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:

In the present work, the concept of Ecological Study was used, this method of epidemiological study helps us to understand how there is a relation between characteristics of the municipalities in Brazil and the COVID-19 contamination and death rate. Considering the 486 cities present in a database fed daily, which had confirmation of the SARS-Cov-2, in addition, we seek to understand the territorial area, demographic density, average age among inhabitants, gender, socioeconomic data and MHDI (Municipal Human Development Index) ) through a database collected for the same municipalities.

For the exploratory data analysis (EDA) and the development of computational solutions, the Python programming language was used, which allows the use of several libraries, specific for this purpose. The data sets were worked in an Anaconda environment (IDE Jupyter Notebook), platforms widely applied in the field of data science, offering users robust and established tools and libraries. Along with the Python language, it was necessary to import different packages and libraries, the most used ones being: Pandas, Numpy and Scipy, with the function of organizing and structuring the data. For statistical calculations, Statsmodels, were imported to analyze the Time Series and the ARIMA/SARIMA models for the forecasts. Matchplotlib and Seaborn, used to generate two-dimensional (2D) graphics. The project can be accessed through the github page, <https://github.com/gfsilveira/covid>.

The records of COVID-19 at the municipal level were obtained through a set of daily information from the Health Departments of the Federative Units compiled by Álvaro Justen and his collaborators until April 22, 2020, available at https://brasil.io/dataset/covid19/caso. Demographic and socioeconomic characteristics publicly available at the municipal level, such as population density, Municipal Human Development Index (MHDI), total area in km², per capita income were obtained from the Brazilian Institute of Geography and Statistics (IBGE) from the demographic census conducted in 2010. data on the average age between inhabitants and gender were obtained from the 2015 census.

A data structure, containing different age groups between 0 and over 80 years old from the IBGE 2015 census database was used. To compare the incidence rate of contamination by SARS-Cov-2, age of people resident in Brazilian municipalities was separated into groups of children (0 to 9 years), teenager (10 to 19 years), yang adults (20 to 29 years), adults (30 to 64 years) and the seniors (65+ years). Declared gender data was also considered, comparing the incidence rate of COVID-19 among men and women residents. In addition to data from the population, the number of cities with COVID-19 cases within each state was analyzed and the percentage of Brazilian municipalities / states of confirmed cases was calculated by April 22, 2020.

For Spearman's correlation tests, data from the IBGE 2010 and 2015 demographic census database was compared with data from municipalities with COVID-19 cases and deaths rate. Correlations were generated using stats.spearmanr imported from the scipy library. To generate the graph, sns.jointplot was imported from the seaborn library. Correlations and graphs were generated between the Demographic Density (hab/km²), Municipal Human Development Index (MHDI), socioeconomic data (per capita income), the different age groups, men and women, each analyzed in isolation with cases confirmed /100k and deaths/population ratio.

The Time Series was generated from the database of confirmed cases by COVID-19. The period analyzed is from February 25, 2020 to April 22, 2020, demonstrating the number of cases in Brazilian municipalities from daily data, verifying the trend, seasonality of the data and presenting the noise that was not incorporated into the series. To generate the decomposition of the Time Series, it was necessary to import the Statsmodels module (statsmodels.tsa.seasonal).

For the development of the predictive model, being imported sm.tsa.statespace.SARIMAX, through the Statemodels, we use the Autoregressive Integrated Moving Average with Seasonality (SARIMA) model that use the parameters p, q and d, where p represents the number of autoregressive terms, q the number of the moving average and the number non-seasonal differences, adds three hyperparameters: SAMIRA (p , d, p) x (P, D, Q) m where m represents the number of time steps for a single seasonal period. All different parameters between (0,0,0)x(0,0,0)30 and (2,2,2)x(2,2,2)30 was tested and the one with the lowest Akaike Information Criterion (AIC) was used. The best adjustment was an AIC = 397,846 with a P>|z| lower than 0.08 for all parameters. The best configuration was a SARIMA (2,2,2) x (1,0,1) 30. The model's adjustments were tested (Figure 01). The forecast model is adjusted within the analyzed period to generate the forecast for the next months, it was the same to generate our time series.

|  |
| --- |
|  |
| **Figure 01. Residual Plots to test the parameters used.** (A) Residual Graph: residual errors are close to an average of zero (line). (B) Histogram graph: this density graph suggests normal distribution of the data from the adjusted model. (C) Normal QQ graph: the points (quantile) are aligned with the red line, testing the model. (D) ACF Graph: shows residual errors that are not automatically correlated. Testing parameter d (number of non-seasonal differences that was used in the model). (E) Adjustment of forecast and data that have been observed. |

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 SARS-Cov-2 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 April 22, 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 April 22, 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 427,765 to 496,005, up to April 22, 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 427,765 to 496,005, up to April 22, 2020.

