Brazilian Journal of Experimental Design, Data Analysis and Inferential Statistics

BJEDIS

Special Edition **Received mm/dd/yyyy. Revised mm/dd/yyyy. Accepted mm/dd/yyyy. Published mm/dd/yyyy.**

Title:

VALIDATION METHOD OF THE MATHEMATICAL MODEL FOR SARS-Cov-2 PANDEMIC FROM DATA MINING AND STATISTICAL ANALYSIS.

Rafael Pereira Santana[1], Anderson Lupo Nunes[1] and Pedro Maia Salomone[1]
rafael.santana@ifrj.edu.br*[1]

[1]Physics Research and Physics Teaching Group, Physics Laboratory Nilton de Souza Medeiros, Federal Institute of Education, Science and Technology of Rio de Janeiro, Campus Duque de Caxias, RJ, Brazil.

Abstract: Nowadays, in 2020, we live the most dangerous global pandemic that has been reported since the Spanish Flu, which occurred between 1918 and 1920. According to World Health Organization (WHO) records, the pandemic caused by the Sars-Cov-2 virus began in December 2019 and is present in all continents and almost all countries, surpassing more than 79 million infected and 1.7 million deaths by December 2020. Several mathematical models applied to Epidemiology have been adopted over time. One of the most widely adopted is the Susceptible-Infected-Recovered (SIR), developed in India by Kermack and Mckendrick in 1927. In our research, a comprehensive collection of data on the SARS-Cov-2 pandemic was made from reports by WHO, Dadax Limited (Chinese data company), and Johns Hopkins University in the United States of America (USA). Facts were collected from many different countries, regarding the number of confirmed, recovered, and death cases. In this article, we constructed a mathematical model that describes the evolution of the pandemic from the similarity with models already adopted in the field of nuclear physics. For the validation of the mathematical model, we chose information from Germany due to the reliability of the available information. Thus, a statistical analysis was executed to qualify the performance of the method and the predictive character of the mathematical model. To date, 11,716 raw data have been collected, of which we performed data mining relevant in order to use in this research.

Keywords: Statistical analysis; Data mining; Epidemiology; Mathematical models; SARS-Cov-2; New coronavirus; Pandemic; Differential equations.

Adherence to the BJEDIS' scope: In this written work, we used mathematical modeling to analyze a very large number of data related to the contamination and spread of a highly contagious virus in human populations. We used spreadsheets for mining relevant data and producing new data to recognize trends and make forecasts. All of us believe that our work can contribute in a relevant way to both the journal and the scientific community, informing, updating researchers, and strengthening the sciences focused on the evolution of epidemics.

*Address correspondence to this author at the Physics Laboratory Nilton de Souza Medeiros, Campus Duque de Caxias, Federal Institute of Education, Science and Technology of Rio de Janeiro, CEP: 25.050-100, Duque de Caxias, RJ, Brazil; Tel: ++55(21)99721-2726; E-mails: rafael.santana@ifrj.edu.br

1. INTRODUCTION

In December 2019, an outbreak of pneumonia of unknown cause was reported by health authorities in Wuhan (China). Laboratory research results identified a new coronavirus as responsible for the outbreak. The new coronavirus was named by the International Committee on Virus Taxonomy (ICTV) as coronavirus 2, SARS-Cov-2. After that, the World Health Organization (WHO) called COVID-19 the disease caused by SARS-Cov-2 (CHAVES, 2020) [1].

The most recent pandemic, which has had greater impacts than the current SARS-COV-2, was the Spanish Flu, it occurred between January 1918 and December 1920. It is estimated that there were 500 million infected, 50 million deaths, and attained, through the first two waves in 1918, the United States, Europe, Asia, Central and South America. According to Goulart (2005) [2], the first news about an epidemic in Spain reached Rio de Janeiro´s newspapers in August 1918. The nickname "Spanish Flu" came because, in this country, information about the epidemic was widely disseminated in the press, unlike other countries where there was an attempt to cover up the disease. This first information about that pandemic was received by the Fluminense population with disdain and disbelief (GOULART, 2020) [2]. Both in previous and current pandemics, this incredulity may hinder the performance of effective procedures to control the spread and contamination of diseases through viruses.

