EXACTA Engenharia de Produção eISSN 1983-9308

Artigos

https://doi.org/10.5585/exactaep.2022.20401

Check for updates

APLICAÇÃO DE REGRESSÃO LINEAR MÚLTIPLA PARA ANALISAR A RELAÇÃO ENTRE BUSCAS POR PALAVRAS-CHAVE NA INTERNET E CASOS DE COVID-19 NO BRASIL

Recebido: 06 jul. 2021
Aprovado: 27 jan. 2022
Versão do autor aceita publicada online: 27 jan. 2022
Publicado online: 16 fev. 2022

Como citar esse artigo - American Psychological Association (APA):

Furletti, L. A., Carvalho, Í. L., Silva, H. B. & Peixoto, L. de C. (2023, out./dez.). Aplicação de regressão linear múltipla para analisar a relação entre buscas por palavras-chave na internet e casos de COVID-19 no Brasil. *Exacta, 21*(4), 892-904. DOI: https://doi.org/10.5585/exactaep.2022.20401

Submeta seu artigo para este periódico

Processo de Avaliação: *Double Blind Review*

Editor: Dr. Luiz Fernando Rodrigues Pinto

Dados Crossmark

EXACTA Engenharia de Produção eISSN 1983-9308

https://doi.org/10.5585/exactaep.2022.20401

Processo de Avaliação: *Double Blind Review*
Editor: Prof. Dr. Luiz Fernando Rodrigues Pinto
Recebido em: 06 jul. 2021
Aprovado em: 27 jan. 2022

Artigos

APLICAÇÃO DE REGRESSÃO LINEAR MÚLTIPLA PARA ANALISAR A RELAÇÃO ENTRE BUSCAS POR PALAVRAS-CHAVE NA INTERNET E CASOS DE COVID-19 NO BRASIL

Letícia Alexandre Furletti[1] Ítalo Lelis De Carvalho[2] Hendrigo Batista Da Silva[3]
Letícia de Castro Peixoto[4]

Resumo: Esforços em diversas áreas do conhecimento científico, sobretudo na de infodemiologia, têm se dedicado a propor soluções para os problemas socioeconômicos provocados pela pandemia da SARS-CoV-2 e do isolamento social decorrente dela. O estudo de *insights* sobre como a COVID-19 está se disseminando em determinadas regiões é uma alternativa eficaz para lidar com este desafio. Este artigo analisa, de forma exploratória, a relação entre os dados publicamente disponíveis de índices relativos de busca por palavras-chave no Google Trends® e os dados governamentais de notificação de infectados pela doença em Minas Gerais, Rio de Janeiro e São Paulo. Em uma análise longitudinal, constata-se que as pesquisas por "Teste de COVID" e "Sintomas de COVID" são explicativas para o número de casos notificados nestas localidades. Este estudo fornece percepções iniciais para auxiliar pesquisas no uso das buscas *online*, relacionadas à doença, como variáveis explicativas adicionais a modelos preditivos, resguardadas suas devidas limitações.

Palavras-chave: COVID-19. Regressão linear múltipla. Ciência de dados. Google Trends.

APPLYING MULTIPLE LINEAR REGRESSION TO ANALYZE THE RELATIONSHIP BETWEEN RELATIVE SEARCH VOLUME AND COVID-19 CASES IN BRAZIL

Abstract: Efforts in multiple areas of scientific knowledge, in especial infodemiology, have been dedicated to propose solutions for the socioeconomic problems caused by the SARS-CoV-2 pandemic and the social isolation measures. Early insights studies into how COVID-19 may be spreading in a particular region is an effective tool to tackle this challenge. This paper analyses in an explanatory way the relation between publicly available data from relative search volume in Google Trends® and government data about COVID-19 infection in Minas Gerais, Rio de Janeiro and São Paulo. Through these analyses, it was possible to infer that "Teste de COVID" and "Sintomas de COVID" could be used as explanatory variables to the number of cases in these places. This study provides interesting insights to support future research in the use of relative search volume as additional explanatory variables in predictive models, regarding its limitations.

Keywords: COVID-19. Multiple linear regression. Data science. Google Trends.

[1] Pontificia Universidade Catolica de Minas Gerais - PUC Minas / Belo Horizonte (MG)
[2] Pontificia Universidade Catolica de Minas Gerais - PUC Minas / Belo Horizonte (MG)
[3] Pontificia Universidade Catolica de Minas Gerais - PUC Minas / Belo Horizonte (MG)
[4] Pontificia Universidade Catolica de Minas Gerais - PUC Minas / Belo Horizonte (MG). Doutora em Ciência da Informação pela UFMG

1 Introdução

O aumento no número de casos de COVID-19, causado pelo vírus SARS-CoV-2, provocou um colapso no sistema de saúde do Brasil a partir do primeiro semestre de 2020 e mudou a percepção da população — que inicialmente acreditava tratar-se de uma doença leve e limitada a determinada faixa etária — para uma pandemia de rápida evolução e que impacta negativamente os setores da economia de cada estado. Segundo McKibbin e Fernando (2021), uma redução de 1,9% a 8% pode ocorrer no Produto Interno Bruto (PIB) a depender do cenário epidemiológico adotado no Brasil. Por isso, torna-se crucial o estudo de *insights* sobre a chegada de novos picos para a devida tomada de decisões visando diminuir os impactos da doença na sociedade.

No cenário pandêmico atual, diferentes áreas da ciência têm demonstrado interesse em encontrar formas de mitigar esse problema. No entanto, apesar de grandes, os esforços são assíncronos e se distanciam de um consenso sobre as estratégias ideais para se alcançar a meta de acabar com a pandemia (ANGUS, 2020). Em contextos passados, durante surtos epidemiológicos, como a influenza e a dengue, temos o surgimento de métodos que utilizam dados baseados na internet como uma alternativa complementar aos sistemas de vigilância tradicionais, com objetivo de estimar a atividade da doença em tempo real, favorecendo a rápida tomada de decisão e com isso reduzindo novos infectados (Liu *et al.,* 2020). Segundo Teng *et al.* (2017), o Google Trends® se mostrou eficaz para previsão de surtos de doenças, sendo uma ferramenta importante para ser considerada.

No que tange à área de ciência dos dados, a detecção de padrões explicativos entre diferentes variáveis pode contribuir para um maior entendimento da situação atual e futura da pandemia em cada localidade, oferecendo *insights* para modelos preditivos de tomada de decisões de saúde pública, resguardadas as devidas limitações de cada abordagem. Siettos e Russo (2013) defendem que modelos matemáticos têm função crucial nos esforços em prever e controlar potenciais surtos de doenças infecciosas.

Assim, o objetivo deste artigo é, a partir da correlação entre o valor relativo de buscas diárias – *Relative Search Volume* (RSV) – disponibilizado pelo Google Trends® e o total de notificações oficiais de infectados pela COVID-19 em determinada região, construir um modelo de regressão linear múltipla (RLM), para analisar o potencial das informações online de busca por algumas palavras-chave em predizer a incidência de casos de COVID-19, auxiliando a tomada de decisão em cada localidade. Tal análise permite compreender em determinado grau uma alteração de comportamento da população em relação a surtos locais da pandemia, sendo refletida no aumento da procura na internet por determinados termos ou palavras-chave, auxiliando assim a tomada pró-ativa de decisões para a contenção de surtos.

O presente estudo está dividido em 5 seções. A seção 2 apresenta as literaturas existentes sobre impactos da COVID-19 na sociedade e sobre a utilização da ciência de dados e da internet para minimizá-los. Já a seção 3 descreve como os dados foram coletados e como um modelo matemático de regressão multivariada foi criado com a ajuda de funções da linguagem R. A seção 4 discute os resultados da regressão linear múltipla criada entre dois termos de pesquisa e o número de infectados pela doença, assim como testes preliminares da qualidade dos ajustes quanto à normalidade dos resíduos e multicolinearidade. Por fim, a seção 5 conclui o estudo sumarizando os principais achados e discutindo implicações e limitações a serem abordadas em pesquisas futuras.

2 Referencial teórico

Detectado pela primeira vez em Wuhan, China, em dezembro de 2019, o vírus SARS-CoV-2 teve rápida evolução e poucos meses depois já havia se espalhado por muitos outros países. Em 30 de janeiro de 2020, a Organização Mundial da Saúde (OMS) emitiu uma Emergência de Saúde Pública Internacional e em 11 de março do mesmo ano, a OMS declarou uma pandemia de Coronavírus (OMS, 2020).

O pouco conhecimento sobre controle preditivo da doença de forma a antever possíveis surtos contribuiu para o aumento exponencial de casos, sobretudo no Brasil. Até junho de 2021 no país o número de infectados totalizou 18 milhões, e o de morte atingiu 500 mil (CONASS, 2021), sendo considerado pela organização como um dos epicentros da doença desde o ano passado, já que representa 10% de todo o número de infectados em todo o mundo. Em decorrência disso, surgiram impactos negativos na sociedade e na economia do país, reforçando o caráter econômico instável, produzindo colapso da produção e contribuindo para o aumento das taxas de desemprego (Ferreira Jr & Rita, 2020). Por isso, enfatiza-se a importância da vigilância precoce de doenças infecciosas somada a uma resposta urgente dos governantes na redução dos efeitos nocivos dos surtos epidemiológicos (Tizzoni *et al.,* 2012).

Essa resposta deve evitar tanto a paralisia do sistema de saúde, como os efeitos negativos do fechamento completo das atividades sobre as famílias e empresas (Ferreira Jr & Rita, 2020). Isso porque existem duas importantes premissas a serem analisadas: o relaxamento antecipado e sem embasamento das medidas de distanciamento social provocariam o surgimento de picos de doença; a tomada de medidas extremas, como fechamento das atividades comerciais, poderia resultar em déficits econômicos. Portanto, análises aprofundadas se fazem úteis para encontrar soluções eficientes no combate da COVID-19.

2.1 Infodemiologia

A infodemiologia, termo amplamente utilizado em estudos epidemiológicos, é definida como a ciência da distribuição de informação em meio eletrônico, sobretudo na internet, com o objetivo de informar a saúde e as políticas públicas, através da previsão de surtos de doenças (Eysenbach, 2009). Essa abordagem foi avaliada como importante para o monitoramento e previsão de surtos e epidemias passadas, como ebola, zika, MERS e influenza (Mavragani & Gkillas, 2020).

Segundo Mavragani e Gkillas (2020), o Google Trends®, plataforma *online* que permite acompanhar a evolução do número de buscas por uma determinada palavra-chave ao longo do tempo, foi identificado como uma das mais populares fontes de infodemiologia. Em estudo realizado por Bakker, Helm e Stevenson (2016), o uso do Google Trends® mostrou-se extremamente eficaz em prever surtos de varicela com até um mês de antecedência, sendo que a relação positiva entre as buscas e o número de casos chegou a atingir 81%. Em outro artigo, Ginsberg et al. (2009) encontrou também relação entre palavras-chave pesquisadas através da plataforma e o número de infectados por influenza nos Estados Unidos, estimando de forma precisa o nível atual de atividade semanal da doença em cada região do país.

A experiência adquirida com epidemias anteriores fornece informações valiosas sobre como tratar as possíveis implicações da COVID-19 (McKibbin & Fernando, 2021). Portanto, encontra-se possibilidade em antever novos surtos da atual pandemia de SARS-CoV-2 no Brasil utilizando a infodemiologia, e sobretudo o Google Trends®. Como contribuição geral tem-se o ganho de tempo das autoridades públicas em organizar esforços decisórios quanto à tomada de medidas extremas, como o fechamento total de atividades, ou brandas como o controle do horário de funcionamento dos estabelecimentos comerciais.

2.2 Regressão linear múltipla (RLM)

É comum a associação da infodemiologia com ferramentas estatísticas, com objetivo de obter melhores resultados nesses estudos. De acordo com Rabajante (2020) modelos matemáticos podem ser um método valioso de "primeiros socorros" durante um surto de doenças.

Para análise de ambientes nos quais diversas variáveis afetam um único resultado, modelos de regressão linear múltipla são apropriados (Lin *et al***.** 2020). Alguns autores fizeram uso dessa regressão para prever a disseminação de casos de COVID-19 no mundo. Um deles é Ayyoubzadeh *et al***.** (2020) que utilizou esse método para prever a incidência da doença no Irã, encontrando relação positiva entre a pesquisa *online* por termos equivalentes no português a “lavagem das mãos” e "antissépticos" e o aumento da SARS-CoV. Kumar, Sinwar e Saini (2021) fizeram um estudo de correlação entre nove variáveis, incluindo densidade populacional e PIB e o número de mortes por COVID-19, os quesitos citados foram considerados como os mais significativos no estudo. No presente artigo, analisou-se a

correlação entre vários termos de busca e o aumento do número de casos da doença e para isso fez uso da regressão linear múltipla.

3 Método de pesquisa

De acordo com Marconi e Lakatos (1996), a pesquisa é fundamental para a obtenção de soluções para problemas coletivos. Portanto, para o desenvolvimento deste estudo utilizou-se uma pesquisa de caráter quantitativo e exploratório objetivando identificar, através de um modelo matemático, uma possível relação de causa e efeito entre o valor relativo de buscas diárias por palavras-chave relacionadas a COVID-19 e o total de notificações oficiais de infectados pela doença em determinadas regiões do Brasil.

Segundo Bregman (1999), a regressão linear múltipla é mais flexível e eficiente para estudo exploratório se comparada a outras técnicas estatísticas. Assim, a pesquisa pautou-se na realização de uma análise multivariada para explorar o grau de explicabilidade de variáveis preditoras sobre a variável resposta a ser estimada (Hair *et al.,* 2009). Essa forma de análise foi escolhida por ser a mais indicada para se obter resultados mais próximos à realidade exigida pela maioria dos ambientes de pesquisa educacional (Henson & Capraro, 2001).

O estudo consistiu em uma análise exploratória de RLM que utilizou como variável preditora dados disponibilizados publicamente na plataforma de buscas Google Trends®, dimensionadas em intervalo relativo de 0 a 100, sendo 100 o dia com mais buscas.

Como principal hipótese para a escolha das palavras-chave, temos a possível relação sequencial entre os termos “Sintomas de COVID” e “Teste de COVID” e o possível surgimento de picos. Isso ocorre porque no primeiro momento uma pessoa, possivelmente contaminada com a doença, começa suas buscas por semelhanças entre o que ela está sentindo e o que o COVID-19 manifesta. Consequentemente, caso essa pessoa perceba um alto índice de semelhança, ela possivelmente começará a buscar por formas de confirmar a suspeita levantada, o que a levaria a pesquisar sobre os testes disponíveis no mercado para detectar a doença. Kurian *et al.* (2020), em um estudo sobre o tema, fez uso de palavras-chave relativas aos principais sintomas e ao diagnóstico da enfermidade, dentre elas estavam “sintomas COVID” e “Centros de testagem para COVID”.

Posto isso, a Tabela 1 mostra os termos escolhidos para análise do dia 1º de abril de 2020 ao dia 29 de março de 2021 em 3 estados brasileiros, os quais foram selecionados a partir da análise de dois parâmetros: tamanho populacional e impacto na economia do país. São Paulo, Minas Gerais, e Rio de Janeiro ocupam 1º, 2º e 3º lugares na lista de unidades federativas com maior número de habitantes (IBGE, 2020) e juntos representam 51,2% do produto interno bruto brasileiro (IBGE, 2018).

Tabela 1

Descrição dos termos utilizados para predizer novos casos de COVID-19

Nome do termo de pesquisa	Descrição
Teste de COVID	O interesse por " Teste de COVID" em Minas Gerais, Rio de Janeiro e São Paulo
Sintomas de COVID	O interesse por "Sintomas de COVID" em Minas Gerais, Rio de Janeiro e São Paulo

Fonte: Elaborada pelos autores.

A variável resposta a ser investigada foi obtida através dos dados das respectivas Secretarias Estaduais de Saúde coletadas por pesquisadores independentes e compilados em um único banco que disponibiliza a quantidade diária de notificações oficiais de infectados para os estados considerados no período em estudo. Buscando minimizar possíveis assimetrias nos resultados de cada estado, utilizou-se o tamanho populacional fornecido pelo Instituto Brasileiro de Geografia e Estatística (IBGE), como variável de controle. Dividindo a quantidade de infectados pela população foi possível obter a incidência de COVID - 19 em cada local.

Posteriormente, com as informações coletadas com a ajuda de pacotes estatísticos obtidos via R-Studio (versão 3.6.1), o qual é um ambiente de software livre para computação e gráficos estatísticos, iniciaram-se as análises preliminares. Foi feita uma média móvel dos resultados da variável resposta para retirar a sazonalidade desses dados, que acontecem principalmente pela disparidade de notificações da quantidade de infectados durante a semana e aos finais de semana. Em seguida, um modelo de predição foi criado. A Equação 1 demonstra a relação entre as variáveis do modelo.

$$y_i = \beta_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \varepsilon_1 + \varepsilon_2 \qquad \text{onde i = 1,2,3...}$$

Equação 1

Onde α_1 e α_2 são os coeficientes angulares, β_0 é o intercepto e corresponde ao valor do estimador de y quando x é zero, ε_1 e ε_2 são os erros de estimação dos coeficientes angulares e x_{1i} e x_{2i} representam os termos de pesquisa "Teste de COVID" e "Sintomas de COVID", e i representa cada valor amostrado da série histórica. A equação é aplicada, individualmente, para cada estado.

A qualidade de ajuste do modelo preditivo proposto é avaliada através do p-valor e o R^2 ajustado, que inferem a significância do modelo e o grau de explicabilidade do mesmo. Adicionalmente, é avaliado o ajuste de qualidade, através do teste de normalidade dos resíduos (Teste de Shapiro Wilk) e de multicolinearidade (Teste de VIF).

