Predictive model of COVID-19 incidence in Brazilian municipalities

Isadora C. R. Carneiro1#, Eloiza D. Ferreira1#, Janaina C. da Silva2#, Guilherme Soares1,3, Guilherme F. Silveira\*1

1. Instituto Carlos Chagas - Fiocruz/PR, Curitiba, PR
2. Laboratório de Biologia de Tumores, Universidade Estadual do Oeste do Paraná, UNIOESTE, Francisco Beltrão-PR.
3. Laboratórios De Referência Em Viroses Emergentes E Reemergentes, Instituto Carlos Chagas - Fiocruz/PR, Curitiba, PR

\* Corresponding author: gfsilveira@gmail.com, guilherme.silveira@fiocruz.br

# These authors contributed equally to the work.

ABSTRACT:

Keywords: COVID-19, COVID-19 in Brazil, Exploratory data analysis, ARIMA, computational model.

INTRODUCTION:

M&M:

M&M – Correção.

Utilizando o conceito de Estudo ecológico, no presente trabalho foi utilizado esse método para compreender como se dá a correlação entre distribuição populacional por municípios brasileiros e a taxa de contaminação de COVID-19 por área geográfica. Considerando as 486 cidades presentes em um banco de dados alimentado diariamente, que tiveram confirmação do novo Coronavírus, sua área territorial, densidade demográfica, idade média entre os habitantes, gênero, dados socioeconômicos e IDHM (Índice de Desenvolvimento Humano Municipal).

Para a análise exploratória de dados (EDA) e o desenvolvimento de soluções computacionais, foi utilizada a linguagem de programação Python, que permite o uso de diversas bibliotecas, específicas para esse fim. Os conjuntos de dados foram trabalhados em um ambiente Anaconda (IDE Jupyter Notebook), plataformas amplamente aplicadas no campo da ciência de dados, oferecendo aos usuários ferramentas e bibliotecas robustas e estabelecidas. Junto da linguagem Python, foi necessário importar diferentes pacotes e bibliotecas, como *Pandas,* *Numpy e Scipy*, utilizadas para organização e estruturação dos dados. Para cálculos estatísticos, foi importado *Statsmodels (statsmodels.tsa.seasonal e* sm.tsa.statespace.SARIMAX)*,* um modulo usado em Python utilizado para gerar as séries temporais e a classe ARIMA para as previsões. M*atchplotlib e Seaborn,* utilizadas para gerar gráficos de duas dimensões (2D). O projeto pode ser acessado através da página do github, <https://github.com/gfsilveira/covid>.

Base de dados: Os registros de casos confirmados do COVID-19 em nível municipal foram obtidos por meio de um conjunto de informações diários dos Departamentos de Saúde das Unidades Federativas compiladas por Álvaro Justen e seus colaboradores disponíveis em <https://brasil.io/dataset/covid19/caso>. Características demográficas e socioeconômicas disponíveis publicamente a nível municipal, como densidade demográfica, Índice Municipal de Desenvolvimento Humano (IDHM) e área total em km² (…) foram obtidas do Instituto Brasileiro de Geografia e Estatística (IBGE) a partir do censo demográfico realizado em 2010. Já os dados referentes da idade média entre os habitantes e gênero, foram obtidas a partir do censo demográfico realizado em 2015.

Correlação: Para os testes de correlação de Spearman, diferentes bibliotecas foram usadas para análise de dados, como pandas, numpy e para geração de gráficos, as bibliotecas seaborn e matplotlib. Os diferentes bancos de dados foram importados para python, usando a biblioteca pandas, no formato de estrutura de dados DataFrame, que continha informações municipais como: número de habitantes na população, casos confirmados, óbitos, casos confirmados para 100 mil habitantes e proporções confirmadas e óbitos , área oficial do município em km², habitantes em dados de densidade populacional, densidade populacional e Índice de Desenvolvimento Humano Municipal (IDHM). (ESPERANDO NOVOS RESULTADOS A PARTIR DAS SUGESTÕES)Medidas utilizadas para descrever o perfil da taxa de incidência de COVID-19: O perfil de incidência foi considerada através da correlação entre municípios de maior população com o número de casos confirmados por covid-19. DESCREVER APOS RESULTADOS

Séries Temporais: A Série Temporal foi gerada a partir do banco de dados de casos confirmados por COVID-19. O período analisado é de 25 de janeiro até o dia 22 de abril de 2020, demonstrando o número de casos em municípios brasileiros a partir de dados diários, verificando a tendência, sazonalidade dos dados e apresentando o ruído que não foi incorporada a série.