DISCUSSION:

In this work, we studied the recent occurrence of severe acute respiratory syndrome, a respiratory disease caused by coronavirus SARS-Cov-2, originated in the city of Wuhan, China and the correlation of transmission and death rates, through confirmed cases, with Brazilian demographic and socioeconomic data. Having knowledge of the demographic distribution of the population and socioeconomic situation becomes significantly when it is aimed to compare transmission and death rates across the country.

Brazil currently has 5,570 municipalities distributed disproportionately over a total area of 8,511,000 km² that has 26 states (IBGE, 2019). As previously described, in the results section, municipalities with a population larger than 295,955 inhabitants showed a positive correlation between the size of the population and the number of confirmed cases of the disease. However, of this total of municipalities, only 95 have a population of over 295,955 inhabitants, including all state capitals and the Federal District. As a percentage, these 95 municipalities represent 1.71% of the country's total cities, with the majority of Brazilian cities, 94.22% (5,245 municipalities), having a population less than or equal to 100,000 inhabitants (IBGE, 2019).

Therefore, based on the results of this article, it is possible to note that the transmission of the disease is more likely to impact in less than 2% of Brazilian municipalities. However, it is essential to reinforce that these 95 cities, together, are home to 83,951,535 inhabitants, which represents 40% of the total population of Brazil, which currently has a demographic density of 209.5 million people (IBGE, 2018). Thus, neglecting the recommendations of the World Health Organization about isolation and social conduct, in a time of pandemic, will be an attack on public health in Brazil.

When analyzing the data in relation to the states of the Federation, it is noted that it follows the same line obtained in relation to the municipalities, with the state of São Paulo being the most populous in the country and in turn the state with the highest number of confirmed cases of COVID -19, 11,043 cases to date, according to the Ministry of Health (2019). Through data studies described in this article, it is also possible to observe, through a more refined analysis, that the new measures of conduct and social isolation employed in the country have been the main form of controlling the pandemic. It is important to note that the actual numbers of cases and deaths from the disease may be different from the official data, taking into account the impact that the delay in notifications has on the estimates and also the fact notifications depend on hospitalization. (FIOCRUZ, 2020).

When we take into account the data obtained, it is possible to say that yes, social isolation is a valid measure to be applied in municipalities that have a resident population larger than 295,000 inhabitants and that in these municipalities, the more intensified the measure, flatter the transmission curve become, thus enabling hospitals and health units to have greater control under the demand of patients who inspire specialized care.

However, it is important to mention that it is not possible to conclude the real importance of social isolation in municipalities with a population below the aforementioned number, however, according to the results, there is a negative correlation between demographic density and the number of cases in these cities, which, in theory, would indicate a lack of connection between these aspects.

According to Hellewell et al. (2012), in a study carried out to evaluate measures to contain the transmission of the disease, social isolation is insufficient to control the outbreak, requiring new interventions to achieve control of the transmission of the disease. However, isolation can contribute to reducing the overall size of an outbreak over a longer period of time (Hellewell et al., 2012). Taking this into account, it becomes possible to assess the importance of measures of social isolation, even for the municipalities with a small population, demonstrating the great importance of such measures, but which should be intensified in the most populous cities and not neglected in the cities with less than 295,000 inhabitants. Therefore, it is interesting to evaluate the average traffic of the Brazilian population mainly in the forms of essential workers, as truck drives, with the supply of basic needs items, primarily carried out in Brazil, through land transportation, often on long distance journeys, between the capitals and other municipalities of the federation, with an average, during the month of march, of 1000 trucks arriving at the Supply Center of the Federal District every Monday and Thursday (CEASA, 2020).

It would be interesting, as a future perspective, to deepen the studies in order to develop a system of equations that would indicate the existence or not of a proportional factor of relationship between population vs the lethality rate, trying to correlate how many times the lethality of a more populous city it is larger than that of a less populous city. In order to show that as you double or triple the number of inhabitants in a city, lethality does not necessarily double or triple, indicating that there is no linearity in the cases.

Therefore, the information reported in this study, allows us to highlight that cities with a higher number of inhabitants, who choose not to comply with social isolation, have a higher risk and probability of infection. However, it is essential to show that contamination by the SARS-Cov-2 is not due to the simple fact of living in a more or less populous city, and that there are different regional characteristics, both geographic and socioeconomic that can influence dispersion, not only for SARS-CoV-2 virus, but for many other pathogens (Mogi and Spijker, 2020; Dowd et al., 2020).

Thus, it is worth emphasizing once again that social isolation has proven to be a efficient measure to control contamination in the most populous cities, which shows that such an approach should indeed not just be stimulated but also amplified, as an option to contain the pandemic, together with other protective measures such as masks and gloves, following the recommendation of use by the World Health Organization, in order to avoid hospital demand above the service capacity, since the saturation of the hospital network, is already being reported by some states, which also will have an impact over the data and the rate of reported cases as well as on lethality rate (FIOCRUZ, 2020).

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: