



Paulo Henrique Couto Simões

**Análise da contribuição das características
associadas à evolução dos óbitos por
COVID-19 nos estados brasileiros
utilizando os valores de Shapley**

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para
obtenção do grau de Mestre pelo Programa de Pós-
Graduação em Engenharia de Produção, do
Departamento de Engenharia Industrial da PUC-Rio.

Orientadora: Prof. Paula Medina Maçaira Louro

Coorientadora: Prof. Fernanda Araujo Baião

Rio de janeiro

agosto 2022



Paulo Henrique Couto Simões

**Análise da contribuição das características
associadas à evolução dos óbitos por
COVID-19 nos estados brasileiros
utilizando os valores de Shapley**

Dissertação apresentada como requisito parcial para
obtenção do grau de Mestre pelo Programa de Pós-
Graduação em Engenharia de Produção, do
Departamento de Engenharia Industrial da PUC-Rio.
Aprovada pela Comissão Examinadora abaixo:

Prof. Paula Medina Maçaira Louro

Orientadora

Departamento de Engenharia Industrial - PUC-Rio

Prof. Fernanda Araujo Baião

Coorientadora

Departamento de Engenharia Industrial - PUC-Rio

Prof. Otavio Tavares Ranzani

FMUSP

Prof. Vinicius Layter Xavier

UERJ

Rio de Janeiro, 24 de agosto de 2022

Todos os direitos reservados. A reprodução, total ou parcial do trabalho, é proibida sem a autorização da universidade, da autora e do orientador.

Paulo Henrique Couto Simões

Graduou-se em Estatística pela Universidade do Estado do Rio de Janeiro (UERJ) em 2015. Trabalha na área de Planejamento e Controle para a Força Aérea Brasileira. É Professor substituto no departamento de Estatística da UERJ. É integrante do projeto de extensão Solução Estatística Junior (UERJ), objetivando formar profissionais nas áreas de Estatística e Ciências Atuariais para o uso de inteligência artificial. É pesquisador no Effectiveness of the COVID-19 vaccination in Brazil using mobile data (EFFECT-Brazil). Sendo este um dos 10 projetos selecionados entre mais de 400 propostas para o Grand Challenge of the International COVID-19 Data Alliance (ICODA) Data Science Initiative e financiado pela Gates Foundation.

Ficha Catalográfica

Simões, Paulo Henrique Couto

Análise da contribuição das características associadas à evolução dos óbitos por COVID-19 nos estados brasileiros utilizando os valores de Shapley / Paulo Henrique Couto Simões; orientadora: Paula Medina Maçaira Louro; coorientadora: Fernanda Araujo Baião. – 2022.

42 f. : il. color. ; 30 cm

Dissertação (mestrado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Industrial, 2022.

Inclui bibliografia

1. Engenharia Industrial – Teses. 2. COVID-19. 3. Medidas de contenção. 4. Cobertura vacinal. 5. XGBoost. 6. Valor de Shapley. I. Louro, Paula Medina Maçaira. II. Baião, Fernanda Araujo. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Industrial. IV. Título.

CDD: 658.5

Agradecimentos

Agradeço aos meus familiares, amigos e professores, pelo carinho e pelas orientações tão essenciais nesta jornada.

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

Resumo

Simões, Paulo Henrique Couto; Louro, Paula Medina Maçaira (Orientadora); Baião, Fernanda Araujo (Coorientadora). **Análise da contribuição das características associadas à evolução dos óbitos por COVID-19 nos estados brasileiros utilizando os valores de Shapley.** Rio de Janeiro, 2022. 42p. Dissertação de Mestrado – Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

Este trabalho propõe um método para hierarquizar a contribuição de diferentes estratégias para conter a evolução da pandemia de COVID-19 em diferentes estados do Brasil, nos períodos pré- e pós-vacinação. O método proposto incluiu o aprendizado automático de modelos de regressão utilizando o algoritmo de aprendizado de máquina XGBoost, e aplicou a teoria dos jogos cooperativos de Shapley para quantificar a contribuição das características analisadas para a variável-alvo. Para interpretar o modelo globalmente, foi usado o *SHapley Additive exPlanations* (SHAP), que é um algoritmo baseado na teoria de Shapley. Os resultados de avaliação do método apontaram a sua eficácia para quantificar a contribuição de cada variável de forma robusta, e revelam que os percentuais de cobertura vacinal de primeira e segunda dose, além do fechamento das escolas, foram as medidas que tiveram maior contribuição na evolução do número de casos e óbitos por COVID-19. A ponderação das variáveis pode ajudar os atores responsáveis na elaboração de políticas públicas para minimizar os efeitos socioeconômicos em suas regiões, dado que o Brasil é um país que possui extrema desigualdade social.

Palavras-chave

COVID-19; medidas de contenção; cobertura vacinal; XGBoost; valor de Shapley.

Abstract

Simões, Paulo Henrique Couto; Louro, Paula Medina Maçaira (Advisor); Baião, Fernanda Araujo (Co-advisor). **Analysis of the contribution of characteristics associated with the evolution of deaths from COVID-19 in Brazilian states using Shapley values.** Rio de Janeiro, 2022. 42p. Dissertação de Mestrado – Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

This work proposes a method to rank the contribution of different strategies to contain the evolution of the COVID-19 pandemic in different states of Brazil, in the pre- and post-vaccination periods. The proposed method included the automatic learning of regression models using the XGBoost machine learning algorithm, and applied Shapley's cooperative game theory to quantify the contribution of the analyzed characteristics to the target variable. To interpret the model globally, the SHapley Additive exPlanations (SHAP) was used, which is an algorithm based on Shapley's theory. The evaluation results point to its efficacy to quantify the contribution of each variable in a robust way, and reveal that the percentages of first and second dose vaccination coverage, in addition to the closing of schools, were the measures that had the greatest contribution in the evolution of the number of cases and deaths due to COVID-19. The weighting of variables can help the actors responsible in the elaboration of public policies to minimize the socioeconomic effects in their regions, since Brazil is a country that has extreme social inequality.

Keywords

COVID-19; containment measures; vaccination coverage; XGBoost; Shapley's value.

Sumário

1. Introdução.....	9
2. Revisão teórica	11
2.1 A pandemia da COVID-19	11
2.2 Estudos com a metodologia de Shapley	14
3. Metodologia.....	18
3.1 Análise estatística	20
4. Evolução dos casos de COVID-19 no Rio Grande do Sul	23
5. Evolução dos óbitos por COVID-19 no Brasil e seus estados	27
6. Discussão.....	34
7. Conclusão	39
8. Referências	40

Lista de figuras

Figura 1 - Fluxograma do processo.	18
Figura 2 - Índice de restrição e média móvel de 7 dias.	24
Figura 3 - Defasagem de restrição e média móvel de 7 dias.	25
Figura 4 - Média dos valores Shapley.	26
Figura 5 - Quadro de séries temporais das médias móveis de sete dias para óbitos nas escalas nacional (BR) e estadual (AC-TO) para o ano de 2021.	28
Figura 6 – Boxplot da defasagem em dias dos 26 estados brasileiros, do Distrito Federal e do Brasil entre variáveis preditoras do seu modelo e os óbitos.	30
Figura 7 - Features relacionadas aos óbitos por COVID-19 em ordem de frequências selecionadas pelos 28 modelos do estudo e separados por cores relacionadas às respectivas dimensões.	31
Figura 8 - Importância das features via valor de Shapley ilustra a classificação das diferentes medidas (cobertura vacinal, medidas de restrição e padrões de mobilidade) em termos de seus impactos médios absolutos no Brasil e nos estados.	32
Figura 9 - O ranking dos boxplots das diferentes features em ordem decrescente de contribuição em termos de seus impactos mediano na evolução do número de óbitos por COVID-19.	33

1. Introdução

A partir de dezembro de 2019, vários casos de pacientes com pneumonia foram associados à COVID-19, uma doença provocada por um novo coronavírus humano (SARS-CoV-2) (ZHU et al., 2020). Tal vírus demonstrou uma capacidade abundante de transmissões inter-humanas, espalhou-se rapidamente pelo mundo e se tornou uma pandemia com milhões de mortes, em março de 2020. Os pacientes infectados apresentaram sintomas variados e seus casos variaram de indivíduos assintomáticos até mesmo ao óbito. A discriminação de pacientes gravemente enfermos daqueles com sintomas leves pode ajudar a compreender as variações individualizadas do prognóstico da COVID-19.

Diversos estudos foram publicados revelando que os especialistas estavam corretos em alertar que as medidas não farmacológicas (distanciamento social, interrupção de atividades não essenciais, higiene respiratória e uso de máscaras) eram as melhores opções para conter a disseminação do vírus (FLAXMAN et al., 2020; SHARMA et al., 2021), em um período no qual as vacinas ainda não haviam sido disponibilizadas para a população. As medidas de contenção foram tomadas em diferentes tempos no Brasil, e como não houve uma centralização dessas medidas por parte do governo federal, cada estado executou sua própria política.

O surgimento da COVID-19 estabeleceu o SARS-CoV-2 como o terceiro coronavírus altamente patogênico em humanos no século XXI, após a síndrome respiratória aguda grave (SARS) em 2003 e a síndrome respiratória do Oriente Médio (MERS) em 2012 (DROSTEN et al., 2003; ZAKI et al., 2012). Estudos apontam que o SARS-CoV-2 está mais relacionado ao coronavírus de morcego semelhante ao SARS do que ao MERS (LU et al., 2020; WU et al., 2020). A mortalidade de indivíduos gravemente enfermos com a COVID-19 é consideravelmente alta. O tempo de sobrevivência dos pacientes que chegam a óbito pode ser entre uma e duas semanas após sua admissão em um leito de UTI (YANG X. et al., 2020).

Uma motivação para o trabalho foi entender a importância de compreender melhor o quanto as diferentes estratégias adotadas em diferentes localidades brasileiras, e os diferentes graus de adesão da população a tais estratégias prescritas, contribuíram para reduzir a disseminação do vírus. Uma mudança significativa

dentro do período pandêmico foi a introdução da vacinação. Sendo assim, esta dissertação foi dividida em dois estudos.

O primeiro estudo tem como objetivo classificar e hierarquizar as medidas de contenção coletadas pelo projeto Oxford COVID-19 Government Response Tracker (OxCGRT) quanto ao seu impacto sobre a evolução dos casos de COVID-19 no estado do Rio Grande do Sul, no período pré-vacinação. O objetivo foi obter evidências de quais medidas surtiram mais efeito e que poderiam ter menor impacto social (visto que o custo econômico de implantar soluções como lockdown é grande), visando auxiliar os gestores públicos. Foram analisadas medidas de contenção de nove categorias, que poderiam explicar as variações nos casos associados à COVID-19.

No segundo estudo, tornou-se oportuno realizar outra análise incluindo a cobertura vacinal e a mobilidade populacional. Deste modo, foram analisadas 10 medidas de contenção, duas medidas de cobertura vacinal e seis medidas de mobilidade quanto a sua contribuição sobre a evolução do número de óbitos relacionados à pandemia no Brasil, em todos seus 26 estados, e no Distrito Federal. Neste caso, os dados sobre vacinação foram extraídos do Sistema de Informação do Programa Nacional de Imunização (SI-PNI) e os dados sobre mobilidade disponibilizados pelo Google (COVID-19 Community Mobility Reports).

Sendo assim, além desta introdução, a presente dissertação revela, no Capítulo 2, uma revisão teórica geral sobre a COVID-19. No Capítulo 3 é apresentado o método proposto no trabalho, que traz as informações sobre as bases de dados, o algoritmo XGBoost e o valor de Shapley. Os Capítulos 4 e 5 contêm, respectivamente, as contribuições que foram implementadas no contexto acadêmico: no Capítulo 4 descreve-se o estudo que foi realizado para o estado do Rio Grande do Sul, onde foi objetivo classificar e hierarquizar a contribuição das medidas de contenção sobre os casos de COVID-19 no início da pandemia, e no Capítulo 5 descreve-se o estudo que foi realizado para o Brasil, seus estados mais o Distrito Federal, onde foi objetivo classificar e hierarquizar a contribuição das medidas de contenção, da cobertura vacinal e da mobilidade populacional sobre a evolução dos óbitos por COVID-19 após o início da vacinação. O Capítulo 6 discute os resultados encontrados nestes estudos e em outros trabalhos relacionados à COVID-19. Por fim, o Capítulo 7 trata da conclusão deste trabalho, com a proposição de trabalhos futuros.

2. Revisão teórica

Nos últimos anos diversos trabalhos e pesquisas foram elaborados relacionados à COVID-19. A crise da pandemia criou uma necessidade sem precedentes e exigiu que as pessoas aprendessem algumas novas habilidades rapidamente (SOHRABI, et al., 2020). Além disso, medidas de contenção como distanciamento social, maiores cuidados com a higiene pessoal, lockdown, entre outras, foram tomadas com intuito de conter a propagação do vírus (CUCINOTTA & VANELLI, 2020).

A doença, que atingiu os cinco continentes do planeta, tornou-se uma pandemia com milhões de mortes. Vacinas foram rapidamente desenvolvidas como a AstraZeneca (Oxford), CoronaVac (Sinovar), Pfizer (BioNTech), Janssen (Johnson & Johnson), Moderna e Sputnik V (vacina russa). Ainda assim, continuam os cuidados necessários para evitar o surgimento de novas variantes do vírus.

2.1 A pandemia da COVID-19

Em dezembro de 2019, um surto local de pneumonia de causa inicialmente desconhecida foi detectado em Wuhan (província de Hubei, China), e foi rapidamente determinado como sendo causado por um novo coronavírus humano, o Coronavírus da Síndrome Respiratória Aguda Grave 2 (SARS-CoV-2). A principal causa da mortalidade deste vírus é a Síndrome do Desconforto Respiratório Agudo (SDRA). Desde então, o surto se espalhou para todas as províncias da China continental, bem como 27 outros países e regiões, com mais de 70.000 casos confirmados até 17 de fevereiro de 2020 (DONG, DU & GARDNER, 2020).

A Organização Mundial da Saúde (OMS), em 11 de março de 2020, declarou o novo surto de coronavírus humano (COVID-19) como uma pandemia global. Em uma entrevista coletiva o Diretor-Geral da OMS, Dr. Tedros Adhanom Ghebreyesus, pediu aos países que tomassem medidas para conter o vírus. Não se sabe quando a pandemia terminará, visto que a mesma atingiu vários países e milhões de pessoas infectadas no mundo (SOHRABI, et al., 2020). Em 12 de março de 2020, a COVID-19 foi confirmada em 125.048 pessoas em todo o mundo,

levando uma mortalidade de aproximadamente 3,7%. Em comparação, a influenza possui uma taxa de mortalidade de menos de 1% (MEHTA et al., 2020).

Devido ao surto do SARS-CoV-2, a população mundial passou a tomar algumas precauções no dia a dia, como evitar contato próximo com pessoas que estão doentes, assegurar medidas preventivas diárias e lavar as mãos frequentemente com água e sabão, especialmente depois de assoar o nariz, tossir ou espirrar, ou depois de ter ido a lugar público. Se não houver água e sabão disponíveis, deve-se usar um desinfetante para as mãos que contenha pelo menos 70% de álcool. Evitar contato em seu rosto, nariz, olhos, etc. Evitar aglomerações, especialmente em espaços mal ventilados. O risco de exposição a vírus respiratórios como o SARS-COV-2 aumenta em ambientes lotados e fechados, com pouca circulação de ar, principalmente se houver pessoas doentes na multidão (CUCINOTTA & VANELLI, 2020).

A transmissão pessoa a pessoa foi demonstrada. Em janeiro de 2020 ocorreu um estudo de caso em Anyang, China. A amostra foi realizada considerando um grupo familiar de seis indivíduos, sendo cinco pacientes com febre e sintomas respiratórios e um membro assintomático. Uma análise detalhada dos prontuários dos pacientes foi realizada e todos foram submetidos à tomografia computadorizada de tórax. Dois pacientes desenvolveram pneumonia grave. As outras infecções foram moderadas. Todos os pacientes sintomáticos apresentavam anormalidades na tomografia computadorizada de tórax. Todos os pacientes sintomáticos tinham níveis aumentados de proteína C reativa e contagem de linfócitos reduzida. A sequência de eventos relatada dentro do estudo sugeriu que o novo coronavírus pode ter sido transmitido pelo portador assintomático. O período de incubação para o paciente 1 foi de 19 dias, o que é longo, mas dentro do intervalo relatado de zero a 24 dias (BAI et al., 2020).

Identificar características comuns de casos de COVID-19 para mais bem compreender quais fatores promovem a transmissão secundária (entre pessoas que não moram juntas) da doença não é muito simples. No Japão, uma amostra de 110 casos foi examinada entre 11 grupos e casos esporádicos, investigando quem adquiriu a infecção de quem. O número de casos secundários gerados por cada caso primário foi calculado usando dados de rastreamento de contato. Dos 110 casos examinados, 24,6% foram casos primários que geraram casos secundários. A probabilidade de um caso primário transmitir COVID-19 em um ambiente fechado

foi 18,7 vezes maior em comparação com um ambiente ao ar livre (NISHIURA et al., 2020).

O principal método de desacelerar a taxa de infecção foi impor regulamentos de distanciamento social, onde as pessoas ficaram restritas a suas próprias casas, exceto as atividades econômicas essenciais. A quarentena contra a COVID-19 retirou automóveis das ruas, diminuiu a produção das fábricas e reduziu a emissão de poluentes em boa parte do mundo (CUCINOTTA & VANELLI, 2020). Em Wuhan, foi feito um estudo em duas etapas nas concentrações de quatro poluentes atmosféricos. Os níveis de concentrações de NO₂ caíram em até 63% pré-bloqueio, enquanto as concentrações de PM₁₀ caíram em uma quantidade semelhante, mas por um período mais curto. O bloqueio não teve impacto perceptível nas concentrações de SO₂ ou CO (COLE, ELLIOTT & LUI, 2020). O grande efeito no NO₂ sugere que as políticas para reduzir as emissões dos veículos, como o incentivo à eletrificação de carros e ônibus, teriam benefícios bastante consideráveis para a saúde.

Uma série de modelos de aprendizado de máquina foi usado a um conjunto de dados sobre a mortalidade de pacientes com COVID-19. O conjunto de dados foi composto por amostras de sangue de 375 pacientes internados em um hospital na região de Wuhan, China, onde 201 sobreviveram à hospitalização e 174 morreram. O foco deste estudo não foi apenas ver qual modelo é capaz de obter a maior precisão absoluta, e sim na interpretação do que os modelos fornecem. Foi descoberto que as variáveis idade, dias no hospital, linfócitos e neutrófilos são preditores importantes e robustos ao prever a mortalidade de um paciente. Além disso, os algoritmos que foram usados permitiram observar o impacto marginal de cada variável em um nível de paciente caso a caso, o que pode ajudar os profissionais a detectar facilmente padrões anômalos (SMITH & ALVAREZ, 2021).

Após o início da cobertura vacinal, surgiram diversos estudos por todo o mundo sobre a efetividade das vacinas (DAGAN et al., 2021; CHEMAITELLY et al., 2021; LOPEZ BERNAL et al., 2021; JARA et al., 2021). A efetividade da Pfizer foi de 72% para os dias entre 14 e 20 após a primeira dose. Contudo, após sete ou mais dias da segunda dose, essa eficácia chegou a 92% para o desenvolvimento grave da doença (DAGAN et al., 2021). A Moderna, contra casos críticos ou fatais, teve efetividade de aproximadamente 82% após a primeira dose e de 96% após a

segunda dose (CHEMAITELLY et al., 2021). Em relação a CoronaVac, a eficácia de duas doses é de aproximadamente 75% entre as pessoas com a variante Alfa e 67% para a variante Delta (LOPEZ BERNAL et al., 2021). Na imunização completa, a estimativa da eficácia das vacinas foi de aproximadamente 66% para a prevenção, 88% para a prevenção de hospitalização, 90% para a prevenção de internações em UTI's, e 86% para a prevenção de óbito (JARA et al., 2021).

Estudos mostraram que indicadores de renda influenciam o acesso às vacinas. Países de extrema pobreza tiveram baixo acesso às vacinas. Já países como Reino Unido, Estados Unidos, China e Israel tiveram prioridade (DE OLIVEIRA et al., 2021). A pandemia estimulou ainda mais as diferenças globais econômicas presentes entre países ricos e pobres. Entre alguns elementos, destacam-se: a importância do fortalecimento de sistemas de saúde universais, da ciência, tecnologia, inovação e das bases econômicas, em países com diferentes graus de desenvolvimento (LIMA & GADELHA, 2021).

2.2 Estudos com a metodologia de Shapley

O conhecimento sobre a disseminação socioeconômica da primeira onda de infecções por COVID-19 na Alemanha foi avaliado em diferentes estudos. Em um deles, os autores Doblhammer, Reinke e Kreft (2022) procuraram saber se as taxas de incidência de COVID-19 seriam distintas entre os municípios de acordo com suas características socioeconômicas usando uma ampla variedade de indicadores. Foram usados 204.217 diagnósticos de COVID-19 na população total alemã, distinguindo cinco períodos distintos entre 1 de janeiro e 23 de julho de 2020. Para cada período, foram calculadas as taxas de incidência padronizadas por idade de diagnósticos de COVID-19 em nível de município. Foram treinados modelos de aumento de gradiente para prever as taxas de incidência padronizadas por idade com as macroestruturas dos municípios e usados os valores *Shapley Additive Explanations* (SHAP) para representar as 20 características mais proeminentes em termos de correlações negativas/positivas com a variável de resultado.

A primeira onda de COVID-19 começou como uma doença em condados rurais ricos no sul da Alemanha e se alastrou para condados urbanos e agrícolas mais pobres durante o curso da primeira onda. A alta incidência padronizada por

idade em condados de baixo nível socioeconômico tornou-se mais pronunciada a partir do segundo período de bloqueio, quando os condados ricos pareciam estar mais protegidos. As características relacionadas com as características econômicas e educativas da população jovem de um concelho desempenharam um papel importante desde o início da pandemia até à segunda fase de confinamento, assim como as características relacionadas com a população residente em lares de idosos. Aqueles relacionados à migração internacional e uma grande proporção de estrangeiros que vivem em um município tornaram-se importantes no período pós-bloqueio. A alta mobilidade de grupos de alto nível socioeconômico pode impulsionar a pandemia no início das ondas, enquanto medidas de mitigação e crenças sobre a gravidade da pandemia, bem como o cumprimento de medidas de mitigação, podem colocar grupos de baixo nível socioeconômico em maior risco posteriormente (DOBLHAMMER, REINKE, KREFT, 2022).

A COVID-19 pode ocorrer de forma assintomática, como uma doença semelhante à gripe, ou como formas mais graves, que caracterizam a Síndrome Respiratória Aguda Grave (SARS). Sua taxa de mortalidade é maior em indivíduos acima de 80 anos e em pessoas com comorbidades, portanto, constituem o grupo de risco para formas graves da doença. Oliveira et al. (2021) analisaram os fatores associados ao óbito em casos confirmados de COVID-19 no estado do Rio de Janeiro. Este estudo transversal avaliou a associação entre variáveis demográficas, clínicas e epidemiológicas individuais e o desfecho (óbito) utilizando dados dos sistemas de informação do Sistema Único de Saúde. Foi usado o modelo *Extreme Boosting Gradient* (XGBoost) para analisar os dados, que usa árvores de decisão ponderadas pela dificuldade de estimativa. Para avaliar a relevância de cada variável independente, utilizou-se a métrica *Shapley Additive exPlanations* (SHAP).

A partir das probabilidades geradas pelo modelo XGBoost, transformamos os dados para o logaritmo de *odds* para estimar a *odds ratio* para cada variável independente. Este estudo mostrou que indivíduos mais velhos de cor da pele negra, cardiopatas ou diabéticos que apresentavam dispneia ou febre tinham maior probabilidade de morrer. A identificação precoce de pacientes que podem evoluir para uma forma mais grave da doença pode ajudar a melhorar o manejo clínico de pacientes com COVID-19 e, portanto, é essencial para reduzir a letalidade da doença (OLIVEIRA et al., 2021).

Vários modelos foram construídos em diversos trabalhos para estudar as características e tendências de disseminação da pandemia. No entanto, devido às informações e fontes de dados limitadas, a compreensão da propagação e do impacto da pandemia de COVID-19 ainda é restrita. Nesse sentido, JING et al. (2022) não usaram apenas dados históricos diários de séries temporais, mas também atributos regionais como fatores geográficos e locais, que podem ter desempenhado um papel importante na confirmação de casos do vírus em certas regiões. A esse respeito, o estudo realizou uma análise transversal abrangente e uma previsão baseada em dados dessa pandemia.

Os recursos críticos, que têm uma influência significativa na taxa de infecção, foram determinados usando o algoritmo *Extreme Gradient Boosting* e *Shapley Additive Explanation*. A comparação foi realizada utilizando o *Random Forest* e modelos *Light Gradient Boosting*. Para prever o número de casos confirmados de COVID-19 com mais precisão, uma rede neural recorrente baseada em atenção de dois estágios (DA-RNN) foi aplicada. Este modelo tem melhor desempenho que o *Support Vector Regression* e a rede codificador-decodificador no conjunto de dados experimental. O desempenho do modelo foi avaliado à luz de três métricas estatísticas (MAE, RMSE e R^2) (JING et al., 2022).

Para identificar de modo quantitativo as medidas de controle ideais para os reguladores minimizarem as taxas de crescimento (taxa G) e mortalidade (taxa D) do COVID-19 no contexto atual, Chew e Zhang (2022) desenvolveram uma abordagem de engenharia multiescala, a qual engloba uma série de análises sistemáticas. A primeira modelagem preditiva é a escala global da taxa G e da taxa D, seguida pela determinação dos fatores de controle mais eficazes que podem minimizar melhor ambos os parâmetros ao longo do tempo por meio de Inteligência Artificial explicável com método SHAP. A segunda modelagem foi a escala continental (mesma previsão preditiva da taxa G e da taxa D em todos os continentes), seguida da realização de análises SHAP explicáveis para determinar os fatores de controle mais eficazes para os respectivos continentes. E a terceira modelagem foi a escala de país, onde foram usados diferentes países em três agrupamentos principais para identificar o conjunto universal de medidas de controle eficazes.

Ao usarem o período histórico entre 2 de maio de 2020 e 1 de outubro de 2021, as pontuações médias do MAPE para prever a taxa G e a taxa D ficaram

abaixo de 10% nas escalas global e continental. Sistemáticamente, foi demonstrado de modo quantitativo que as três principais medidas de controle mais eficazes para os reguladores minimizarem a taxa G universalmente são a *COVID-CONTACT-TRACING*, a *PUBLIC-GATHERING-RULES* e a *COVID-STRINGENCY-INDEX*. Já os fatores de controle relacionados a taxa D dependem especificamente do cenário de modelagem (CHEW & ZHANG, 2022).

3. Metodologia

O procedimento de construção e as etapas pertinentes ao método proposto por este trabalho podem ser visualizados na Figura 1. O fluxograma apresenta os passos utilizados desde a fonte e preparação dos dados até a elaboração dos modelos finais que estimam a contribuição das variáveis.

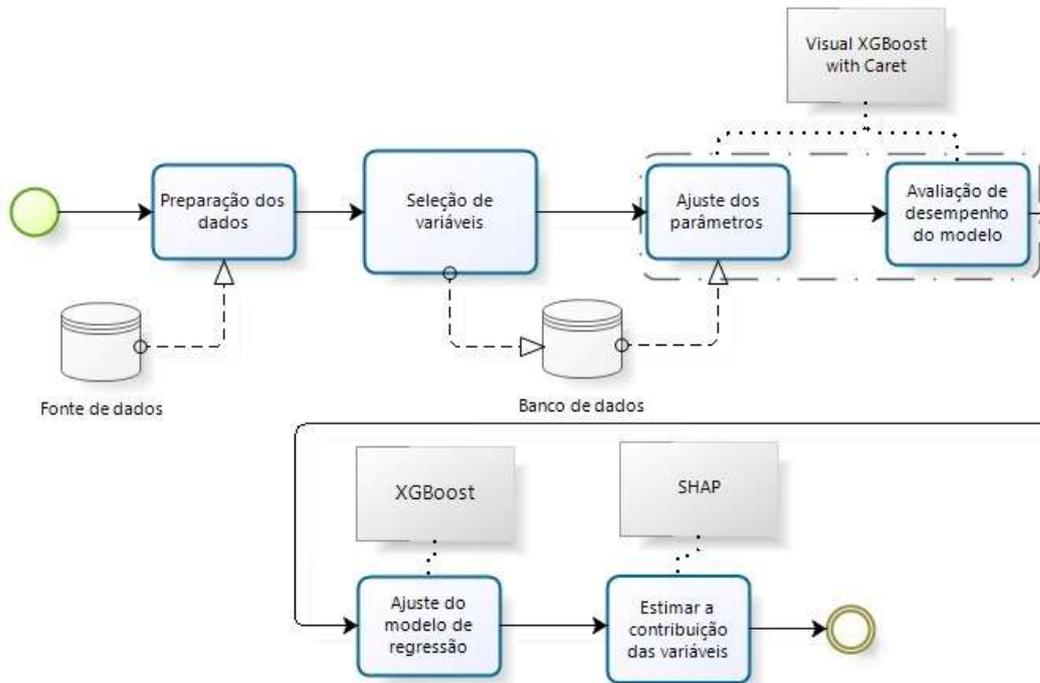


Figura 1 - Fluxograma do processo.

A iniciativa Oxford COVID-19 Government Response Tracker (OxCGRT) da Universidade de Oxford é responsável por monitorar e relatar as diferentes respostas governamentais à COVID-19, que são codificadas em 23 indicadores, como fechamento de escolas, restrições de viagem e política de vacinação. Essas políticas de medidas de contenção são dimensionadas para refletir a extensão da ação do governo e as pontuações são agregadas em um conjunto de índices de políticas. Na primeira parte deste trabalho, foi utilizado o conjunto de dados do relatório do projeto OxCGRT - BSG-WP-2020-036-PT, pois trata das respostas governamentais à pandemia de COVID-19 no Brasil, incluindo medidas de contenção (fechamento de escolas e locais de trabalho, *lockdown*, restrição de viagens), campanhas de informação para a população, dentre outras (HALE et al., 2020).

O projeto brasileiro dentro da iniciativa OxCGRT é realizado em parceria com a EBAPE-FGV do Rio de Janeiro, USP, e a Universidade do Estado do Pará (HALE et al., 2020). Os dados oferecem uma maneira sistemática de rastrear as respostas do governo à COVID-19 no país e nas jurisdições estaduais. Esse banco de dados fornece informações sobre os 26 estados, o distrito federal e 52 cidades, incluindo 26 capitais e as cidades com maior número de habitantes em cada estado que não são capitais. O banco de dados inclui 12 indicadores sobre políticas de fechamento, contenção e saúde seguidas pelo número de novos casos diários de COVID-19 entre 1 de janeiro e 14 de outubro de 2020.

A base de dados estadual do Brasil é apresentada em diferentes formatos. Foram usados dados que capturam o número total de apólices em vigor em uma determinada jurisdição. Inclui medidas adotadas pelos mais altos níveis de governo que podem substituir as políticas locais, como a proibição de entrada de estrangeiros pelo governo federal, que também se aplica a todas as unidades federativas.

Os nove indicadores utilizados na primeira parte deste trabalho são os mesmos utilizados para a construção da base de dados “OxCGRT Índice de Restrições”: C1- fechamento de escolas, C2- fechamento de locais de trabalho, C3- cancelamento de eventos públicos, C4- restrições a aglomerações, C5- restrições a transporte público, C6- confinamento domiciliar, C7- restrições à circulação interna, C8- restrições de viagens internacionais e H1- campanha de informação ao público.

Na segunda parte deste trabalho, o estudo foi feito analisando 10 medidas de contenção (as mesmas nove anteriores mais a H6- coberturas faciais), duas medidas de cobertura vacinal (D1- cobertura vacinal de primeira dose e D2- cobertura vacinal de segunda e dose única) e seis medidas de mobilidade (M1- varejo e recreação, M2- mercearia e farmácia, M3- parques, M4- estações de trânsito, M5- locais de trabalho e M6- residencial).

Cada uma destas 3 dimensões de análise (medidas de contenção, cobertura vacinal e mobilidade) tem um papel importante na compreensão como um todo:

- Medidas de contenção: tem a atribuição de mensurar as respostas políticas que cada governo tomou. Registra essas políticas em uma escala para refletir a extensão da ação do governo no combate ao avanço da COVID-19;

- Cobertura vacinal: é a medida percentual da quantidade de pessoas vacinadas em uma determinada região (país ou estado);
- Mobilidade: os relatórios de mobilidade têm como objetivo fornecer informações sobre como o padrão de circulação da população mudou em determinada região geográfica (que pode ser uma cidade, um estado ou um país) em cada dia em relação à circulação média da população aferida antes do início da pandemia.

Esses indicadores foram relacionados com a evolução do número de óbitos à pandemia no Brasil e todos seus 26 estados mais o Distrito Federal. Neste caso, os dados sobre vacinação foram retirados do Sistema de Informação do Programa Nacional de Imunização (SI-PNI) (MINISTÉRIO DA SAÚDE, 2022) e os dados sobre mobilidade retirados do Google (COVID-19 *Community Mobility Reports*) (AKTAY et al., 2022). Os indicadores foram uniformizados entre 0 e 1 (3.1).

$$Xp_i = \frac{X_i}{MAX(X_i)}, i = 1,2,3, \dots, n \quad (3.1)$$

onde X_i é o valor diário de cada uma das variáveis utilizadas.

O *software* utilizado para as análises em ambos os estudos foi o R (R CORE TEAM, 2019).