Modelo (S)ARIMA : Para o desenvolvimento do modelo preditivo, usamos os modelos ARIMA (*Autoregressive Integrated Moving Average*) que utiliza os parâmetros p, q e d, em que p representa o número de termos auto-regressivos, q o número da média móvel e d o número de diferenças não sazonais, porém, não obtivemos bons resultados ao testar diferentes parâmetros, portanto, foi incorporado a Sazonalidade da série temporal (SAMIRA), que adiciona novos três hiperparâmetros: SAMIRA(p,d,p)x(P,D,Q)m em que m representa número de etapas de tempo para um único período sazonal. Foi gerado uma lista de diferentes parâmetros e filtrado o de menor valor AIC (Critério de Informação Akaike), que nos resultou o melhor ajuste de AIC = 397.846 e o valor de P>|Z| (..descrever..) foi igual a zero (0.00). O parâmetro utilizado com melhor ajuste foi SARIMA (2,2,2)x(1,0,1)30 em que 30 representa o intervalo de dias. Após determinar os parâmetros, foi testado os ajustes do modelo (Figura 01). O modelo de previsão está ajustado dentro do período analisado para gerar a previsão para os próximos meses foi o mesmo para gerar nossa série temporal. Já o período reservado para avaliar a previsão com IC95% foi a partir do dia 23 de macho a 20 de abril de 2020.



**Figure 01. Parcelas Residuais para testar os parâmetros utilizados**. (A) Gráfico de Resíduos: os erros residuais estão próximos de uma média de zero (linha). (B) Gráfico de Histograma: este gráfico de densidade sugere distribuição normal dos dados a partir do modelo ajustado. (C) Gráfico QQ Normal: os pontos (quantil) estão alinhados com a linha vermelha, testando o modelo. (D) Gráfico ACF: apresenta os erros residuais que não são correlacionados automaticamente. Testando o parâmetro d (número de diferenças não sazonais que foi utilizado no modelo). (E) Ajuste da previsão e dos dados que foram observados.

RESULTS:

1. The demographic density (hab/km²), the MHDI and the per capita income of the municipalities with cases of COVID-19 are above the national average.

Since COVID-19 is a pathology caused by coronavirus that is transmitted directly from human to human, we seek to understand whether the characteristics of the affected cities can help in understanding the pandemic. The analyzed database has 486 municipalities (8.7% of the cities in Brazil) of the 26 states of the federation, plus the Federal District, which until April 22, 2020 had at least 1 confirmed infection, totaling 44,397 ( 0.02% of the Brazilian population) cases of COVID-19. COVID-19 cases are reported in all states, where up to 35% of the municipalities have reports (Figure 02).

|  |
| --- |
| Texto preto sobre fundo branco  Descrição gerada automaticamente |
| **Figure 02. All Brazilian states have confirmed cases of COVID-19.** (A) Distribution of confirmed cases of COVID-19 by state. Percentage of cities with at least 1 case of positive infection. |

Seeking to understand whether the characteristics of the affected municipalities can influence viral transmission, characteristics such as the age of the inhabitants, declared gender, area (km²), demographic density (hab/km²), Municipal Human Development Index (MHDI) and per capita income. In the analysis of age, Brazilian municipalities have a relative homogeneity in the distribution range every 5 years of life, up to 35 years, when there is a gradual reduction in the numbers (Figure 2A). In the cities that present cases of COVID-19, the distribution of age groups is the same as the general distribution in Brazil (Figure 2B). In the present study, the age ranges of the population in the affected cities were grouped (Figure 2C), in order to obtain elderly patients over 65 and adults (who represent a greater likelihood of comorbidity) between 30 and 64 years, with a higher risk, and children aged 0 to 9, adolescents aged 10 to 19 and young adults aged 20 to 29, with lower risk. The percentages of each age group (Figure 2D) will be used for the analysis of correlation with the COVID-19 case rate. It is possible to observe that there is no difference between the age distribution in the total Brazilian municipalities (Figure 3A) and in the affected municipalities (Figure 3B).

|  |
| --- |
| Uma imagem contendo screenshot, texto  Descrição gerada automaticamente |
| **Figure 03. There is no significant difference between the age distribution comparing cities with cases of COVID-19 and all country.** (A) Number of inhabitants by age group (5 years) in Brazil. (B) Number of inhabitants by age group in cities with cases of COVID-19. Stratification of age groups by age in the municipalities studied, represented by number of inhabitants, absolute (C) and percentage for each city (D). |

Another demographic information analyzed was the declared gender of the inhabitants. In Brazil, the percentage distribution is 51.7% for women and 48.3% for men. In cities with cases of COVID-19, we observed a similar distribution, with 50.29% (4,618,650) women and 49.71% (4,581,800) men (Figure 4).

|  |
| --- |
| Uma imagem contendo screenshot, texto  Descrição gerada automaticamente |
| **Figure 04. Declared gender of inhabitants in cities with cases of COVID-19 is similar to average in Brazil.** (A) Number and (B) percentage of men and women living in cities that present cases of COVID-19. |

Since SARS-Cov-2 are transmitted from person-to-person, we look for characteristics that can influence the displacement and agglomeration of people, such as area (km²) and demographic density (hab / km²) in the municipalities. As expected, in cities with COVID-19 reports, we observed that the demographic density (hab/km²) for different quantiles, 0.39 to 55.085 (Figure 5A); 55.085 to 167.315 (Figure 5B); 167,315 to 602,5475 (Figure 5C); 602.5475 to 13024.56 (Figure 5D), obeys the relationship between number of inhabitants (population) and area (km²). The cities analyzed, therefore, are more densely populated than the national average, which is 23.8 hab/km².

|  |
| --- |
| Mapa colorido com texto preto sobre fundo branco  Descrição gerada automaticamente |
| **Figure 05. The demographic density in cities with cases of COVID-19 is above the national average.** Number of inhabitants (Population) by area (km²) resulting in demographic density (hab / km²) in four quantiles, 0.39 to 55.085 (A); 55.085 to 167.315 (B); 167.315 to 602.5475 (C); 602.5475 to 13024.56 (D). |

The Municipal Human Development Index (MHDI) is related to three main factors, health, education and income. Where: healthy and long life (measured by life expectancy); education, access to knowledge (calculated using the average schooling of adults and the expected years of schooling for children of school age); and income, standard of living (measured by Gross Domestic Income per capita). The global Brazilian HDI for 2013 was 0.744, the 79th position in the world ranking among the 187 countries and territories recognized by the United Nations. In the Global HDI for HDR 2014, the three dimensions have the same weight and the human development ranges are fixed, being: Low Human Development - less than 0.550; Average - 0.550 and 0.699; High - 0.700 and 0.799; and Very High - above 0.800. The per capita income, number of minimum wages [R $ 975.00] per month per capita for formal workers, in Brazil in 2017 (last year measured), was 1.48 (R$ 1,443.10). In this context, the MHDI was compared to per capita income, for the different population ranges, in cities with cases of COVID-19 (Figure 06).

|  |
| --- |
| Texto preto sobre fundo branco  Descrição gerada automaticamente |
| **Figure 06. The Municipal Human Development Index (MHDI) and per capita income in cities with cases of COVID-19, are above the Brazilian average.** (A) MHDI and income (number of minimum wages [R $ 975.00] per month for formal workers) in cities with cases of COVID-19. |

The results show that for both MHDI and per capita income, most cities with positive cases for COVID-19 are above the national average. With the set of data exposed, we seek to determine whether there is a correlation between the indexes analyzed and the numbers of confirmed cases of COVID-19 for every 100,000 inhabitants (Confirmed-100k) (Figure 07).

|  |
| --- |
| Uma imagem contendo texto, mapa, chuva  Descrição gerada automaticamente |
| **Figure 07. There is no correlation between the demographic indices analyzed and the number of confirmed cases / 100k inhabitants and the death rate, in cities with cases of COVID-19.** Spearman correlation in cities with cases of COVID-19. Confirmed cases / 100,000 inhabitants (A to J) and death rate / 100,000 inhabitants (K to T) in relation to demographic density (hab / km²) (A and K), MHDI (B and L), per capita income (C and M),% of men (D and N),% of women (E and O),% of children (F and P)% of adolescents (G and Q)% of young adults (H and R)% of adults (I and S)% of elderly (J and T). Since correlation models are strongly influenced by extreme values, the 3 municipalities with more than 200 confirmed cases for each 100k inhabitants were removed from this analysis. |

Once determined some of the characteristics of the cities that present cases of COVID-19, using the time series of confirmed cases, deaths and the confirmed index for 100k inhabitants, we seek to determine a model for predicting the infection.

1. Until 05/10/2020, the evolution model allows predicting 56,829 to 70,447 confirmed cases.

The results presented suggest that municipal characteristics should be considered regarding the current epidemiological condition. However, due to the current level of infection in the cities analyzed, the scarcity of data does not allow the development of a robust predictive model for cases confirmed at the municipal level. Seeking to understand the condition of the infection at the national level, we analyzed the time series of accumulated data for confirmed cases (blue), deaths (orange) and confirmed for 100k inhabitants (green) from February 25, 2020 to 04/10/2020 (Figure 8A). The decomposition of the time series allows to perceive a clear tendency of increase in the number of confirmed cases and deaths, however, still a stationary condition for confirmed for 100k inhabitants (Figure 8B). As for seasonality (Figure 8C) and the random component (Figure 8D), the variation in the number of confirmed cases is significantly greater than the other data analyzed, as expected.

|  |
| --- |
|  |
| **Figure 08. The number of confirmed cases and ratio of confirmed / 100k inhabitants shows a clear upward trend**. Decomposition of the time series of the daily values of number of confirmed cases (blue), deaths (orange) and ratio of confirmed / 100k inhabitants (green), in components (A) raw data, (B) trends, (C) seasonality and (D) randomness. |

There was a clear upward trend in the number of cases. In order to determine a computational model to predict the evolution of COVID-19 in Brazil, we use computational modeling in the time series. The best adjusted model for the forecast was seasonal ARIMA, where we reached the forecast of 56,829 to 70,447, up to 05/10/2020, with 95% confidence (Figure 9).

|  |
| --- |
|  |
| **Figure 09. Average estimate of 63,638 confirmed cases in 30 days**. SARIMA model of forecast of confirmed cases until 05/10/2020. Confirmed cases (blue), forecast (orange), model fit analysis (green) and forecast interval with 95% confidence (gray). Up to the end date, between 56,829 and 70,447 cases are expected. |

CONCLUSION:

Together, the analyzed data from 486 cities with at least 1 case of COVID-19 until April 22, 2020 shows that the average age group of the inhabitants remains the same as the average age group of the Brazilian population, with the most populous cities of Brazil then included. The average of women and men in the cities studied also agrees with the national average. The demographic density, the MHDI and the per capita income of the municipalities with cases of COVID-19 are above the national average. However, there is no correlation between the indexes analyzed and the numbers of confirmed CONVID-19 cases or deaths per 100,000 inhabitants. In addition, if conditions are maintained, our model predicts 5000 to 7000 cases by June 10, 2020.

DISCUSSION:

ACKNOWLEDGMENTS:

We would like to thank the Creative Commons Attribution ShareAlike, and Álvaro Justen and collaborators from Brasil.IO for confirmed case data base. Scholarship funding agencies, CAPES and Fiocruz.

BIBLIOGRAPHY: