# Evolução do número de casos de contágios por COVID-19 no Brasil de acordo com diferentes variáveis demográficas

Germán D. Buitrago-Salazar*, Guilherme A. de Oliveira†, Christianne O. Dalforno‡, and Diego B. Santi§
Instituto de Computação - IC.
Universidade Estadual de Campinas - UNICAMP, Campinas, Brasil
Email: *gedabusa@gmail.com, ‡profcod@gmail.com
Universidade Tecnológica Federal do Paraná - UTFPR, Curitiba, Brasil
Email: †guilhermeoliveira.2019@alunos.utfpr.edu.br, §diegobsanti@hotmail.com

***Resumo*—Este documento apresenta os resultados das análises da evolução do número de casos de COVID-19 no Brasil, considerando variáveis demográficas como idade, nível de escolaridade, sexo, etnia e densidade populacional. As bases de dados da SIVEP-Gripe e *datasets* disponibilizados pelo e-SUS foram usadas nas análises e foram propostos vários modelos que relacionavam cada variável. Os resultados obtidos mostram como o foi a afetação da população brasileira desde a visão demográfica.**

***Index Terms*—COVID-19, variáveis demográficos, serie temporal**

## I. Introdução

Em 2020, a Organização Mundial de Saúde catalogou como pandemia o novo surto de vírus que se propagou por todos os países. Este vírus, que é uma mutação do coronavírus, é o agente causador da Síndrome Respiratória Aguda Grave (nomeado SARS-CoV2) que tem afetado mais de 7.450.000[1] de habitantes no Brasil. Para entender a incidência dessa doença e sua evolução desde a visão sociocultural no Brasil, é essencial estudar e analisar diferentes indicadores demográficos em função do número de casos registrados para cada zona.

Isto motivou a realização deste trabalho em que foi realizada uma análise da incidência de número de casos de pacientes que foram registrados com COVID-19 ao longo do tempo, baseado em variáveis demográficas. A análise tem como objetivo entender se há relações entre a idade, nível de escolaridade, sexo e etnia dos pacientes, além da densidade populacional do lugar onde foram identificados os casos, com a evolução do número de casos no Brasil.

O documento está dividido em 5 partes. A seção 2 contém trabalhos correlatos usados como base durante as análises. A seção 3 descreve as bases de dados, processamentos aplicados sobre estas e modelos considerados. Na 4 são apresentados os resultados bem como a metodologia aplicada para obtê-los. Finalmente, na seção 5 são relatadas as limitações e os trabalhos futuros.

[1]Número de casos reportados no portal do Ministério de Saúde e de Datasus no dia 26 de dezembro de 2020

## II. Trabalhos relacionados

Diversos trabalhos foram realizados para determinar a relação entre os casos de COVID-19 e algumas variáveis demográficas. Em [1], os autores visaram caracterizar os pacientes hospitalizados com Síndrome Respiratória Aguda (SRAG) por COVID-19 quanto a variáveis demográficas e comorbidades. O estudo se limitou até a 21º semana epidemiológica. Identificou e classificou os grupos com maior propensão à hospitalização. Para realizar a pesquisa, utilizaram a base do SIVEP-Gripe e bases auxiliares da projeção da população do Brasil e suas unidades federativas por sexo e idade, bases domiciliárias do IBGE, e informações de nascidos vivos.

Dessa maneira, analisaram variáveis tais como região de residência, sexo, faixa etária, raça e existência de comorbidades e gestantes segmentados por idade do paciente. O resultado dessa pesquisa demonstrou que há maior relação de hospitalização por COVID-19 para pessoas com idades mais avançadas ou que pertencem à faixa de 40 a 59 anos de idade, do sexo masculino e que apresente algum tipo de comorbidade.

Em [2] foi relatado um estudo que estimou a desigualdade no número de casos de COVID-19 e as mortes por esta doença entre etnias distintas. O trabalho mostrou que grupos de etnia africana-americana são os mais afetados pela doença e possuem a maior taxa de morte. Também foi observado a desproporcionalidade na participação do número de pessoas das diferentes etnias nos municípios estudados com o número de casos e mortes registrados, associadas com a doença.

