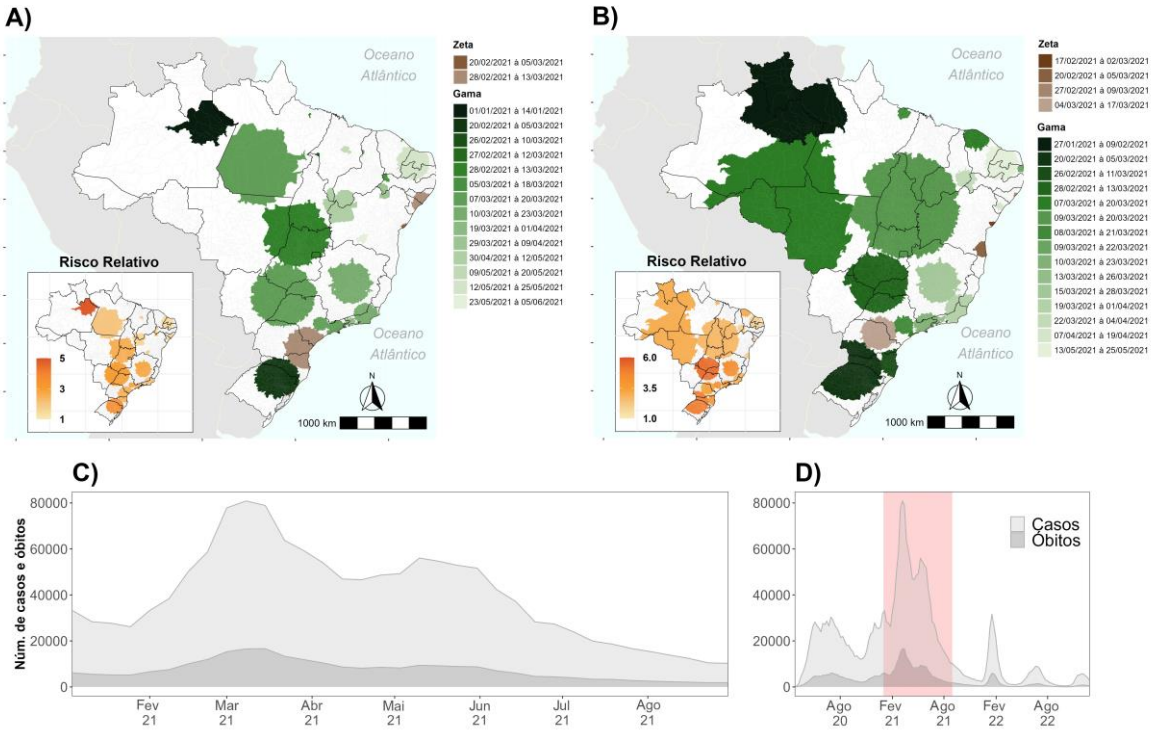
	12/02/2021	28/02/2021	Zeta	94	3.807.348	114,9	173,86	1410	821,18	1,72	1,72
	15/02/2021	28/02/2021	Zeta	569	19.233.517	453,29	669,91	5756	3416,26	1,68	1,69
	17/02/2021	28/02/2021	Gama	8	343.181	41,16	44,46	134	52,25	2,56	2,57
	18/02/2021	28/02/2021	Gama	1	1.506.420	0	85,85	427	210,23	2,03	2,03
Óbitos											
	12/12/2020	31/12/2020	Zeta	1	1.661.017	0	20,78	147	82,07	1,79	1,79
	13/12/2020	02/01/2021	Zeta	90	9.927.611	172,54	149,7	953	515,04	1,85	1,86
	04/01/2021	24/01/2021	Zeta	199	2.603.180	143,62	22,45	220	135,05	1,63	1,63
	10/01/2021	30/01/2021	Zeta	595	18.140.969	283,67	753,32	2349	941,15	2,5	2,54
	11/01/2021	31/01/2021	Zeta	145	10.821.209	117,11	462,87	1419	561,4	2,53	2,56
	27/01/2021	16/02/2021	Gama	81	5.400.009	967,02	2527,66	2147	280,15	7,66	7,85
	08/02/2021	28/02/2021	Gama	19	600.853	86,72	47,77	100	31,17	3,21	3,21
	08/02/2021	28/02/2021	Zeta	742	21.242.738	441,51	1608,26	3436	1102,07	3,12	3,21
	10/02/2021	28/02/2021	Zeta	7	914.716	28,74	37,4	111	42,94	2,59	2,59
	10/02/2021	28/02/2021	Zeta	57	1.605.768	114,57	34,36	158	75,37	2,1	2,1
	18/02/2021	28/02/2021	Gama	928	20.862.910	425,85	660,17	1625	566,95	2,87	2,91
	20/02/2021	28/02/2021	Gama	85	3.440.129	160,51	33,88	159	76,49	2,08	2,08

De acordo com a tabela 3, a variante Gama teve o seu primeiro *cluster* identificado no último dia de dezembro, sendo a responsável por maximizar o risco relativo neste período. Os riscos relativos mínimos e máximos dos *clusters* de casos e óbitos por SRAG-Covid foram, respectivamente, 1,69 (Zeta) e 6,86 (Gama); 1,79 (Zeta) e 7,85 (Gama).

3.3 Fase 3: 01/01/2021 à 31/08/2021

As variantes predominantes neste período continuaram sendo as mesmas vistas anteriormente, entretanto, a variante Gama prevaleceu sob a Zeta, atingindo a maioria dos municípios Brasileiros, como aponta a figura 1. A Zeta passou a perder força para a Gama em fevereiro de 2021 (SE 5) e a partir de então, os números de casos e óbitos tiveram um abrupto aumento. Os casos de SRAG-Covid chegaram a mais de 60 mil (SE 9), enquanto o de óbitos para mais de 15 mil (SE 9).

Figura 4 – Distribuição espaço-temporal dos *clusters*, seus respectivos riscos relativos para casos (A) e óbitos (B), recorte da série temporal (C) e série temporal completa de casos e óbitos (D) por SRAG-Covid durante o período 01/01/2021 à 31/08/2021.



A variante Zeta apresentou *clusters* apenas até início de abril (SE 13), depois disso, todos os *clusters* identificados foram exclusivamente atribuídos à Gama. Como visto na figura 3, o mesmo conjunto de municípios foi identificado como um *cluster* de casos, diferenciando-se apenas na magnitude do risco relativo, que antes foi de 6,86 e agora passou a ser 3,09. Entende-se que essa diferença se deve ao aumento de casos de SRAG-Covid em relação ao período anterior. Nesse período foram identificados 24 *clusters* (3 referentes à Zeta e 21 à Gama) para casos e 15 para óbitos (2 referentes à Zeta, 13 à Gama).

Tabela 4 – *Clusters* de casos e óbitos de SRAG-Covid para o período 01/01/2021 à 31/08/2021.

SRAG	Início	Fim	Variante	Núm. de municípios	População	Raio	LLR	Observado	Esperado	ODE	Risco Relativo
Casos											
	08/01/2021	28/01/2021	Gama	20	2.908.639	254,22	3236,77	5259	1354,24	3,88	3,9
	24/02/2021	16/03/2021	Zeta	6	3.722.838	28,74	419,6	3071	1733,33	1,77	1,77
	26/02/2021	18/03/2021	Zeta	3	88.075	34,79	31,05	101	41,01	2,46	2,46
	28/02/2021	20/03/2021	Gama	754	21.289.903	312,48	8434,68	25312	9912,44	2,55	2,59
	28/02/2021	20/03/2021	Gama	3	2.989.412	15,15	1727,74	4103	1391,85	2,95	2,95
	28/02/2021	20/03/2021	Gama	9	1.035.594	51,37	323,32	1140	482,17	2,36	2,37
	28/02/2021	20/03/2021	Gama	31	942.223	180,71	149,67	848	438,69	1,93	1,93
	28/02/2021	20/03/2021	Zeta	2	1.045.639	15,98	202,47	994	486,84	2,04	2,04
	01/03/2021	21/03/2021	Gama	1	825.796	0	517,91	1171	384,49	3,05	3,05
	02/03/2021	22/03/2021	Gama	2	116.404	24,85	31,19	122	54,2	2,25	2,25
	05/03/2021	25/03/2021	Gama	34	21.280.449	44,02	7641,11	24455	9908,04	2,47	2,5

05/03/2021	25/03/2021	Gama	1	259.980	0	65,27	267	121,04	2,21	2,21
12/03/2021	01/04/2021	Gama	476	21.264.638	428,03	4821,64	21100	9900,68	2,13	2,15
15/03/2021	04/04/2021	Gama	7	1.768.457	70,07	378,93	1731	823,38	2,1	2,1
19/03/2021	08/04/2021	Gama	1	1.661.017	0	99,32	1197	773,36	1,55	1,55
20/03/2021	09/04/2021	Gama	304	21.309.469	209,06	2277,76	17324	9921,55	1,75	1,76
21/03/2021	10/04/2021	Gama	4	983.876	20,74	329,8	1109	458,09	2,42	2,42
03/04/2021	18/04/2021	Gama	1	359.372	0	28,08	221	127,48	1,73	1,73
27/04/2021	16/05/2021	Gama	3	67.478	15,34	54,25	103	29,92	3,44	3,44
09/05/2021	29/05/2021	Gama	12	561.648	30,17	92,2	510	261,5	1,95	1,95
14/05/2021	01/06/2021	Gama	12	462.178	26,57	84,9	403	194,69	2,07	2,07
17/05/2021	06/06/2021	Gama	6	58.020	32	61,87	103	27,01	3,81	3,81
21/05/2021	10/06/2021	Gama	41	511.313	64,48	34,84	378	238,06	1,59	1,59
23/05/2021	12/06/2021	Gama	635	21.178.336	296,78	7368,4	24079	9860,49	2,44	2,47
Óbitos										
27/01/2021	16/02/2021	Gama	43	3.977.107	542,99	1281,41	1856	433,2	4,28	4,31
26/02/2021	18/03/2021	Zeta	7	914.716	28,74	98,85	270	99,63	2,71	2,71
09/03/2021	29/03/2021	Gama	15	425.318	58,65	155,28	211	46,33	4,55	4,56
11/03/2021	31/03/2021	Gama	898	21.207.344	749,76	4641,61	8260	2310	3,58	3,66
14/03/2021	02/04/2021	Zeta	602	16.382.586	562,59	128,05	2398	1699,49	1,41	1,41
15/03/2021	31/03/2021	Gama	19	572.308	39,95	23,89	107	50,46	2,12	2,12
19/03/2021	08/04/2021	Gama	785	21.304.422	577,11	5991,93	9278	2320,57	4	4,11
20/03/2021	09/04/2021	Gama	323	20.463.249	183,73	3683,71	7338	2228,95	3,29	3,36
22/03/2021	11/04/2021	Gama	134	4.574.087	198,36	346,07	1191	498,23	2,39	2,4
24/03/2021	13/04/2021	Gama	15	4.132.657	45,92	652,91	1411	450,15	3,13	3,15
25/03/2021	14/04/2021	Gama	494	18.537.238	259,7	2947,68	6322	2019,16	3,13	3,18
02/04/2021	15/04/2021	Gama	52	4.734.106	224,04	187,03	760	343,77	2,21	2,21
04/04/2021	24/04/2021	Gama	13	986.937	44,95	211,26	383	107,5	3,56	3,57
06/04/2021	26/04/2021	Gama	30	753.343	42,11	24,39	153	82,06	1,86	1,87
31/05/2021	18/06/2021	Gama	327	4.526.604	191,77	85,92	750	446,1	1,68	1,68

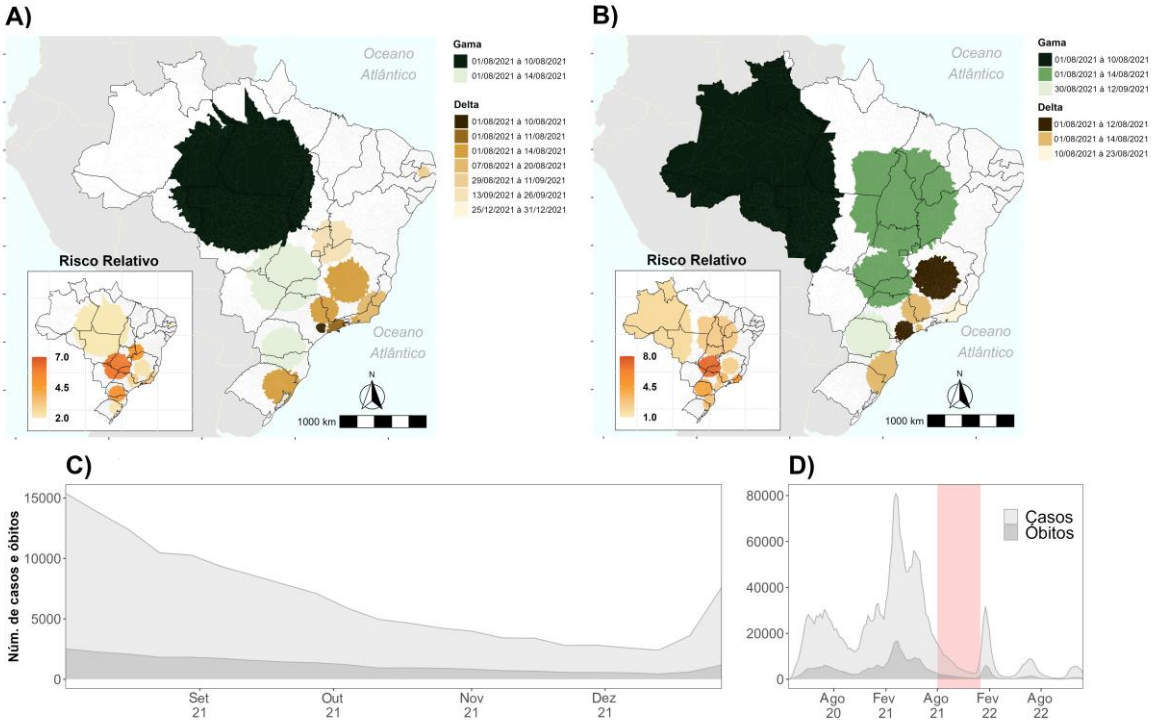
Assim como visto na seção anterior (seção 3.2) a variante Gama teve o seu primeiro *cluster* para casos e óbitos ocorrendo na região Norte, entre os estados de Roraima, Amazonas e Pará. Em seguida, o segundo *cluster* ocorreu entre a divisa do Pará com Tocantins e o terceiro *cluster* ocorreu no extremo oposto, na região Sul do país, abrangendo os estados do Rio Grande do Sul e Santa Catarina. Apenas no Acre e em Rondônia não foram observados *clusters* para casos (e óbitos) neste período. Os riscos relativos mínimos e máximos para casos e óbitos por SRAG-Covid foram, respectivamente, 1,55 (Gama) e 3,90 (Gama); 1,41 (Zeta) e 4,56 (Gama).

3.4 Fase 4: 01/07/2021 à 31/12/2021

Embora o período analisado seja predominado por *clusters* de casos e óbitos atribuídos a Gama, em agosto de 2021 ocorreu a emergência de uma nova variante, a Delta. Esta variante ganhou o espaço até então ocupado pela Gama, porém apresentando riscos relativos menores que a Gama.

Além disso, os *clusters* da variante Gama alcançaram quase 2 mil km para casos e o risco relativo chegou a 6,45 para óbitos ainda em julho. Como ilustrado na figura 5, houve uma redução nos números de casos e óbitos iniciada antes da emergência do primeiro *cluster* da Delta, que se estendeu até a segunda quinzena de dezembro.

Figura 5 – Distribuição espaço-temporal dos *clusters*, seus respectivos riscos relativos para casos(A) e óbitos (B), recorte da série temporal (C) e série temporal completa de casos e óbitos (D) por SRAG-Covid durante o período 01/07/2021 à 31/12/2021.



De acordo com a tabela 5, devido agosto apresentar os valores mais elevados, todos *clusters* de casos e óbitos, com exceção do último, iniciaram no dia primeiro dia de julho (SE 26) e duraram entre 6 e 20 dias. Apenas um *cluster* foi atribuído a Delta e o seu alcance abrangeu 135 municípios apresentando um risco estimado de 3,32. Por fim, foram identificados 9 *clusters* para casos (8 referentes à Gama e 1 à Delta) e 9 para óbitos, todos referentes à Gama.

Tabela 5 – *Clusters* de casos e óbitos de SRAG-Covid para o período 01/07/2021 à 31/12/2021.

SRAG	Início	Fim	Variante	Núm. de municípios	População	Raio	LLR	Observado	Esperado	ODE	Risco Relativo
Casos											
	01/07/2021	07/07/2021	Gama	377	9.893.621	363,69	31,26	546	381,62	1,43	1,43

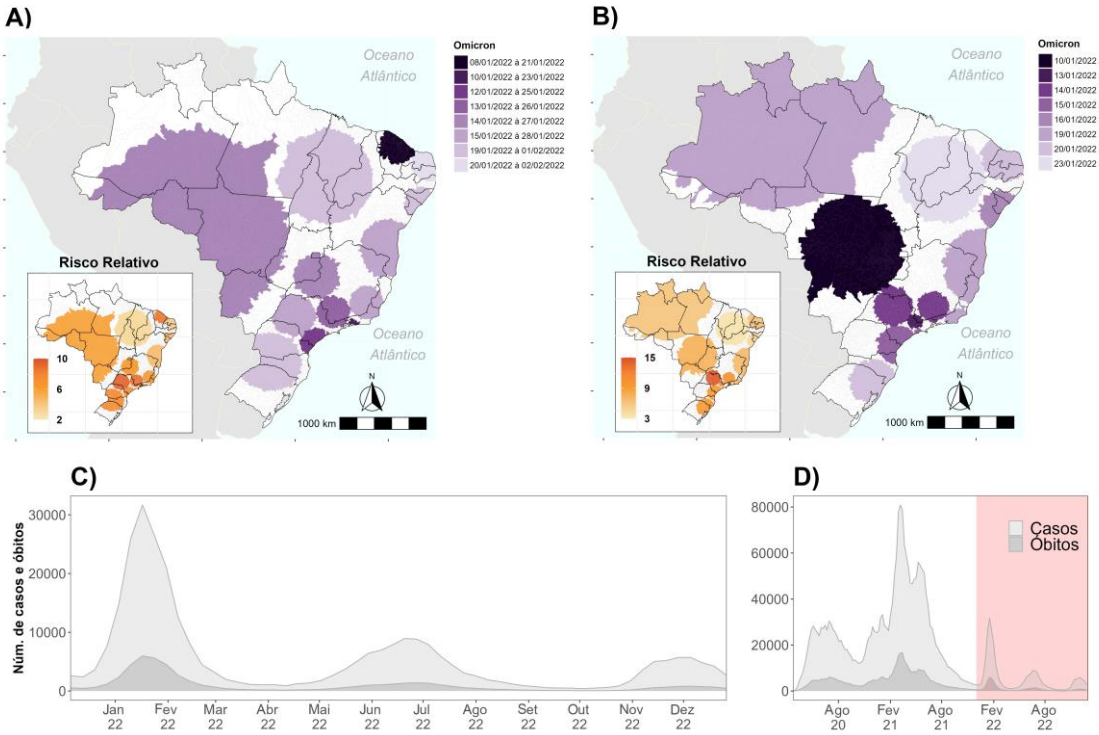
	01/07/2021	08/07/2021	Gama	310	13.487.931	297,31	108,16	987	594,58	1,66	1,66
	01/07/2021	15/07/2021	Gama	350	5.875.464	220,33	73,07	775	485,64	1,6	1,6
	01/07/2021	20/07/2021	Gama	482	21.155.836	1888,4	398,27	3811	2331,51	1,63	1,65
	01/07/2021	21/07/2021	Gama	780	21.320.801	519	11995,4	13382	2467,18	5,42	5,72
	01/07/2021	21/07/2021	Gama	326	19.581.044	186,78	5444,13	8755	2265,86	3,86	3,98
	01/07/2021	21/07/2021	Gama	890	20.046.271	344,67	4547,64	8194	2319,69	3,53	3,63
	01/07/2021	21/07/2021	Gama	653	21.083.456	485,19	2184,35	6334	2439,71	2,6	2,64
	01/08/2021	21/08/2021	Delta	135	17.698.556	201,73	3276,2	6647	2048,02	3,25	3,32
Óbitos											
	01/07/2021	12/07/2021	Gama	163	5.443.309	265,01	19,47	166	98,08	1,69	1,69
	01/07/2021	16/07/2021	Gama	551	18.217.504	486,16	151,33	847	437,67	1,94	1,95
	01/07/2021	19/07/2021	Gama	410	7.014.482	242,7	76,79	399	200,12	1,99	2
	01/07/2021	21/07/2021	Gama	617	21.133.298	280,47	4027,11	4051	666,38	6,08	6,45
	01/07/2021	21/07/2021	Gama	1031	21.313.772	632,59	1479,87	2496	672,07	3,71	3,83
	01/07/2021	21/07/2021	Gama	702	19.923.009	335,16	873,35	1925	628,22	3,06	3,13
	01/07/2021	21/07/2021	Gama	63	10.303.965	115,47	431,63	981	324,91	3,02	3,05
	01/07/2021	21/07/2021	Gama	12	3.076.052	21,73	185,21	341	97	3,52	3,53
	25/07/2021	14/08/2021	Gama	667	19.763.955	843,45	1301,78	2263	623,2	3,63	3,74

Como apresentado na tabela 5, os riscos relativos variaram entre 1, 43 (Gama) e 5, 72 (Gama) para casos; 1,69 (Gama) à 6, 45 (Gama) para óbitos.

3.5 Fase 5: 01/12/2021 à 31/12/2022

Diferente do que visto nos resultados anteriores, este período de análise contempla apenas uma única variante, a Ômicron. A Ômicron emergiu em dezembro de 2021 e seus primeiros *clusters* de casos e óbitos foram detectados no estado de São Paulo no mês seguinte. As áreas com os maiores riscos relativos para casos e óbitos por SRAG-Covid neste período se situam entre os estados de São Paulo e Minas Gerais.

Figura 6 – Distribuição espaço-temporal dos *clusters*, seus respectivos riscos relativos para casos (A) e óbitos (B), recorte da série temporal (C) e série temporal completa de casos e óbitos (D) por SRAG-Covid durante o período 01/12/2021 à 31/12/2022.



A maior onda de casos e óbitos durante esse período ocorreu de janeiro a março conforme apresentado pela figura 6 C. Consequentemente, todos os *clusters*, tanto para casos quanto para óbitos por SRAG-Covid, concentraram-se nesse período, conforme apresentado na tabela 6. Todos os *clusters* tiveram duração de 20 dias.

Tabela 6 – *Clusters* de casos e óbitos de SRAG-Covid para o período 01/12/2021 à 31/12/2022.