3.1 Análise estatística

Para modelar o impacto dos indicadores na redução do número de casos diários de COVID-19, foi construído um modelo de árvore para regressão de Poisson (3.2) utilizando o *Extreme Gradient Boosting* (XGBoost). A regressão de Poisson é indicada para casos de contagens discretas (MCCULLAGH & NELDER, 1989).

$$\text{Log}(E(y)) = \beta_0 + \beta_1 C_1 + \dots + \beta_9 H_1 \quad (3.2)$$

Para modelar o impacto dos indicadores na evolução do número de óbitos diários de COVID-19, foi construído um modelo de árvore para regressão de

múltipla utilizando o *Extreme Gradient Boosting* (XGBoost) (3.3). Foi usado também o método de seleção de variáveis *stepwise*. O XGBoost é usado para resolver problemas de classificação, regressão e ranqueamento. Para o Brasil e cada estado mais o Distrito Federal, foi selecionado um conjunto de variáveis e uma defasagem em relação a variável resposta (HASTIE, TIBSHIRANI, FRIEDMAN, 2009).

$$L(t) = -\frac{1}{2} \sum_{j=1}^T \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \quad (3.3)$$

onde I é conjunto de índices das amostras contidas na folha da árvore de decisão, g_i é o gradiente (derivada de primeira ordem) e h_i é o hessiano (derivada de segunda ordem).

O XGBoost pertence à categoria de modelos de *Ensemble Learning*, que se baseiam no treinamento de vários modelos simples para produzir um modelo final mais robusto. O algoritmo constrói árvores de decisão para cada iteração, onde os modelos não são mais treinados de forma independente, mas sequencial, com base no ajuste dos modelos previamente treinados (CHEN et al., 2015).

Para otimizar os hiperparâmetros do modelo XGBoost, foi usada uma implementação visual do pacote *caret* do R. Este é um processo de otimização dos hiperparâmetros do XGBoost. Ele se concentra em avaliar visualmente as etapas do processo enquanto constrói uma lógica, simples e gradual, para ajustar o modelo. Isso usa cinco etapas para otimizar os hiperparâmetros: na primeira etapa, são estabelecidos a taxa de aprendizagem η e o número de iterações; na segunda etapa, define-se a profundidade máxima de uma árvore e a soma mínima do peso da instância. Na terceira etapa do processo, a razão de subamostra das colunas e a razão de subamostra das linhas são otimizadas. Na quarta etapa, são realizados experimentos com diferentes valores de gama. E, por fim, na quinta etapa, a taxa de aprendizado é otimizada (HASTIE, TIBSHIRANI, FRIEDMAN, 2009).

A ideia é usar inicialmente uma taxa de aprendizado mais alta para ajustar os hiperparâmetros. Esses parâmetros serão usados para ajustar o modelo final com maior número de árvores e menor taxa de aprendizado. Outra obrigação nesta etapa do processo é usar a validação cruzada, pois uma simples reamostragem aleatória

da série temporal não é a melhor maneira de reamostrar os dados. Foram aplicadas técnicas de bloco para dividir o conjunto de dados original em conjuntos de treinamento e teste ao longo da série temporal, usando a função *timeslices* do pacote *caret*.

A regressão de Poisson foi escolhida para este estudo por utilizar variáveis de contagem. O algoritmo XGBoost ajusta esta regressão de Poisson através da construção de uma floresta aleatória (ou seja, um conjunto de árvores de decisão). Para resolver a falta de inteligibilidade deste modelo de floresta aleatória, o *Shapley Additive Explanations* (SHAP) (3.4) foi aplicado para calcular os valores de contribuição das características que compõem o conjunto de dados de entrada (ou seja, os indicadores) (LUNDBERG et. al., 2019; MOLNAR, 2020).

$$g(z') = \phi_0 + \sum_{j=1}^M \phi_j z'_j \quad (3.4)$$

onde g é o modelo de explicação, $z' \in \{0,1\}^M$ é o vetor de coalizão, M é o tamanho máximo da coalizão e ϕ_j é a atribuição de característica para uma característica j , os valores de Shapley.

Especificamente, o valor SHAP é caracterizado pela contribuição marginal média entre todas as combinações possíveis dos fatores que compõem o modelo e representam uma distribuição justa da importância total entre eles (ALGABA, FRAGNELLI, SÁNCHEZ-SORIANO, 2019).

4. Evolução dos casos de COVID-19 no Rio Grande do Sul

Este estudo de caso foi enviado e aceito para publicação no LIV Simpósio Brasileiro de Pesquisa Operacional (SBPO, 2022). O objetivo foi inferir qual medida surtiu mais efeito no contexto de auxiliar os atores responsáveis pela tomada de decisão, e que poderia ter menor impacto social.

Desta forma, o presente trabalho almeja mensurar o efeito das medidas de contenção na redução do número de novos casos de COVID-19 utilizando informações do estado do Rio Grande do Sul, onde o primeiro caso foi identificado em 29 de fevereiro de 2020 e na Semana Epidemiológica 41 havia 34.337 casos confirmados, sendo o Rio Grande do Sul o estado que mais chamou a atenção das autoridades nacionais pela rápida expansão do número de casos na respectiva semana epidemiológica (BOLETIM EPIDEMIOLÓGICO COVID-19, 2020).

A série temporal utilizada corresponde ao número de novos casos diários de COVID-19 no estado do Rio Grande do Sul entre janeiro de 2020 e 10 de dezembro de 2020 (345 dias) - período anterior ao início da vacinação. Foi aplicada uma média móvel de 7 dias (7MM) para suavizar a série temporal de casos de COVID-19, com o objetivo de minimizar a variabilidade da taxa de notificação em cada semana, incluindo as altas taxas de subnotificação nos finais de semana. A fonte de dados usada foi conjunto de dados do relatório do projeto OxCGRT - BSG-WP-2020-036-PT.

Os nove indicadores usados foram: C1- fechamento de escolas, C2- fechamento de locais de trabalho, C3- cancelamento de eventos públicos, C4- restrições a aglomerações, C5- restrições a transporte público, C6- confinamento domiciliar, C7- restrições à circulação interna, C8- restrições de viagens internacionais e H1- campanha de informação ao público.

A Figura 2 mostra o índice de restrição e a média móvel de 7 dias dos casos. É possível observar um aumento muito rápido das restrições impostas pelo governo do Rio Grande do Sul no período inicial da pandemia e, em seguida, uma diminuição dessas restrições, enquanto o número de casos continuou aumentando.

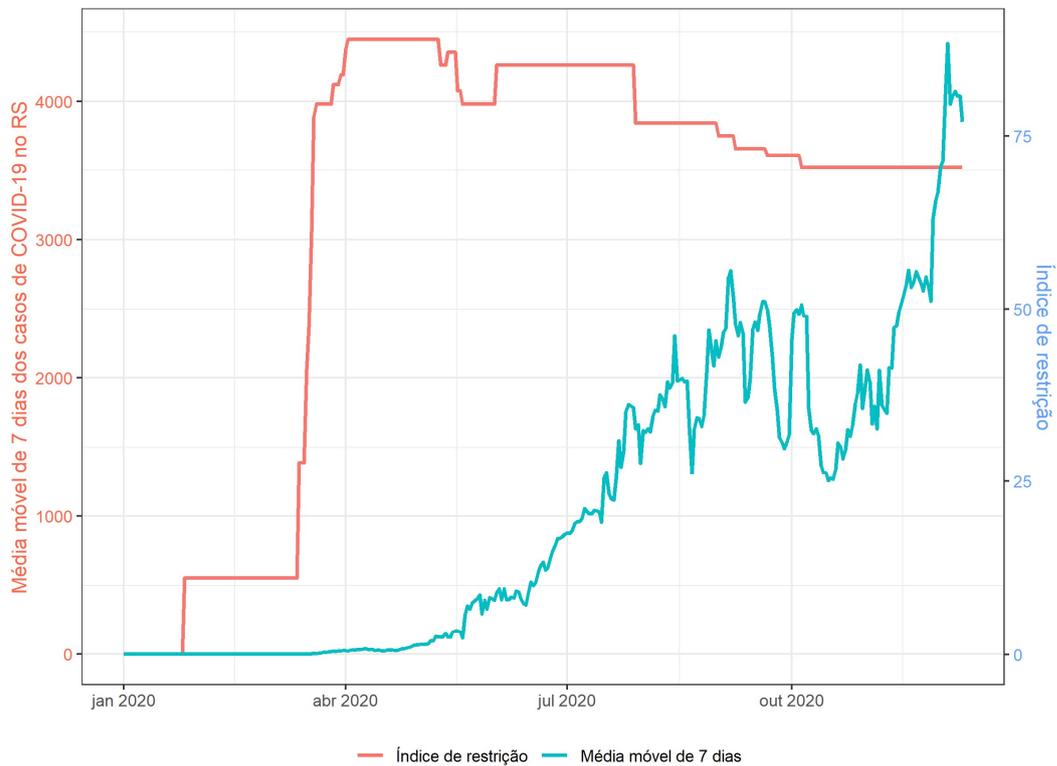


Figura 2 - Índice de restrição e média móvel de 7 dias.

Uma vez que as medidas de contenção não têm efeito imediato, mas sim entre 7 a 14 dias após serem impostas, foi avaliada a adequação do modelo de regressão de Poisson construído pelo algoritmo XGBoost, aplicando uma defasagem entre 1 e 21 dias. A defasagem que minimizou a raiz quadrada do erro médio (RMSE) foi de 13 dias, conforme mostrado na Figura 3. Isso significa que, em média, as medidas de contenção são mais efetivas 13 dias após sua implementação.

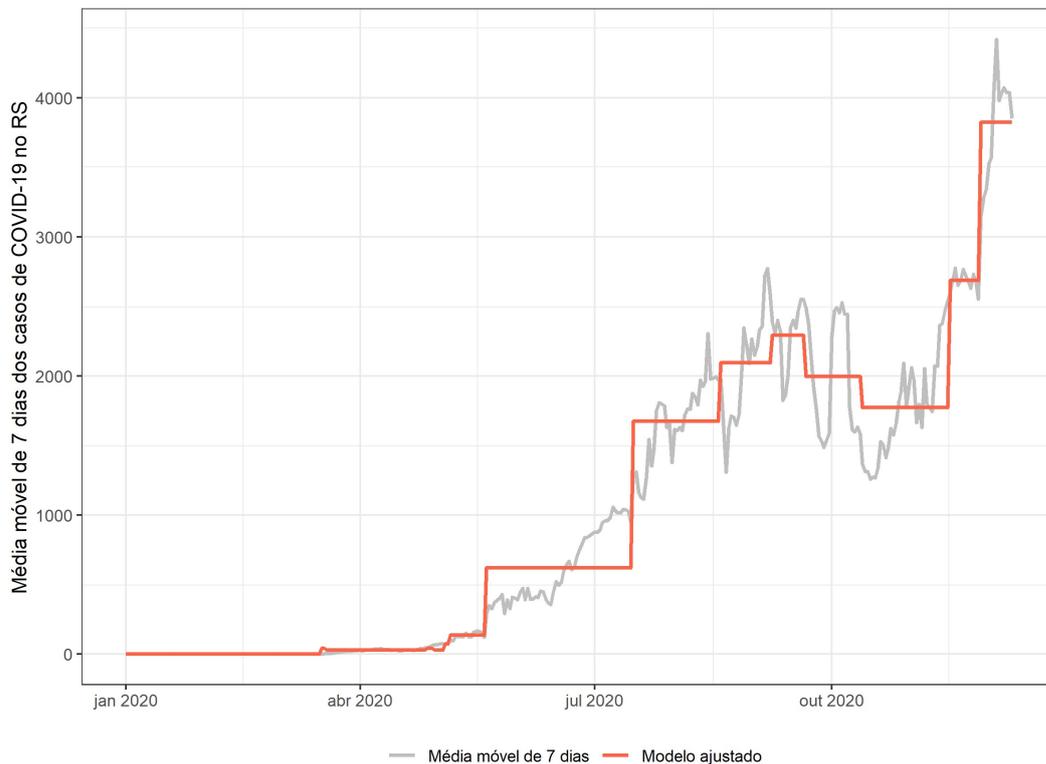


Figura 3 - Defasagem de restrição e média móvel de 7 dias.