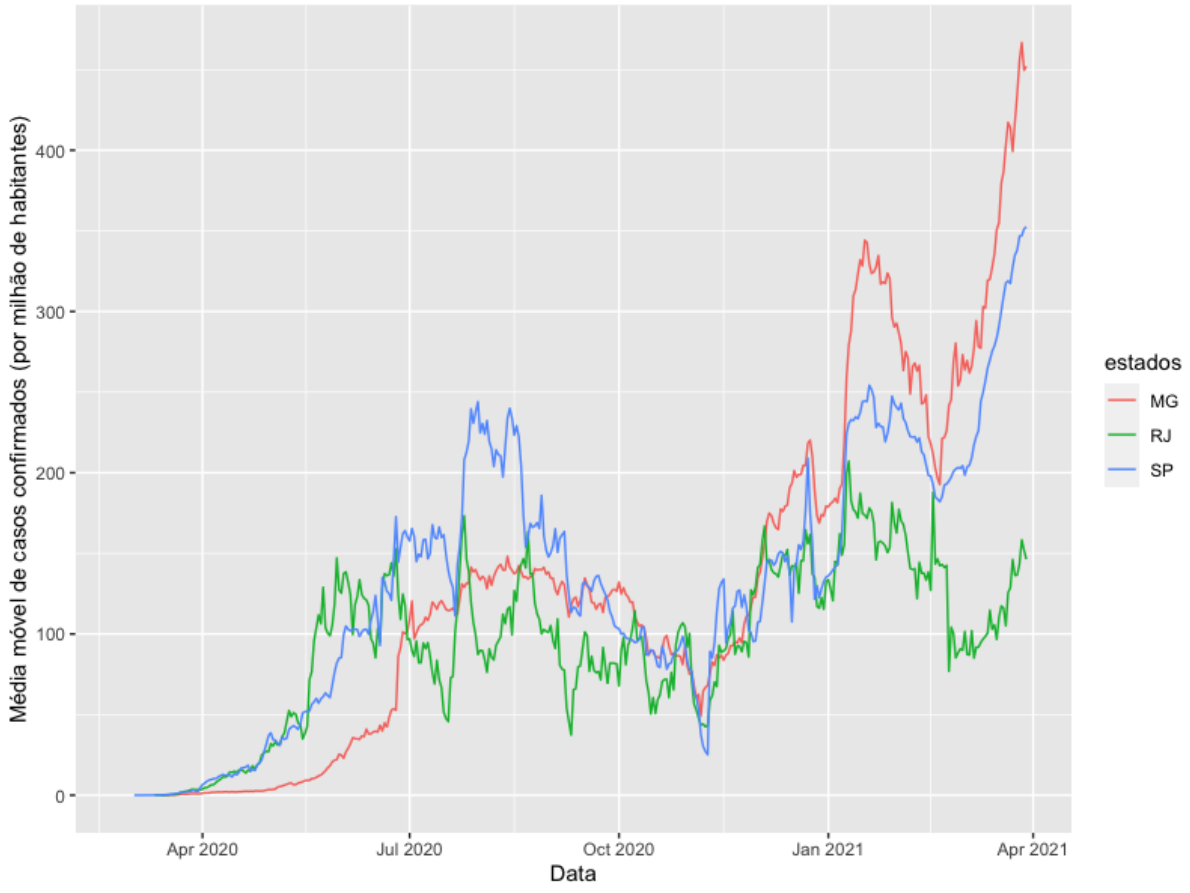
O teste de Shapiro Wilk é considerado o mais comum para análise de normalidade na estatística de software (Razali & Wah, 2011). Sua função é avaliar a aderência dos resíduos à distribuição normal através do p-valor, quando ele é inferior a 0,05 a hipótese de normalidade é rejeitada. Em relação ao teste de VIF o fator de inflação da variância (VIF) é inversamente proporcional à tolerância, a qual se refere a quanto uma variável é não explicada pelas demais variáveis independentes (Hair *et al.*, 2009). Segundo Draper e Smith (1998) valores de VIF maiores do que 10 podem causar problemas na estimação dos coeficientes de regressão, confirmando a presença da multicolinearidade.

4 Discussão dos resultados

Essa seção é destinada a apresentar os resultados do estudo quantitativo realizado ao longo do desenvolvimento da pesquisa. Para compreensão de como a pandemia desenvolveu-se nos estados de Minas Gerais, Rio de Janeiro e São Paulo do dia 1º de abril de 2020 ao dia 29 de março de 2021 elaborou-se o Gráfico 1, o qual apresenta a média móvel do número de infectados por COVID - 19.

Gráfico 1

Média móvel de casos confirmados por milhão de habitantes



Fonte: Elaborado pelos autores.

Nota-se a existência de picos que ocorrem simultaneamente nos três estados: o primeiro em agosto de 2020 e o segundo em fevereiro do ano seguinte. Ademais, nota-se a possibilidade de um terceiro ponto de ápice no número de casos confirmados no início de abril de 2021. Portanto, percebe-se a importância de se antever, com a ajuda de modelos preditivos, o desenvolvimento da doença para que decisões sejam tomadas a fim de mitigar ou impedir novos picos nesta situação de ressurgência constante de surtos.

4.1 Análise da regressão linear múltipla

Para investigação de seus resultados da regressão linear multivariada foi feita uma exposição comparativa entre os estados levando em conta o coeficiente, simbolizado por a1 e a2, que correspondem às palavras chave "Teste de COVID" e "Sintomas de COVID", respectivamente. A Tabela 2 oferece uma visão macro da análise deste estudo. Nos tópicos posteriores à tabela, tal visão será detalhada. Como apoio à tabela e a análise dos resultados tem-se o Gráfico 2, o qual contém projeções da regressão linear múltipla para cada estado.

Tabela 2

Resultados da análise de regressão linear

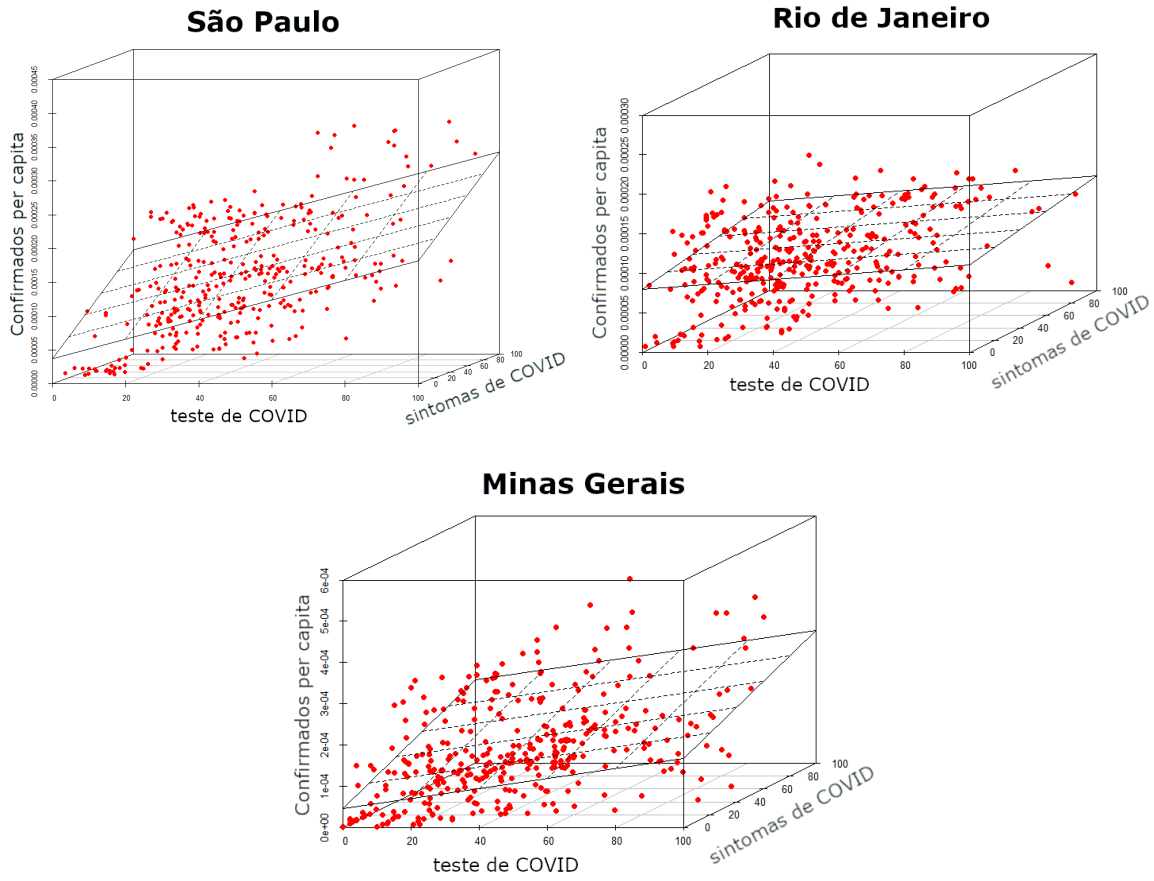
Estado	a (teste)	a (sintomas)	p-valor (teste)	p-valor (sintomas)	R^2
Minas Gerais	$1{,}663e^{-6}$	$2{,}127e^{-6}$	$9{,}022e^{-6}$	$1{,}446e^{-11}$	0,238
Rio de Janeiro	$3{,}279e^{-7}$	$4{,}828e^{-7}$	0,131	0,029	0,019
São Paulo	$1{,}623e^{-6}$	$1{,}257e^{-6}$	$3{,}1792e^{-10}$	$2{,}297e^{-6}$	0,228

Fonte: Elaborado pelos autores.

Assim, com a análise da Tabela 2 percebe-se que quando a busca por "Teste de COVID" aumenta em Minas Gerais, Rio de Janeiro e São Paulo, o número de infectados cresce em $1.663e^{-6}$, $3{,}279e^{-7}$, $1{,}623e^{-6}$ por unidade de aumento, respectivamente. De forma análoga, o acréscimo no número de pesquisas por "Sintomas de COVID" está relacionado a um aumento de $2.127e^{-6}$, $4{,}828e^{-7}$, $1{,}257e^{-6}$ na quantidade de infectados em MG, RJ e SP. Portanto, foi possível perceber a existência de conexão entre as variáveis.

Gráfico 2

Regressão linear múltipla para os diferentes estados



Fonte: Elaborado pelos autores.

O Gráfico 2 possui cinco componentes a serem observados: os eixos X e Y, que apresentam o valor relativo de buscas *online* diárias por "Teste de COVID" e por "Sintomas de COVID"; eixo Z, que contém a quantidade diária de notificações oficiais de infectados pela doença; a dispersão em vermelho representa como a série histórica da quantidade de pesquisas se converte no número de casos confirmados de COVID-19 em dados reais; já o plano permite obter uma projeção da quantidade de casos em dependência do RSV das palavras chaves, de acordo com os coeficientes obtidos pelo modelo de regressão linear.

Assim, a análise gráfica auxilia na visualização da conexão entre as variáveis citadas anteriormente, ao qual observa-se que o aumento de buscas reflete em um cenário de aumento de casos confirmados de COVID-19 e uma diminuição das buscas em redução de casos confirmados.

4.2 Qualidade do ajuste da regressão

Como apresentado anteriormente, aplicou-se uma regressão linear múltipla para verificar como esta abordagem pode auxiliar na predição de novos ápices da doença em abordagens futuras. No entanto, para validar a qualidade estatística dessa regressão foram aplicados os testes de normalidade e multicolinearidade.

Após realização do teste de normalidade dos resíduos, foram encontrados valores baixos para o p-valor de Minas Gerais, Rio de Janeiro e São Paulo, sendo eles 4,03e-9, 2,88e-5 e 0,024 respectivamente. Assim, a hipótese de normalidade dos resíduos acabou sendo rejeitada e concluiu-se que os erros não são normalmente distribuídos. Isto mostra que alguns fatores que podem impactar na validade do modelo de regressão devem ser considerados para melhoria deste modelo em abordagens futuras. Análise da tendência e sazonalidade poderiam ser verificadas no tratamento prévio da série de dados, por exemplo, para se buscar atingir a normalidade dos resíduos.

Entretanto, com o teste de VIF, os valores encontrados para o fator de inflação da variância foram 1,053, 1,217, 1,124 para MG, RJ e SP, respectivamente. Assim são valores baixos, o que descarta a hipótese de multicolinearidade, e garante que as variáveis preditoras do modelo não estão correlacionadas estatisticamente entre si, mas apenas à variável resposta. Isto pode indicar que o uso das palavras-chave propostas neste artigo podem ser utilizadas de forma conjunta sem adicionar redundância entre si considerando os dados destas regiões.

5 Conclusões para a teoria, prática e sociedade

Este estudo exploratório investigou uma possível relação entre o valor relativo de buscas diárias por palavras-chave relacionadas à COVID-19 e o total de notificações oficiais de infectados pela doença nos estados de Minas Gerais, Rio de Janeiro e São Paulo. Através da aplicação de um modelo de regressão múltipla percebeu-se conexões significativas entre as variáveis estudadas com grau de explicabilidade aceitável, assim como a ausência de multicolinearidade entre as palavras sugeridas.

Convém ressaltar que o método exposto nesta análise não foi projetado para substituir as redes de vigilância tradicionais ou suplantar a necessidade de diagnósticos laboratoriais e vigilância. Ele apresenta *insights* iniciais sobre a relação entre as buscas na internet e o aumento de casos de COVID em determinada região. Uma das limitações da abordagem deste artigo é que ela não apresenta uma amostra estatística de toda uma população, já que os dados do Google Trends®, retratam apenas uma parcela da população que possui inclusão digital. Também é importante salientar que os dados não refletem a evolução da doença em si, e sim a evolução das notificações, que podem apresentar atrasos e subnotificação.

Para trabalhos futuros, pode-se verificar a validade da relação das palavras-chave levantadas com casos notificados de COVID-19 para as semanas seguintes, ou seja, deslocando alguns dias na variável resposta. Com isso, é possível observar quão forte é a relação entre as buscas e as notificações alguns dias depois, com o intuito de auxiliar gestores locais de saúde em decisões estratégicas através de interfaces amigáveis via regressão linear múltipla, como uma ferramenta adicional às ferramentas já existentes de controle da pandemia. Além disso, análises de séries temporais podem ser exploradas, como análise de estacionariedade, buscando atingir a normalidade dos resíduos para a validação estatística das premissas desta abordagem, assim como testes de homocedasticidade para verificação de variância constante.

Referências

Angus, D. C. (2020). Optimizing the trade-off between learning and doing in a pandemic. Jama, 323(19), 1895-1896. https://doi:10.1001/jama.2020.4984

Ayyoubzadeh, S. M., Ayyoubzadeh, S. M., Zahedi, H., Ahmadi, M., & Kalhori, S. R. N. (2020). Predicting COVID-19 incidence through analysis of google trends data in iran: data mining and deep learning pilot study. JMIR public health and surveillance, 6(2), e18828. https://doi.org/10.2196/18828

Bakker, K. M., Martinez-Bakker, M. E., Helm, B., & Stevenson, T. J. (2016). Digital epidemiology reveals global childhood disease seasonality and the effects of immunization. Proceedings of the National Academy of Sciences, 113(24), 6689-6694. https://doi.org/10.1073/pnas.1523941113

Brasil, & Brasil. (2021). Conselho Nacional dos Secretários de Saúde (CONASS). Atenção Primária e promoção da Saúde.

Bregman, J. I. (1999). Environmental impact statements. CRC Press.

Buheji, M., da Costa Cunha, K., Beka, G., Mavric, B., De Souza, Y. L., da Costa Silva, S. S., ... & Yein, T. C. (2020). The extent of covid-19 pandemic socio-economic impact on global poverty. a global integrative multidisciplinary review. American Journal of Economics, 10(4), 213-224. https://doi.org/10.5923/j.economics.20201004.02

Draper, N. R., & Smith, H. (1998). Applied regression analysis (Vol. 326). John Wiley & Sons.

Eysenbach, G. (2009). Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. Journal of medical Internet research, 11(1), e11. https://doi.org/10.2196/jmir.1157

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. Nature, 457(7232), 1012-1014. https://doi.org/10.1038/nature07634

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2009). Análise multivariada de dados. Bookman editora.

Henson, R. K., Capraro, R. M., & Capraro, M. M. (2001). Reporting Practice and Use of Exploratory Factor Analysis in Educational Research Journals.

Instituto Brasileiro de Geografia, & Estatística. (2020). Censo demográfico 2020: características da população e dos domicílios: resultados do universo. Ministério de Planejamento, Orçamento e Gestão, Instituto Brasileiro de Geografia Estatística-IBGE.

Instituto Brasileiro de Geografia, & Estatística. Coordenação de Contas Nacionais. (2018). Produto interno bruto dos municípios. IBGE.

Junior, R. R. F., & Santa Rita, L. P. (2016). Impactos da Covid-19 na Economia: limites, desafios e políticas.Cadernos de Prospecção, vol. 13, n. 2, 2020. http://dx.doi.org/10.9771/rf.v1i7.37324

Kumar, A., Sinwar, D., & Saini, M. (2020). Study of several key parameters responsible for COVID-19 outbreak using multiple regression analysis and multi-layer feed forward neural network. Journal of Interdisciplinary Mathematics, 1-23. https://doi.org/10.1080/09720502.2020.1833443

Kurian, S. J., Alvi, M. A., Ting, H. H., Storlie, C., Wilson, P. M., Shah, N. D., ... & Bydon, M. (2020, November). Correlations Between COVID-19 Cases and Google Trends Data in the United States: A State-by-State Analysis. In Mayo Clinic Proceedings (Vol. 95, No. 11, pp. 2370-2381). Elsevier. https://doi.org/10.1016/j.mayocp.2020.08.022

Lakatos, E. M., & Marconi, M. D. A. (1996). Técnicas de pesquisa. São Paulo: Atlas, 205, 88.

Lin, S., Fu, Y., Jia, X., Ding, S., Wu, Y., & Huang, Z. (2020). Discovering Correlations between the COVID-19 Epidemic Spread and Climate. International Journal of Environmental Research and Public Health, 17(21), 7958. https://doi.org/10.3390/ijerph17217958

Liu, D., Clemente, L., Poirier, C., Ding, X., Chinazzi, M., Davis, J. T., ... & Santillana, M. (2020). A machine learning methodology for real-time forecasting of the 2019-2020 COVID-19 outbreak using Internet searches, news alerts, and estimates from mechanistic models. arXiv preprint arXiv:2004.04019.

Mavragani, A., & Gkillas, K. (2020). COVID-19 predictability in the United States using Google Trends time series. Scientific reports, 10(1), 1-12. https://doi.org/10.1038/s41598-020-77275-9

McKibbin, W., & Fernando, R. (2021). The global macroeconomic impacts of COVID-19: Seven scenarios. Asian Economic Papers, 20(2), 1-30. https://doi.org/10.1162/asep_a_00796

Rabajante, J. F. (2020). Insights from early mathematical models of 2019-nCoV acute respiratory disease (COVID-19) dynamics. arXiv preprint arXiv:2002.05296.

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. Journal of statistical modeling and analytics, 2(1), 21-33.

Siettos, C. I., & Russo, L. (2013). Mathematical modeling of infectious disease dynamics. Virulence, 4(4), 295-306. https://doi.org/10.4161/viru.24041

Teng, Y., Bi, D., Xie, G., Jin, Y., Huang, Y., Lin, B., ... & Tong, Y. (2017). Dynamic forecasting of Zika epidemics using Google Trends. PloS one, 12(1), e0165085. https://doi.org/10.1371/journal.pone.0165085

Tizzoni, M., Bajardi, P., Poletto, C., Ramasco, J. J., Balcan, D., Gonçalves, B., ... & Vespignani, A. (2012). Real-time numerical forecast of global epidemic spreading: case study of 2009 A/H1N1pdm. BMC medicine, 10(1), 1-31. https://doi.org/10.1186/1741-7015-10-165

WHO. Statement on the second meeting of the international health regulations (2005) emergency committee regarding the outbreak of novel coronavirus (2019-ncov). URL https://www.who.int/news-room/detail/. [Online; accessed 9-March2021].