SRAG	Início	Fim	Variante	Núm. de municípios	População	Raio	LLR	Observado	Esperado	ODE	Risco Relativo
Casos											
	08/01/2022	28/01/2022	Omicron	41	21.273.329	68,22	8129,24	7732	1171,47	6,6	6,8
	10/01/2022	30/01/2022	Omicron	559	16.140.102	459,72	2953,41	4017	888,8	4,52	4,58
	13/01/2022	02/02/2022	Omicron	324	21.320.135	212,77	6093,18	6632	1174,05	5,65	5,79
	14/01/2022	03/02/2022	Omicron	598	21.308.601	295,63	9473,15	8423	1173,41	7,18	7,42
	14/01/2022	03/02/2022	Omicron	894	21.288.743	351,31	6466,68	6838	1172,32	5,83	5,99
	14/01/2022	03/02/2022	Omicron	601	21.279.968	1392,1	4095,46	5421	1171,84	4,63	4,72
	17/01/2022	06/02/2022	Omicron	625	21.320.828	483,01	3922,59	5314	1174,09	4,53	4,61
	18/01/2022	07/02/2022	Omicron	333	14.110.666	310,24	1160,76	2462	777,04	3,17	3,19
Óbitos											
	19/01/2022	08/02/2022	Omicron	17	3.449.396	26,08	320,92	296	42,62	6,94	6,98
	22/01/2022	11/02/2022	Omicron	598	21.308.601	295,63	4219,28	2829	263,29	10,74	11,33
	22/01/2022	11/02/2022	Omicron	601	21.279.968	1392,1	1153,45	1362	262,94	5,18	5,3
	24/01/2022	13/02/2022	Omicron	60	9.736.943	113,64	1321,2	1035	120,31	8,6	8,76
	25/01/2022	14/02/2022	Omicron	673	21.330.636	429,12	1594,87	1615	263,56	6,13	6,3
	26/01/2022	15/02/2022	Omicron	494	21.066.482	235,63	1748,77	1688	260,3	6,48	6,68
	28/01/2022	17/02/2022	Omicron	758	21.314.436	513,62	744,41	1101	263,36	4,18	4,25

29/01/2022	18/02/2022	Omicron	708	21.142.430	588,81	695,19	1062	261,24	4,07	4,13

O risco relativo dos *clusters* de casos por SRAG-Covid variou entre 3,19 e 7,42, enquanto os de óbitos ficaram entre 4,13 e 11,33.

4 Discussão

A evolução temporal dos casos e óbitos pela COVID-19 foi caracterizada por alguns picos bem definidos, porém analisar apenas na dimensão "tempo" pode omitir as interações espaciais existentes entre os municípios. Os municípios Brasileiros são independentes, desta forma, durante a pandemia, impuseram medidas de restrição e relaxamento em diferentes momentos, facilitando a circulação do vírus (Kortessis *et al*., 2020).

De acordo com Ministério da Saúde (2021a), São Paulo foi o primeiro estado a apresentar um caso (26 de fevereiro, SE 9) e um óbito (17 de março, SE 12) confirmados pela COVID-19. Na semana epidemiológica seguinte, foram notificados casos confirmados na Bahia, no Distrito Federal, no Espírito Santo e no Rio de Janeiro. Roraima foi o último estado a apresentar um caso (22 de março, SE 13) e Tocantins foi o último a notificar um óbito confirmado por COVID-19 (15 de abril, SE 16). Embora Castro *et al*. (2021) evidencie que em todos os estados o primeiro óbito tenha ocorrido em menos de um mês do primeiro caso, os *clusters* só foram identificados em abril, sendo o primeiro *cluster* de óbitos ocorrendo apenas um dia após o primeiro *cluster* de casos.

Candido *et al*. (2020) aponta que a epidemia no Brasil foi impulsionada principalmente pelas variantes B.1.1.28 e B.1.1.33, corroborando com o apresentado na 1 (A). Em maio de 2020 (SE 19), o Brasil ultrapassou a marca de 100 mil casos de COVID-19 e mais de 20 mil mortes por SRAG-Covid, se tornando o terceiro país com mais casos confirmados no mundo, atrás apenas dos Estados Unidos e da Rússia. Segundo Castro *et al*. (2021) a resposta do Brasil diante a pandemia foi uma combinação de erros, o qual envolve a falta de uma estratégia nacional coordenada e promoção de tratamento medicamentoso sem evidências científicas, como foi o caso da cloroquina. Neste período, os primeiros *clusters* atribuídos à variante B.1.1.33 estavam na emergindo região Norte e Nordeste, enquanto a região Sul e Sudeste, eram mais afetadas pela B.1.1.28.2.

Em julho, o número de casos passou a cair e permaneceu praticamente estável até início de novembro de 2020 (SE 27 a 42), onde a variante Zeta, originada da variante B.1.1.28, começou a elevar o risco de internação e óbitos, especialmente na região Sudeste, como mostra a figura 2.

Posteriormente, passou a co-circular outra variante de origem nacional, a Gama (WHO, 2021), porém esta tornou-se predominante no país, apenas em dezembro, como mostra a figura 1.

Segundo o Fiocruz (2021a), ao longo das 44 semanas epidemiológicas de 2020 foram contabilizados 7.714.819 casos e 195.742 óbitos. Conforme os dados apresentados pelo Ministério da Saúde (2021), nesta última semana (SE 53), foi identificado o primeiro *cluster* atribuídos a variante Gama. Faria *et al*. (2021), Oliveira, Lippi e Henry (2021), Souza *et al*. (2021) apontam que a emergência desta variante resultou em uma nova onda de hospitalizações e óbitos, refletindo o que é mostrado na figura 3. Essa nova onda de transmissão coincidiu com o relaxamento de medidas de restrição, período de férias e festas de fim de ano (Fiocruz, 2022).

Em janeiro de 2021, sob predominância da variante Gama, a situação da pandemia no Brasil continuou a piorar, alcançando os números mais elevados de hospitalizações e óbitos por SRAG-Covid para todo período em estudo. Corroborando com os resultados obtidos, o *cluster* de casos identificado entre os dias 31/12/2020 e 13/01/2021, além de possuir um risco relativo de quase 9 vezes às demais regiões, algumas semanas depois resultou no *cluster* de óbitos com área ainda maior. Um fator que contribuiu para inflacionar esses números foi à crise causada pela falta de oxigênio, resultando em mortes por asfixia nos hospitais.

Embora a vacinação de pessoas idosas tenha sido iniciada em janeiro, apenas em março o processo de vacinação ganhou mais velocidade (Fiocruz, 2022). Ainda em março, o número de casos e óbitos alcançaram os patamares mais elevados durante toda a pandemia, os riscos relativos dos *clusters* de casos ficaram entre 1,86 e 3,03 e os *clusters* de óbitos entre 1,80 e 4,35. O mês de junho apresentou novamente um crescimento no número de casos e óbitos. Como visto na seção 3.3, os *clusters* identificados foram atribuídos à Gama. Dessa vez, diferente do que se observado até então, com a retomada das atividades presenciais e o efeito protetor da vacinação, a proporção de idosos acometidos pela COVID-19 diminuiu em relação àqueles que estão em idade economicamente ativa (Fiocruz, 2022).

Conforme constatado pela Castro *et al*. (2021), a falha em evitar essa a propagação da Gama facilitou o surgimento de novas variantes. Assim, após causar muitas internações e óbitos, a variante Gama foi sucedida pela Delta. Por mais que essa nova variante estivesse emergindo e alguns *clusters* apresentem riscos relativos elevados comparados às demais regiões (tabela 5), o país estava passando por uma redução quanto à gravidade dos casos, consequentemente, redução no número de casos por SRAG-Covid.

O último período analisado compete exclusivamente a uma única variante, a Ômicron. Corroborando com Fiocruz (2022), a transmissão dessa variante começou no último mês de 2021, novamente coincidindo com o período de festividades e férias de final de ano. Os números máximos de casos e óbitos por SRAG-Covid causados pela Ômicron foram parecidos com os valores máximos observados nos primeiros meses da pandemia, porém a curva foi mais acentuada e decresceu rapidamente, ou seja, ocorreu em um intervalo de tempo menor que as "ondas" anteriores. A Ômicron ainda apresentou dois outros momentos de aumento de casos (e óbitos), junho/julho (BA.2) e dezembro, cada uma delas atribuída a diferentes mutações da mesma variante, porém, em intensidade menor.

Apesar da metodologia *SaTScan*^{TM} ser consolidada na epidemiologia, não foi encontrado nenhum artigo em acervo público, que aplicou essa técnica alinhada com a espacialização das diferentes variantes circulantes visando a compreensão da dinamicidade da pandemia no Brasil, bem como a identificação de áreas de alto risco para casos e óbitos por SRAG-Covid.

5 Limitações

Uma das limitações da estatística *Scan* de *Kulldorff* é utilização de janelas circulares para a detecção dos *clusters*, podendo, dessa forma, agrupar municípios com baixo riscos que estão cercados ou próximos aos municípios com alto risco. Visando minimizar esse problema, foi considerado um limite para o tamanho que os *clusters* pudessem alcançar.

Outra limitação encontrada foi, ao considerar como *cluster* apenas o agrupamento de municípios com maior risco, descartou-se a possibilidade desses mesmos municípios voltarem a formar novos *clusters* para o mesmo período da análise. Para mitigar esse problema, as análises foram separadas em 5 fases distintas, correspondendo aos períodos de prevalência das variantes. Apesar dos benefícios dessa abordagem, em casos em que ao menos uma variante apresenta múltiplos picos, a análise *Scan* tende a encontrar apenas o "pico mais alto", desconsiderando os possíveis *clusters* que seriam formados pelos "picos menores". Um exemplo disso ocorreu com a variante Ômicron, os *clusters* identificados são apenas do período em que foram observados os números de casos e óbitos mais altos.

Outra limitação é quanto as amostras de materiais genéticos coletados para a identificação das variantes que estão em circulação. Municípios que carecem de infraestrutura adequada para coleta, armazenamento e análise dessas amostras, além de ficarem dependentes de outros municípios, podem simplesmente deixarem de coletar. Ainda como as informações disponibilizadas pela Rede

Genômica utiliza sequências depositadas nos servidores *GISAID*, podem não haver pesquisadores em todas as regiões do país, como o Tocantins, que tendo o seu primeiro caso registrado no mês de março de 2020, a sua primeira amostra sequenciada ocorreu apenas 4 meses depois. Assim, Fiocruz (2021b) ressalta que pode haver viés de seleção com a inclusão de investigação genômica de casos inusitados, rastreio de contactantes e seleção de amostras através de protocolo de inferência de RT-PCR em tempo real para a detecção de potenciais variantes de preocupação.

Por fim, é importante considerar que essa metodologia envolve na definição prévia de alguns parâmetros que ao serem modificados podem produzir resultados distintos.

6 Conclusão

O Brasil é um país de contrastes e possui dimensões continentais onde múltiplas epidemias de COVID-19 aconteceram simultaneamente. Além disso, a falta do emprego coordenado das políticas públicas de saúde e sanitárias que como resultado favoreceram a co-circulação de diferentes variantes. Essas múltiplas epidemias puderam ser vistas através do uso da estatística *Scan*, que não apenas conseguiu identificar as ocorrências dos *clusters* espaço-temporais, mas estimou os riscos relativos às demais regiões que não pertencem ao foco epidêmico. De modo geral, este método se mostrou eficaz e apropriado para detectar áreas de altos riscos de casos e óbitos por SRAG-Covid, independentemente da variante circulante. Uma das principais limitações encontrada é que os municípios que formaram *clusters* só podem voltar formar novos *clusters* se forem analisados em períodos distintos. De toda forma, diante das rápidas mutações do vírus Sars-Cov-2, é importante e necessário o aprimoramento contínuo da rede de vigilância genômica e das técnicas de sequenciamento genético, pois estas podem contribuir para a detecção precoce dos vírus circulantes, para o diagnóstico, tratamento e diretrizes de enfrentamento à pandemia. Ademais, o uso de dados contribui para políticas públicas, facilita o processo de decisão por gestores, bem como, possibilita compartilhamento de informações de modo mais amplo entre a comunidade científica.

Referências

BURKI, T. **Who ends the COVID-19 public health emergency**. The Lancet Respiratory Medicine, Elsevier, 2023.

CANDIDO, D. S. et al. **Evolution and epidemic spread of Sars-Cov-2 in Brazil**. Science, American Association for the Advancement of Science, v. 369, n. 6508, p. 1255–1260, 2020.

CASTRO, M. C. et al. **Spatiotemporal pattern of COVID-19 spread in Brazil**. Science, American Association for the Advancement of Science, v. 372, n. 6544, p. 821–826, 2021.

DWASS, M. **Modified randomization tests for nonparametric hypotheses**. The Annals of Mathematical Statistics, JSTOR, p. 181–187, 1957.

FARIA, N. R. et al. **Genomics and epidemiology of the p. 1 Sars-Cov-2 lineage in Manaus, Brazil**. Science, American Association for the Advancement of Science, v. 372, n. 6544, p. 815–821, 2021.

FIOCRUZ. **Boletim observatório COVID-19 Fiocruz: Balanço da pandemia no Brasil em 2020**. Fiocruz, 2021.

FIOCRUZ. **Dashboard Rede Genômica**. Vigilância genômica do Sars-Cov-2 no Brasil. Fiocruz, 2021. Disponível em: http://www.genomahcov.fiocruz.br/dashboard/. Online; Acesso em: 02 fev. 2023.

FIOCRUZ. **Boletim observatório COVID-19 Fiocruz - boletim especial: Balanço de dois anos da pandemia COVID-19**. Fiocruz, 2022.

IBGE. **Estimativas da população. 2021**. Disponível em: https://www.ibge.gov.br/estatisticas/sociais/populacao/9103-estimativas-de-populacao.html?=&t=resultados. Online; Acesso em: 18 dez. 2022.

KORTESSIS, N. et al. **The interplay of movement and spatiotemporal variation in transmission degrades pandemic control**. Proceedings of the National Academy of Sciences, National Acad Sciences, v. 117, n. 48, p. 30104–30106, 2020.

KULLDORFF, M. **Spatial scan statistics: models, calculations, and applications**. [S.l.]: Springer, 1999.

KULLDORFF, M. S. f. v. . **SaTScan**, 2022.

MINISTÉRIO DA SAÚDE. **Boletim epidemiológico 5 - Doença pelo Coronavírus 2019 Ampliação da Vigilância, Medidas não Farmacológicas e Descentralização do Diagnóstico Laboratorial**. Ministério da Saúde, Secretaria de Vigilância em Saúde. – Brasília: Ministério da Saúde: Secretaria de Vigilância em Saúde, 2020. v. 05.

MINISTÉRIO DA SAÚDE. **Boletim epidemiológico 17 - COE COVID-19**. Ministério da Saúde, Secretaria de Vigilância em Saúde. – Brasília: Ministério da Saúde: Secretaria de Vigilância em Saúde. Centro de Operações e de Emergência em Saúde Pública | Doença pelo novo Coronavírus 2019, 2021. Atualizado em 07-05-2021.

MINISTÉRIO DA SAÚDE. B**oletim epidemiológico especia**l. Semana Epidemiológica 53, v. 53, p. 76, Ministério da Saúde, Secretaria de Vigilância em Saúde. – Brasília: Ministério da Saúde. 2021.

MINISTÉRIO DA SAÚDE. **Guia de vigilância epidemiológica Emergência de saúde pública de Importância nacional pela Doença pelo coronavírus 2019 – COVID-19**. Ministério da Saúde, Secretaria de Vigilância em Saúde. – Brasília: Ministério da Saúde: Secretaria de Vigilância em Saúde. Departamento de Análise em Saúde e Doenças não Transmissíveis, 2021.

OLIVEIRA, M. H. S. de; LIPPI, G.; HENRY, B. M. **Sudden rise in COVID-19 case fatality among young and middle-aged adults in the south of Brazil after identification of the novel B.1.1.28.1 (P.1) Sars-Cov-2 strain: analysis of data from the state of Parana**. MedRxiv, Cold Spring Harbor Laboratory Press, p. 2021–03, 2021.

SOUZA, U. et al. **Detection of potential new Sars-Cov-2 gamma-related lineage in Tocantins shows the spread and ongoing evolution of P.1 in Brazil**. bioRxiv, Cold Spring Harbor Laboratory, p. 2021–06, 2021.

WHO. **Epidemiological update: Occurrence of variants of Sars-Cov-2 in the Americas** - 20 January 2021.

WHO. **WHO Coronavirus (COVID-19) Dashboard. 2023**. Online; Acesso 17 jun. 2023.

7 ARTIGO 2 - CENÁRIOS CONTRAFACTUAIS EM ESTUDOS ECOLÓGICOS DE SÉRIESTEMPORAIS UTILIZANDO A REDE NEURAL *LSTM*

RESUMO

Background. Os estudos tradicionais referentes aos modelos de respostas potenciais na epidemiologia apresentam como cenário ótimo a hipotética e lúdica ideia de se observar o efeito da exposição por um fator de interesse simultaneamente com a não exposição desse mesmo fator. Diante da impossibilidade deste cenário, algumas abordagens foram desenvolvidas para que o efeito causal possa ser mensurado de modo mais plausível. Em estudos experimentais e em estudos observacionais analíticos as metodologias já se encontram bem definidas. Entretanto, em estudos ecológicos, ainda há muito que se evoluir. Visto que políticas públicas são geralmente projetadas ao nível de população, faz-se necessário o desenvolvimento de abordagens para que a estimação de efeitos contrafactuais possa ser utilizada com a plausibilidade, respeitando o histórico do evento de interesse. Assim, esse artigo propõe um método para a estimação de cenários contrafactuais em estudos ecológicos de séries temporais usando a rede convolucional *LSTM (Long Short Term Memory)*.

Métodos. Foi utilizado a rede *LSTM* para estimação do efeito contrafactual. Para isso, definiu-se como exemplo de estudo a mensuração da introdução da variante Gama no número de hospitalizações por SRAG-Covid no Brasil baseando-se na distribuição temporal das variantes que a antecederam. Foram utilizados dados de casos hospitalizados para o período de 26/02/2020 a 31/07/2021 por Síndrome Respiratória Aguda Grave (SRAG) com classificação final de SRAG por COVID-19 no Brasil. da. Os valores estimados foram subtraídos dos valores observados a fim de se obter a carga de hospitalizações.

Resultados. O uso da rede neural *LSTM* alinhada às técnicas estatísticas resultou na criação de um grupo de comparação gerado pelo mesmo processo estocástico que à curva observada, superando assim o problema de simultaneidade e equivalência observacional. O modelo contrafactual estimou uma curva que se distancia significativamente das taxas de hospitalizações observadas, indicando que, embora a introdução de uma nova linhagem causasse um esperado aumento de casos hospitalizados, a variante Gama fez isso em uma escala muito maior.

Conclusões. Essa metodologia é particularmente útil para cálculos de cargas de doenças, pois possibilita a comparação de um cenário observado com um cenário hipotético de repetição de padrões já vivenciados.

Palavras-chave: SRAG-Covid, Contrafactual, Modelo de respostas potenciais, *LSTM*.

1 Introdução

Estudos não-randomizados que propunham realizar estimações de efeitos causais foram muito criticados, pois se acreditava que apenas experimentos randomizados poderiam produzir estimações úteis, mesmo que muitos achados científicos tenham sido produzidos sem uso de randomizações (Rubin, 1986). Além disso, há razões éticas que impedem o uso de randomização em tratamentos, pois isso poderia implicar em prejuízo a saúde dos participantes do experimento. Outro fato relevante é que os delineamentos de estudos tradicionais da epidemiologia como caso-controle e coortes, apesar de serem capazes de apresentar altos níveis de evidência científica, pecam para fazer o mesmo a níveis populacionais Soumerai, Starr e Majumdar (2015).

Diante dessas limitações e da importância de se mensurar efeitos causais ao nível populacional na epidemiologia, como exemplos, a introdução de uma nova doença, campanhas de vacinação ou ainda práticas não-farmacêuticas, parece não haver uma conformidade quanto as metodologias empregadas para a estimação das relações causais. Ainda, as metodologias mais utilizadas esbarram na ausência de uma efetiva capacidade de determinar o grupo controle, ou seja, o grupo de comparação não-exposto ao evento de interesse. Especificamente em dados ordenados no tempo, séries temporais, concluir que a mudança observada na trajetória da curva se deu somente pela ocorrência do evento em estudo é uma afirmação muito delicada de ser feita, pois além da dificuldade de se obter um pareamento para comparação, isolar os efeitos externos é muitas vezes impossível. Assim, uma das possíveis abordagens, é a utilização de cenários contrafactuais.

A ideia de contrafactualidade não é nova e remete-se aos tempos de Aristóteles. Desde lá, esse conceito evoluiu a ponto de receber diferentes interpretações. Apenas no século XVII, por John Graunt, através de um estudo sobre diferenças no número de óbitos entre sexo, cidade e campo, que o pensamento contrafactual passou a ser aplicado na epidemiologia (Barata, 1997). No século seguinte, Hume (2013) expandiu o conceito o de pensamento contrafactual adicionando a ideia de que se a causa não tivesse acontecido, o evento também não aconteceria. Assim, Hume (2013) ao abranger um aspecto de condicionalidade, propiciou a outros pensadores, a possibilidade de relacionar noções de probabilidade à contrafactualidade. Williamson (2007) resume, em duas principais regras, como a abordagem contrafactual pode ser vista como condicional por meio da redução das relações de causas e efeitos à condicionais subjuntivas em que X é uma causa direta de Y se e só: se X ocorre, Y ocorre (ou a sua probabilidade de ocorrência aumenta significativamente); e se X não ocorrer, Y não ocorre (ou a sua probabilidade de ocorrência diminui significativamente). Holland (1986) propôs um modelo, que ao assegurar que tudo tivesse permanecido constante, incluindo espaço e tempo, um fator torna-se uma causa se o resultado não tivesse ocorrido na ausência desse mesmo fator. Nesse modelo, Y mensura o efeito de uma causa observada após a

exposição, fazendo isso através da diferença de vetores (Y_t, Y_c) do efeito causal t (tratamento ou exposição) e c (controle ou não-exposição) para cada unidade observacional.

Apesar de muitas abordagens sobre efeitos causais, a significância estatística de uma correlação é insuficiente para determinar uma relação causal (Szwarcwald; Castilho, 1992). Ainda Durlauf e Blume (2016) evidencia a existência de três principais desafios quanto a interpretação causal a partir da análise empírica dos dados, são eles: simultaneidade, equivalência observacional e da identificação. O primeiro problema se refere ao fato de que ao menos uma variável explicativa no modelo de regressão linear múltipla é determinada em conjunto com a variável dependente; o segundo problema reflete a dificuldade de identificar a direção do efeito causal, visto que os dados em si não revelam qual é a direção causal correta (Fonseca; Sánchez-Rivero, 2020); o terceiro problema é estimar relações causais a partir de um número menor de observações empíricas (Fonseca; Sánchez-Rivero, 2020).

Em vista de enfrentar esses desafios, este artigo propõe o emprego de uma técnica enquadrada no contexto de inteligência artificial, a rede neural *Long Short Term Memory* (*LSTM*). Essa rede tem a capacidade de "aprender" dependências de curto e longo prazo. Assim, ao "ensinar" padrões que antecedem ao evento de interesse (causa), a rede, se bem treinada, é capaz de produzir estimativas semelhantes às quais fora apresentada, construindo não apenas um grupo de comparação equivalente, ou seja, respeitando a ideia de não alteração do tempo e espaço mencionados por Holland (1986), mas também constrói um cenário contrafactual onde a causa não teria acontecido ou se tivesse acontecido, o seu desfecho não ocorreria em magnitudes diferentes das observadas no histórico desse evento.

Em frente a tamanhos desafios, este artigo visa apresentar uma metodologia para estimação de efeitos causais por meio do pensamento contrafactual, carregando consigo o peso de minimizar os problemas de simultaneidade, equivalência observacional e estimação. A metodologia proposta abrange conceitos de epidemiologia, filosofia, estatística temporal e inteligência artificial. Este último, base para a construção de um grupo de comparação, também tem como responsabilidade, mostrar que técnicas mais recentes podem e devem ser aplicadas a problemas antigos da epidemiologia.

2 Materiais e métodos

2.1 Série temporal

Uma série temporal é qualquer conjunto de observações ordenadas no tempo (Morettin; Tokoi, 2006). As séries podem ser discretas ou contínuas e a análise da série temporal pode ser feita no

domínio do tempo ou no domínio das frequências. Uma série temporal pode ser empregada visando investigar o processo gerador da série, prever valores a curto ou longo prazo, descrever comportamentos como a verificação de tendência, sazonalidade e/ou ciclos e buscar periodicidades (Morettin; Toloi, 2006). Um modelo clássico para séries temporais supõe que {Z1 ..., Zn} pode ser escrita como

$$Z_t = T_t + S_t + a_t = 1, 2, \ldots, n$$

em que Z_t representa a série temporal, T_t um componente de tendência, S_t um componente de sazonalidade e a_t é um componente aleatório. Trazendo para o contexto da epidemiologia, uma série temporal pode ser usada para observação de um evento de saúde ao longo do tempo.

2.2 Rede neural artificial

O uso de redes neurais tem sido amplamente adotado, e entre as várias arquiteturas disponíveis, o perceptron multi-camadas (MLP) se destaca como uma das bases fundamentais para diversas aplicações. Proposto por Rosenblatt (1961), o algoritmo perceptron consiste em um conjunto de neurônios conectados formando uma rede, onde cada neurônio recebe sinais de entrada e gera sinais de saída em resposta. Proposto por McCulloch e Pitts (1943), o primeiro modelo matemático para um neurônio pode ser expresso por:

$$u_k = \sum_{j=1}^{m} w_{kj} x_j,$$

e

$$y_k = \rho(u_k + b)$$

em que $x_1, x_2, \ldots x_m$ são as entradas, $w_1, w_2, \ldots w_m$ os pesos sinápticos correspondentes do neurônio *k*, u_k é a combinação linear das entradas com os pesos sinápticos, b_k é a tendência, $\rho(.)$ é função de ativação e y_k é a saída do neurônio.

2.3 Rede *LSTM*

A rede *LSTM* é um tipo de rede neural recorrente (RNR) capaz de aprender dependências de longo e curto prazo através da retenção de informações passadas, em outras palavras, é capaz de incorporar a dependência temporal entre as observações (Hochreiter; Schmidhuberr, 1997). Quanto ao funcionamento da rede *LSTM*, Géron (2019) descreve que o vetor de entrada atual x_t e a memória de curto prazo anterior $h_{(t-1)}$ são fornecidas para quatro diferentes camadas que estão conectadas, porém, servem a um propósito diferente:

- Camada principal $g_{(t)}$: responsável por analisar as entradas de x_t e a memória de curto prazo $h_{(t-1)}$;
- Forget Gate $f_{(t)}$: responsável por definir as informações que serão retidas;
- Input Gate $i_{(t)}$: responsável por definir as informações que incorporadas a memória de longo prazo;
- Output Gate $o_{(t)}$: responsável por controlar as partes do estado de longo prazo devem ser lidas tanto para $h_{(t)}$ e $y_{(t)}$ e exibidas no intervalo de tempo *t*

A rede é composta de uma camada de entrada, de uma ou mais camadas intermediárias e a camada de saída. Não há necessariamente uma configuração única para as redes neurais, sendo o tamanho da sequência de entrada, de saída, quantos dias a frente o modelo estimará, bem como os tipos de camadas de neurônios e suas respectivas quantidades, dependentes, principalmente, do objetivo proposto para o estudo.

2.4 Dados

Os dados utilizados no presente estudo se referem ao número de casos hospitalizados para o período de 26/02/2020 a 31/07/2021 por Síndrome Respiratória Aguda Grave (SRAG) com classificação final de SRAG por COVID-19. Os casos são notificados compulsoriamente no Sistema de Informação da Vigilância Epidemiológica da Gripe (SIVEP-Gripe) e estão disponibilizados na plataforma *OpenDataSUS*. São considerados como casos de hospitalização por COVID-19 qualquer registro hospitalar com a classificação final de SRAG por COVID-19, sendo essa classificação sendo feita baseada em critérios laboratoriais, clínico-epidemiológicos, clínicos ou clínico-imagem para encerramento do registro. Foram também utilizados a estimativa do tamanho populacional informada pelo IBGE e as variantes predominantes foram obtidas no Fiocruz (2021). Por fim, o modelo recebe como variável de entrada (e saída) a taxa de hospitalização diária por SRAG-Covid por 100 mil pessoas.

2.5 Construção do cenário contrafactual

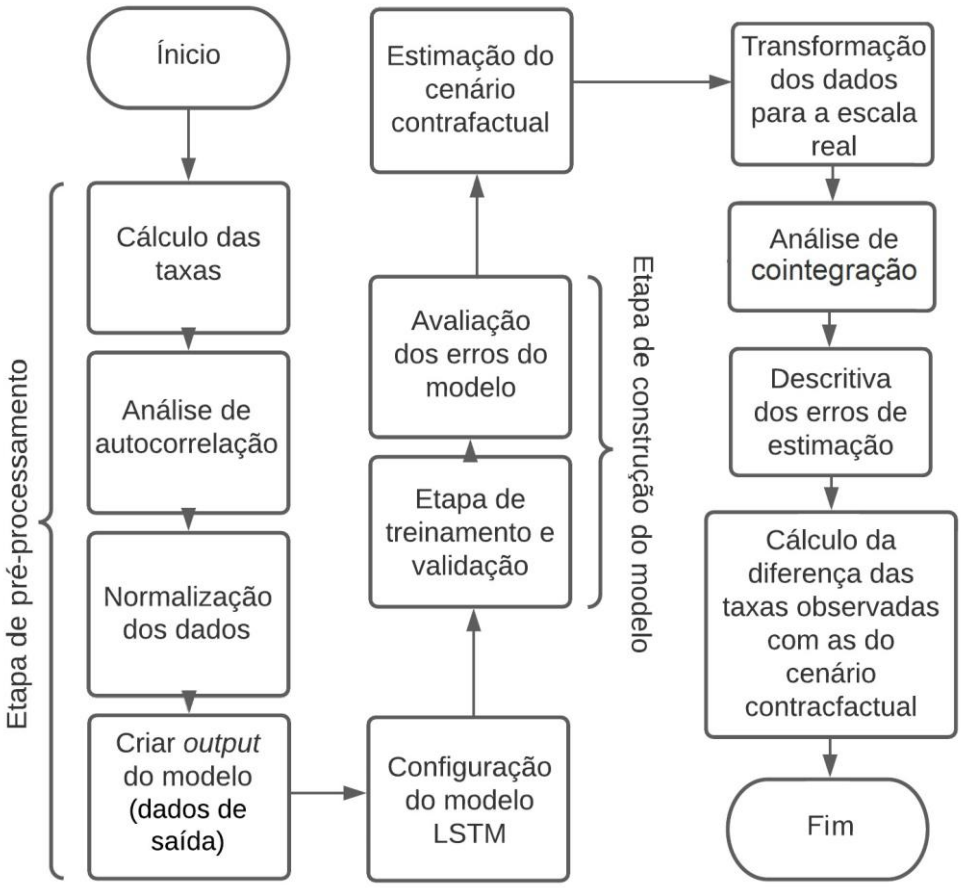
Segundo Rubin (1974), o experimento ideal de um modelo de respostas potenciais é que o objeto de estudo, geralmente um indivíduo, precisa ser observado sob exposição do fator de interesse, bem como a não exposição no mesmo tempo. Desta forma, um fator é uma causa se o resultado não ocorresse na ausência desse fator, assegurando que as demais condições, incluindo espaço e o tempo, sejam mantidas. Matematicamente essa ideia pode ser expressa como:

Não há efeito causal: $Y^{a=1} \neq Y^{a=0}$

Há efeito causal: $Y^{a=1} = Y^{a=0}$

A figura 1 segmenta em 8 passos a construção de um cenário contrafactual em um estudo ecológico de séries temporais, são eles:

Figura 1 – Fluxograma para estimação contrafactual usando a rede *LSTM*.



Passo 1: Preparação dos dados

É necessário que os dados estejam disponíveis para o período que antecede o evento de interesse, bem como durante o seu acontecimento. Apesar de não haver um número mínimo de observações, é importante que o comprimento da série temporal seja grande suficiente para, ao ser dividida em treino e validação, mantenha as características originais da série, como exemplo, a autocorrelação. O particionamento dos dados é feito, primeiramente, separando o período para o qual serão geradas estimativas contrafactuais do restante dos dados, essa partição recebe o nome de conjunto de teste. Os dados que não compõe o conjunto de teste formam outros dois grupos, de treinamento e outro de validação, que são mutuamente exclusivos e separados aleatoriamente, geralmente nas proporções 60%/40%, 70%/30% e 80%/20%. As observações atribuídas ao conjunto de teste são omitidas das etapas de treinamento e validação, pois a rede *LSTM* precisa captar apenas os padrões anteriores ao evento de interesse. Em modelos preditivos, o conjunto teste desempenha a

função de verificar se o modelo proposto apresenta boa capacidade de generalização mesmo em padrões desconhecidos. Como esta metodologia visa a estimação do efeito causal, não são necessariamente esperadas pequenas diferenças entre os dados observados e as estimativas contrafactuais. Todavia, é importante definir previamente a hipótese do cenário contrafactual.

Passo 2: Transformação dos dados

Conforme Nayak, Misra e Behera (2014), a efetividade de qualquer algoritmo de aprendizado depende fortemente do método de normalização utilizado. Sabe-se que as observações dispostas em séries temporais podem assumir um amplo leque de valores, portanto é importante que os valores das entradas sejam proporcionais aos limites das funções de ativação para acelerar o processo de aprendizagem (Bhanja; Das 2018). Existem diversos métodos de transformação de dados, alguns deles são:

- **Normalização MiniMax:** Esse método limita a variação dos dados ao intervalo [0, 1] ou [−1, 1] e converte o valor x em x_{norm} e é expressa pela fórmula abaixo:

$$x_{norm} = \frac{(x - x_{min})}{(x_{max} - x_{min})}$$

em que x_{norm} é o valor normalizado, x é o valor da entrada, x_{min} é o valor mínimo da variável de entrada e x_{max} é o valor máximo das variáveis de entrada. Os valores mínimos e máximos são referentes a todo período de estudo.

```
minimax <- function(x) {
        min_val <- min(x, na.rm = T)
        max_val <- max(x, na.rm = T)
        transformed <- (x - min_val)/(max_val - min_val)
        return(transformed)
}
```

- **Normalização do escore Z:** Esse método converte as observações para um intervalo entre 0 e o desvio padrão do atributo. Nesse método, o valor x é convertido em x_{norm} usando a fórmula:

$$x_{norm} = \frac{x - \mu(x)}{\sigma(x)}$$

onde $\mu(x)$ e $\sigma(x)$ é a média e o desvio-padrão das observações, respectivamente.

```
z_score <- function(x) {
        mean_val <- mean(x, na.rm = T)
```

```
        std_dev <- sd(x, na.rm = T)
        transformed <- (x - mean_val)/std_dev
        return(transformed)
}
```

- **Normalização pela mediana:** Nessa técnica as observações são normalizadas pela mediana. O valor normalizado x_{norm} pode ser calculado da seguinte fórmula:

$$x_{norm} = \frac{x}{mediana(x)}$$

```
mediana <- function(x) {
        median_val <- median(x, na.rm = T)
        transformed <- x/median_val
        return(transformed)
}
```

- **Normalização Sigmoide:** Essa abordagem limita a variação dos dados ao intervalo [0,1]. A conversão do valor x em:

$$x_{norm} = \frac{x}{1 - e^{-x}}$$

```
sigmoid <- function(x) {
        transformed <- 1/(1 + exp(-x))
        return(transformed)
}
```

- **Normalização pela Tangente Hiperbólica:** Essa transformação limita a variação dos dados ao intervalo [−1,1] e é expressa pela seguinte fórmula:

$$x_{norm} = 0{,}5\left[tanh\left[\frac{0{,}01(x - \mu(x))}{\sigma(x)}\right] + 1\right]$$

onde $\mu(x)$ e $\sigma(x)$ é a média e o desvio-padrão das observações, respectivamente.

```
tanhip <- function(x) {
        mean_val <- mean(x, na.rm = T)
        std_dev <- sd(x, na.rm = T)
        transformed <- 0.5*(tan(0.01*(x - mean_val)/std_dev)+1)
        return(transformed)
}
```

Passo 3: Definir a função de ativação

De acordo com Haykin (2009), a função de ativação $\varphi(v)$ tem por objetivo limitar a amplitude da camada de saída, geralmente, definido em um intervalo fechado [0,1] ou [−1,1]. Ainda,

Gomes e Ludermir (2021) enfatiza que a função de ativação influencia tanto na complexidade, no desempenho e na convergência do algoritmo de aprendizagem. Algumas das funções de ativação mais conhecidas são:

- **Limiar ou degrau:** A saída da última camada pode assumir apenas valores binários.

$$\varphi(v) = \begin{cases} 1 \; se \; v > 0 \\ 0 \; se \; v < 0 \end{cases}$$

- **Sigmóide ou logística:** Uma das funções de ativações mais utilizadas, assume valores entre 0 e 1.

$$\varphi(v) = \frac{1}{1 + e^{-av}}$$

  sendo a o parâmetro de inclinação da curva. Uma importante característica dessa função de ativação é que ela é diferenciável, enquanto a função de ativação limiar não é.

- **Tangente hiperbólica:** Semelhante à função de ativação sigmóide, porém com intervalo [-1,1].

$$\varphi(v) = \frac{1 - e^{-2v}}{1 + e^{-2v}}$$

- **ReLU:** Essa função se diferencia da função de ativação linear φ(v) = v que retorna o próprio valor, por retornar 0 para valores negativos.

$$\varphi(v) = \begin{cases} v \; se \; v > 0 \\ 0 \; se \; v < 0 \end{cases}$$

O bloco abaixo apresenta o código para a construção do modelo com duas camadas *LSTM* usando a função de ativação sigmóide:

```
model <- keras_model_sequential()
# arquitetura da rede
model %>%
        layer_LSTM(
          name = "LSTM1",
          units = unit_LSTM1,
          input_shape = c(timesteps, dim(df_train_x)[3]),
          dropout = dropout_rate1,
          recurrent_dropout = recurrent_dropout1,
          return_sequences = TRUE
        ) %>%
        layer_LSTM(
```

```
    name = "LSTM2",
    units = unit_LSTM2,
    dropout = dropout_rate2,
    recurrent_dropout = recurrent_dropout2,
    return_sequences = TRUE
  ) %>%
  layer_dense(
    name = "output",
    activation = "sigmoid",
    units = 1
  )
```

onde, o termo *unit* refere-se ao número de unidades de memória (neurônios) presentes na camada. Esses neurônios têm conexões recorrentes, permitindo que as informações fluam por várias etapas de tempo. Essa característica é crucial para o funcionamento das *LSTM*, pois lhes permite capturar dependências sequenciais em dados de séries temporais ou em outras tarefas que envolvam sequências. Além disso, o *df_train_x* representa o tensor contendo os dados de entrada, *dropout_rate* (1 e 2) as taxas de *dropout* aplicado apenas nas camadas densas que conectam a saída da *LSTM* às camadas de saída da rede e *recurrent_dropout* (1 e 2) são as taxas de *dropout* aplicadas nas conexões recorrentes da *LSTM* durante o treinamento. O *dropout* é uma técnica de regularização. Goodfellow (2016) define "regularização" como o ato de modificar o algoritmo de aprendizagem objetivando minimizar os erros de generalização, porém sem afetar o erro de treinamento. O *dropout* é o ato da rede desligar aleatoriamente alguns neurônios durante o processo de treinamento, sendo esse desligamento dado pelo hiperparâmetro *dropout rate*. Esse hiperparâmetro expressa a probabilidade p de um neurônio ser ficar ligado, sendo $p = 1$ a ausência de *dropout*. Conforme p se aproxima de zero, maior é a probabilidade de cada neurônios seja desligado (Srivastava *et al*., 2014). Um baixo valor para p requer um n grande, o que torna etapa de treinamento mais demorada e causa também falta de ajuste (*underfitting*), enquanto um p grande pode não gerar *dropouts* suficientes para evitar o *overfitting* (Srivastava *et al*., 2014). Embora seja necessário determinar cuidadosamente a taxa de *dropout*, é ainda um método simples, rápido e computacionalmente barato de ser implementado (Frau *et al*., 2021).

Passo 4: Treinamento, validação e otimização

Ao se utilizar modelos de aprendizagem de máquina, dois tipos de treinamento podem ser utilizados, não-supervisionado ou supervisionado. Para esta metodologia, se utiliza a aprendizagem supervisionada para treinamento e validação do modelo. Este método consiste na comparação da saída atual com a saída desejada. Os pesos de cada neurônio são inicialmente aleatorizados e

ajustados através do algoritmo de aprendizagem a cada iteração, sendo que esse ajuste depende do valor esperado e do sinal atual de saída (Anderson; Mcneill, 1992).

O processo de aprendizagem pode ser otimizado utilizando alguns algoritmos que ajustam a taxa de aprendizagem em tempo de execução, visando uma melhor convergência ou mesmo evitando *overfitting* (Ruder, 2016). A taxa de aprendizagem é um parâmetro que varia no intervalo (0, 1), sendo valores muito baixos tornam as mudanças nos pesos sinápticos menores de uma iteração para outra e as trajetórias no espaço definido pelos pesos serão suavizadas, enquanto a taxa de aprendizado muito alta provoca oscilações no treinamento e dificultando o processo de aprendizado (Ballini, 2000). Alguns dos principais algoritmos de otimização são:

- **Estimativa de Momento Adaptável (ADAM)**

  As principais vantagens desse método envolvem eficiência computacional, baixo requerimento de memória, boa capacidade de adaptação para problemas ruidosos ou com gradientes esparsos, entre outras. Esse algoritmo utiliza taxas de aprendizado para cada parâmetro, adaptadas separadamente durante o treinamento. Conforme Kingma e Ba (2014), e esse algoritmo reúne vantagens de dois outros métodos:

  - *AdaGrad*: adapta a taxa de aprendizado de cada parâmetro do modelo de acordo com sua importância relativa para a convergência do algoritmo. Isso permite que o algoritmo se adapte automaticamente a diferentes taxas de aprendizado para cada parâmetro, melhorando a eficiência e o desempenho do treinamento. Uma desvantagem é que a soma acumulada dos gradientes ao quadrado pode crescer indefinidamente durante o treinamento, o que pode reduzir a taxa de aprendizado a ponto de tornar-se as atualizações quase insignificantes (Duchi; Hazan; Singer, 2011).
  - *RMSProp*: utiliza média ponderada dos gradientes ao quadrado anteriores, enquanto o *AdaGrad* acumula os gradientes ao quadrado sem ponderação. Isso significa que o *RMSProp* dá mais importância aos gradientes recentes, o que pode ajudar a evitar que a taxa de aprendizado se torne muito pequena à medida que o treinamento progride.

O *ADAM* utiliza os parâmetros β_1 e β_2 que se referem aos coeficientes de decaimento do momento de primeira ordem e do momento de segunda ordem, respectivamente. Esses coeficientes controlam a influência das estimativas anteriores dos momentos na atualização dos parâmetros durante o treinamento. No R, os valores padrões para β_1, β_2 e ϵ são, respectivamente, 0,9, 0,999 e 10-08. Em outras palavras, tem-se que a estimativa do momento de primeira ordem considera 90% da informação dos gradientes anteriores, enquanto a estimativa do momento de segunda ordem considera 99,9% das informações anteriores dos gradientes ao quadrado. Por fim, o parâmetro ϵ é

um valor pequeno adicionado ao denominador nas fórmulas de cálculo dos momentos de primeira e segunda ordem, feito para evitar a divisão por zero e melhorar a estabilidade numérica do algoritmo (Kingma; Ba, 2014).

```
model %>%
        compile(
          optimizer = optimizer_adam(
          learning_rate = 0.01, beta_1 = 0.9,
          beta_2 = 0.999, epsilon = 1e-08),
          loss = "mse"
        )
```

Passo 5: Avaliação das etapas de treino e validação

A avaliação do modelo durante a etapa de treinamento e validação se dá visando minimizar a função de perda, sendo as mais utilizadas:

- **Erro Quadrático Médio (EQM):** Calcula a média da diferença ao quadrado entre o valor real com o predito conforme a fórmula abaixo:

$$EQM = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2$$

- **Raiz do Erro Quadrático Médio (REQM):** É a raiz quadrada do erro quadrático médio, deixando o resultado na mesma escala dos dados originais. Sua fórmula é apresentada abaixo:

$$EQM = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}$$

- **Erro Absoluto Médio (MAE):** Calcula a média do módulo da diferença entre o valor real com o predito, sua fórmula é dada abaixo:

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|$$

- **Erro Percentual Absoluto Médio (MAPE):** Indica a porcentagem de erro em relação aos valores reais e é expresso pela fórmula abaixo:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|y_t - \hat{y}_t|}{\max(\epsilon, |y_t|)}$$

Tem-se que n é o número total de observações (dias) no conjunto de dados, y_t é o valor observado no tempo t e $\hat{y}_t$ é o valor estimado para a observação t. Assim, para cada época, período correspondente a uma passagem completa pelo conjunto de treinamento, é calculado o erro conforme a função de perda definida. Espera-se que ao longo da etapa de treinamento, o erro se aproxime de zero e se estabilize.

```
# erro de treino
erro_train <- model %>%
        evaluate(x = df_train_x, y = df_train_y)
# erro de validacao
erro_val <- model %>%
        evaluate(x = df_val_x, y = df_val_y)
```

onde, *df_train_x* e *df_train_y* são, respectivamente, os tensores contendo os dados de entrada e saída do conjunto de treino, enquanto *df_val_x* e *df_val_y* são os tensores contendo os dados de entrada e saída do conjunto de validação. Erros de treinamento e validação próximos a zero indicam que o modelo conseguiu produzir estimativas semelhantes aos valores reais ao qual foi treinado.

Passo 6: Obtenção dos intervalos de incertezas

A estimação dos intervalos de incertezas se dá através da obtenção dos percentis das saídas do modelo. Para isso, utiliza-se o conjunto de dados reais adicionados de um ruído. A adição dos ruídos se faz necessária para criar uma variabilidade em cada observação, gerando estimações distintas a cada iteração. Isso é feito para o conjunto de dados do período não-contrafactual quanto para o período contrafactual. Assim como empregado na abordagem bayesiana utilizando o método de simulação de Monte Carlo (MC), o método proposto utiliza-se de simulações para obtenção dos intervalos. O código abaixo realiza *n_sim* predições para cada dia e, a partir desses valores, obtém os percentis de 2,5% e 97,5% como intervalos de incertezas.

```
# Periodo nao-contrafactual
df_train_pred <- NULL

for(i in 1:n_sim){
        ruido <- runif(1, 0.0001, 0.1)
        temp <- predict(model_LSTM, jitter(df_train2_x, amount = ruido)) %>%
        rev.minimax(., min_history, max_history) %>%
        data.frame()

        colnames(temp) <- NULL
        df_train_pred <- bind_cols(df_train_pred, temp)
}
df_train_pred <- df_metrics_row(df_train_pred, "Estimado")
```

```
# Periodo contrafactual
df_test_pred <- NULL
for(i in 1:n_sim){
        ruido <- runif(1, 0.0001, 0.1)
        temp <- predict(model_LSTM, jitter(df_test2_x, amount = ruido)) %>%
        rev.minimax(., min_history, max_history) %>%
        data.frame()

        colnames(temp) <- NULL
        df_test_pred <- bind_cols(df_test_pred, temp)
}
df_test_pred <- df_metrics_row(df_test_pred, "Contrafactual")
```

Passo 7: Verificação equivalência observacional

A construção do cenário contrafactual foi baseada no conceito do modelo de respostas potenciais proposto por Holland (1986), o qual considera a ideia de que um fator se torna uma causa se o resultado não tivesse ocorrido na ausência desse mesmo fator, assegurando que tudo tivesse permanecido constante, incluindo espaço e tempo. Assim, ao construir o modelo apresentando apenas dados que antecedem ao evento de interesse, a rede *LSTM* pode ser capaz de produzir estimativas semelhantes aos padrões ao qual foi treinada. Assim, é necessário verificar se os valores gerados para o período não-contrafactual são estatisticamente iguais aos valores observados para o mesmo período. Para isso, é utilizado o teste de cointegração proposto por Engle e Granger (1987). Seja $y_t \equiv [y_{t1}, .., y_{tN}]^T$ denotado pela *t*-ésima observação em *N* séries temporais não-cointegradas *(I(1))*. Se essas séries são cointegradas, existirá então um vetor α tal que o processo estocástico com uma observação $z_t \equiv [1 y_t]^T > \alpha$. Caso não sejam cointegradas, não existirá nenhum vetor α com essa propriedade, e qualquer combinação linear de y_1 até y_N e uma constante ainda será *I(1)*. Para implementar a forma original do teste Engler-Granger, primeiro é necessário realizar a regressão de cointegração $y_t = \alpha_1 + \sum_{j=2}^{N} \alpha_j y_{tj} + u_t$, para uma amostra de tamanho *T + 1*, assim obtendo o vetor de coeficientes $\hat{\alpha} \equiv [1 - \hat{\alpha}_1, \ldots, -\hat{\alpha}_N]^T$, então calcula-se:

$$\hat{z}_t = [1, y_t]^T \hat{\alpha} = y_{1t} - \hat{\alpha}_1 - \hat{\alpha}_{2yt2} \ldots - \hat{\alpha}_{NytN}$$

e testa para verificar se $\hat{z}_t$ é cointegrada usando um procedimento essencialmente o mesmo (exceto pela distribuição da estatística de teste) que o teste de Dickey-Fuller. Se rejeitarmos a hipótese nula, concluímos que y_1 até y_N estão cointegrados (Mackinnon, 1991). No R, a função *coint.test* do pacote *aTSA* realiza 3 tipos de teste de cointegração de Engler-Granger (sem tendência, com tendência linear e com tendência quadrática). A inclusão de termos de tendência linear ou quadrática permite capturar esses possíveis padrões e verificar se eles exercem um impacto significativo nas relações de cointegração entre as séries temporais.

```
aTSA::coint.test(observado, contrafacto, d = 0, output = TRUE)
```

onde, *d* é a quantidade de diferenciações necessárias para que as séries se tornem estacionárias. Assim, caso verificado que as séries são cointegradas, ao fazer as previsões para o período de ocorrência do evento de interesse, está sendo geradas estimativas contrafactuais mantendo a condição de que não houve mudanças ao longo tempo.

Passo 8: Estimação do impacto contrafactual

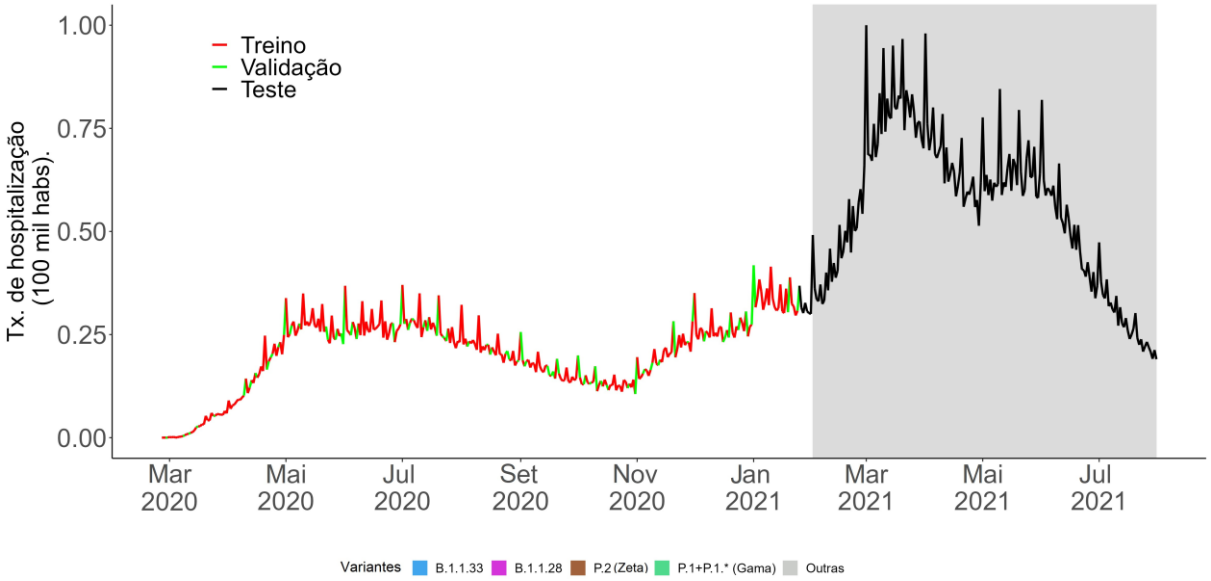
Rubin (1974) propôs que a estimação do efeito causal entre dois tratamentos na unidade *i* é dado pela comparação $Y_i(1) - Y_i(2)$ ou $Y_i(1)/Y_i(2)$. De modo semelhante, porém estendendo essas diferenciações ao longo do tempo, (Wing; Simon; Bello-Gomez, 2018) estima relações causais simplesmente obtendo a diferença entre os valores observados e hipotéticos. Considerando ambas as metodologias, a estimação do impacto causado pela divergência entre o real e o contrafactual é dada pela saída do modelo $Y_t(contrafacto)$ com os valores $Y_t(observado)$, sendo *t* cada tempo da série temporal. Ao final, tem-se um vetor de diferenças que possibilita a estimação do efeito causal, bem como o cálculo do número de casos atribuídos. O código abaixo apresenta, respectivamente, para a média e mediana, o cálculo dos erros entre as estimativas e as taxas observadas e os números de casos atribuídos para cada tempo *t*.

```
dados_final <- dados_final %>%
        mutate(
        erro_media = observado/pop * 10e4 - media/pop * 10e4,
        erro_mediana = observado/pop * 10e4 - mediana/pop * 10e4,
        casos_atribuiveis_medias = (observado - media) * pop/10e4,
        casos_atribuiveis_mediana = (observado - mediana) * pop/10e4
)
```

Resultados

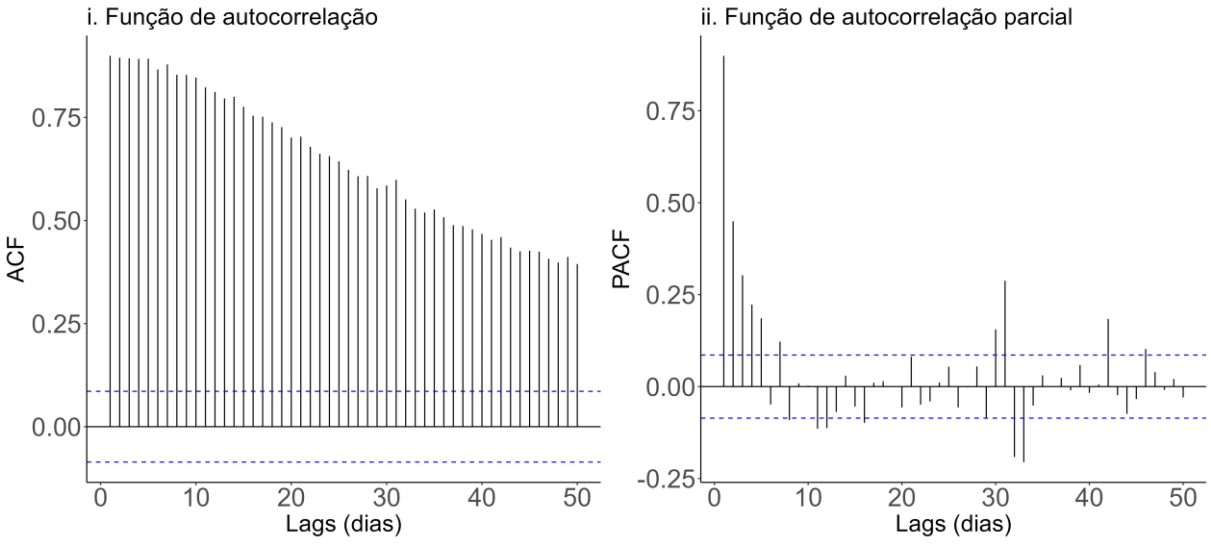
A definição do período contrafactual se deu considerando o fato da variante Gama ter causado a maior taxa de hospitalização para todo período do estudo. Dessa forma, a concepção do cenário contrafactual se deu pela ideia de que se a Gama fosse semelhante às variantes que a antecederam, o número de hospitalizações seriam menores. A figura 2 apresenta a taxa de hospitalização por SRAG-Covid e o particionamento dos dados nos grupos de treino, validação e teste. As observações até 31/01/2021 foram aleatoriamente divididas em treino e validação, respectivamente, nas proporções de (80%) e (20%).

Figura 2 – Particionamento dos dados em treino, validação e teste.



A separação entre as observações pertencentes ao grupo de treino e validação não ultrapassaram a distância de 50 dias, valor referente ao último *lag* significativo apresentado na figura 3, pois se entende que taxas de incidência em um intervalo inferior ou igual ao *lag* máximo observado são autocorrelacionadas, ou seja, dependentes.

Figura 3 – Função de autocorrelação e autocorrelação parcial para taxa de hospitalização por SRAG-Covid no Brasil.



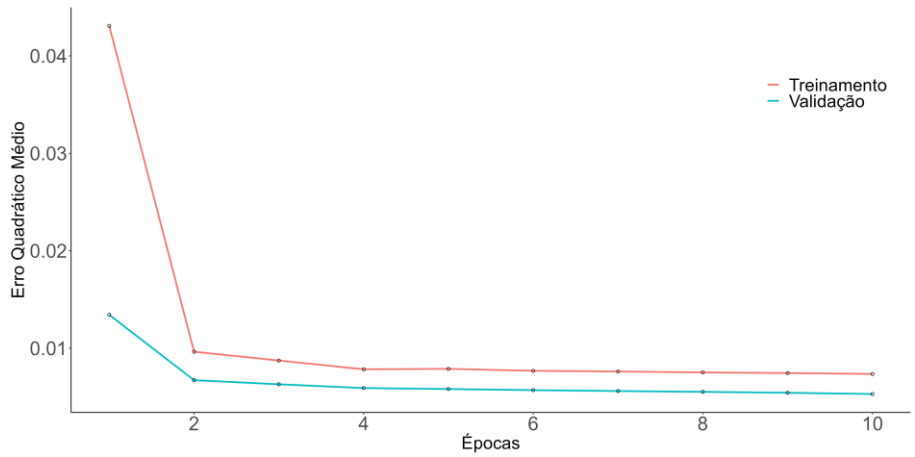
A rede utilizada no presente estudo possui a camada de entrada, duas camadas *LSTM* e a camada de saída, sendo essa última utilizando a função de ativação sigmoide. Os valores de saída do modelo (*output*) foram obtidos atrasando os dados de entrada em 7 dias, além disso, receberam a adição de ruídos (valores aleatórios de pequena magnitude) que visam dificultar que a rede neural

"decore" os padrões apresentados. Assim, o modelo recebe como entrada uma sequência de 7 valores para realizar a predição do oitavo. Ainda, quanto a configuração da rede, utilizaram-se os seguintes parâmetros: otimizador *ADAM* com taxa de aprendizagem em 0,01, mantendo os valores padrões para β_1, β_2 e ϵ; época e o tamanho do lote iguais a 50; 512 neurônios com taxa de *dropout* e *recorrent dropout* de 0,25 e 0,10 para as duas camadas *LSTM*.

Como a rede *LSTM* é capaz de aprender padrões, desta forma, apresentando apenas os valores que antecederam a variante Gama tem-se como resultados predições que se distribuem semelhantemente aos padrões a qual a rede foi treinada. Visando confirmar que a rede *LSTM* está produzindo valores que seguem os padrões apresentados durante a etapa de treinamento e validação, foi utilizado o teste de cointegração de Engler-Granger para os valores observados e estimados até 31/01/2021. Como resultado, rejeitou-se a hipótese de nulidade (sem tendência), ou seja, as séries são geradas pelo mesmo processo estocástico (valor-p < 0, 01).

A avaliação do modelo durante a etapa de treinamento e validação foi feita visando minimizar o erro quadrático médio (EQM). Como visto na figura 4, o erro do modelo vai diminuindo à medida que se aumenta o número de épocas, entretanto, a partir da quarta época de treinamento, o erro quadrático médio praticamente se estabiliza, indicando que o modelo alcançou um ponto de convergência ou aprendizado suficiente em relação aos dados de treinamento. Esse resultado corrobora com a cointegração das séries, indicando que a diferença entre as duas curvas é próxima a zero.

Figura 4 – Evolução dos erros de treinamento e validação do modelo *LSTM* para taxa de hospitalização por SRAG-Covid no Brasil.

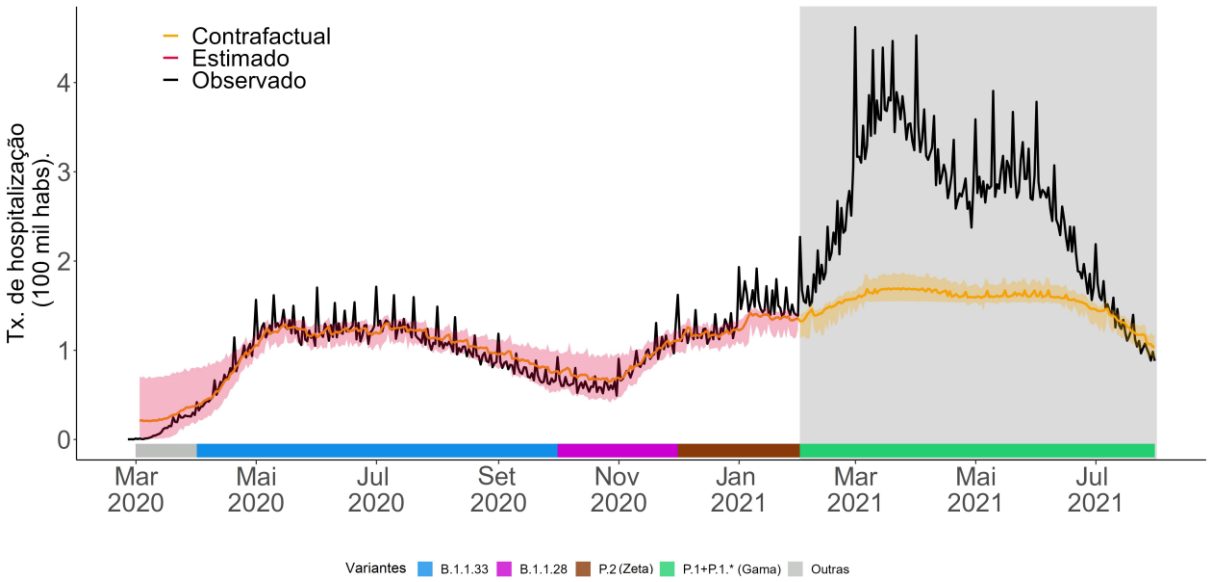


Conforme o apresentado na figura 5, são observadas a predominância de três linhagens, B.1.1.33, B.1.1.28 e a P.2. Essas linhagens resultaram em dois momentos de aumento da taxa de

hospitalizações até que a Gama se tornar prevalente. O período que a B.1.1.33 foi prevalente, início de abril ao final de setembro, chegou a atingir uma taxa de 1,74 hospitalizações por 100 mil habitantes. Em janeiro de 2021, período de prevalência ainda atribuído à variante P.2, máxima para a taxa de hospitalizações foi de 1,37 por 100 mil habitantes.

Embora as linhagens que antecederam a Gama tenham causado sérios problemas de saúde pública, a Gama, por ser ainda mais transmissível, resultou em taxas diárias ainda mais elevadas. Como visto na figura 5, a partir de fevereiro de 2021, nota-se um vertiginoso aumento da taxa de hospitalização. Março, o segundo mês de predominância da Gama, resultou na maior taxa de hospitalização, chegando a 4,62 por 100 mil habitantes. Apenas três meses depois, a partir do início de junho, que esses valores passaram a diminuir. O modelo contrafactual estimou uma curva que se distancia significativamente das taxas de hospitalizações observadas, indicando que, embora a introdução de uma nova linhagem causasse um esperado aumento de casos hospitalizados, a variante Gama fez isso em uma escala muito maior.

Figura 5 – Modelo contrafactual para taxa de hospitalização por SRAG-Covid no Brasil.



A tabela 1 apresenta as medidas descritivas para as hospitalizações atribuídas a Gama, isso é, a diferença das taxas esperadas para o cenário contrafactual com as taxas, observado vezes o tamanho da população por 100 mil. Nota-se que em março, o total de hospitalizações foi superior a 3 vezes o ocorrido no mês anterior. Apesar dos meses de abril e maio apresentarem números de hospitalizações totais bem semelhantes, observa-se que o número máximo diário de hospitalizações diminuiu aproximadamente 20%. Junho, em relação ao mês anterior, apresentou redução de 44% na média diária de hospitalizações. Por fim, julho apresenta uma mínima negativa, indicando que o

número de hospitalizações observadas atribuídas a Gama foi menor que o estimado no cenário contrafactual, representando assim o fim da ascensão da variante Gama.

Tabela 1 – Número de hospitalizações atribuídas à variante Gama da Sars-Cov-2 em 2021.

Mês	Min	Max	Média	DP	Total
Fev	164,56	3.135,66	1.431,93	780,79	40.094,03
Mar	3.132,71	6.545,66	4.223,41	851,47	130.925,84
Abr	1.593,26	6.051,37	2.980,61	878,03	89.418,22
Mai	2.161,26	4.862,08	2.868,86	654,32	88.934,55
Jun	204,31	4.600,10	1.630,57	954,36	48.917,13
Jul	-380,02	1.517,56	88,03	405,82	2.728,84

Discussão

Esse estudo apresenta uma proposta para a estimação de medidas de impacto em estudos ecológicos de séries temporais. Sua aplicação reúne um conjunto de metodologias que juntas possibilitam estimar efeitos causais. Sua aplicabilidade, no contexto da epidemiologia, se dá em mensurar o efeito de um evento de interesse em uma população ao longo do tempo. Neste caso, entende-se por evento de interesse uma possível intervenção como campanhas de vacinação, o surgimento de novas doenças ou mesmo de novas variantes, como o caso do exemplo utilizado, a introdução da variante Gama.

Delineamentos de estudos como coorte e caso-controle possuem limitações quando se trata de estimar efeitos causas em populações. Além disso, muitas vezes é necessário avaliar retrospectivamente as intervenções que já foram implementadas, muitas vezes, sem randomização ou para toda a população e, portanto, sem qualquer controle (Bonell *et al*., 2011). Diversos autores buscam estimar o efeito do impacto de intervenções em populações através da construção de cenários contrafactuais. Duas metodologias muito utilizadas em estudos ecológicos são elas, séries temporais interrompidas e diferenças em diferenças.

Séries temporais interrompidas consiste em comparar a continuidade de uma tendência (cenário contrafactual) com o ocorrido após a intervenção. Bernal, Cummins e Gasparrini (2017) aplicou esse método para estimar o efeito da proibição de fumar na Sicília, Itália. Na Coreia do Sul, Mun, Yang e Chang (2021) estimou o efeito de intervenções não-farmacêuticas como o uso de máscara e o distanciamento social. Silva, Filho e Fernandes (2020) usando o mesmo método, estimou o efeito do lockdown durante a pandemia de COVID-19 em 4 capitais situadas no nordeste

do Brasil. Entretanto, esse método não apresenta um grupo controle que não tenha sido exposto à intervenção, dificultando determinar se a mudança observada na série temporal é exclusivamente atribuível à intervenção ou se outros fatores também podem estar influenciando os resultados.

Diferenças em diferenças é um método que compara as mudanças ao longo do tempo entre um grupo de tratamento (exposto à intervenção) e um grupo de controle (não exposto à intervenção), baseando-se na suposição de que, na ausência da intervenção, a tendência dos grupos de tratamento e controle seria a mesma ao longo do tempo. Goodman-Bacon e Marcus (2020) utilizou essa metodologia para verificar se políticas de intervenções não-farmacêuticas como lockdown, fechamentos de escolas e proibição de eventos públicos reduziriam o espalhamento da COVID-19. O mesmo constatou que a COVID-19 apresenta características que dificultam a aplicação dessa metodologia, como atrasos entre exposição e infecções registradas, não linearidades que surgem da transmissão de pessoa para pessoa e a probabilidade de que as políticas tenham efeitos diferentes ao longo do tempo (Goodman-Bacon; Marcus, 2020). Além dessas limitações, fatores como a não-aleatorização da intervenção, influência de fatores não observados e a impossibilidade de controlar todos os confundidores dificultam estimação mais confiável das medidas de impacto.