Após a construção do modelo de regressão, o SHAP é aplicado para avaliar a contribuição das medidas de contenção na trajetória do número de casos de COVID-19 no estado do Rio Grande do Sul. Um valor de Shapley foi calculado para cada dia e cada medida de contenção. O valor final do SHAP é uma média dos 345 dias.

Na Figura 4, os resultados negativos estão diretamente relacionados à redução do número de casos. Quanto menor o valor de Shapley, mais significativa é sua contribuição para a redução de casos de COVID-19. Assim, fechamento de escolas (C1) foi a medida que mais contribuiu para a redução do número de casos com $-1,01$, seguida de fechamento de locais de trabalho (C2) com $-0,77$ e restrições a aglomerações (C4) com $-0,73$. As campanhas de informação pública (H1) tiveram valor SHAP zero, o que significa que não contribuíram para a redução dos casos de COVID-19.

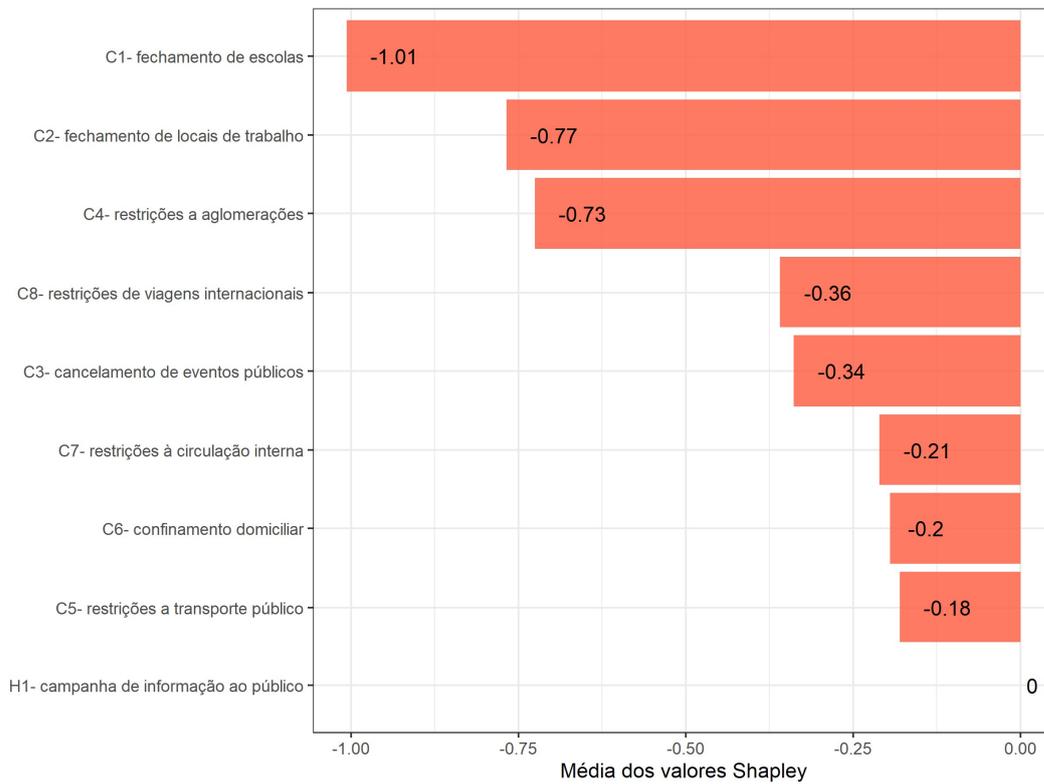


Figura 4 - Média dos valores Shapley.

Este estudo de caso implementou um modelo para mensurar e hierarquizar as medidas adotadas no estado do Rio Grande do Sul que mais contribuíram para a redução do número de casos de COVID-19 em 2020 (pré-vacinação). O modelo final conseguiu revelar a contribuição marginal de cada medida de contenção, ou seja, inferir qual delas teve o impacto mais significativo durante o primeiro ano da pandemia, sendo o fechamento das escolas a medida que mais colaborou.

5. Evolução dos óbitos por COVID-19 no Brasil e seus estados

Este trabalho será enviado para apreciação e possível publicação na Revista Socio-Economic Planning Sciences. Neste segundo estudo, tornou-se oportuno incluir a cobertura vacinal e os padrões de mobilidade, com o intuito de revelar qual medida contribuiu para evolução de casos e óbitos por COVID-19.

O estudo foi feito analisando 10 medidas de contenção (as mesmas nove anteriores do Capítulo 4 mais a H6- coberturas faciais), duas medidas de cobertura vacinal (D1- cobertura vacinal de primeira dose e D2- cobertura vacinal de segunda e dose única) e seis medidas de mobilidade (M1- varejo e recreação, M2- mercearia e farmácia, M3- parques, M4- estações de trânsito, M5- locais de trabalho e M6- residencial). Neste caso, os dados sobre vacinação foram retirados do Sistema de Informação do Programa Nacional de Imunização (SI-PNI) (MINISTÉRIO DA SAÚDE, 2022) e os dados sobre mobilidade disponibilizados pelo Google (COVID-19 *Community Mobility Reports*) (AKTAY et al., 2022).

A evolução dos óbitos por COVID-19 no ano de 2021 no Brasil e seus estados tiveram características bem variadas. A Figura 5 ilustra 28 perfis de séries temporais nos diferentes cenários da pandemia, onde picos e vales são apresentados em instantes diferentes de tempo para estados diferentes.

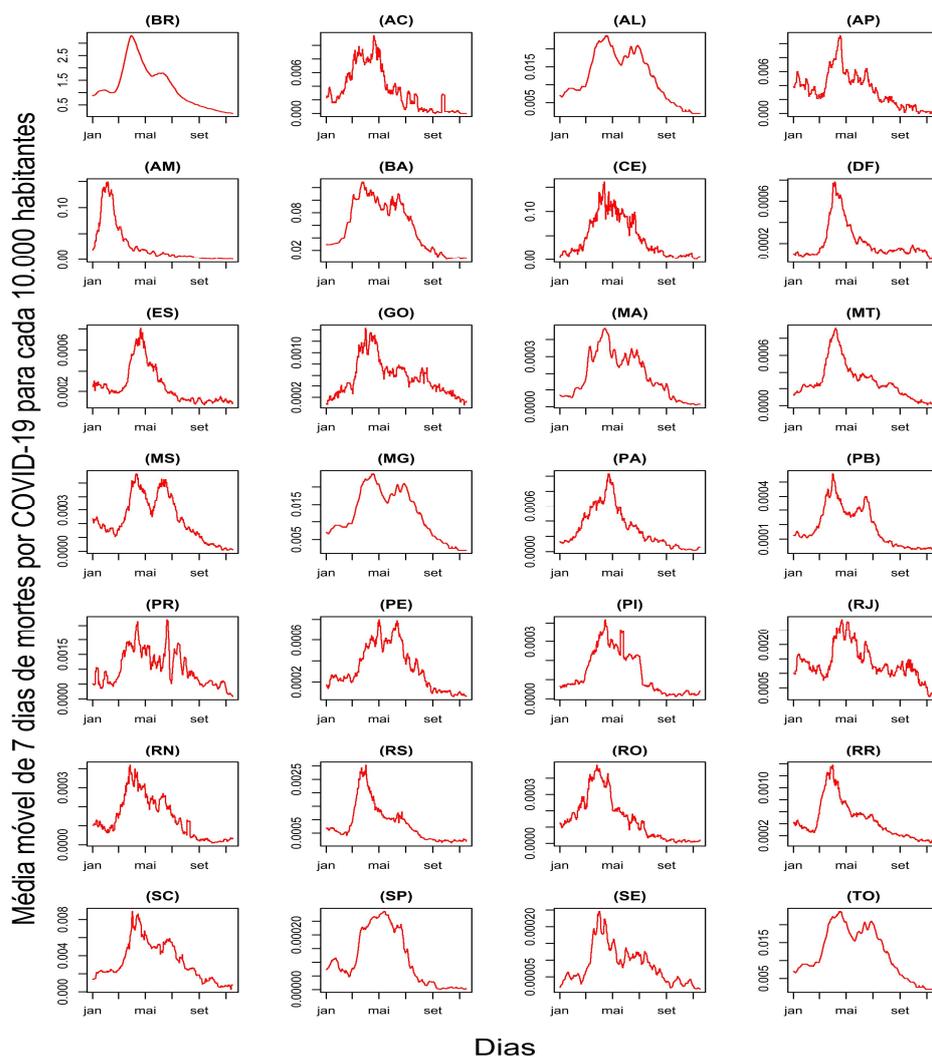


Figura 5 - Quadro de séries temporais das médias móveis de sete dias para óbitos nas escalas nacional (BR) e estadual (AC-TO) para o ano de 2021.

A complexidade reforça a dificuldade enfrentada pelos gestores públicos em implementar as medidas de controle mais adequadas para melhor combater a disseminação da COVID-19. Portanto, a questão permanece sobre como as diferentes localidades, nas escalas nacional, estadual e/ou municipal, podem implementar políticas de controle mais adequadas como por exemplo, medidas de restrição, estratégias de distribuição de vacinas, critérios de priorização de vacinação, políticas de máscaras faciais, e outros.

O pico do país ocorreu no mês de abril/2021 quando se obteve a maior taxa de mortes, decaindo então até o fim de maio/2021. Porém, em junho/2021 os óbitos voltaram a crescer tendo, a partir daí, uma redução gradual até o final de 2021.

Como dito antes, os padrões para cada estado não foram similares. Contudo, todos os estados exibiram aumento no início do ano, com destaque para o Amazonas, que apresentou o maior pico dentre todos os estados brasileiros neste período. O Amazonas também teve seu pico antes dos demais (em fevereiro). Os estados do Amapá, Paraná e Rio de Janeiro mostraram três ondas ao longo desse período. Outros estados tiveram duas ondas (assim como o Brasil, um pico maior e um segundo menor): Alagoas, Bahia, Goiás, Maranhão, Mato Grosso do Sul, Minas Gerais, Paraíba, Pernambuco, Piauí, Rio Grande do Norte, Rio Grande do Sul, Santa Catarina, São Paulo, Sergipe e Tocantins. Os demais, incluindo o Distrito Federal, apresentaram somente um pico em 2021, ocorrido no primeiro semestre. Em números absolutos, São Paulo foi o recordista. O estado registrou 104.632 mortes ocasionadas pelo vírus.

Para o Brasil e para cada um dos 27 estados foi aplicado o método de *stepwise* que é uma técnica usada para selecionar as *features* estatisticamente significativas a partir de um conjunto de variáveis preditoras (18 *features*), inserindo e removendo estas *features* no modelo até que não haja razão estatisticamente válida para adicionar ou remover mais nenhuma *feature*. O objetivo do *stepwise* é construir um modelo de regressão que inclua todas as variáveis preditoras estatisticamente significativas relacionadas à variável resposta (óbitos). Conjuntamente, foram testadas as defasagens em dias de cada um dos modelos e a variável resposta (óbitos) dos respectivos estados de forma que também fossem validadas as defasagens combinadas com cada um dos modelos. Desta forma, foi possível selecionar os modelos com os menores critérios de informação (BIC) e com todas as *features* estatisticamente significativas para cada um dos estados e para o Brasil.

Dentre os valores de defasagem que podemos destacar no *boxplot* da Figura 6, tem-se o Brasil que apresentou 39 dias de defasagem entre variáveis preditoras do seu modelo e os óbitos. Também se destacam os estados do Rio Grande do Sul, Rondônia, Santa Catarina com 40 dias de defasagem, sendo este o maior valor apresentado neste estudo. A mediana do *boxplot* da Figura 6 foi de 29 dias de defasagem sendo referente aos estados do Amapá, Piauí e Roraima. E com a menor defasagem obtida neste estudo temos o estado do Pará com apenas 6 dias. No geral, tem-se um *boxplot* com assimetria negativa que aponta uma maior concentração de valores acima da mediana (29 dias) e sem a presença de *outliers*.

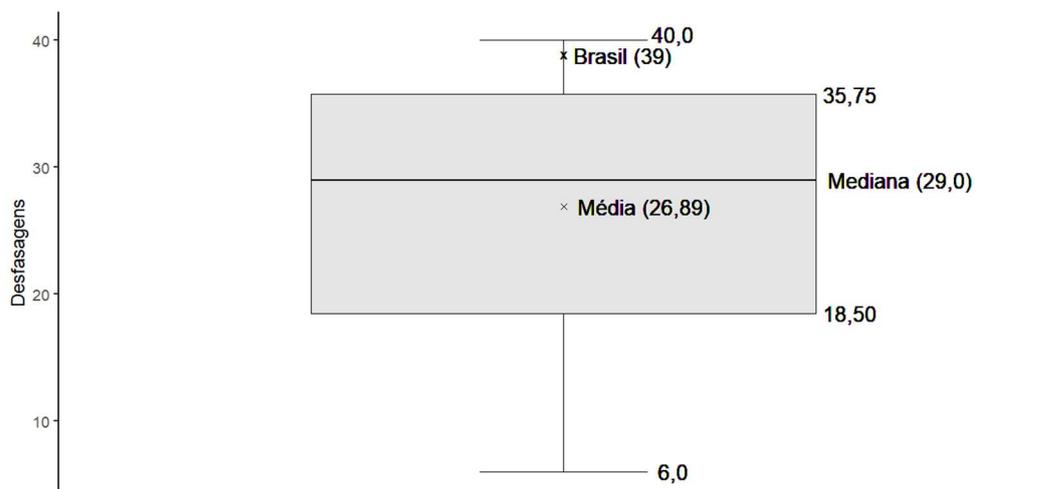


Figura 6 – Boxplot da defasagem em dias dos 26 estados brasileiros, do Distrito Federal e do Brasil entre variáveis preditoras do seu modelo e os óbitos.

Na Figura 7, a *feature* C1 (fechamento de escolas) foi selecionada como significativa em todos os cenários (Brasil e estados). A *feature* D1 (cobertura vacinal de primeira dose) não foi significativa apenas nos estados de Alagoas (AL) e da Bahia (BA). A *feature* D2 (cobertura vacinal de segunda e dose única) nos estados de Amazonas (AM), Roraima (RR), Sergipe (SE), Piauí (PI), Mato Grosso do Sul (MS), Rondônia (RO), Ceará (CE), Acre (AC) também não foram selecionadas por não ser estatisticamente significativa na seleção das variáveis para o melhor modelo. As *features* C7 (restrições à circulação interna) e H1 (campanha pública de informação) não foram selecionadas em nenhum cenário (a nível nacional e a nível estadual) por não serem estatisticamente significativas na seleção das variáveis aplicadas aos melhores modelos.

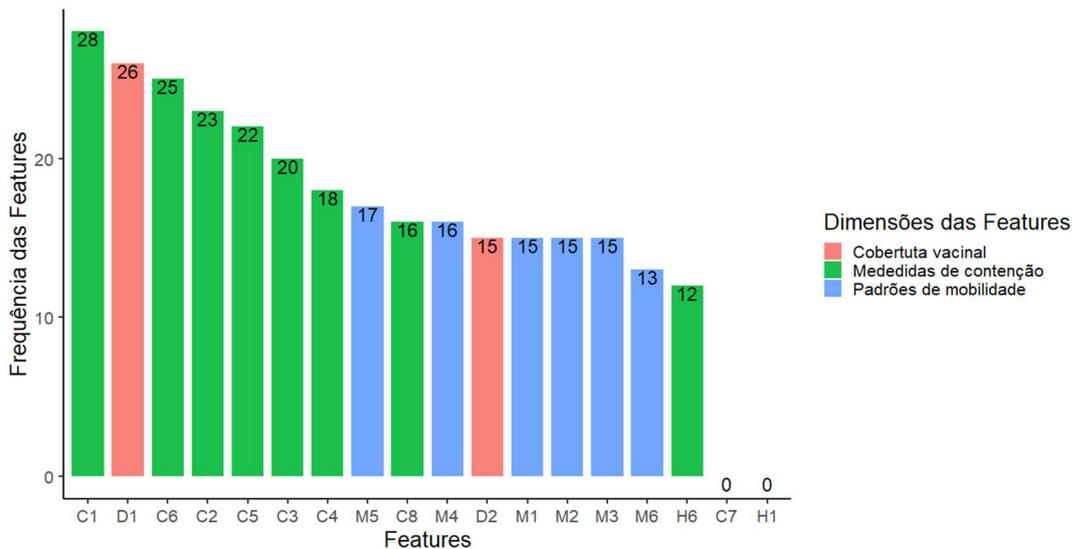


Figura 7 - Features relacionadas aos óbitos por COVID-19 em ordem de frequências selecionadas pelos 28 modelos do estudo e separados por cores relacionadas às respectivas dimensões.

A contribuição quantitativa de cada *feature* é medida por sua contribuição absoluta do valor de Shapley. Para quantificar as contribuições de cada *feature* processada pelo algoritmo XGBoost na análise foram calculados os valores SHAP.

Para a análise de importância das *features* via SHAP, a Figura 8 ilustra a classificação das diferentes medidas (*Cobertura vacinal*, *Medidas de contenção* e *Padrões de mobilidade*) em termos de seus impactos médios absolutos no Brasil e nos estados, respectivamente. A análise foca em determinar os recursos mais influentes que podem melhor interferir na evolução do parâmetro alvo (os óbitos).

Em relação ao Brasil, a Figura 8 apresenta as *features* D1 (cobertura vacinal de primeira dose) e D2 (cobertura vacinal de segunda e dose única) com uma maior contribuição absoluta de valor de SHAP. Dessa forma, são elas duas que possuem uma maior influência na evolução dos óbitos. Em contrapartida, tem-se as *features* C6 (confinamento domiciliar) e M2 (mercearia e farmácia) com as menores contribuições absolutas de valores de SHAP, tendo assim a menor influência na evolução dos óbitos para o Brasil.

Em relação aos estados brasileiros, a Figura 8 indica de forma geral que D1 (cobertura vacinal de primeira dose) e D2 (cobertura vacinal de segunda e dose única) são as duas *features* que tem as maiores contribuições absolutas nos estados. Ainda podemos citar as *features* C8 (controle de viagens internacionais) e H6 (uso de máscara) com as menores contribuições absolutas de valores de SHAP para todos os 27 estados.

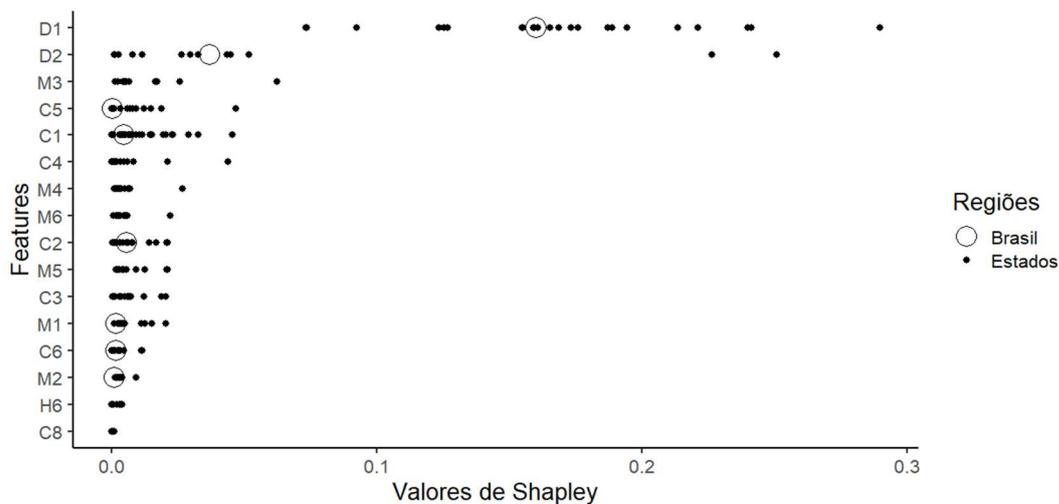


Figura 8 - Importância das features via valor de Shapley ilustra a classificação das diferentes medidas (cobertura vacinal, medidas de restrição e padrões de mobilidade) em termos de seus impactos médios absolutos no Brasil e nos estados.

A Figura 9 mostra o eixo y ordenado de modo crescente de mediana de cima para baixo. Pode-se ver que a *feature* D1 (cobertura vacinal de primeira dose) tem a maior contribuição de mediana negativa, enquanto a *feature* C2 (fechamento de locais de trabalho) tem a maior contribuição mediana positiva.

Para a análise de importância dos valores médios (positivos ou negativos) de SHAP ao nível nacional e estadual, A Figura 9 ilustra o ranking dos boxplots das diferentes features em ordem decrescente de contribuição em termos de seus impactos mediano na evolução do número de óbitos por COVID-19. Observa-se que o valor de SHAP negativo indica que o recurso específico tem uma correlação negativa com o parâmetro alvo (óbitos), e vice-versa para valores de SHAP positivos.

Quanto maior o comprimento da caixa, maior a distância entre o primeiro e o terceiro quartil, o que implica que o intervalo de variação das distribuições é maior. Assim, podemos comparar o comprimento dos boxplots das features para determinar qual das distribuições possui uma variação maior.

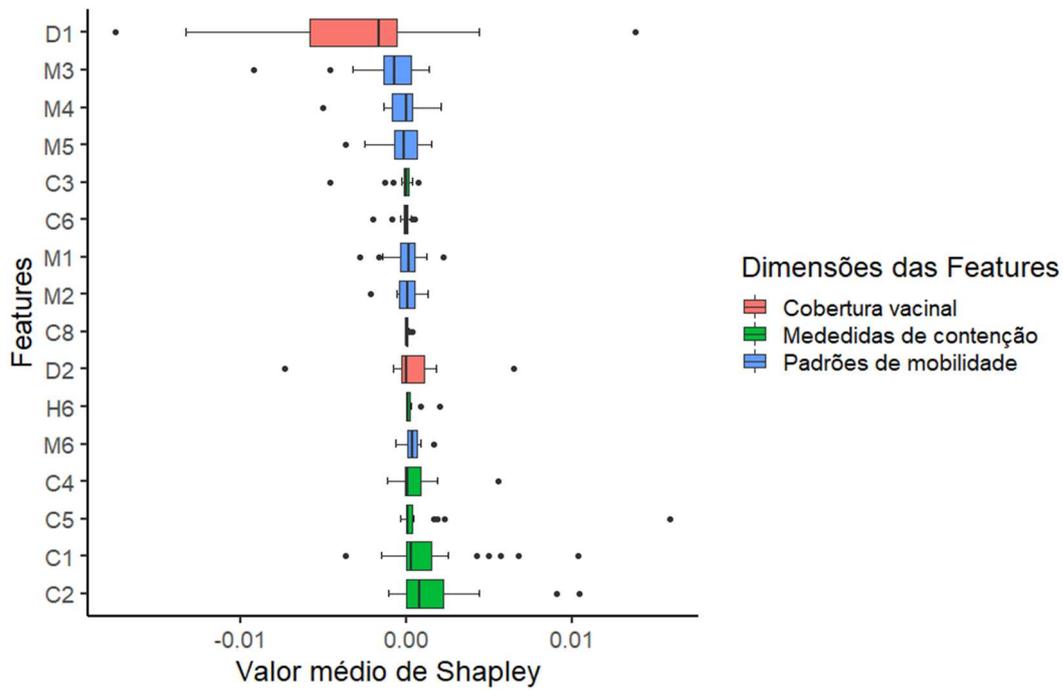


Figura 9 - O ranking dos boxplots das diferentes features em ordem decrescente de contribuição em termos de seus impactos mediano na evolução do número de óbitos por COVID-19.

Este segundo estudo implementou a contribuição quantitativa de cada medida para o país como um todo, os estados e o Distrito Federal. Os resultados demonstraram que a cobertura vacinal teve maior influência na evolução dos óbitos por COVID-19.

6. Discussão

O primeiro estudo apresenta um estudo de caso sobre as nove medidas de contenção (C1 a C8) e a campanha de informação pública (H1) da OxCGRT. O objetivo foi hierarquizar tais características quanto ao seu impacto no número de casos de COVID-19 no estado do Rio Grande do Sul. Deste modo, foi aplicado o algoritmo XGBoost para construir um modelo de regressão de Poisson que se ajustasse à curva de casos diários de COVID-19, utilizando as medidas de controle como variáveis explicativas.

A metodologia de Shapley foi aplicada ao modelo final para atribuir um valor de contribuição a cada uma das medidas de contenção, classificando-as quanto ao seu impacto na variação do número de casos de COVID-19. Com base nos resultados obtidos para o período estudado (ano de 2020), podemos afirmar que o fechamento das escolas teve o impacto mais significativo na redução do número de novos casos de COVID-19 no RS, seguido pelo fechamento de locais de trabalho e restrições às aglomerações, que também foram muito impactantes na redução dos casos. Por outro lado, as campanhas de informação pública não reduziram o número de casos de COVID-19.

Provavelmente, o efeito observado no estudo está diretamente ligado ao fator humano ao lidar com restrições e fatores socioeconômicos (FERGUSON et al., 2020). Quando aplicados a países de baixa e média renda como o Brasil, esses elementos dificultam a implementação de medidas como o confinamento domiciliar (GANEM et al., 2020; CANABARRO et al., 2020). De acordo com os resultados preliminares do IBGE (2019), havia 133.021 domicílios em situação de aglomeração subnormal em o estado do Rio Grande do Sul, o que representa 3,5% das casas do estado. Essa informação corrobora com um dos resultados deste estudo, de que as exigências de permanência em casa (C6- confinamento domiciliar) pouco contribuíram para as medidas de contenção.

Outra questão abordada no primeiro estudo foi a estimativa do intervalo de tempo desde a implementação da medida de contenção até sua efetiva efetividade na redução do número de casos de COVID-19 no RS em 2020. Tendo em vista que a defasagem de 13 dias resultou no menor erro, pode-se inferir que este foi o tempo necessário para que as medidas de contenção afetassem principalmente a curva de casos do Rio Grande do Sul.

Ademais, o resultado encontrado neste primeiro estudo de caso que compõe esta dissertação corrobora com a literatura existente, pois em Portugal, a demora durou cerca de duas semanas (14 dias) depois da implementação medidas de contenção para reduzir os novos casos e o número de pacientes em hospitais e unidades de cuidados intensivos (PEIXOTO et al., 2020).

Realizada esta primeira parte, que serviu como uma sustentação do método proposto, o segundo estudo objetivou analisar a evolução dos óbitos por COVID-19, aplicando o método proposto. Nesse sentido, o estudo trabalhou com dados sobre medidas de contenção, cobertura vacinal e variação da mobilidade populacional sobre a evolução do número de óbitos por COVID-19 no Brasil e em todos seus 26 estados mais o Distrito Federal, totalizando 28 regiões.

Foi aplicado o método *stepwise*, técnica utilizada para seleção das variáveis com objetivo de construir um modelo de regressão a partir de um conjunto de features preditores com todas as variáveis estatisticamente significantes. Ao todo, foram usadas 18 variáveis, sendo 10 sobre medidas de contenção, duas sobre cobertura vacinal e seis sobre mobilidade. A intenção foi construir um modelo que incluía as variáveis estatisticamente significativas relacionadas aos óbitos por COVID-19. As defasagens (em dias) também foram validadas para cada região e, ao final, foram obtidos os modelos com menores Critérios de Informação Bayesiano (BIC).

O boxplot revelou alguns destaques nas defasagens em dias como os estados do Rio Grande do Sul, Rondônia e Santa Catarina (40 dias) e o próprio Brasil (39 dias). A menor defasagem ocorreu no Pará (6 dias). Este resultado específico diverge completamente dos demais trabalhos que apresentam uma defasagem de 30 a 40 dias após a aplicação das medidas para que ocorra um efeito observável das mesmas na evolução do número de óbitos. Um dos possíveis motivos para uma amplitude de 34 dias pode vir dos diferentes cenários de evolução do número de óbitos por COVID-19, refletindo o comportamento da população de cada um dos estados e do Brasil em relação as medidas de contenção implementadas pelas autoridades competentes. Os resultados obtidos foram consistentes com a literatura vigente, mostrando que em geral a evolução do quadro dos pacientes infectados pelo vírus até o óbito leva em torno de 30 a 40 dias (FAGUNDES & DUARTE, 2022). E também a rapidez com que evoluiu a cobertura vacinal em cada

uma das 28 regiões pode ter influenciado a diferença nas defasagens entre os estados brasileiros (DE CASTRO-NUNES & DA ROCHA RIBEIRO, 2022).

Em todos os cenários, a variável com maior frequência foi o fechamento de escolas, estando presente em todas as 28 regiões estudadas, seguida da cobertura vacinal em primeira dose, estando presente em 26 regiões. A cobertura vacinal em primeira dose não foi estatisticamente significativa na construção do modelo em Alagoas e na Bahia. Seis estados se destacaram com valor de Shapley maior que 0,2: Sergipe (0,290), Piauí (0,241), Maranhão (0,240), Mato Grosso do Sul (0,221), Rondônia (0,221) e Ceará (0,213).

A cobertura vacinal de segunda dose e dose única não foi estatisticamente significativa na construção do modelo nos estados do Amazonas, Roraima, Sergipe, Piauí, Mato Grosso do Sul, Rondônia, Ceará, Acre. Dois estados se destacaram com valor de Shapley maior que 0,2: Bahia (0,251) e Alagoas (0,226). Uma possível hipótese para esses resultados é a proximidade entre as datas da primeira e segunda doses da vacinação. Na Bahia, o início da primeira dose foi em 19/01/2021 e a segunda dose em 16/02/2021. Em Alagoas, o início da primeira dose foi em 19/01/2021 e da segunda dose em 11/02/2021.

Restrições à circulação interna e campanha pública de informação não foram selecionadas em nenhum cenário. Especificamente, essas duas medidas foram as únicas que se mantiveram constantes durante todo ano de 2021 em todos os estados, desta forma os modelos não captaram nenhuma contribuição das mesmas para a evolução do número de óbitos.

De forma geral, devido à complexidade da evolução dos óbitos a nível nacional e estadual, cada região adotou medidas de contenção com diferentes rigores e, adicionalmente, com índices de mobilidade populacional distintos para as mesmas medidas devido ao comportamento heterogêneo da população brasileira (GALINDO, SILVA, PEDREIRA JUNIOR, 2022). Outra medida relevante foi a cobertura vacinal que, apesar de ter iniciado quase ao mesmo tempo em todos os estados, sua evolução ao longo dos meses de 2021 foi bastante desigual na comparação entre os estados (FLEURY & FAVA, 2022). Devido a essa variabilidade de cenários os modelos propostos para o Brasil e estados foram bem diversificados tanto em relação as medidas quanto em defasagem em dias.

Por tratar-se de uma regressão múltipla, uma preocupação pertinente era a multicolinearidade, que se refere a preditores correlacionados entre si, como

possivelmente a cobertura vacinal de primeira dose - D1 e cobertura vacinal de segunda e dose única - D2. A multicolinearidade ocorre quando o modelo inclui vários fatores correlacionados não apenas à sua variável de resposta, mas também uns aos outros.

De fato, essa seria uma preocupação pertinente caso o modelo de regressão fosse estimado utilizando o método de mínimos quadrados, porém o algoritmo Xgboost é baseado em árvore de decisão, que utiliza uma estrutura de Gradient Boosting que não sofre influência da multicolinearidade em suas estimativas.

A contribuição quantitativa de cada medida foi avaliada por sua contribuição absoluta do valor de Shapley. Em relação ao Brasil e aos estados, foi a cobertura vacinal que mais contribuiu para a evolução dos óbitos (ORELLANA et al., 2022). Em contrapartida, para o Brasil, o confinamento domiciliar e o deslocamento de pessoas para mercearia e farmácias foram as medidas que menos influenciaram. E para os estados, foi o controle de viagens internacionais e o uso de máscaras que menos influenciaram.

O primeiro estudo selecionou o estado do Rio Grande do Sul como foco e aplicou o método proposto para avaliar a contribuição das medidas de contenção na evolução do número de casos de COVID-19 no período pré-vacinação. O segundo estudo é uma expansão do primeiro, selecionando 28 regiões (26 estados brasileiros, Distrito Federal e Brasil) e analisando também a contribuição da cobertura vacinal (que é uma medida farmacológica) e do índice de variação da mobilidade populacional disponibilizado pelo Google (AKTAY et al., 2022) para a evolução do número de óbitos no período pós-vacinação, até Dezembro/2021. O intuito foi verificar a hipótese de que a contribuição da cobertura vacinal se destacaria em relação às demais medidas, e essa hipótese foi confirmada.

Conforme a cobertura vacinal foi aumentando, as medidas de contenção foram sendo reduzidas e a mobilidade aumentando gradativamente. Possivelmente, a sensação de segurança mediante a aplicação da vacina fez com que as pessoas diminuíssem o distanciamento social. Sendo assim, essas medidas passaram a contribuir menos na evolução de casos e óbitos por COVID-19.

Embora não tenha sido considerada no escopo do presente trabalho, Sugere-se como trabalhos futuros a utilização de uma regressão do tipo lasso na seleção das variáveis. Seria útil, pois a seleção das mesmas não dependeria assim de segundo critério (BIC) para a escolhas dos melhores modelos para cada um dos

estados brasileiros. Além disso, sugere-se a implementação da metodologia aqui proposta para classificar e hierarquizar as medidas de contenção, a cobertura vacinal e a mobilidade nos demais países. Acredita-se que esta aplicação possa ser um relevante tema de pesquisa no contexto dos sistemas de mensuração internacional dessas medidas, uma vez que os resultados dos indicadores se apresentam se condizentes com a realidade brasileira que certamente diverge dos outros países e pode trazer um novo olhar sobre a dinâmica da evolução da pandemia.

7. Conclusão

Foi objetivo deste trabalho mensurar quais medidas de contenção mais contribuíram antes e depois do início da vacinação em relação à evolução de casos e óbitos durante a pandemia de COVID-19 no Brasil. Foram integradas como fontes de dados o repositório gerado pelo projeto de Oxford (HALE et al., 2020), o índice de mobilidade populacional disponibilizado pelo Google (AKTAY et al., 2022) e os percentuais de cobertura vacinal de 1ª e 2ª doses disponibilizados pelo SI-PNI (MINISTÉRIO DA SAÚDE, 2022). Através do valor de Shapley, foi possível ranquear a contribuição dessas variáveis em dois estudos.

O primeiro estudo implementou um modelo para mensurar e hierarquizar as medidas de contenção que mais contribuíram para a redução do número de casos de COVID-19 no estado do RS, Brasil. O resultado revelou a contribuição marginal de cada medida de contenção, ou seja, quantificou qual delas teve o impacto mais significativo durante o primeiro ano da pandemia. O fechamento de escolas foi a medida que mais contribuiu para a redução do número de casos, seguida de fechamento de locais de trabalho e restrições a aglomerações.

No segundo estudo, avaliou-se a contribuição quantitativa dentre as medidas de contenção, o progresso da vacinação e a mobilidade da população, com relação a curva de óbitos por COVID-19, no Brasil e em cada estado brasileiro, desde o início da vacinação até Dezembro/2021. Os resultados evidenciaram que tanto no Brasil quanto em seus estados o progresso da vacinação teve a maior influência na evolução dos óbitos.

Contudo, o confinamento domiciliar e o deslocamento de pessoas para mercearias e farmácias foram as que menos influenciaram no Brasil. Para os estados e o Distrito Federal, o controle de viagens internacionais e o uso de máscaras foram as que menos influenciaram. Os resultados obtidos com este estudo podem e devem aprimorar as políticas públicas de combate à pandemia de COVID-19, principalmente no cenário brasileiro, que é de grande desigualdade social.

Dentro de um contexto social, a ponderação das medidas de contenção pode e deve ajudar os atores responsáveis na elaboração de políticas públicas que minimizem os efeitos socioeconômicos em suas regiões. O uso de técnicas científicas é necessário para que a abrangência social sirva de apoio em todas as classes, principalmente as menos favorecidas.

8. Referências

ALGABA, Encarnación; FRAGNELLI, Vito; SÁNCHEZ-SORIANO, Joaquín (Ed.). Handbook of the Shapley value. CRC Press, 2019.

AKTAY, Ahmet et al. **Google COVID-19 community mobility reports: anonymization process description**, 2022.

BAI, Yan et al. Presumed asymptomatic carrier transmission of COVID-19. **Jama**, v. 323, n. 14, p. 1406-1407, 2020.

BOLETIM EPIDEMIOLÓGICO COVID-19. Available: <https://coronavirus.rs.gov.br/informe-epidemiologico>. cited 14 Oct 2020.

CANABARRO, Askery et al. Data-driven study of the COVID-19 pandemic via age-structured modelling and prediction of the health system failure in Brazil amid diverse intervention strategies. **Plos one**, v. 15, n. 7, p. e0236310, 2020.

CHEMAITELLY, Hiam et al. Eficácia da vacina mRNA-1273 COVID-19 contra B. 1.1. 7 e B. 1.351 variantes e doença grave de COVID-19 no Catar. **Medicina da natureza**, v. 27, n. 9, pág. 1614-1621, 2021.

CHEN, Tianqi et al. Xgboost: extreme gradient boosting. **R package version 0.4-2**, v. 1, n. 4, p. 1-4, 2015.

CHEW, Alvin Wei Ze; ZHANG, Limao. Data-driven multiscale modelling and analysis of COVID-19 spatiotemporal evolution using explainable AI. **Sustainable cities and society**, v. 80, p. 103772, 2022.

COLE, Matthew A.; ELLIOTT, Robert JR; LIU, Bowen. The impact of the Wuhan Covid-19 lockdown on air pollution and health: a machine learning and augmented synthetic control approach. **Environmental and Resource Economics**, v. 76, n. 4, p. 553-580, 2020.

CUCINOTTA, Domenico; VANELLI, Maurizio. WHO declares COVID-19 a pandemic. **Acta Bio Medica: Atenei Parmensis**, v. 91, n. 1, p. 157, 2020.

DAGAN, Noa et al. BNT162b2 mRNA Covid-19 vaccine in a nationwide mass vaccination setting. **New England Journal of Medicine**, 2021.

DE CASTRO-NUNES, Paula; DA ROCHA RIBEIRO, Gizele. Equidade e vulnerabilidade em saúde no acesso às vacinas contra a COVID-19. **Revista Panamericana de Salud Pública**, v. 46, 2022.

DE OLIVEIRA, Brigitte Renata Bezerra et al. Determinants of access to the SARS-CoV-2 vaccine: a preliminary approach. **International journal for equity in health**, v. 20, n. 1, p. 1-11, 2021.

DOBLHAMMER, Gabriele; REINKE, Constantin; KREFT, Daniel. Social disparities in the first wave of COVID-19 incidence rates in Germany: a county-scale explainable machine learning approach. **BMJ open**, v. 12, n. 2, p. e049852, 2022.

DONG, Ensheng; DU, Hongru; GARDNER, Lauren. An interactive web-based dashboard to track COVID-19 in real time. **The Lancet infectious diseases**, v. 20, n. 5, p. 533-534, 2020.

DROSTEN, Christian et al. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. **New England journal of medicine**, v. 348, n. 20, p. 1967-1976, 2003.

FAGUNDES, Gabriel Lima Miranda Gonçalves; DUARTE, Luciana Gaspar Melquíades Duarte. A PROPORCIONALIDADE DAS MEDIDAS DE CONTENÇÃO DA PANDEMIA DE COVID-19: ANÁLISE EMPÍRICA SOBRE O BRASIL. **Revista do Direito**, n. 66, p. 64-90, 2022.

FLAXMAN, Seth et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. **Nature**, v. 584, n. 7820, p. 257-261, 2020.

FERGUSON, Neil M. et al. **Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand**. 2020.

FLEURY, Sonia; FAVA, Virgínia Maria Dalfior. Vacina contra Covid-19: arena da disputa federativa brasileira. **Saúde em Debate**, v. 46, p. 248-264, 2022.

GALINDO, Ernesto Pereira; SILVA, Sandro Pereira; PEDREIRA JUNIOR, Jorge Ubirajara. **Impactos fatais da COVID-19 nos trabalhadores brasileiros**. 2022. Repositório do Conhecimento do IPEA. Disponível em: <http://repositorio.ipea.gov.br/handle/11058/11084>. Acesso em: 23/06/2022.

GANEM, Fabiana et al. The impact of early social distancing at COVID-19 Outbreak in the largest Metropolitan Area of Brazil. **MedRxiv**, 2020.

HALE, Thomas et al. Oxford COVID-19 government response tracker (OxCGRT). **Last updated**, v. 8, p. 30, 2020.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. Unsupervised learning. In: **The elements of statistical learning**. Springer, New York, NY, 2009. p. 485-585.

JARA, Alejandro et al. Effectiveness of an inactivated SARS-CoV-2 vaccine in Chile. **New England Journal of Medicine**, v. 385, n. 10, p. 875-884, 2021.

JING, Nan et al. Cross-sectional analysis and data-driven forecasting of confirmed COVID-19 cases. **Applied Intelligence**, v. 52, n. 3, p. 3303-3318, 2022.

LIMA, Nisia Trindade; GADELHA, Carlos Grabois. The COVID-19 Pandemic: Global Asymmetries and Challenges for the Future of Health. **China CDC Weekly**, v. 3, n. 7, p. 140, 2021.

LOPEZ BERNAL, Jamie et al. Effectiveness of Covid-19 vaccines against the B. 1.617. 2 (Delta) variant. **New England Journal of Medicine**, 2021.

LU, Roujian et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. **The lancet**, v. 395, n. 10224, p. 565-574, 2020.

LUNDBERG, Scott M. et al. Explainable AI for trees: From local explanations to global understanding. **arXiv preprint arXiv:1905.04610**, 2019.

MCCULLAGH, P.; NELDER, J. A. Generalized linear models. 2 ed. Flórida, EUA: Chapman & Hall, 1989.

MEHTA, Puja et al. COVID-19: consider cytokine storm syndromes and immunosuppression. **Lancet (London, England)**, v. 395, n. 10229, p. 1033, 2020.

MINISTÉRIO DA SAÚDE (BR). **Sistema de Informação do Programa Nacional de Imunizações SI-PNI**. 2022.

MOLNAR, Christoph. Interpretable machine learning. Lulu. com, 2020.

NISHIURA, Hiroshi et al. Closed environments facilitate secondary transmission of coronavirus disease 2019 (COVID-19). **MedRxiv**, 2020.

OLIVEIRA, Marcella Cini et al. Fatores associados ao óbito em casos confirmados de COVID-19 no estado do Rio de Janeiro. **BMC infectious diseases**, v. 21, n. 1, p. 1-16, 2021.

ORELLANA, Jesem Douglas Yamall et al. Mudanças no padrão de internações e óbitos por COVID-19 após substancial vacinação de idosos em Manaus, Amazonas, Brasil. **Cadernos de Saúde Pública**, v. 38, p. PT192321, 2022.

PEIXOTO, Vasco Ricoca Freire Duarte et al. “**Timing**”, **adesão e impacto das medidas de contenção da COVID-19 em Portugal**. 2020.

R CORE TEAM. R: A language and environment for statistical computing. 2019.

SHARMA, Mrinank et al. Understanding the effectiveness of government interventions against the resurgence of COVID-19 in Europe. **Nature communications**, v. 12, n. 1, p. 1-13, 2021.

SMITH, Matthew; ALVAREZ, Francisco. Identifying mortality factors from Machine Learning using Shapley values—a case of COVID19. **Expert Systems with Applications**, v. 176, p. 114832, 2021.

SOHRABI, Catrin et al. World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). **International journal of surgery**, v. 76, p. 71-76, 2020.

WU, Chaomin et al. Risk factors associated with acute respiratory distress syndrome and death in patients with coronavirus disease 2019 pneumonia in Wuhan, China. **JAMA internal medicine**, v. 180, n. 7, p. 934-943, 2020.

YANG, Xiaobo et al. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. **The Lancet Respiratory Medicine**, v. 8, n. 5, p. 475-481, 2020.

ZAKI, Ali M. et al. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. **New England Journal of Medicine**, v. 367, n. 19, p. 1814-1820, 2012.

ZHU, Na et al. A novel coronavirus from patients with pneumonia in China, 2019. **New England journal of medicine**, 2020.