Também existem autores que centralizaram sua pesquisa em torno da ligação de contagio por COVID-19 e o índice de escolaridade dos pacientes. Em [3] destacaram que além dos indicadores anteriores, consideraram fatores de risco como a idade, comorbidades e baixa renda, entre outros, que exacerbaram cada grupo de risco. Assim, destacaram que fatores de riscos relacionados a comorbidades são distribuídos de forma desproporcional segundo a escolaridade dos indivíduos, sendo os mais afetados aqueles que frequentaram só o ensino fundamental. Isto contribui para uma maior gravidade da

doença, bem como da necessidade de internação.

Um dos primeiros estudos a avaliar a influência do gênero nos casos de SARS-CoV2 revelou que a susceptibilidade de incidência da doença é similar entre homens e mulheres. Porém pessoas do sexo masculino têm maior tendência a desenvolver casos mais graves e ter como consequência a morte. Dentre os indivíduos estudados, a maioria tinha idade superior a 59 anos, enquanto poucos casos foram encontrados entre crianças e nenhuma mortalidade. Convém destacar que o estudo realizado teve limitações causadas por uma pequena quantidade de dados ou falta de dados detalhados sobre os pacientes. Ressalta-se que a falta de homogeneidade dos dados usados podem causar resultados tendenciosos [4].

Além das variáveis avaliadas nos estudos citados anteriormente, determinar a distribuição e crescimento dos números de casos dentro de uma área é importante dentro do estudo das variáveis demográficas. Para colaborar na tomada de decisão assertiva e divulgar dados epidemiológicos, [5] apresentou um histórico da difusão da COVID-19 no Estado de Santa Catarina através de uma coleção de mapas temporais. O trabalho documentou historicamente mapas de fluxo de novos casos causados por indivíduos que chegam de outras zonas. Os resultados finais foram as evoluções de novos casos produzidos dentro do estado e como se espalhou temporalmente desde os centros com maior densidade populacional até os municípios mais interiores.

## III. Dados e modelos

O estudo foi realizado com datasets disponibilizados publicamente por entidades governamentais do Brasil. A primeira base usada foi a *Sivep-Gripe*, originalmente criada para documentar os número de casos durante a pandemia de influenza H1N1 ocorrida em 2009. Hoje, esta base passou a registrar também os casos de Síndrome Respiratória Aguda Grave causados pelo vírus SARS-CoV-2 para todos os municípios que possui cadastro dentro da plataforma [1].

A segunda base utilizada é composta por datasets disponibilizados pelo *e-SUS*, cujos dados estão fracionados por estado [6]. A base registra os casos de Síndrome Gripal causados por COVID-19 diariamente e tem uma maior volumetria que a base anterior. Convém destacar que ambas as bases não estão completas devido a que alguns estados e municípios utilizam seu próprio sistema de registro de casos, gerando divergência entre as informações. Para as análises, os dados usados contêm registros até novembro de 2020.

As duas bases são preenchidas com assistência humana, provocando que haja registros nulos, inconsistências nos dados, falta de padronização das bases e formatação errada das colunas, entre outros. Assim, o processo de limpeza das bases começou com a eliminação de outliers implementando o método do escore padrão, rejeitando os valores que não estejam dentro do range $\pm 3$.

Logo, registros com informações inconsistentes ou faltantes foram removidas. O próximo passo foi a padronização das datas aplicando expressões regulares e deixando em uma formatação consistente. Finalmente, as palavras com acentos são normalizadas para ter um maior índice de assertividade no momento de cruzar informações desses campos.

Após a limpeza, os dados foram filtrados para considerar apenas os registros com casos confirmados de COVID-19. A respeito do preenchimento dos campos a serem estudados, aqueles que tinham valores como 'indefinido' ou 'ignorado' foram desconsiderados para evitar resultados anômalos. A etapa final deste pré-processamento consistiu na normalização dos valores numéricos usando os dados populacionais projetados para 2020 pelo IBGE, aplicando-os por estado.

Depois do pré-processamento da informação, a modelagem aplicada consistiu no uso da regressão linear para entender a correlação existente entre os dados, a clusterização para dividir, classificar e extrair as informações, e o modelo ANOVA para validar as afirmações propostas para algumas das variáveis demográficas.

## IV. Resultados

As análises realizadas com as informações extraídas estão divididas segundo a variável demográfica a ser estudada. Cada uma visava conferir a veracidade de uma hipótese formulada, que relacionava a variável com o número de casos em função do tempo.

### A. Análise por escolaridade

Para análise dos dados de escolaridade foram removidos os valores irrelevantes, sendo eles 51,86% dos registros com campo de escolaridade preenchidos como 'ignorado'. Após a filtragem, fez-se uma normalização por estados com uma taxa de 100 mil habitantes. A normalização foi aplicada agrupando e quantificando o número de infectados por estado e por escolaridade e, em seguida, dividido pelo número de habitantes do mesmo estado com a respectiva escolaridade.

Para verificar o comportamento dos dados normalizados, foram utilizadas algumas medidas estatísticas como média, quartis e desvio padrão aplicadas às taxas de infecção de todos os estados. Observando essas medidas na tabela I, as taxas de infecção de pessoas com ensino fundamental, médio e superior apresentam medidas muito próximas. Já as medidas das taxas de pessoas sem escolaridade diverge significativamente das demais.

Tabela I: Indicadores estatísticos das taxas de infecção de escolaridade.

| **medida** | **fund** | **med** | **sem** | **sup** |
|---|---|---|---|---|
| count | 27 | 27 | 27 | 27 |
| mean | 19.58 | 19.19 | 32.76 | 21.73 |
| std | 24.05 | 22.39 | 57.53 | 25.79 |
| min | 1.6 | 3.91 | 2.37 | 0.86 |
| 25% | 6.07 | 6.8 | 8.09 | 7.7 |
| 50% | 10.07 | 10.81 | 13.78 | 10.57 |
| 75% | 23.73 | 21.87 | 25.24 | 22.73 |
| max | 118.96 | 89.92 | 285.38 | 119.41 |

Foi aplicado o modelo ANOVA para verificar a proximidade das escolaridades pelas taxas de infecção. A hipótese nula do ANOVA considera que as médias são iguais entre os valores de

duas variáveis categóricas. Com o teste de Tukey foi possível comparar as escolaridades duas a duas como mostra a tabela II. Ainda na tabela II pode-se observar que todos os valores de $p$ confirmam a hipótese nula, porém, observa-se que o valor de p é menor e há uma diferença percentual maior nas comparações de 'sem escolaridade' com as outras categorias. Já a diferença entre as escolaridades de ensino fundamental, médio e superior é muito baixa.

Tabela II: Comparação das taxas de infecção de escolaridades com o modelo ANOVA.

| group1 | group2 | Diff | Lower | Upper | q-valuq | p-value |
|---|---|---|---|---|---|---|
| fund | med | 0.39 | -24.87 | 25.65 | 0.06 | 0.9 |
| fund | sem | 13.18 | -12.08 | 38.45 | 1.93 | 0.52 |
| fund | sup | 2.15 | -23.11 | 27.41 | 0.31 | 0.9 |
| med | sem | 13.58 | -11.69 | 38.84 | 1.98 | 0.5 |
| med | sup | 2.54 | -22.72 | 27.8 | 0.37 | 0.9 |
| sem | sup | 11.04 | -14.23 | 36.3 | 1.61 | 0.65 |

Foi verificado também se os pressupostos do modelo ANOVA foram atendidos. Primeiramente, foi testado se os resíduos têm uma distribuição normal com o teste de Shapiro-Wilk. A hipótese nula desse teste afirma que os dados apresentam distribuição normal. Aplicando o teste para os resíduos do modelo obteve-se que $p = 4.3 \times 10^{-16}$ e portanto rejeita a hipótese nula. Porém o modelo ANOVA não é muito sensitivo com a violação desse pressuposto [7].

Outro pressuposto verificado foi se as populações tinham variâncias iguais, pelo qual foi realizado o teste de Levene. O valor de p para o teste de Levene's foi igual a 0.49, por conseguinte validou a hipótese nula e concluiu-se que as populações apresentaram igual variância.

Além disso, foi verificado como os casos se comportaram ao longo do tempo. Para visualizar isso, fez uma média das taxas de infecção de todos os estados pelo o tempo para cada escolaridade. A figura 1 mostra a evolução nos casos do Brasil com a média dos estados. Observe que pacientes sem um grau de escolaridade têm sido as mais afetadas pelo coronavirus. Dentre os pacientes que cursar pelo menos um grau de escolaridade, pessoas com nível superior são os mais contagiados.

### *B. Análise por etnia*

A análise de etnia foi semelhante a de escolaridade. Inicialmente foram filtrados os dados retirando os dados de etnia preenchidos como 'ignorado' correspondendo a 20.58% dos registros. Após, foi feita uma normalização por estados usando dados do IBGE. O procedimento consistiu em fazer uma taxa de infectados por 100 mil habitantes separado por etnia e estado.

Com os dados normalizados, uma breve análise com indicadores estatísticos foi feita como mostra a tabela III. Analisando a tabela III, percebe-se que a etnia branca apresenta a menor média, variância, mediana, valor mínimo e valor máximo. Os dados indicam que pessoas de etnia branca apresentam as menores taxas de infecção no país todo. Além disso nota-se

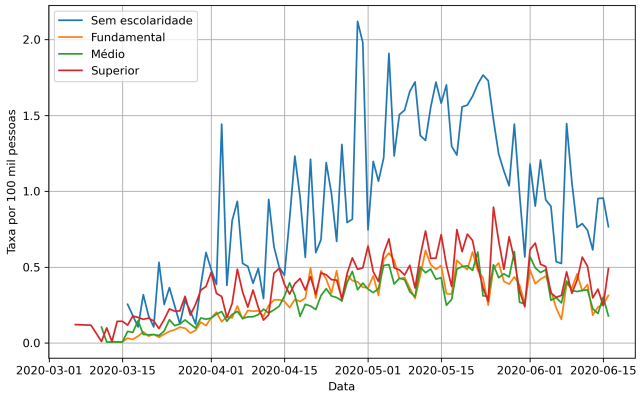


Figura 1: Médias das taxas de infecção dos estados para diferentes escolaridade ao longo do tempo.

na mesma tabela, que a etnia amarela apresenta as maiores taxas. Logo em seguida as etnias parda e indígena apresentam taxas parecidas, porém a etnia parda com maior variância nas taxas.

Tabela III: Indicadores estatísticos das taxas de infecção de etnia.

| medida | amarela | branca | indígena | parda | preta |
|---|---|---|---|---|---|
| count | 25 | 27 | 21 | 27 | 27 |
| mean | 41.57 | 19.06 | 30.88 | 32.82 | 21.18 |
| std | 27.61 | 16.70 | 23.87 | 30.86 | 20.63 |
| min | 7.23 | 3.49 | 4.34 | 4.92 | 3.92 |
| 25% | 20.44 | 10.18 | 14.59 | 11.57 | 8.61 |
| 50% | 32.23 | 13.85 | 28.11 | 20.00 | 14.14 |
| 75% | 55.90 | 18.23 | 32.65 | 39.56 | 22.74 |
| max | 97.70 | 62.94 | 94.35 | 145.35 | 92.73 |

### *C. Análise por sexo*

Na análise por sexo primeiramente verificou-se a quantidade de infectados de cada sexo. Os dados de casos confirmados de COVID-19 na base SIVEP eram de 57.04% homens, 42.93% de mulheres e 0.03% preenchido como indefinido. Os dados mostraram inicialmente que homens eram mais infectados que mulheres. Para melhor examinar isso, fez-se uma normalização por estados com os dados do IBGE fazendo uma taxa de 100 mil habitantes. Analisando os dados normalizados, somente o estado do Mato Grosso apresentou uma maior taxa de infecção em mulheres do que em homens com uma pequena diferença como pode ser observado na figura 2. Isso pode indicar que homens estão sendo mais infectados do que mulheres no Brasil.

### *D. Análise por faixa etária*

A análise feita nessa variável mostrou que os casos de COVID-19 prevaleceram em pessoas na faixa etária de 30 a 50 anos com a maioria dos casos entre as pessoas na faixa de 30 a 40 anos. Na primeira fase, a idade dos pacientes foi arrendondada para a dezena anterior e os registros foram ranqueados e quantificados. Na normalização foi aplicado o

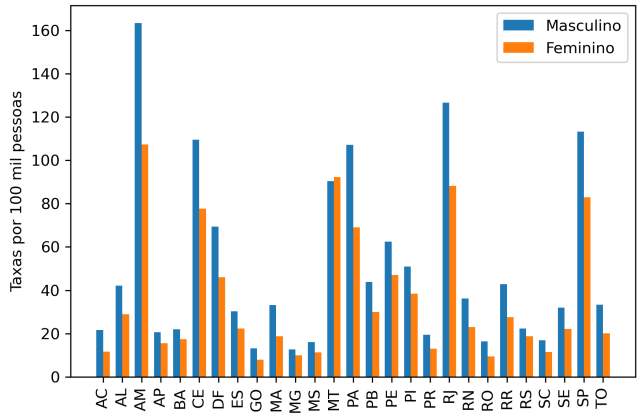


Figura 2: Taxa de infecção por sexo em todos estados brasileiros.

índice *per capita* por 1 milhão de habitantes e se formulou sua série temporal. Para evitar que as quedas no número de casos, causadas pelo atraso no registro dos dados, refletisse no gráfico e dificultasse sua visualização, aplicou-se a média móvel considerando 50 amostras por vez. Além disso, considerou-se o intervalo de confiança de 95%.

A figura 3 plota o comportamento de novos casos por dia. Cada linha representa uma faixa etária diferente. Observe que a velocidade de crescimento no começo da pandemia foi similar, mas o maior número de casos reportados encontra-se para pacientes com idades entre 30 e 50 anos. O terceiro grupo com maior número é o das pessoas entre 50 e 60 anos, tendo valores próximos dos anteriores.

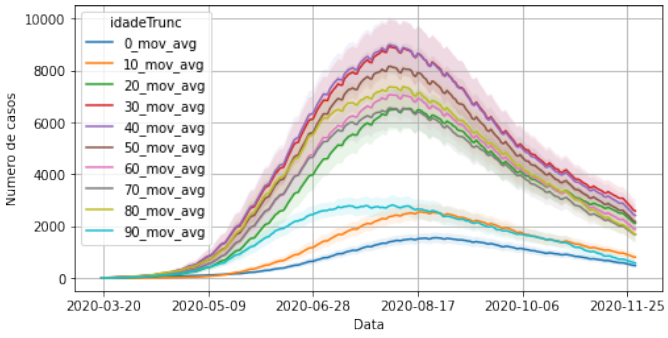


Figura 3: Número de casos em relação à faixa etária ao longo do tempo.

Convém destacar que grupos mais vulneráveis (pessoas da terceira idade e crianças) têm sido os menos afetados pelo contágio. Também pôde-se observar que o pico de novos casos entre as crianças aconteceu após o pico de novos casos nos demais grupos.

### *E. Análise por densidade demográfica*

Para entender o relacionamento dos novos casos com a densidade demográfica, a análise foi dividida para entender em qual momento da pandemia foram registrados o maior número de casos, qual foi a velocidade de crescimento de novos casos segmentado em regiões de alta e baixa densidade populacional e a correlação existente entre densidade demográfica e registro de casos. Para cada abordagem, desconsiderou-se dados de municípios que não tivessem mais de 1000 casos confirmados, que possuíssem uma população estimada inferior que 100000 habitantes e que sua densidade populacional fosse menor que 10 hab/km$^2$, dado que podiam gerar inconsistências no modelo por ser outliers.

Como realizado no processamento da faixa etária (seção IV-D), os dados normalizaram-se com o índice *per capita* por 1 milhão de habitantes. A divisão das cidades entre alta e baixa densidade populacional se realizou com uma distribuição normal, classificando-os entre valores maiores e menores que 0.

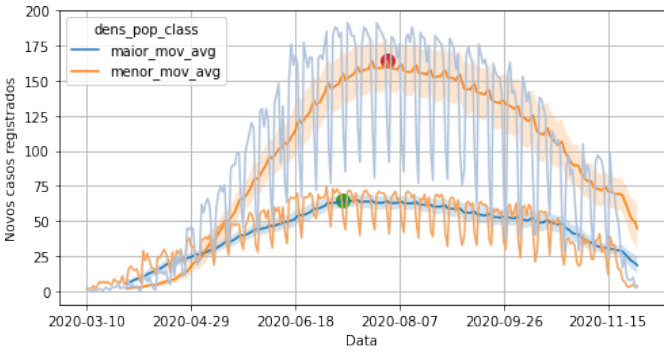


Figura 4: Séries temporais do número de novos casos de COVID-19 divididos pela densidade populacional.

A figura 4 mostra como foi o número de casos por dia. O modelo aplicou a média móvel usando 20 amostras por vez para amortecer a mudança brusca nos dados registrados. Observe que no começo, o número de casos reportados era maior nas cidades de densidade populacional alta. Porém, com o decorrer do tempo, a pandemia foi se espalhando para as cidades com menor densidade, subindo rapidamente o número de casos e ultrapassando os registros do outro grupo.

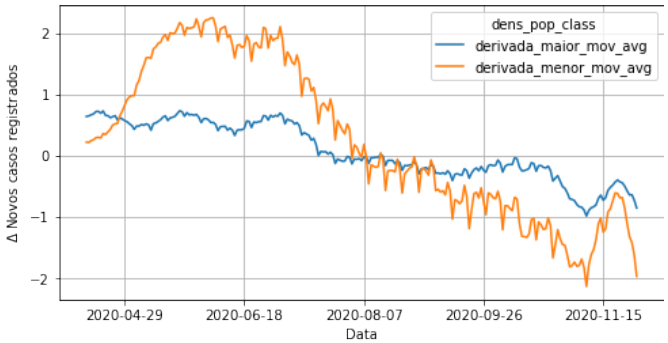


Figura 5: Taxas de crescimento de número de casos divididos pela densidade populacional.

Em adição ao anterior, o dia com maior número de casos foi atingido primeiro nas cidades maiores, mantendo um número constante de casos antes de decrescer velozmente. No caso

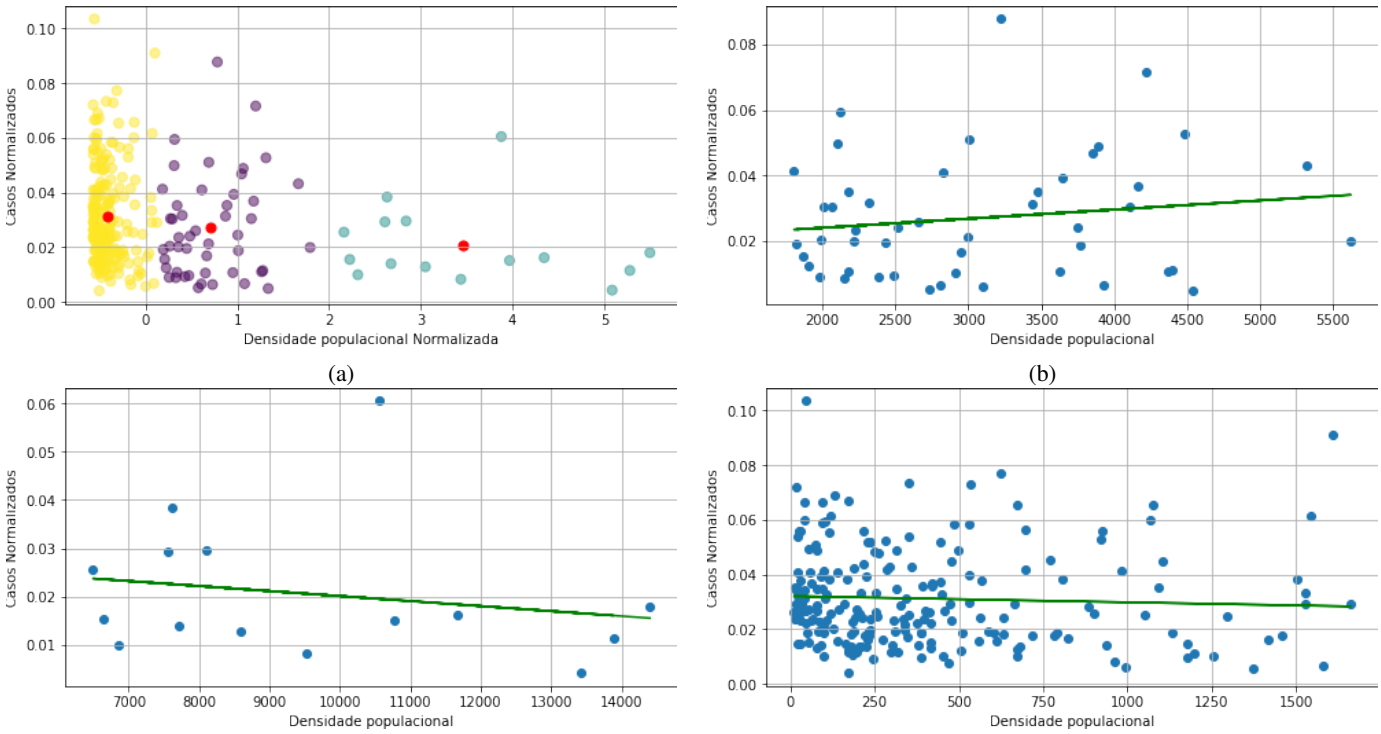


Figura 6: Resultados da análise correlacional considerando a densidade populacional. (a) Clusterização dos número de casos. (b) Regressão linear realizado no cluster 0. (c) Regressão linear realizado no cluster 1. (d) Regressão linear realizado no cluster 2.

dos outros municípios, depois de alcançar o máximo de novos casos, a taxa de decrescimento foi maior. A figura 5 expõe a taxa de crescimento dos casos como um diferencial de novos casos relacionado com o $\Delta$ de tempo.

Para encontrar um modelo adequado que evidencie com maior precisão a correlação entre a densidade populacional e os casos de COVID-19, aplicou-se uma clusterização utilizando o método de K-Means. O algoritmo criou três clusters como observado na figura 6a. Para o cluster 0, o range de densidade populacional dos municípios está entre 1700 e 5600 hab/km$^2$ e equivalem a 17.7% das amostras usadas. O cluster 1 representa todas as cidades cuja densidade populacional seja superior de 5600 hab/km$^2$, sendo 5.53% das amostras usadas, enquanto o cluster 2 representa todas as cidades com densidade demográfica menor de 1800 hab/km$^2$.

Para cada cluster foi realizado uma regressão linear para ajustar os dados numa reta, como observados nas figuras 6b, 6c, 6d. Para todos os cluster, o coeficiente de determinação ($r^2$) foi próximo do zero, entendendo que o modelo linear no possui nenhuma variabilidade dos seus dados com relação à sua média e, portanto, a correlação é mínima. Os valores são observados na tabela IV. Também, foi aplicada a correlação de Pearson para entender se a clusterização melhorava a relação existente entre densidade populacional e número de casos. Quando foi analisada a partir dos casos não normalizados, somente o cluster 0 teve uma correlação maior do que o conjunto completo. Já considerando os casos normalizados, para todos os clusters, seu valor foi maior, porém, não superando o limiar de 20%.

Tabela IV: Análise correlacional das variáveis para os clusters.

| **Grupo** | **r$^2$** | **Correlação de Pearson da Densidade Demográfica normalizada** | |
|---|---|---|---|
| | | ***Número de casos*** | ***Número de casos normalizados*** |
| Dados completos | - | 0.36 | 0.024 |
| Cluster 0 | 0.0217 | 0.444 | 0.1473 |
| Cluster 1 | 0.0385 | -0.237 | -0.1963 |
| Cluster 2 | 0.0025 | 0.176 | -0.0504 |

## V. Limitações e Trabalhos Futuros

As limitações foram principalmente relacionadas com as bases usadas. Um caso específico foi no momento de normalizar os dados de escolaridade usando a base populacional do IBGE, que não diferencia se o nível de escolaridade indicado foi concluído ou não. Para contornar isso somou-se a quantidade de pessoas com ensino completo e incompleto daquela respectiva categoria. Contudo, é possível que esse contorno seja um erro de interpretação, pois a documentação da base SIVEP não deixa claro em que categoria estariam as pessoas com ensino incompleto.

No caso do *e-SUS*, os dados apresentavam inconsistências, valores zerados ou com formatação errada. Foi realizado um

pré-processamento para limpar, ajustar e normalizar os dados para ter um percentual maior de confiabilidade. Como a base do *e-SUS* discriminava por estado, foi necessário padronizar todas as tabelas para um único modelo que satisfizesse os requerimentos das análises.

Dentre os trabalhos futuros a serem realizados, destacam-se a análise de mais variáveis que melhorem o modelo correlacional; a definição de perfis com a combinação das variáveis para entender a probabilidade de uma pessoa, com uma determinada combinação de características, ser infectada pelo coronavírus e a implementação de um modelo de AI (*Artificial Inteligence*) que relacione melhor todas as variáveis demográficas.

## AGRADECIMENTOS

Os autores agradecem aos professores Luiz Celso Gomes Junior, André Santanché, Paula Dornhofer e Ricardo Luders pelas orientações dadas para o desenvolvimento do projeto.

## REFERÊNCIAS

[1] R. Niquini, R. Lana, A. Pacheco, O. Cruz, F. Coelho, L. Carvalho, D. Villela, M. Gomes, and L. Bastos, "Srag por covid-19 no brasil: descrição e comparação de características demográficas e comorbidades com srag por influenza e com a população geral," in *Caderno de Saúde Pública 2020*. CSP - Cadernos de Saúde Pública, 2020.

[2] U. V. Mahajan and M. Larkins-Pettigrew, "Racial demographics and COVID-19 confirmed cases and deaths: a correlational analysis of 2886 US counties," *Journal of Public Health*, vol. 42, no. 3, pp. 445–447, 05 2020. [Online]. Available: https://doi.org/10.1093/pubmed/fdaa070

[3] L. Carvalho, L. Nassif Pires, and L. de Lima Xavier, "Covid-19 e desigualdade no brasil," 04 2020.

[4] J.-M. Jin, P. Bai, W. He, F. Wu, X.-F. Liu, D.-M. Han, S. Liu, and J.-K. Yang, "Gender differences in patients with covid-19: Focus on severity and mortality," *Frontiers in Public Health*, vol. 8, p. 152, 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpubh.2020.00152

[5] V. d. S. C. Reginato, E. Werneck, P. F. Meliani, S. d. S. Fernandez, and A. F. Bozio, "Coleção de mapas temporais como auxílio na representação da difusão da covid-19 no estado de santa catarina – histórico entre 12/03/2020 e 11/05/2020," *Metodologias e Aprendizado*, vol. 3, 2020. [Online]. Available: https://doi.org/10.21166/metapre.v3i0.1335

[6] SUS, "Notificações de síndrome gripal," 2020. [Online]. Available: https://opendatasus.saude.gov.br/ne/dataset/casos-nacionais

[7] M. Spiegel, J. Schiller, and R. A. Srinivasan, *Schaum's Outline of Theory and Problems of Probability and Statistics*, 2nd ed. McGraw-Hill, 2007, ch. 9. Análisis de la varianza, pp. 335–371.