A incapacidade interpretativa dos parâmetros de muitos modelos de aprendizagem de máquinas tem, possivelmente, conduzido muitos epidemiologistas a elaborações de propostas, sejam elas conceituais ou metodológicas, que adotam estratégias mais tradicionais e fundamentadas em teorias epidemiológicas já consolidadas. Molnar (2021) ressalta que a falta de interpretabilidade dos modelos de inteligências artificiais tem tornado o próprio modelo como fonte de informação ao invés dos dados. Na contramão desse comportamento, este artigo mostrou que o uso da rede neural *LSTM*, alinhada com outras técnicas estatísticas e com o raciocínio embasado nos conceitos de epidemiologia, que é possível produzir conhecimento que, segundo Molnar (2021), é o objetivo da ciência.

Muitos autores utilizaram a rede *LSTM* para predição de dados epidemiológicos em estudos ecológicos de séries temporais. Polyzos, Samitas e Spyridou (2021) utilizou a rede *LSTM* para estimar a redução no número de turistas chineses indo para a Austrália e para os EUA. Para a Índia, Chandra, Jain e Chauhan (2022) observou que a ocorrência de uma nova onda de casos seria baixa em outubro e novembro de 2021. Alassafi, Jarrah e Alotaibi (2022) realizaram predições para o número de casos e óbitos na Malásia, Marrocos e Arábia Saudita. Leite *et al*. (2021) também realizou predições para casos e óbitos confirmados, porém, para o Brasil, Índia e EUA, destacando que o modelo pode ser usado como ferramenta para o auxílio na pandemia.

Através das etapas de treinamento e validação da rede *LSTM*, respeitando a autocorrelação temporal dos dados, o modelo proposto conseguiu criar um grupo de comparação gerado pelo mesmo

processo estocástico que à curva observada, superando assim o problema de simultaneidade e equivalência observacional mencionados por Durlauf e Blume (2016). O problema de identificação, o qual Fonseca e Sánchez-Rivero (2020) descreve, talvez seja ainda um problema não resolvido, pois por mais que a direção causal, no exemplo utilizado, esteja clara, ou seja, a introdução da variante Gama causou um aumento no número de hospitalizações, é difícil isolar outros fatores que também possam atuar para o aumento desses números, como exemplo, flexibilização do distanciamento.

Não foi encontrado em acervo público nenhum artigo que, utilizando rede neural *LSTM* ou outra similar, propusesse documentar passos para o uso da abordagem contrafactual em estudos de séries temporais na epidemiologia. Além disso, também não foi encontrado nenhum artigo que, aplicando a rede *LSTM* ou ainda outra similar, realizassem estimações contrafactuais visando estimar efeitos causais para um evento de saúde. A maioria dos artigos consiste apenas em gerar predições, comparando os valores preditos com o esperado visando que o modelo tenha alta acurácia.

Limitações

A construção de uma rede neural com boa capacidade de generalização baseia-se primordialmente na qualidade de dados ofertados para o seu treinamento e validação. Embora não exista uma regra que especifique a quantidade mínima necessária de observações a serem fornecidas para essas etapas, é importante que a série original, ao ser dividida em duas (treino e validação), mantenha a dependência temporal. Isso se dá, não necessariamente, por uma limitação da rede neural, mas por manter uma característica que é inerente aos dados que estão sendo analisados. Logo, a utilização de redes *LSTM* em estudos de séries temporais com poucas observações pode ser inviabilizado pela incapacidade do modelo reconhecer padrões, visto que o conjunto de treinamento é muito pequeno. Além disso, em problemas mais complexos ou com grande variabilidade, podem ser necessários um conjunto de dados muito maior.

Outra limitação é quanto a configuração da rede *LSTM*, pois esta pode influenciar na qualidade do ajuste e consequentemente nas estimativas geradas para o cenário contrafactual. Uma das possíveis abordagens para a minimização desse problema é a adição de ruídos nos dados de entrada para a construção dos intervalos de incertezas e a utilização de *dropouts* para evitar o *overfitting* nas etapas de treinamento e validação.

Conclusão

Rubin (1974) ao dizer sobre os cuidados de estimar os efeitos causais em ensaios randomizados e não-randomizados, enfatizou a importância de o pesquisador estar ciente dos

possíveis efeitos causados pelas variáveis não-observadas. Da mesma forma, uma intervenção ou ainda algum outro evento pode exercer em uma população, modificações nos efeitos causais. No caso da introdução da variante Gama, exemplo utilizado neste artigo, a flexibilização do distanciamento social, o período de férias escolares e festividades de final de ano podem ter impulsionado a velocidade de transmissão dessa variante do vírus. Assim, a metodologia proposta, não é capaz de distinguir, muito menos quantificar o quanto da diferença entre os valores observados e os estimados se devem a outros fatores. Admitindo que os fatores não observados contribuíram para o aumento dos casos de hospitalizações, é muito mais sensato compreender que o impacto atribuído possa ser superestimado. Em outras palavras, reconhecendo o contexto em que a metodologia está sendo aplicada, a rede *LSTM* pode produzir estimativas plausíveis, apresentando baixo erro nas etapas de treinamento e validação e demonstrando ser uma proposta viável para a estimação de cenários contrafactuais em estudos ecológicos que utilizam dados dispostos periodicamente ao longo do tempo.

Referências

ALASSAFI, M. O.; JARRAH, M.; ALOTAIBI, R. **Time series predicting of COVID-19 based on deep learning. Neurocomputing**, Elsevier BV, v. 468, p. 335–344, jan. 2022.

ANDERSON, D.; MCNEILL, G. **Artificial neural networks technology.** Kaman Sciences Corporation, v. 258, n. 6, 1992.

BALLINI, R**. Análise e previsão de vazões utilizando modelos de séries temporais, redes neurais e redes neurais nebulosas. Universidade Estadual de Campinas**: Tese de Doutorado em Engenharia Elétrica, Faculdadede Engenharia Elétrica e de Computação, 2000.

BARATA, R. B. **Causalidade e epidemiologia**. FapUNIFESP (SciELO), v. 4, n. 1, p. 31–49, jun. 1997.

BERNAL, J. L.; CUMMINS, S.; GASPARRINI, A. **Interrupted time series regression for the evaluation of public health interventions: a tutorial**. International journal of epidemiology, Oxford University Press, v. 46, n. 1, p. 348–355, 2017.

BHANJA, S.; DAS, A. **Impact of data normalization on Deep Neural Network for Time Series Forecasting**. arXiv preprint arXiv:1812.05519, 2018.

BONELL, C. P. et al. **Alternatives to randomisation in the evaluation of public health interventions**: design challenges and solutions. Journal of Epidemiology & Community Health, BMJ Publishing Group Ltd, v. 65, n. 7, p. 582–587, 2011.

CHANDRA, R.; JAIN, A.; CHAUHAN, D. S. **Deep learning via LSTM models for COVID-19 infection forecasting in India**. PloS one, Public Library of Science San Francisco, CA USA, v. 17, n.1, p. e0262708, 2022.

DUCHI, J.; HAZAN, E.; SINGER, Y. **Adaptive subgradient methods for online learning and stochastic optimization**. Journal of machine learning research, v. 12, n. 7, 2011.

DURLAUF, S.; BLUME, L. E. **The new Palgrave dictionary of economics**. Springer, 2016.

ENGLE, R. F.; GRANGER, C. W. **Co-integration and error correction: representation, estimation, and testing**. Econometrica: journal of the Econometric Society, JSTOR, p. 251–276, 1987.

FIOCRUZ. **Dashboard Rede Genômica. Vigilância genômica do Sars-Cov-2 no Brasil**. Fiocruz, 2021. Disponível em: http://www.genomahcov.fiocruz.br/dashboard/. Acesso em: 01 nov. 2021.

FONSECA, N.; SÁNCHEZ-RIVERO, M. **Causalidade em economia com séries temporais: uma visita guiada desde a antiguidade clássica**. FapUNIFESP (SciELO), v. 30, n. 3, p. 999–1027, dez. 2020.

FRAU, L. et al. **Uncertainty estimation for machine learning models in multiphase flow applications**. Informatics, v. 8, n. 3, 2021. ISSN 2227-9709.

GÉRON, A. **Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to build Intelligent Systems**. Sebastopol, CA: O'Reilly Media, Inc, 2019. ISBN 978-1492032649.

GOMES, G.; LUDERMIR, T**. Otimização de pesos e funções de ativação de redes neurais aplicadas na previsão de séries temporais**. arXiv preprint arXiv:2107.14370, 2021.

GOODFELLOW, I. Deep Learning/Goodfellow I., Bengio Y. and Courville A. Cambridge MA: MIT Press [2017]–777 pages, 2016.

GOODMAN-BACON, A.; MARCUS, J. **Using difference-in-differences to identify causal effects of COVID-19 policies**. DIW Berlin Discussion Paper, 2020.

HAYKIN, S. **Neural networks and learning machines**, 3/E. Pearson Education India, 2009.

HOCHREITER, S.; SCHMIDHUBER, J. **Long Short Term Memory**. MIT Press- Journals, v. 9, n. 8, p. 1735–1780, nov. 1997.

HOLLAND, P. W. **Statistics and causal inference**. Journal of the American Statistical Association, Taylor & Francis, v. 81, n. 396, p. 945–960, 1986.

HUME, D. **Investigação sobre o entendimento humano**. Leya, 2013.

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

LEITE, S. J. O. et al. **Predição de séries temporais da COVID-19: uma avaliação de redes neurais com células LSTM**. 2021.

MACKINNON, J. **Critical values for cointegration tests. Long-run economic relationships**, Oxford University Press, v. 13, 1991.

MCCULLOCH, W. S.; PITTS, W. **A logical calculus of the ideas immanent in nervous activity**. The bulletin of mathematical biophysics, Springer, v. 5, n. 4, p. 115–133, 1943.

MOLNAR, C. **Interpretable Machine Learning: A Guide for Making Black Box Models Explainable**. Leanpub; 2019. 2021.

MORETTIN, P.; TOLOI, C. de C. **Análise de séries temporais**. Edgard Blucher, 2006. (ABE - Projeto Fisher). ISBN 9788521203896.

MUN, S.-K.; YANG, B. R.; CHANG, M. **Changes in respiratory diseases in south korea during the COVID-19 pandemic: An interrupted time series study**. BMJ global health, BMJ Specialist Journals, v. 6, n. 12, p. e006912, 2021.

NAYAK, S.; MISRA, B. B.; BEHERA, H. S. **Impact of data normalization on stock index forecasting**. International Journal of Computer Information Systems and Industrial Management Applications, v. 6, n. 2014, p. 257–269, 2014.

POLYZOS, S.; SAMITAS, A.; SPYRIDOU, A. E. **Tourism demand and the COVID-19 pandemic: An LSTM approach**. Tourism Recreation Research, Taylor & Francis, v. 46, n. 2, p. 175–187, 2021.

ROSENBLATT, F. **Principles of neurodynamics. perceptrons and the theory of brain mechanisms**. 1961.

RUBIN, D. B. **Estimating causal effects of treatments in randomized and nonrandomized studies**. Journal of educational Psychology, American Psychological Association, v. 66, n. 5, p. 688, 1974.

RUBIN, D. B. **Statistics and causal inference: Comment: Which ifs have causal answers**. Journal of the American Statistical Association, v. 81, n. 396, p. 961–962, 1986.

RUDER, S. **An overview of gradient descent optimization algorithms**. ArXiv preprint arXiv:1609.04747, 2016.

SILVA, L.; FILHO, D. F.; FERNANDES, A. **The effect of lockdown on the COVID-19 epidemic in Brazil: evidence from an interrupted time series design**. Cadernos de Saúde Pública, SciELO Public Health, v. 36, p. e00213920, 2020.

SOUMERAI, S. B.; STARR, D.; MAJUMDAR, S. R. **How do you know which health care effectiveness research you can trust? a guide to study design for the perplexed**. Preventing chronic disease, Centers for Disease Control and Prevention, v. 12, 2015.

SRIVASTAVA, N. et al. Dropout: **A Simple Way to Prevent Neural Networks from Overfitting**. J. Machine Learning Res., v. 15, p. 1929–1958, 2014.

SZWARCWALD, C. L.; CASTILHO, E. A. d. **Os caminhos da estatística e suas incursões pela epidemiologia**. Cadernos de Saúde Pública, SciELO Brasil, v. 8, p. 5–21, 1992.

WILLIAMSON, J. **Causality**. In: Springer Netherlands, 2007. p. 95–126.

WING, C.; SIMON, K.; BELLO-GOMEZ, R. A. **Designing difference in difference studies: best practices for public health policy research**. Annual review of public health, Annual Reviews, v. 39, p. 453–469, 2018.

7 Anexo

Tabela 2 – Número de hospitalizações por SRAG-Covid por ano e mês

Ano	Mês	Taxa observada				Total
		Min	Máx	Média	DP	
	Fev	5	7	6,25	0,96	25
	Mar	11	644	284,13	228,87	8808
	Abr	696	2462	1499,73	541,65	44992
	Mai	2242	3451	2666,03	294,79	82647
	Jun	2288	3635	2677,40	297,70	80322
2020	Jul	2144	3655	2655,06	345,96	82307
	Ago	1729	3179	2150,97	302,77	66680
	Set	1323	2531	1639,43	255,00	49183
	Out	1049	1970	1318,35	179,57	40869
	Nov	1412	2788	1987,30	360,76	59619
	Dez	2291	3461	2606,52	257,83	80802
2021	Jan	2925	4125	3289,58	339,22	101977

Fev	3199	6513	4478,46	906,35	125397
Mar	6620	9861	7787,48	836,53	241412
Abr	5069	9663	6513,20	908,57	195396
Mai	5668	8337	6360,97	647,05	197190
Jun	3336	8077	4998,57	1053,16	149957
Jul	1881	4672	2755,97	640,34	85435

Tabela 3 – Soma dos erros do modelo das etapas de treinamento e validação do modelo *LSTM* para as taxas de hospitalização.

Ano	**Mês**	**Min**	**Max**	**Média**	**DP**	**Total**
2020	Mar	0,01	0,30	0,14	0,11	4,11
	Abr	0,33	1,15	0,70	0,25	21,08
	Mai	1,05	1,62	1,25	0,14	38,73
	Jun	1,07	1,70	1,25	0,14	37,64
	Jul	1,00	1,71	1,24	0,16	38,57
	Ago	0,81	1,49	1,01	0,14	31,24
	Set	0,62	1,19	0,77	0,12	23,04
	Out	0,49	0,92	0,62	0,08	19,15
	Nov	0,66	1,31	0,93	0,17	27,94
	Dez	1,07	1,62	1,22	0,12	37,86
2021	Jan	1,37	1,93	1,54	0,16	47,79

8 ARTIGO 3 - MODELO *LSTM* PARA ESTIMAÇÃO CONTRAFACTUAL DA CARGA DE HOSPITALIZAÇÕES E ÓBITOS POR SRAG-COVID ATRIBUÍVEL À VARIANTEGAMA EM PESSOAS A PARTIR DE 60 ANOS

RESUMO

Background. Este artigo se apresenta como um estudo retrospectivo que visa estimar o número de hospitalizações e óbitos por síndrome respiratória aguda grave (SRAG) em decorrência da COVID-19 (SRAG-Covid) atribuídos à variante Gama em pessoas acima dos 60 anos por meio da construção de um cenário contrafactual em que esta variante apresentasse padrões similares aos observados nas variantes anteriores. Buscou-se comparar as curvas geradas para o cenário contrafactual com a curva observada no mesmo período a fim de obter-se a carga atribuída de casos e óbitos à variante Gama.

Métodos. Os dados utilizados no presente estudo se referem ao número de casos hospitalizados e óbitos em 2020 e 2021 por SRAG-Covid disponibilizados pelo Ministério da Saúde na plataforma *OpenDataSUS* para o Brasil e os estados de São Paulo, Rio de Janeiro e Amazonas. Foi utilizada a rede neural *LSTM* devido a sua capacidade de incorporar a dependência temporal de longo e curto prazo, propiciando boa estimação de valores sequenciais.

Resultados. Os resultados apontam que aproximadamente 50 mil hospitalizações e mais de 30 mil óbitos por SRAG-Covid são atribuídas à variante Gama no Brasil apenas para um único mês. Dentre os estados analisados, São Paulo apresentou o maior número de internações e óbitos. Rio de Janeiro já havia sofrido ondas semelhantes com a B.1.1.33 e P2, logo o cenário contrafactual foi próximo ao observado. Amazonas apresentou um rápido aumento de casos, fazendo com que a metodologia utilizada, não se mostrasse ideal para este caso.

Conclusões. O modelo *LSTM* apresentou boa capacidade de generalização, resultando em erros de estimação próximos a zero. Para o Brasil e o estado de São Paulo as curvas contrafactuais que se distanciaram dos valores observados indicando que a Gama foi mais severa do que se tivesse repetido os padrões das variantes que a antecederam; Rio de Janeiro apresentou estimativas próximas dos valores observados, pois a introdução das variantes B.1.1.33 e P.2 (Zeta) alcançaram magnitudes até maiores que à Gama. Possivelmente, devida a rápida evolução do número de hospitalizações e óbitos já nas primeiras semanas da pandemia, Amazonas apresentou erros mais elevados quando comparado aos demais cenários estudados.

Palavras-chave: *SRAG-Covid*, Modelo Contrafactual, LSTM.

1 Introdução

Um dos principais problemas resultantes do surgimento de novas variantes é a capacidade de serem mais transmissíveis, gerar casos mais graves e/ou ser capaz de infectar indivíduos que já tiveram contatos com outras variantes ou até mesmo vacinados. No Brasil, durante a maior parte de 2020, as variantes B.1.1.28 e B.1.1.33 foram predominantes. A partir delas, mutações no domínio de ligação da proteína Spike (S) deram origem a outras duas novas variantes, a P.1 e a P.2 (Zeta), que se espalharam rapidamente (Faria *et al*., 2021).

A variante P.1, posteriormente denominada Gama, foi primeiramente detectada no Amazonas. Em sua capital, Manaus, a taxa de mortalidade foi, aproximadamente, três vezes maior do que à observada no pico da primeira onda. Em cerca de dois meses, a Gama tornou-se a variante mais prevalente no Brasil. Os estados de São Paulo e do Rio de Janeiro também foram severamente afetados com suas taxas de hospitalizações e óbitos por COVID-19 alcançando os mais elevados níveis até então observados.

Conforme o boletim epidemiológico publicado pelo Ministério da Saúde (2021a), aproximadamente 75% dos óbitos por síndrome respiratória aguda grave (SRAG) ocorreram em pessoas com 60 anos ou mais. O risco de um caso evoluir para óbito por COVID-19 aumenta com a idade, consequentemente, tem-se que a predominância de óbitos ocorrem em idosos, principalmente aqueles com doenças crônicas (Hammerschmidt; Santana, 2020).

Neste contexto, este artigo visa apresentar a carga de hospitalizações e óbitos por SRAG-Covid em pessoas a partir dos 60 anos atribuídas à variante Gama por meio da diferença entre os valores observados contra os valores estimados para o cenário contrafactual no Brasil e nos estados de São Paulo, Rio de Janeiro e Amazonas para o período de 01/03/2020 a 30/06/2021. Também é objetivo do estudo apresentar a rede neural recorrente (RNN) *LSTM* como proposta para modelagem de dados dispostos em séries temporais, tendo em vista que este tipo de modelo é capaz de lidar com a dependência temporal. Por fim, este artigo tem como finalidade ampliar o leque de metodologias utilizadas no contexto de saúde pública, principalmente no que tange o emprego de métodos de aprendizagem de máquina.

2 Materiais e métodos

2.1 Dados

No Brasil é obrigatório a notificação de casos suspeitos ou confirmados de COVID-19. Os dados utilizados no presente estudo se referem ao número de casos hospitalizados e óbitos de pessoas acima dos 60 anos para o período de 01/03/2020 à 31/07/2021 por SRAG-Covid disponibilizados na

plataforma *OpenDataSUS*. São considerados casos hospitalizados por COVID-19 todo registro hospitalar com a classificação final de SRAG por COVID-19, o que equivale a casos com registros de encerramento baseados em critério laboratorial, clínico-epidemiológico, clínico ou clínico-imagem (Ministério da Saúde, 2021b).

Conforme as diretrizes do Ministério da Saúde (2021b), os casos de SRAG são definidos pela apresentação simultânea de quatro critérios: (i.) febre (mesmo que autorreferida), (ii.) dispneia/desconforto respiratório ou pressão persistente no tórax ou saturação de O2 menor que 95%, (iii.) dor de garganta ou tosse e (iv.) que tenham sido hospitalizados ou evoluído a óbito independentemente de hospitalização prévia.

Informações referentes às populações estimadas foram obtidas pelo IBGE e as informações referentes às variantes predominantes pela Fiocruz (2021).

2.2 Análises estatísticas

A fim de mensurar o número de hospitalizações e óbitos para as quatro regiões de estudo, foram desenvolvidos 8 modelos *LSTM*. Na etapa de pré-processamento calculou-se as taxas de hospitalização e óbitos diários por SRAG-Covid por 100 mil pessoas. Em seguida, visando prevenir os problemas de maior demora para aprendizagem nas primeiras camadas neurais quando comparadas as camadas mais profundas (problema de gradiente), as taxas foram normalizadas utilizando a metodologia minimax, conforme expresso na fórmula abaixo:

$$y = \frac{(x - x_{min})}{(x_{max} - x_{min})}$$

em que y é o valor normalizado, x é o valor da entrada, x_{min} é o valor mínimo da variável de entrada e x_{max} é o valor máximo das variáveis de entrada. A normalização também serve para manter os valores das entradas proporcionais aos limites das funções de ativação logística sigmoide, cujos valores são limitados pelo intervalo [0,1].

A duração da dependência temporal foi estimada por meio da função de autocorrelação. Para o Brasil, São Paulo e Rio de Janeiro, tanto para hospitalização quanto óbito, as dependências temporais foram de respectivamente, 50, 50 e 29 dias. No Amazonas, a dependência temporal para as taxas de hospitalização foi de 40 dias e de óbito 36 dias.

A rede *LSTM* é um tipo de RNN capaz de aprender dependências de longo e curto prazo através da retenção de informações passadas, em outras palavras, é capaz de incorporar a dependência temporal entre as observações (Hochreiter; Schmidhuber, 1997). Quanto ao

funcionamento da rede *LSTM*, Géron (2019) descreve que o vetor de entrada atual x_t e a memória de curto prazo anterior $h_{(t-1)}$ são fornecidas para quatro diferentes camadas que estão conectadas, porém, servem a um propósito diferente:

- Camada principal $g_{(t)}$: responsável por analisar as entradas de x_t e a memória de curto prazo $h_{(t-1)}$;
- Forget Gate $f_{(t)}$: responsável por definir as informações que serão retidas;
- Input Gate $i_{(t)}$: responsável por definir as informações que incorporadas a memória de longo prazo;
- Output Gate $o_{(t)}$: responsável por controlar as partes do estado de longo prazo devem ser lidas tanto para $h_{(t)}$ e $y_{(t)}$ e exibidas no intervalo de tempo *t*

A rede utilizada apresenta a camada de entrada, três camadas *LSTM* e a camada de saída, sendo essa última utilizando a função de ativação sigmoide, que apresenta espaço de variação [0,1]. Para a criação da saída do modelo, o output, os dados de entrada foram atrasados em 7 dias, de modo que o modelo receberá como entrada uma sequência de 7 valores e fará a predição do oitavo.

Foi utilizado aprendizagem supervisionada para treinamento do modelo. Este método consiste na comparação da saída atual com a saída desejada. Os pesos de cada neurônio que foram inicialmente aleatorizados, são ajustados através do algoritmo de aprendizagem a cada iteração, sendo que esse ajuste depende do valor esperado e do sinal atual de saída (Anderson; Mcneill, 1992).

Antes da construção dos modelos, as séries de hospitalização e óbito foram divididas separando os períodos pré-prevalência e prevalência da variante Gama. Diferente de uma modelagem estatística, os modelos de aprendizagem de máquina possuem duas etapas de execução: a etapa de treinamento e de validação. Considerando o conjunto de dados para o período pré-prevalência da Gama, 80% das observações foram aleatoriamente utilizadas para treinamento e as 20% restantes para validação. O processo de aleatorização foi controlado de modo que a distância entre as observações não fosse superior ao valor da dependência temporal de cada série. Das quatro regiões abrangidas por este estudo, apenas para o Amazonas teve seu período de pré-prevalência da variante Gama no final de novembro de 2020. Para as demais regiões, o período de pré-prevalência da variante Gama foi no final de janeiro de 2021. O período de prevalência da Gama foi usado para comparação das taxas observadas com as estimativas contrafactuais obtidas pelos modelos.

Como a rede *LSTM* gera por padrão apenas estimativas pontuais, as estimações das incertezas foram obtidas utilizando o *dropout*. O *dropout* consiste no ato da rede desligar aleatoriamente alguns neurônios durante o processo de treinamento, sendo esse desligamento dado pelo hiperparâmetro *dropout rate*. Esse hiperparâmetro expressa a probabilidade *p* de um neurônio ser mantido ligado,

sendo $p = 1$ a ausência de *dropout*. Conforme p se aproxima de zero, maior é a probabilidade de cada neurônios seja desligado (Srivastava *et al*., 2014). Um baixo valor para p requer um n grande, o que torna etapa de treinamento mais lento e causa falta de ajuste (*underfitting*), enquanto um p grande pode não gerar *dropouts* suficientes para evitar o *overfitting* (Srivastava *et al*., 2014). Embora seja necessário determinar cuidadosamente a taxa de *dropout*, isso é um método simples, rápido e computacionalmente barato de se implementar (Frau *et al*., 2021). Khosravi *et al*. (2011) ressalta que enquanto os intervalos de confiança assumem apenas as incertezas inerentes ao modelo, os intervalos de predição levam consideram também a variância do ruído dos dados.

A construção do cenário contrafactual foi baseada no conceito do modelo de respostas potenciais proposto por Holland (1986), o qual considera a ideia de que um fator se torna uma causa se o resultado não tivesse ocorrido na ausência desse mesmo fator, assegurando que tudo tivesse permanecido constante, incluindo espaço e tempo. O emprego da rede *LSTM* se dá em aprender os padrões que antecederam a Gama e gerar predições para o período de prevalência dessa variante. A saída do modelo $Y_{\text{contrafacto}}(t)$ é comparada às taxas observadas $Y_{\text{observado}}(t)$, sendo t cada dia do período fora da etapa de treinamento e validação, ou seja, o período de predominância da variante Gama. Ao final, tem-se um vetor de diferenças de taxas e considerando a população suscetível é possível obter o número de hospitalizações e óbitos para cada tempo t. A avaliação do modelo durante a etapa de treinamento e validação se deu visando minimizar o erro quadrático médio (EQM).

3 Resultados

A tabela 1 apresenta a soma por mês dos erros do modelo *LSTM* para as taxas de hospitalização e óbito para o período pré-prevalência da variante Gama enquanto a tabela 2 apresenta o número de hospitalizações e óbitos atribuídos à variante Gama diante do cenário contrafactual proposto. Nota-se que na tabela 1, a soma dos erros se concentrou muito próximo de zero para o Brasil, São Paulo e Rio de Janeiro, refletindo a boa capacidade de generalização dos modelos.

Tabela 1 – Soma dos erros do modelo das etapas de treinamento e validação do modelo *LSTM* para as taxas de hospitalização e óbito.

Ano	Mês	Brasil		São Paulo		Rio de Janeiro		Amazonas	
		Hospitalização	Óbito	Hospitalização	Óbito	Hospitalização	Óbito	Hospitalização	Óbito
2020	Abr	-0,0172	0,0110	0,0224	0,0412	0,9009	0,6902	326,504	331,218
	Mai	0,0336	0,0229	0,1960	0,0693	12,424	0,6436	50,306	-52,771
	Jun	0,0317	0,0126	0,1266	0,0500	0,1473	0,0962	-74,244	0,6349
	Jul	0,0379	0,0165	0,1413	0,0531	-0,1787	-0,0238	-70,045	-0,9907
	Ago	0,0147	0,0056	0,0490	-0,0249	-0,0100	-0,0348	-140,686	-57,569
	Set	0,0051	0,0072	0,0503	-0,0201	0,1035	-0,0533	-16,521	-29,846

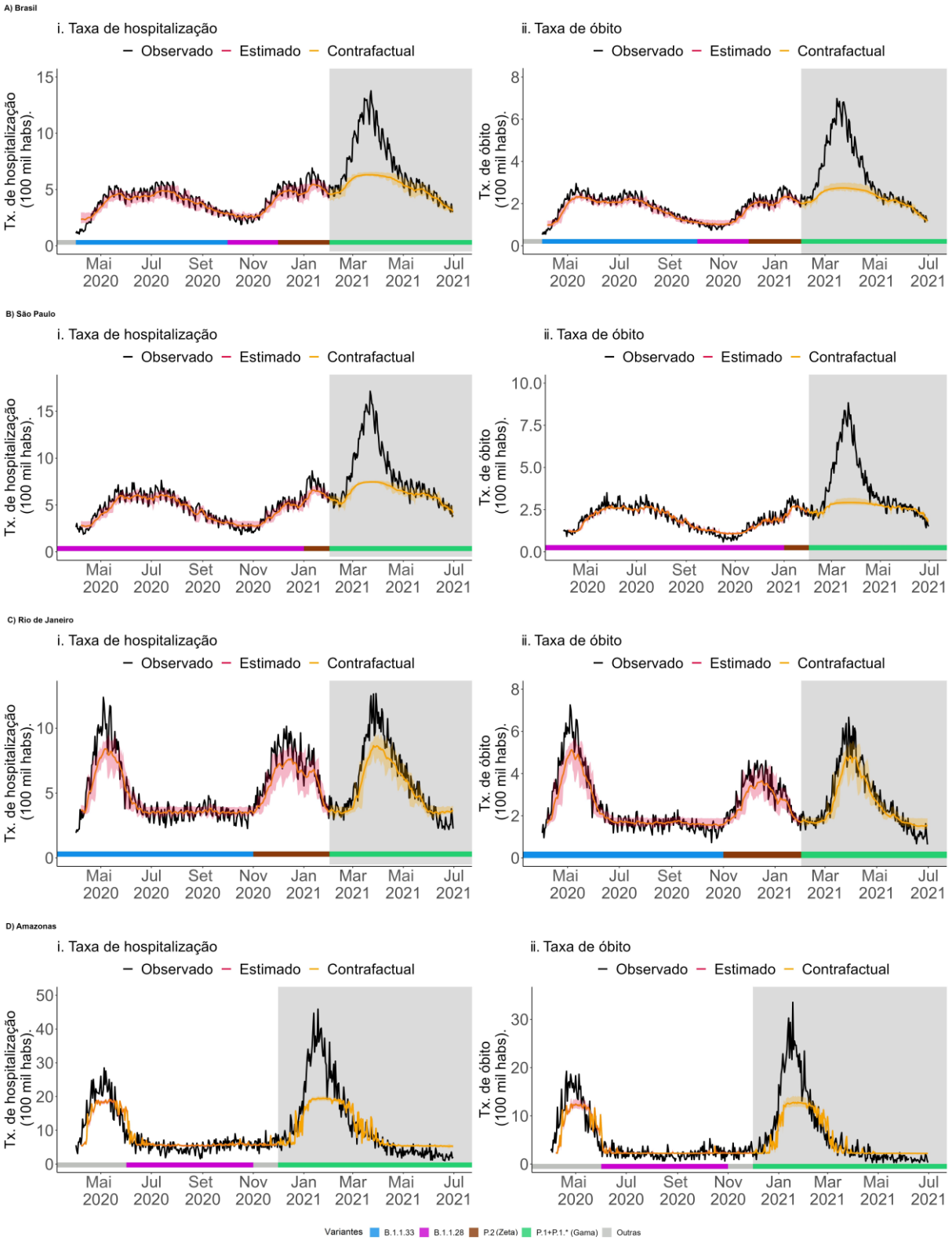
	Out	-0,0205	-0,0100	-0,1422	-0,0947	-0,2825	-0,2415	23,144	25,203
	Nov	0,0280	0,0134	0,0845	0,0168	0,9432	0,4477	20,960	35,871
	Dec	0,0397	0,0123	0,1517	0,0380	12345	0,4715	-	-
2021	Jan	0,0484	0,0177	0,2668	0,0542	0,3468	0,1733	-	-

Tabela 2 – Número de hospitalizações e óbitos atribuídas à variante Gama da Sars-Cov-2.

Ano	Mês	Brasil		São Paulo		Rio de Janeiro		Amazonas	
		Hospitalização	Óbito	Hospitalização	Óbito	Hospitalização	Óbito	Hospitalização	Óbito
2020	Dec	-	-	-	-	-	-	195,04	155,60
2021	Jan	-	-	-	-	-	-	1467,39	1031,42
	Fev	8162,18	4696,19	1329,78	740,88	259,39	68,64	152,18	-30,52
	Mar	49989,44	30184,31	14282,07	9122,90	1790,58	774,05	-83,40	10,02
	Abr	19265,12	10253,38	4569,74	2720,19	1132,46	390,14	-108,54	-26,59
	Mai	3279,37	1074,27	1091,56	391,82	280,72	151,52	-203,99	-95,14
	Jun	920,15	424,73	80,65	-205,07	-515,64	-302,40	-289,87	-119,38

A figura 1 apresenta no eixo horizontal o tempo em dias e no eixo vertical a taxa observada por 100 mil de hospitalizações (i.) e óbitos (ii.). As áreas em cinza correspondem ao período de prevalência da linha Gama, ou seja, são observações desconhecidas durante as etapas de treinamento e validação do modelo. Abaixo das curvas das taxas está presente uma reta no eixo horizontal que correspondem às variantes predominantes para os respectivos meses. Os valores estimados para o cenário contrafactual revelam a discrepância do que seria esperado caso as ondas anteriores fossem "revividas" no lugar do que foi a variante Gama.

Figura 1 – Modelo contrafactual para taxa de hospitalização e óbito em pessoas a partir de 60 anos no Brasil.



3.1 Brasil

Observa-se duas principais ondas que antecedem a predominância da Gama, correspondendo, respectivamente, aos períodos em que as variantes B.1.1.33 e P.2 (Zeta) foram predominantes. Durante o início de abril ao final de setembro, intervalo em que a variante B.1.1.33 foi a mais

prevalente, o número de hospitalizações e óbitos diários por 100 mil habitantes alcançaram os valores de 5,92 e 2,98, respectivamente.

Com a predominância da variante P.2 (Zeta), esses valores chegaram a 7,16 para taxa de internação e 2,88 para a taxa de óbito em janeiro de 2021. Considerando o pico da Gama (março de 2021), a taxa de hospitalização diária chegou a 14,25, sendo 2,40 vezes maior que a máxima observada pela variante B.1.1.33 e 1,73 vezes maior que o da P.2 (Zeta). A máxima diária observada para a taxa de óbito foi de 7, 05, o que corresponde a 2,98 vezes a taxa de óbito da variante B.1.1.33 e 1,78 vezes a da P.2 (Zeta).

Conforme o apresentado na figura 1 (A), há um descolamento da curva estimada pelo modelo contrafactual com o observado para as taxas de hospitalizações e óbitos, mostrando que, apesar do esperado pela introdução da variante Gama fosse um aumento desses desfechos, os mesmos ocorreriam em uma magnitude bem inferior. Além disso, o modelo se demonstra capaz de produzir bons ajustes, estimando valores próximos aos observados. Durante os meses de fevereiro a abril houve um vertiginoso aumento, não visto até então, sendo apenas o mês de março totalizando 49.989 hospitalizações. A soma do número óbitos entre fevereiro à maio foi de 46.208, ou seja, a diferença entre o previsto pelo modelo é inferior ao observado, consequentemente, há um excesso de óbito atribuído a variante Gama, sendo março apresentando uma média diária de quase mil óbitos. Nota-se que a soma dos erros para todo o período de treinamento e validação concentrou-se muito próximo de zero, refletindo a boa capacidade de generalização do modelo.

3.2 São Paulo

As taxas de hospitalizações e óbitos se distribuem no tempo de modo semelhante às curvas do Brasil e isso se dá em virtude de que São Paulo ser o estado que apresenta o maior número de casos de COVID-19. Consequentemente, também é observado a presença de duas ondas que antecederam a variante Gama, entretanto, a primeira corresponde à B.1.1.28 e a segunda à P.2 (Zeta). A P.2 (Zeta) foi a variante mais prevalente apenas durante um único mês, janeiro de 2021, sendo que no mês anterior, foi o período em que se observou as taxas mais elevadas para a B.1.1.28. Os valores observados estão muito acima dos previstos pelo modelo para os períodos de março a maio, evidenciando a capacidade de variante Gama produzir casos graves da COVID-19. Além disso, nota-se que o modelo foi capaz de produzir bons ajustes, pois os valores estimados foram próximos aos observados para o período que antecedeu a introdução da variante Gama, como pode ser visto na figura 1 (B).

A máxima diária para a taxa de hospitalização causada pela variante B.1.1.28 foi de 7,68, enquanto para a P.2 (Zeta) foi de 8,70. A máxima diária da taxa de hospitalização da variante Gama

equivale a 2,24 vezes a taxa de hospitalização da B.1.1.28 e de quase 2 vezes a da P.2 (Zeta) (1,98 vezes). Segundo os dados divulgado pelo Governo de são Paulo em 23 de junho de 2021, período ainda sob predominância da circulação da variante Gama, a cobertura vacinal para as faixas etárias de pessoas acima dos 75 anos, com duas doses contra a COVID-19 é de 100%, nas pessoas de 80 a 84 anos é de 94,49% e de 70 a 74 anos, 97,39% e para pessoas de 60 a 69 anos, a cobertura atual é menor que 50% (Vidale, 2021). A máxima diária para taxa de óbito referente a variante B.1.1.28 foi 3, 42 e para a P.2 (Zeta) foi 3,24. A variante Gama apresentou 2,51 vezes a máxima da taxa de óbito diária da B.1.1.28 e 2,65 vezes a máxima da P.2 (Zeta).

O modelo produziu valores ajustados próximos aos observados e com alta precisão. Vê-se que as somas mensais dos erros para todo o período de treinamento e validação foram próximas a zero, indicando a boa capacidade do modelo produzir estimativas precisas.

3.3 Rio de Janeiro

Observa-se que as curvas para as taxas de hospitalizações e óbitos diários em pessoas acima dos 60 anos apresentam três picos bem definidos, correspondendo cronologicamente às variantes B.1.1.33, P.2 (Zeta) e a Gama. O modelo teve boa capacidade de generalização, acompanhando a dinamicidade dos valores observados, o que reflete a capacidade do modelo produzir estimativas verossímeis. Confrontando o observado com o cenário contrafactual, durante os meses de fevereiro a maio os números de hospitalizações foram superiores ao esperado, principalmente em março, chegando a ter 1.790 a mais do que previsto pelo modelo. Para os meses de fevereiro a maio, o cenário contrafactual apresenta até 774 óbitos a menos que o observado. A soma mensal dos erros do modelo é próxima de zero, indicando boa qualidade de ajuste do modelo.

3.4 Amazonas

O estado apresenta apenas uma onda bem definida que antecede a emergência da Gama. Nota-se que em abril, a taxa de hospitalização saltou de 3,08 para 25,03, acelerada alta que refletiu fortemente na curva da taxa de óbitos. Os valores observados estão muito acima dos previstos pelo modelo para os períodos de março a maio, evidenciando a capacidade de variante Gama produzir casos graves da COVID-19. Como a emergência da variante Gama aconteceu no Amazonas, o pico de hospitalizações e óbitos antecedeu as demais áreas em estudo. Em janeiro, identificou-se aproximadamente 1.500 hospitalizações atribuídas à Gama. Em janeiro, mês em que o estado apresentou o maior número de casos, a taxa de óbito por SRAG-Covid atribuída à variante Gama foi superior a 1.000.

As estimativas contrafactuais estimadas para os meses a partir de maio foram superiores às taxas de hospitalizações observadas, parte disso, explicada pelo modelo não ser capaz de gerar estimativas tão baixas. Diferente dos modelos anteriores, as somas mensais dos erros foram muito distantes de zero, indicando que para esse caso, o modelo *LSTM* não apresentou estimações tão boas. Para o mês de abril, a soma dos erros, tanto para a taxa de hospitalização quanto para óbito foram superiores a 30.

4 Discussão

Conforme apontado por Candido *et al*. (2020), inicialmente a epidemia no Brasil foi impulsionada principalmente pelas variantes B.1.1.28 e B.1.1.33. Como visto na figura 1, em 2020, pelo menos uma dessas duas variantes se tornou dominante para as quatro regiões em estudo. Posteriormente, passaram a circular duas variantes de origem nacional, P.1 (Gama) e P.2 (Zeta), originadas da variante B.1.1.28 (WHO, 2021). Em Manaus-AM, por exemplo, a variante Gama correspondeu a 52,2% (n = 35/67) dos casos tipificados de Sars-Cov-2 em dezembro e no mês seguinte essa proporção saltou para 85,4% (n = 41/48) (WHO, 2021). A rápida introdução dessa variante refletiu em uma nova onda de hospitalizações e óbitos, não se restringindo apenas ao Amazonas (Faria *et al*., 2021; Oliveira; Lippi; Henry, 2021; Souza *et al*., 2021).

Embora seja unânime os impactos da introdução da variante Gama nas taxas de hospitalizações e óbitos, outra característica é também evidenciada na figura 1, a heterogeneidade da epidemia de COVID-19 no Brasil. Nota-se apenas que as curvas das taxas de hospitalizações e óbitos por SRAG-Covid no Brasil e em São Paulo são semelhantes, reflexo do fato do Estado representar proporcionalmente a maior parte do "protagonismo" da epidemia no Brasil. Em contrapartida, mesmo sendo vizinho do estado de São Paulo e também localizado na região sudeste, o estado do Rio de Janeiro apresentou outra distribuição, marcada por dois picos bem definidos causados pela B.1.1.28 e P.2 (Zeta) antes de serem atingidos pela variante Gama. O Amazonas, diferenciando-se ainda das outras três regiões de estudo, teve por duas vezes um vertiginoso aumento de casos, seguido por períodos com pouca variabilidade.

Funk *et al*. (2021) identificou maior necessidade de internação em pacientes acometidos por COVID-19 oriundos da variante Gama do que pelas variantes introduzidas anteriormente, indo de encontro aos achados desse estudo em que todas as regiões apresentaram taxas de hospitalização mais elevadas durante o mesmo período como mostra a figura 1. Em relação à taxa de óbitos, apenas o estado do Rio de Janeiro apresentou taxas menores durante o domínio da variante Gama em

comparação a primeira onda. Orellana, Marrero e Horta (2021) constatou o mesmo achado para o município do Rio de Janeiro.

Embora o Brasil seja um país de dimensões continentais onde múltiplas epidemias de COVID-19 acontecem simultaneamente, para Delgado *et al*. (2020), a pandemia de COVID-19 pode ser tudo, menos imprevisível. Partindo desse pensamento, a utilização da rede *LSTM* pode ser vista como uma das possíveis soluções tecnológicas para a "predição", mesmo que de um cenário que nunca ocorreu, o cenário contrafactual. Dessa forma, como visto em outros estudos, a utilização da rede *LSTM* se mostra adequada em diversas situações cujo objetivo principal era predição de algum cenário para a COVID-19. Leite, Oliveira e Campos (2021) implementaram este modelo para a predição de casos e óbitos por COVID-19 no Brasil, Índia e EUA e Vianna e Busana (2021) utilizou-a para a predição de casos em Santa Catarina. Além da mesma rede neural, ambos os trabalhos se assemelham também na métrica para ajustes dos pesos na etapa de treinamento e validação, pois utilizam *RMSE*, enquanto este trabalho o EQM, valor ao quadrado do RMSE. Embora sejam notadas algumas semelhanças com os estudos citados, não foi encontrado em nenhum acervo público ou artigo que empregasse a modelagem *LSTM* para o caso em específico que este trabalho discorre, a geração de estimativas para um cenário contrafactual visando mensurar o impacto da introdução de uma nova variante da COVID-19.

Como visto na figura 1, tanto para o Brasil quanto para o estado de São Paulo, as taxas de hospitalização quanto as de óbitos assumiram valores muito próximo dos registrados durante a predominância das variantes anteriores. O "descolamento" entre as taxas do cenário contrafactual das taxas observadas indica que caso a variante Gama "revivesse" as métricas observadas anteriormente, ambas as taxas de hospitalização quanto as de óbitos seriam bem menores. Em outras palavras, os resultados apresentados para essas regiões indicam que após a introdução da variante Gama houve um aumento de hospitalizações e óbitos, assim como visto em outros estudos (Oliveira; Lippi; Henry, 2021; Banho *et al*., 2021).

Para o estado do Rio de Janeiro, a rede *LSTM* apresentou uma qualidade de ajuste inferior aos modelos para o Brasil e São Paulo, como pode ser visto na tabela 1. Assim, o modelo não foi capaz de capturar os pontos mais altos das taxas observadas e consequentemente, ao gerar os valores para o cenário contrafactual, suas estimativas foram possivelmente menores do que seriam se a qualidade ajuste tivesse sido melhor. De toda forma, diferente de todos os outros cenários, as taxas estimadas para o cenário contrafactual foram as que mais se aproximaram das taxas observadas. Esse comportamento se explica pelo fato do estado do Rio de Janeiro já ter passado por uma onda até mais violenta causada pela variante B.1.1.33 do que a causada pela Gama.

Por último, o estado do Amazonas que apresentou até o período analisado, duas ondas claramente bem especificadas. No histórico da pandemia do coronavírus, o Amazonas se encontra como um dos estados mais devastados pela doença. Consoante a figura 1, a taxa de hospitalização chegou a quase 50 pessoas hospitalizadas por 100 mil, número muito maior que o observado ao nível nacional, aproximadamente 15 pessoas por 100 mil. Isso é visto para a taxa de óbito, acima de 30 pessoas por 100 mil enquanto para o Brasil, essa taxa não alcançou 8 pessoas por 100 mil. Quanto ao cenário contrafactual obtido, dentre as regiões estudadas, ambos os modelos para taxa de hospitalização quanto de óbito apresentaram erros bem mais elevados. É plausível pensar que isso tenha ocorrido em decorrência do abrupto crescimento dessas taxas já no início da pandemia, dificultando a capacidade de generalização do modelo.

5 Limitações

Apesar dos idosos terem sido um dos primeiros grupos a receberem a vacina, os modelos propostos aqui não consideraram os efeitos da cobertura vacinal. Desta forma, as estimativas geradas pelos modelos podem estar sendo superestimadas por não considerar este evento. Por outro lado, mesmo com a vacinação em idosos tendo começado em janeiro de 2021, se observou picos mais elevados tanto para hospitalização quanto para óbitos durante a prevalência da variante Gama quando comparado as ondas anteriores.

Outra possibilidade é que as taxas observadas para o período de prevalência da Gama tenham sido "atenuadas" pela vacinação e a comparação com o cenário contrafactual, que foi construído utilizados dados de um período que antecede a vacinação, fosse ainda mais discrepante se as vacinas não tivessem sido aplicadas.

Ainda, a combinação das duas possibilidades citadas anteriormente pode estar ocorrendo de modo simultâneo, ou seja, os valores observados podem ter sido atenuados pela vacinação, bem como os valores gerados para o cenário contrafactual podem ter sido subestimados por não considerarem o efeito da vacinação.

Outra limitação observada é que a qualidade dos ajustes das redes *LSTM* pode estar associada às velocidades de crescimento das taxas, visto que em cenários onde se houve um aumento mais lento (Brasil e São Paulo), a rede conseguiu apresentar boa capacidade de generalização, enquanto em cenários onde as taxas cresceram mais rapidamente, os ajustes apresentaram erros maiores.

Além disso, é importante ressaltar que por esse ser um estudo de modelagem, a configuração da rede *LSTM* pode influenciar na qualidade do ajuste e consequentemente nas estimativas geradas

para o cenário contrafactual. Para minimizar este efeito nos resultados, foram adicionados ruídos nos dados de entrada a fim de criar intervalos de incertezas e foram utilizados *dropouts* visando evitar *overfitting* das redes.

6 Conclusão

O número de hospitalizações e óbitos por SRAG-Covid durante o período de prevalência da variante Gama alcançou os patamares mais elevados para todo o período observado. Aplicando a diferença entre os valores observados contra os estimados pelo modelo, viu-se que ocorreram aproximadamente 50 mil internações e mais de 30 mil óbitos em todo o Brasil apenas março de 2021. Para o mesmo período, São Paulo teve quase 15.000 e 10.000 internações e óbitos, respectivamente. Devido ao modelo aprender através da apresentação de padrões já vivenciados, as estimativas contrafactuais para o Estado do Rio de Janeiro foram muito próximas dos valores observados, pois a introdução das variantes B.1.1.33 e P.2 (Zeta) alcançaram magnitudes até maiores que à Gama. Ou seja, essas variantes apresentaram elevadas taxas de hospitalizações e óbitos, portanto, era plausível esperar que uma situação similar também ocorresse no cenário contrafactual. De modo geral, o modelo *LSTM* apresentou bons ajustes, resultando em erros de estimação próximos a zero. Dentre as localidades estudadas, apenas o Amazonas apresentou piores ajustes, que pode ser explicado pela rápida evolução do número de hospitalizações e óbitos já nas primeiras semanas da pandemia, dificultando a capacidade de generalização do modelo.

Referências

ANDERSON, D.; MCNEILL, G. **Artificial neural networks technology**. Kaman Sciences Corporation, v. 258, n. 6, 1992.

BANHO, C. A. et al. Effects of Sars-Cov-2 p. 1 **Introduction and the impact of COVID-19 vaccination on the epidemiological landscape of São José do Rio Preto**, Ml. medRxiv, Cold Spring Harbor Laboratory Press, p. 2021–07, 2021.

CANDIDO, D. S. et al. **Evolution and epidemic spread of Sars-Cov-2 in Brazil**. Science, American Association for the Advancement of Science, v. 369, n. 6508, p. 1255–1260, 2020.

DELGADO, I. F. et al. **Drug repurposing clinical trials in the search for life-saving COVID-19 therapies; research targets and methodological and ethical issues**. Vigilância Sanitária em

Debate: Sociedade, Ciência & Tecnologia, Instituto Nacional de Controle e Qualidade em Saúde, v. 8, n. 2, p. 39–53, 2020.

FARIA, N. R. et al. **Genomics and epidemiology of the P.1 Sars-Cov-2 lineage in Manaus, Brazil**. Science, American Association for the Advancement of Science, v. 372, n. 6544, p. 815–821, 2021.

FIOCRUZ. **Dashboard Rede Genômica**. Vigilância genômica do Sars-Cov-2 no Brasil. Fiocruz, 2021. Disponível em: http://www.genomahcov.fiocruz.br/dashboard/. Acesso em: 01 nov. 2021.

FRAU, L. et al. **Uncertainty estimation for machine learning models in multiphase flow applications**. Informatics, v. 8, n. 3, 2021. ISSN 2227-9709.

FUNK, T. et al. **Characteristics of Sars-Cov-2 variants of concern B.1.1. 7, b. 1.351 or P.1: Data from seven eu/eea countries, weeks 38/2020 to 10/2021**. Eurosurveillance, European Centre for Disease Prevention and Control, v. 26, n. 16, p. 2100348, 2021.

GÉRON, A. **Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems**. Sebastopol, CA: O'Reilly Media, Inc, 2019. ISBN 978-1492032649.

HAMMERSCHMIDT, K. S. de A.; SANTANA, R. F. **Saúde do idoso em tempos de pandemia COVID-19**. Universidade Federal do Paraná, v. 25, abr. 2020.

HOCHREITER, S.; SCHMIDHUBER, J. Long Short Term Memory. MIT Press - Journals, v. 9, n. 8, p. 1735–1780, nov. 1997.

HOLLAND, P. W. **Statistics and causal inference**. Journal of the American Statistical Association, Taylor & Francis, v. 81, n. 396, p. 945–960, 1986.

KHOSRAVI, A. et al. **Comprehensive review of neural network-based prediction intervals and new advances**. IEEE Transactions on Neural Networks, v. 22, n. 9, p. 1341–1356, 2011.

LEITE, S. J. O.; OLIVEIRA, R. C. L. de; CAMPOS, L. M. L. de. **Predição de séries temporais da COVID-19: Uma avaliação de redes neurais com células LSTM**. 2021.

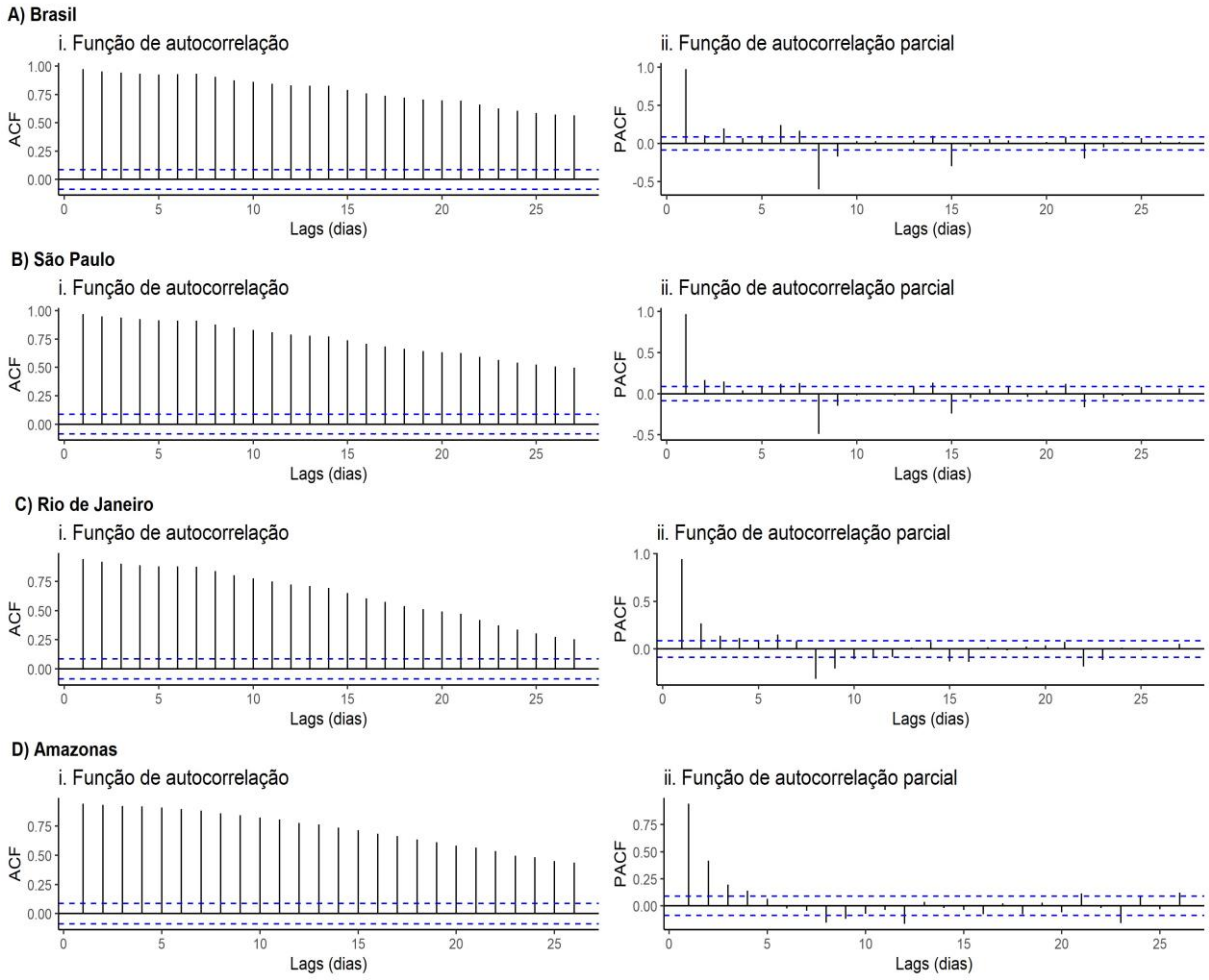
MINISTÉRIO DA SAÚDE. **Boletim epidemiológico especial 44 - Doença pelo Coronavírus COVID-19**. Ministério da Saúde, Secretaria de Vigilância em Saúde. – Brasília: Ministério da Saúde: Secretaria de Vigilância em Saúde. Departamento de Análise em Saúde e Doenças não Transmissíveis. 2021.

MINISTÉRIO DA SAÚDE. **Guia de vigilância epidemiológica Emergência de saúde pública de Importância nacional pela Doença pelo coronavírus 2019 – COVID-19**. Ministério da Saúde, Secretaria de Vigilância em Saúde. – Brasília: Ministério da Saúde: Secretaria de Vigilância em Saúde. Departamento de Análise em Saúde e Doenças não Transmissíveis. 2021.

OLIVEIRA, M. H. S. de; LIPPI, G.; HENRY, B. M. **Sudden rise in COVID-19 case fatality among young and middle-aged adults in the south of Brazil after identification of the novel b. 1.1. 28.1 (P.1) Sars-Cov-2 strain: analysis of data from the state of Parana**. MedRxiv, Cold Spring Harbor Laboratory Press, p. 2021–03, 2021.

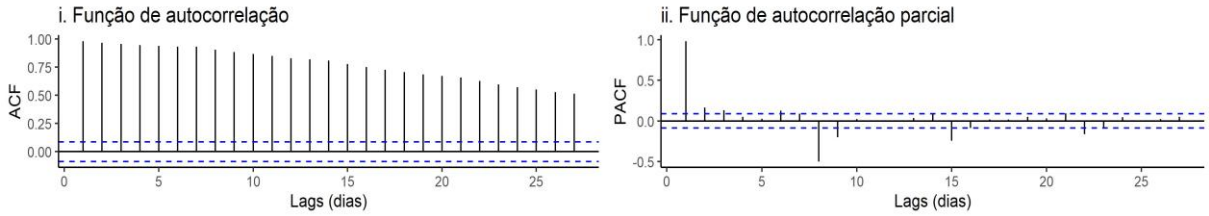
7 Anexo

Função de autocorrelação e autocorrelação parcial para série temporal de taxa hospitalização por SRAG-Covid

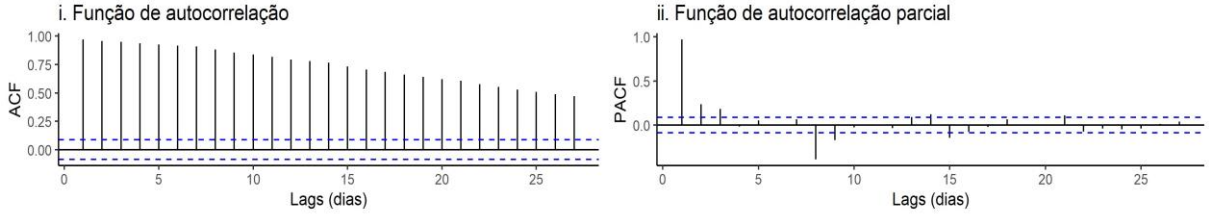


Função de autocorrelação e autocorrelação parcial para série temporal de taxa de óbito por SRAG-Covid

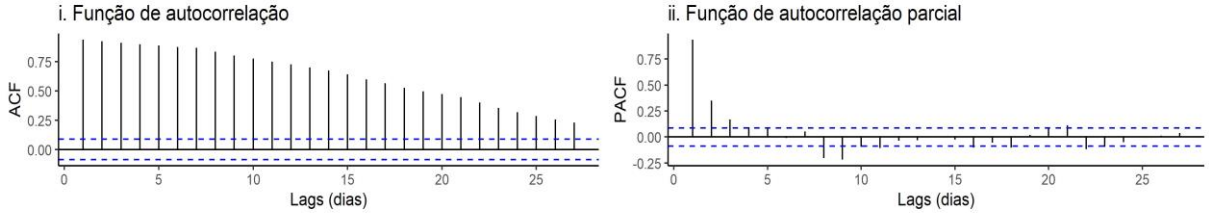
A) Brasil



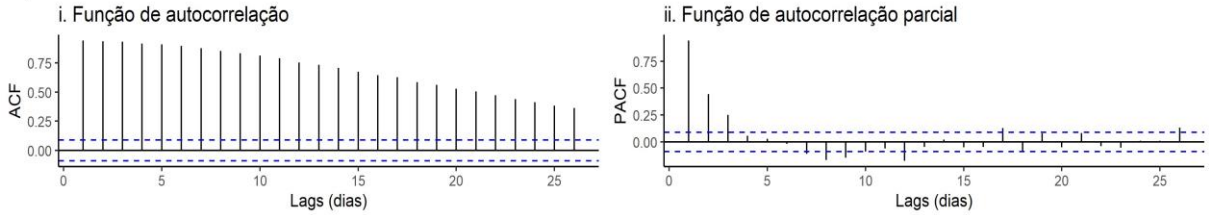
B) São Paulo

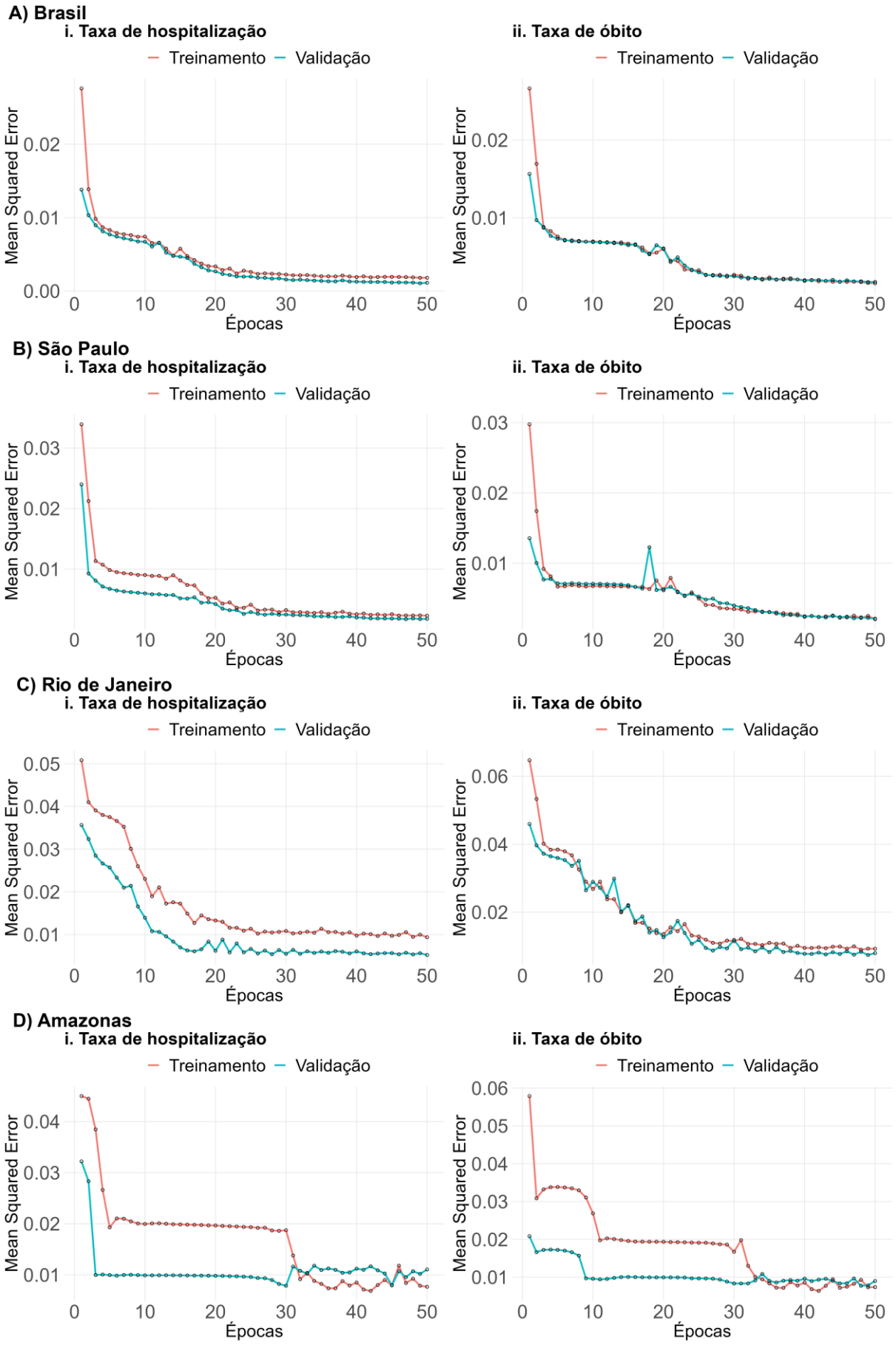


C) Rio de Janeiro



D) Amazonas




A) Brasil
i. Taxa de hospitalização
ii. Taxa de óbito
B) São Paulo
i. Taxa de hospitalização
ii. Taxa de óbito
C) Rio de Janeiro
i. Taxa de hospitalização
ii. Taxa de óbito
D) Amazonas
i. Taxa de hospitalização
ii. Taxa de óbito
Treinamento
Validação
Mean Squared Error
Épocas

9 CONSIDERAÇÕES FINAIS

Esta tese explorou a dinamicidade da COVID-19 nas dimensões do espaço-tempo para os dois primeiros anos da pandemia no Brasil e não se limitando apenas ao uso de metodologias já consolidadas, mas foi também um espaço de idealização, teste, construção e validação de uma nova abordagem metodológica. Como parte fundamental para a construção do conhecimento, além de uma extensa revisão bibliográfica, essa tese teve suas concepções baseadas na elaboração de informações baseadas em dados. Diante disso, o primeiro artigo dessa tese se preocupou em mostrar a velocidade de propagação, alcance, bem como o risco associado de se pertencer a *clusters* de casos e óbitos. Mostrou-se que à falta de estratégias coordenadas e políticas públicas eficazes voltadas para a mitigação do espalhamento do vírus da COVID-19 resultaram na coexistência de múltiplas epidemias, que surgiram e findaram-se em diferentes momentos e lugares, criando polos que se destacaram por apresentarem níveis mais elevados de riscos relativos. A combinação de mapas e intensidades de cores, permitiu expressar a movimentação do vírus. Outro aspecto importante do primeiro artigo foi utilizar uma técnica já consolidada, o *SaTScan*^{TM}, inovou em relacionar o *clusters* com as linhagens circulantes.

O segundo artigo visou trazer os conceitos de contrafactualidade muito discutidos em delineamentos como caso-controle e coortes para os estudos ecológicos. Neste artigo foi proposto uma nova forma de estudar causalidade ao nível populacional. Mostrou que é possível superar o problema da simultaneidade ao construir um modelo que consegue gerar estimativas produzidas pelo mesmo processo estocástico que antecedeu o evento de interesse. Dessa forma, ao gerar predições através desse modelo, esses valores, ao serem comparados aos valores observados para o evento de interesse, representam a ideia de que, se tudo tivesse mantido inalterado, o cenário observado seria algo próximo ao construído pelo modelo contrafactual. A materialização dessa proposta foi posta em prática utilizando dados de SRAG-Covid no Brasil. Como resultado, a rede *LSTM* se mostrou capaz de aprender os padrões vivenciados e de repetir esse mesmo comportamento durante o período de ocorrência da linhagem Gama.

O terceiro e último artigo apostou na aplicabilidade da metodologia construída no segundo artigo, demonstrando em 4 cenários distintos aspectos positivos e negativos dessa nova proposta de avaliação do efeito causal em estudos ecológicos. Assim como no artigo 2, os modelos foram treinados apenas com os próprios valores do evento de interesse, sendo no artigo 3, número de casos hospitalizados e óbitos por SRAG-Covid em pessoas acima dos 60

anos. Como resultado, evidenciou que essa proposta metodológica cumpre o seu papel e dessa forma, se mostra eficaz na estimação de medidas de impactos.

Essa compilação de análises evidencia que os estudos ecológicos se demonstram capazes de contextualizar a dinamicidade da pandemia da COVID-19 no Brasil. Além disso, os conceitos de contrafactualidade até então empregados na epidemiologia são passiveis de novas aplicações, sendo estes renovados e inovados pelas recentes técnicas de aprendizagem de máquina.

Cabe ao epidemiologista, não apenas intervir no problema de saúde, mas também identificar os métodos que por ele são utilizados, sendo aqueles, aos quais não atendam da forma adequada ao que se foram propostos, novamente, intervir, norteando-se pela metodologia cientifica, na reformulação ou proposição de novas técnicas. Isso deve ser visto como algo além de um compromisso inerente a formação, deve ser visto como um ato de devolução à sociedade pelos investimentos aplicados através dos impostos, deve ser visto como uma oportunidade de tornar a sociedade menos desigual, destinando que as políticas públicas alcancem, dessa vez, de forma desigual, na medida das suas desigualdades, aqueles que as mais necessitam.

REFERÊNCIAS

ABELLAN, J. J.; RICHARDSON, S.; BEST, N. Use of space–time models to investigate the stability of patterns of disease. Environmental Health Perspectives, **National Institute of Environmental Health Science**, v. 116, n. 8, p. 1111, 2008.

ADAMI, E. R.; IMIG, D. C.; RIBAS, J. L. Covid 19: revisão, relato de caso e perspectivas. **Revista UNIANDRADE**, v. 21, n. 1, p. 36–48, 2020.

ADAMOSKI, D. et al. Sars-Cov-2 delta and omicron variants surge in Curitiba, southern Brazil, and its impact on overall COVID-19 lethality. **Viruses**, MDPI, v. 14, n. 4, p. 809, 2022.

ALCANTARA, L. C. J. et al. Sars-Cov-2 epidemic in Brazil: how the displacement of variants has driven distinct epidemic waves. **Virus Research**, Elsevier, v. 315, p. 198785, 2022.

ANDERSON, D.; MCNEILL, G. Artificial neural networks technology. **Kaman Sciences Corporation**, v. 258, n. 6, 1992.

ARAÚJO, L. F. S. C. D. et al. **Causas e modelos causais em psiquiatria**. 2013.

ARORA, S. et al. Literature review of omicron: a grim reality amidst COVID-19. **Microorganisms**, MDPI, v. 10, n. 2, p. 451, 2022.

BAILEY, T. C.; GATRELL, A. C. Interactive spatial data analysis. **Longman Scientific & Technical Essex**, v. 413, 1995.

BARATA, R. B. Causalidade e epidemiologia. **FapUNIFESP** (SciELO), v. 4, n. 1, p. 31–49, jun. 1997.

BARBOSA, I. R. et al. Incidence of and mortality from COVID-19 in the older Brazilian population and its relationship with contextual indicators: an ecological study. **FapUNIFESP** (SciELO), v. 23, n. 1, 2020.

BARCELLOS, C.; VILLELA, D. A. M. **COVID-19 no Brasil:** cenários epidemiológicos e vigilância em saúde. Série Informação para ação na COVID-19 Fiocruz, 2021.

BILLAH, M. A.; MIAH, M. M.; KHAN, M. N. Reproductive number of coronavirus: A systematic review and meta-analysis based on global level evidence. **PloS one, Public Library of Science**, v. 15, n. 11, p. e0242128, 2020.

BLANGIARDO, M. et al. Spatial and spatio-temporal models with r-inla**. Spatial and spatio-temporal epidemiology**, Elsevier, v. 4, p. 33–49, 2013.

BORDALLO, B.; BELLAS, M.; CORTEZ, A. **Severe COVID-19:** what have we learned with the immunopathogenesis? DOI: https://doi.org/10.1186/s42358-020-00151-7. Advances in Rheuma-tology, BMC, 2020.

CABRERO, G. R. La crisis del coronavirus y su impacto en las residencias de personas mayores en España. **FapUNIFESP** (SciELO), v. 25, n. 6, p. 1996–1996, jun. 2020.

CDC. The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) — China, 2020. **Chinese Center for Disease Control and Prevention**, v. 2, n. 8, p. 113–122, 2020.

CDC Centers for Disease Control and Prevention. **Sars-Cov-2 Variant Classifications and Definitions**. CDC, 2023. Disponível em: https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications. html. Acesso em: 16 jun. 2023.

CHEN, A. T. et al. **COVID-19 CG enables Sars-Cov-2 mutation and lineage tracking by locations and dates of interest**. eLife Sciences Publications, Ltd, v. 10, fev. 2021.

CHERIAN, S. et al. Sars-Cov-2 spike mutations, l452r, t478k, e484q and p681r, in the second wave of COVID-19 in maharashtra, india. **Microorganisms**, MDPI, v. 9, n. 7, p. 1542, 2021.

CHMIELEWSKA, A. M. et al. Immune response against Sars-Cov-2 variants: the role of neutralization assays. **Npj Vaccines**, Nature Publishing Group UK London, v. 6, n. 1, p. 142, 2021.

CORRÊA, P. R. L. et al. A importância da vigilância de casos e óbitos e a epidemia da COVID-19 em Belo Horizonte, 2020. **Revista Brasileira de Epidemiologia**, SciELO Brasil, v. 23, 2020.

COVID, E. E. Principais variantes do Sars-Cov-2 notificadas no Brasil. **RBAC**, v. 53, n. 2, p. 109–116, 2021.

CZERESNIA, D.; ALBUQUERQUE, M. d. F. M. d. **Modelos de inferência causal:** análise crítica da utilização da estatística na epidemiologia. [S.l.]: SciELO Brasil, 1995.

CZERESNIA, D.; ALBUQUERQUE, M. de Fátima Militão de. Modelos de inferência causal: análise crítica da utilização da estatística na epidemiologia. **FapUNIFESP** (SciELO), v. 29, n. 5, p. 415–423, out. 1995.

DORIGATTI, I. et al. **Report 4**: Severity of 2019-novel coronavirus (ncov). Imperial College London, 2020.

DURLAUF, S.; BLUME, L. E. **The new Palgrave dictionary of economics**. [S.l.]: Springer, 2016.

DWASS, M. **Modified randomization tests for nonparametric hypotheses**. The Annals of Mathematical Statistics, JSTOR, p. 181–187, 1957.

EDARA, V.-V. et al. Infection and vaccine-induced neutralizing-antibody responses to the Sars-Cov-2 b. 1.617 variants. **New England Journal of Medicine**, Mass Medical Soc, v. 385, n. 7, p. 664–666, 2021.

ESPENHAIN, L. et al. Epidemiological characterisation of the first 785 Sars-Cov-2 omicron variant cases in denmark, december 2021. **Eurosurveillance**, European Centre for Disease Prevention and Control, v. 26, n. 50, p. 2101146, 2021.

FARIA, N. R. et al. Genomics and epidemiology of the P.1 Sars-Cov-2 lineage in Manaus, Brazil. **Science, American Association for the Advancement of Science**, v. 372, n. 6544, p. 815–821, 2021.

FERNANDES, M. A. C.; NETO, A. D. D.; BEZERRA, J. B. **A neural network model applied to the detection of digital signals**. In: ITS'98 Proceedings. SBT/IEEE International Telecommunications Symposium (Cat. No. 98EX202). IEEE, 1998. p. 279-283.

FIOCRUZ. Dashboard Rede Genômica. **Vigilância genômica do Sars-Cov-2 no Brasil**. [S.l.]: Fiocruz, 2023. Disponível em: https://www.genomahcov.fiocruz.br/dashboard-pt/. Acesso em: 13 jun. 2023.

FONSECA, N.; SÁNCHEZ-RIVERO, M. Causalidade em economia com séries temporais: uma visita guiada desde a antiguidade clássica. **FapUNIFESP** (SciELO), v. 30, n. 3, p. 999–1027, dez. 2020.

FRAU, L. et al. Uncertainty estimation for machine learning models in multiphase flow applications. **Informatics**, v. 8, n. 3, 2021. ISSN 2227-9709.

FREITAS, C. M. d. et al. **Boletim observatório fiocruz COVID-19:** Boletim especial: balanço de dois anos da pandemia COVID-19: janeiro de 2020 a janeiro de 2022. Fiocruz, 2022.

GAL, Y.; GHAHRAMANI, Z. **Dropout as a Bayesian Approximation:** Representing Model Uncertainty in Deep Learning. 2016.

GÉRON, A. **Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow:** concepts, tools, and techniques to build intelligent systems. Sebastopol, CA: O'Reilly Media, Inc. ISBN 978-1492032649. 2019.

HAMMERSCHMIDT, K. S. de A.; SANTANA, R. F. **Saúde do idoso em tempos de Pandemia COVID-19**. Universidade Federal do Parana, v. 25, abr. 2020.

HOCHREITER, S.; SCHMIDHUBER, J. Long Short Term Memory. **MIT Press - Journals**, v. 9, n. 8, p. 1735–1780, nov. 1997.

HOLLAND, P. W. Statistics and causal inference. **Journal of the American Statistical Association**, Taylor & Francis, v. 81, n. 396, p. 945–960, 1986.

HUME, D. **Investigação sobre o entendimento humano**. [S.l.]: Leya, 2013.

HUSSAIN, M. et al. Structural variations in human ACE2 may influence its binding with Sars-Cov-2 spike protein. **Wiley**, v. 92, n. 9, p. 1580–1586, abr. 2020.

IMAI, N. et al. Report 3: **Transmissibility of 2019-ncov**. Imperial College London, 2020.

KANNAN, S.; ALI, P. S. S.; SHEEZA, A. Omicron (B.1.1.529) - variant of concern-molecular profile and epidemiology: a mini review. **Eur. Rev. Med. Pharmacol**. Sci, v. 25, n. 24, p. 8019–8022, 2021.

KEELING, M. J.; ROHANI, P. **Modeling Infectious Diseases in Humans and Animals**. Princeton University Press. ISBN 9781400841035. 2011.

KHOSRAVI, A. et al. Comprehensive review of neural network-based prediction intervals and new advances. **IEEE Transactions on Neural Networks**, v. 22, n. 9, p. 1341–1356, 2011.

KIM, J.; KORMAN, D. Z.; SOSA, E. **Metaphysics**: An Anthology, 2nd Edition. [S.l.]: Wiley-Blackwell, 2011.

KULLDORFF, M. **Spatial scan statistics:** models, calculations, and applications. [S.l.]: Springer, 1999.

KULLDORFF, M. S. f. v. . SaTScan, 2022. 2022.

LAWSON, A. B. **Bayesian disease mapping:** hierarchical modeling in spatial epidemiology. [S.l.]: Chapman and Hall/CRC, 2013.

LI, J. et al. Sars-Cov-2 omicron BA.1.1 is highly resistant to antibody neutralization of convalescent serum from the origin strain. **Virus Research**, Elsevier, v. 332, p. 199131, 2023.

LI, Q. et al. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. **Massachusetts Medical Society**, v. 382, n. 13, p. 1199–1207, mar. 2020.

LIMA, C. M. A. de O. Information about the new coronavirus disease (COVID-19). **FapUNIFESP** (SciELO), v. 53, n. 2, p. V–VI, abr. 2020.

LIMA-COSTA, M. F. Envelhecimento no Brasil e coronavírus: iniciativa ELSI- COVID-19. **FapUNIFESP** (SciELO), v. 36, n. suppl 3, 2020.

LIU, C. et al. Reduced neutralization of Sars-Cov-2 B.1.617 by vaccine and convalescent serum. **Cell, Elsevier**, v. 184, n. 16, p. 4220–4236, 2021.

LUIZ, R. R.; STRUCHINER, C. J. **Inferencia causal em epidemiologia:** o modelo de respostas potenciais. Rio de Janeiro, Brazil: Editora FIOCRUZ, 2002. ISBN 85-7541-010-5.

MACHADO, C. J. et al. Estimativas de impacto da COVID-19 na mortalidade de idosos institucionalizados no Brasil. **FapUNIFESP** (SciELO), v. 25, n. 9, p. 3437–3444, set. 2020.

MACKIE, J. L. Causes and conditions. **American philosophical quarterly**, JSTOR, v. 2, n. 4, p. 245–264, 1965.

MAZUMDER, H.; HOSSAIN, M. M.; DAS, A. Geriatric care during public health emergencies: Lessons learned from novel corona virus disease (COVID-19) pandemic. **Informa UK Limited**, v. 63, n. 4, p. 257–258, mar. 2020.

MEDEIROS, E. A. S. A luta dos profissionais de saúde no enfrentamento da COVID-19. **Acta Paulista de Enfermagem**, v. 33, 2020.

MENEZES, M. Portal Fiocruz - **Pesquisa sugere maior risco de reinfecção pela variante Delta**. 2021.

MINISTÉRIO DA SAÚDE. **Boletim epidemiológico 1 - Doença pelo Coronavírus COVID-19**. Ministério da Saúde, Secretaria de Vigilância em Saúde. – Brasília: Ministério da Saúde: Secretaria de Vigilância em Saúde, 2020. v. 51.

MINISTÉRIO DA SAÚDE. **Boletim epidemiológico 5 - Doença pelo Coronavírus 2019 Ampliação da Vigilância, Medidas não Farmacológicas e Descentralização do Diagnóstico Laboratorial**. Ministério da Saúde, Secretaria de Vigilância em Saúde. – Brasília: Ministério da Saúde: Secretaria de Vigilância em Saúde, 2020. v. 05.

MINISTÉRIO DA SAÚDE. **Boletim epidemiológico especial 44 - Doença pelo Coronavírus COVID-19**. Ministério da Saúde, Secretaria de Vigilância em Saúde. – Brasília: Ministério da Saúde: Secretaria de Vigilância em Saúde. Departamento de Análise em Saúde e Doenças não Transmissíveis, 2021.

MINISTÉRIO DA SAÚDE. **Guia de vigilância epidemiológica Emergência de saúde pública de Importância nacional pela Doença pelo coronavírus 2019 – COVID-19**. Ministério da Saúde, Secretaria de Vigilância em Saúde. – Brasília: Ministério da Saúde: Secretaria de Vigilância em Saúde. Departamento de Análise em Saúde e Doenças não Transmissíveis, 2021.

MINISTÉRIO DA SAÚDE. **Guia de Vigilância Epidemiológica COVID-19:** Emergência de Saúde Pública de Importância Nacional pela Doença pelo Coronavírus 2019. [S.l.]: Ministério da Saúde, 2022. Disponível em: https://www.gov.br/saude/pt-br/coronavirus/publicacoes-tecnicas/ guias-e-planos/guia-de-vigilancia-epidemiologica-COVID-19/view. Acesso em: 16 jun. 2023.

MINISTÉRIO DA SAÚDE. **Atendimento e fatores de risco**. [S.l.]: Ministério da Saúde, 2023.
Disponível em: https://www.gov.br/saude/pt-br/coronavirus/atendimento-tratamento-e-fatores-de-risco. Acesso em: 17 jun. 2023.

MORAES, E. N. de et al. COVID-19 nas instituições de longa permanência para idosos: estratégias de rastreamento laboratorial e prevenção da propagação da doença. **FapUNIFESP** (SciELO), v. 25, n. 9, p. 3445–3458, set. 2020.

MOREIRA, F. R. et al. **Epidemic spread of Sars-Cov-2 lineage B.1.1.7 in Brazil**. MDPI AG, v. 13, n. 6, p. 984, maio 2021.

MORETTIN, P.; TOLOI, C. de C. **Análise de séries temporais**. [S.l.]: Edgard Blucher, 2006. (ABE - Projeto Fisher). ISBN 9788521203896.

NAVECA, F. et al. **Sars-Cov-2 reinfection by the new variant of concern (voc) P.1 in amazonas, Brazil**. Virol Org, 2021.

NAVECA, F. G. et al. COVID-19 in Amazonas, Brazil, was driven by the persistence of endemic lineages and P.1 emergence. **Springer Science and Business Media LLC**, v. 27, n. 7, p. 1230–1238, maio 2021.

NEUBERG, L. G. Causality: Models, reasoning, and inference, by Judea Pearl. Cambridge university press, 2000. **Econometric Theory**, Cambridge University Press, v. 19, n. 4, p. 675–685, 2003.

NIQUINI, R. P. et al. SRAG por COVID-19 no Brasil: descrição e comparação de características demográficas e comorbidades com SRAG por influenza e com a população geral. **FapUNIFESP** (SciELO), v. 36, n. 7, 2020.

NOGUEIRA, H. **Os lugares e a saúde.** [S.l.]: Imprensa da Universidade de Coimbra/Coimbra University Press, 2008. v. 6.

ONU. **Shared responsibility, global solidarity:** responding to the socio-economic impacts of COVID-19. BMC, 2020. Visitada em 21-10-2021.

PEARCE, T. et al. **High-Quality Prediction Intervals for Deep Learning:** A Distribution-Free, Ensembled Approach. 2018.

PEREZ-GUZMAN, P. N. et al. **Clinical characteristics and predictors of outcomes of hospitalized patients with coronavirus disease 2019 in a multiethnic london national health service trust:** A retrospective cohort study. Oxford University Press (OUP). Ago. 2020.

PESSOA, F. **À dúvida.** In: Mensagem. [S.l.]: Edições Ática, 1942. p. 77.

PLESSIS, L. du et al. Establishment and lineage dynamics of the Sars-Cov-2 epidemic in the UK. **American Association for the Advancement of Science** (AAAS), v. 371, n. 6530, p. 708–712, fev. 2021.

PORCHEDDU, R. et al. Similarity in case fatality rates (CFR) of COVID-19/SARS-COV-2 in Italy and China. **Journal of Infection in Developing Countries**, v. 14, n. 02, p. 125–128, fev. 2020.

PORTO, E. F. et al. Mortalidade por COVID-19 no Brasil: perfil sociodemográfico das primeiras semanas. **Research, Society and Development**, v. 10, n. 1, p. e34210111588, jan. 2021.

REMUZZI, A.; REMUZZI, G. **COVID-19 and Ialy:** what next? Elsevier BV, v. 395, n. 10231, p. 1225–1228, abr. 2020.

RESENDE, P. C. et al. **Identification of a new b.1.1.33 Sars-Cov-2 variant of interest (voi) circulating in Brazil with mutation e484k and multiple deletions in the amino (n)-terminal domain of the spike protein**. 2021.

RESENDE, P. C. et al**.** A potential Sars-Cov-2 variant of interest (voi) harboring mutation e484k in the spike protein was identified within lineage B.1.1.33 circulating in Brazil. **Viruses**, Multidisciplinary Digital Publishing Institute, v. 13, n. 5, p. 724, 2021.

ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. **American Psychological Association** (APA), v. 65, n. 6, p. 386–408, 1958.

ROTHMAN, K.; GREENLAND, S.; LASH, T. **Epidemiologia Moderna**-3ª Edição. [S.l.]: Artmed Editora, 2016.

RUBIN, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. **American Psychological Association** (APA), v. 66, n. 5, p. 688–701, 1974.

SABINO, E. C. et al. Resurgence of COVID-19 in manaus, Brazil, despite high seroprevalence. **Elsevier BV**, v. 397, n. 10273, p. 452–455, fev. 2021.

SANTOS, H. G. dos et al. Machine learning para análises preditivas em saúde: exemplo de aplicação para predizer óbito em idosos de São Paulo, Brasil. **FapUNIFESP** (SciELO), v. 35, n. 7, 2019.

SIDDIQI, H. K.; MEHRA, M. R. COVID-19 illness in native and immunosuppressed states: A clinical–therapeutic staging proposal. **The Journal of Heart and Lung Transplantation**, Elsevier, v. 39, n. 5, p. 405, 2020.

SILVA, M. S. da et al. Early detection of Sars-Cov-2 p.1 variant in southern Brazil and reinfection of the same patient by P.2. **FapUNIFESP** (SciELO), v. 63, 2021.

SRIVASTAVA, N. et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. **J. Machine Learning Res**., v. 15, p. 1929–1958, 2014.

TANG, X. et al. On the origin and continuing evolution of Sars-Cov-2. **Oxford University Press** (OUP), v. 7, n. 6, p. 1012–1023, mar. 2020.

TAUBES, G. **Epidemiology faces its limits**. Science, v. 269, p. 164–169, 1995.

TEGALLY, H. et al. Detection of a Sars-Cov-2 variant of concern in South Africa. **Nature, Nature Publishing Group,** v. 592, n. 7854, p. 438–443, 2021.

UZUNIAN, A. **Coronavirus Sars-Cov-2 and COVID-19**. GN1 Genesis Network, 2020.

VALLBO, E. **A comparison between bootstrap and dropout for uncertainty estimates of time series forecast using a convolutional neural network**. In: . [S.l.: s.n.], 2019.

VILLELA, D. A. M. The value of mitigating epidemic peaks of COVID-19 for more effective public health responses. **FapUNIFESP** (SciELO), v. 53, 2020.

VOLZ, E. et al. Assessing transmissibility of Sars-Cov-2 lineage B. 1.1. 7 in England. **Nature, Nature Publishing Group**, v. 593, n. 7858, p. 266–269, 2021.

WERNECK, G. L.; CARVALHO, M. S. **A pandemia de COVID-19 no Brasil**: crônica de uma crise sanitária anunciada. [S.l.]: SciELO Public Health, 2020.

WHO. **WHO Coronavirus (COVID-19) Dashboard**. [S.l.]: WHO, 2023. Disponível em: https://COVID-19.who.int/. Acesso em: 13 jun. 2023.

WILLIAMSON, J. **Causality**. In: . Springer Netherlands, 2007. p. 95–126.

WOOLDRIDGE, J. **Introductory econometrics:** a modern approach. Mason, OH: Thomson/South-Western, ISBN 9780324323481. 2006.

WU, J. T.; LEUNG, K.; LEUNG, G. M. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. **Elsevier BV**, v. 395, n. 10225, p. 689–697, fev. 2020.

WU, Z.; MCGOOGAN, J. M. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China. **American Medical Association** (AMA), v. 323, n. 13, p. 1239, abr. 2020.

ZHANG, Z. et al. Interval prediction method based on long-short term memory networks for system integrated of hydro, wind and solar power. **Elsevier BV**, v. 158, p. 6176–6182, fev. 2019.

ZHOU, R. et al. Vaccine-breakthrough infection by the Sars-Cov-2 omicron variant elicits broadly cross-reactive immune responses. **Clinical and translational medicine**, Wiley-Blackwell, v. 12, n. 1, 2022.

ZHOU, T. et al. Preliminary prediction of the basic reproduction number of the Wuhan novel coronavirus 2019-ncov. **Journal of Evidence-Based Medicine**, Wiley Online Library, v. 13, n. 1, p. 3–7, 2020.

ZHOU, W. et al. Effects of media reporting on mitigating spread of COVID-19 in the early phase of the outbreak. **Mathematical Biosciences and Engineering**, v. 17, n. 3, p. 2693–2707, 2020.

ZHU, L.; LAPTEV, N. **Deep and confident prediction for time series at uber. 2017 IEEE International Conference on Data Mining Workshops** (ICDMW), IEEE, Nov 2017.