According to the 12/29/2020 report of the World Health Organization (WHO, 2020) [3], by December 2020, the largest number of dead and infected with SARS-Cov-2, is the USA, with 18,648,989 infected and 328,014 dead. Second is Brazil, with 7,448,560 infected and 190,488 dead. The seriousness of the health situation in our country would be enough to justify research on the spread of the virus.

Luiz (2012) [4] defines the word epidemic as an infectious disease that spreads in large proportions, with a huge number of deaths in a very short time. The mathematical modeling in the field of epidemiology is done through the study of equations, usually ordinary differential equations, which describe the interaction between the population and the environment, resulting in a detailed analysis of the infectious illness.

A mathematical model to describe the evolution of the pandemic was used in this research article. Consequently, we can highlight the impacts that the spread of the SARS-Cov-2 virus has imposed on our planet, due to the immense number of deaths and the devastating reflections on people's routine, as well as on the global economy (COSTA, 2020) [5] and (MACEDO, 2020) [6].

This analysis is indispensable, not only because of the topicality and relevance of the subject but also for the possibility of benefits arising from the construction and improvement of mathematical models applied to epidemiology, that may be used in future epidemics. For this reason, the more knowledge about the disease and how it spreads, the greater effective methods adopted to prevent its transmission and the study of preventive actions, when available.

As a result of the vast amount of information, we need to mine the relevant data from which we use mathematical modeling. From this modeling, equations were added to a spreadsheet of calculations to produce new data for future statistical analysis to recognize trends and make provisions.

As BJEDIS is a journal aiming to carry out the broad promulgation of research related to analysis and data mining, we believe our work can contribute in a relevant way to both the journal and the community, informing and updating researchers and strengthening the sciences focused on the study of epidemics´ evolution using mathematical modeling and data mining.

1.1. Mathematical Model (SIR)

Several mathematical models applied in epidemiology field have been adopted over time in the scientific literature. Cristovão (2015) [7], Okhuese (2020) [8], Costa (2020) [9], Ciufolini (2020) [10], and Fokas (2020) [11] are just some examples of more recent mathematical models that are used to describe any epidemic or, more specifically, the current SARS-Cov-2 pandemic.

One of the most extensively adopted mathematical models in epidemiology is the SIR (Susceptible, Infected and Recovered). Developed in India by Kermack and McKendrick (1927) [12], it is the first successful epidemiological model. On the diseases that spread in the population, he proposed a division into disjoint classes, denoted by:

a) Susceptible, $\boldsymbol{S}(\boldsymbol{t})$, representing the class of healthy individuals, that is, those who are exposed to possible infection.

b) Infected, $\boldsymbol{I}(\boldsymbol{t})$, representing the class of individuals who are infected and who are likely to cause new infections, i.e., infectious individuals.

c) Recovered, $\boldsymbol{R}(\boldsymbol{t})$, representing individuals recovered from diseases, thus becoming immune to a new infection.

According to Oliveira (2018) [13], his postulates are:

i. Every person who is part of the population and has not yet been infected is Susceptible, $\boldsymbol{S(t)}$.
ii. When a susceptible contract the disease, it becomes an Infected, $\boldsymbol{I(t)}$.
iii. The individual who evolves to the cure or who dies becomes a Removed, $\boldsymbol{R(t)}$.
iv. The birth and mortality rates are equal, which implies that the total population is constant.
v. The total population, $\boldsymbol{N(t)}$, is the sum of the Susceptible, Infected and Removed.

$$\boldsymbol{N(t) = S(t) + I(t) + R(t)} \qquad \text{(eq.1)}$$

Then, in agreement with Cristóvão (2015) [7] and Oliveira (2018) [13], the equations that relate the time evolution of the number of susceptible, infected, and removed are:

$$\boldsymbol{\frac{dS}{dT} = -\beta \cdot S(t) \cdot I(t)} \qquad \text{(eq.2)}$$

$$\boldsymbol{\frac{dI}{dT} = \beta \cdot S(t) \cdot I(t) - \gamma \cdot I(t)} \qquad \text{(eq.3)}$$

$$\boldsymbol{\frac{dR}{dT} = \beta \cdot S(t) \cdot I(t) - \gamma \cdot I(t)} \qquad \text{(eq.4)}$$

In which equation (eq.2) corresponds to the rate's variation of susceptible people´s number, equation (eq.3) corresponds to the rate´s variation of the infected number, and equation (eq.4) corresponds to the rate´s variation of the removed number.

The coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ correspond respectively to:

$\boldsymbol{\beta}$ → Transmission coefficient
$\boldsymbol{\gamma}$ → Recovery rate

We should note that this model can be greatly enhanced from data more consistent with our reality.

The Spanish flu was caused by the influenza virus and has become a disease that experts call endemic, to wit, it remains an infectious disease that affects a large number of individuals, but with a very low lethality, without ever ceasing to exist (GOULART, 2005) [2].

There is a high probability that the SARS-Cov-2 pandemic will also become endemic. Therefore, the use of these mathematical models may continue to be relevant for a long time. Until an effective vaccine has been developed to prevent the evolution of COVID-19 disease in its most lethal form, the only way to protect individuals is social isolation combined with a lot of information about the pandemic, just like the WHO (2020) [3] recommends.

The mathematical models in epidemiology bring a solid knowledge about the contagion and allow us to make predictions of future scenarios of the epidemiological situation in a population, imposing preventive actions, increasing or decreasing the social distance. This kind of article is perfectly justifiable and essential because of the theme´s relevance. That is why there are many publications related to mathematical modeling for the study of the new coronavirus pandemic´s advances, such as Cruz (2021) [14], Neto (2021) [15], Kamrujjaman (2020) [16], and

Mallapaty (2020) [17] of recognized relevance. In this way, we will present a statistical mathematical model that, using spreadsheets, can be another contribution towards the same objective.

1.2. Proposed mathematical model

The model proposed in this article is inspired by the theory of point kinetics of a nuclear reactor. This theory is described in some literature projects, such as Duderstadt (1987) [18], Hertrick (1971) [19], Akcasu (1971) [20], Nunes (2015) [21] and Palma (2016) [22]. It is a study in the temporal variation of the neutron concentration in the nucleus of the nuclear reactor, a fundamental parameter and directly proportional to the generated power and the internal temperatures of the reactor. Table 1 is comparing the point kinetic model of a nuclear reactor and the mathematical model for the SARS-Cov-2 pandemic. It aims to show similarities between the models, justifying their use.

Table 1: Comparison between two models: point kinetic and Sars-Cov-2

Point Kinetics of Nuclear Physics	Mathematical Model for SARS-Cov-2
For nuclear fission to occur, the uranium core (U 235) absorbs a neutron.	For SARS-Cov-2 disease to occur, the individual has to be contaminated by the virus.
Each neutron is much smaller than the core of U235 in the ratio of 1/235.	Each virus is much smaller than the cell of the infected individual in the ratio of 125 mm to 30,000 mm, that is, in the ratio of 1/240.
By absorbing the neutron, the U235 core becomes unstable and can undergo nuclear fission by releasing 2 or 3 neutrons.	An infected individual contaminates on average between 2 and 3 individuals.
Without nuclear reactor control measures, such as control rods and the addition of neutron-absorbing reagents, the number of neutrons grows exponentially.	Without the measures of prevention, social isolation, use of masks, and constant cleaning of our hands, the number of infected increases exponentially.

The point kinetics model is about the long-run variation of neutron concentration in the nucleus of a nuclear reactor (DUDERSTADT, 1987) [18] and (PALMA, 2016) [22]. In this model, the location of the neutron is not pertinent, however, it is crucial to know whether the neutron is generated immediately after nuclear fission (the so-called ready neutrons) or generated with a relatively long time (the so-called delayed neutrons) after fission, due to the radiative decay of fission fragments. The point kinetic equations in a nuclear reactor have as its only variable the time and are presented below:

$$\frac{dn(t)}{dT} = \frac{[\rho(t)-\beta(t)]}{\Lambda} \cdot n(t) + \sum_{i=1}^{6} \lambda_i \cdot C_i(t) \qquad \text{(eq.5)}$$

$$\frac{dC_i(t)}{dT} = \frac{\beta_i}{\Lambda} \cdot n(t) - \lambda_i \cdot C_i(t) \qquad \text{(eq.6)}$$

In which $n(t)$ is the density of ready neutrons, id est, neutrons generated in the act of nuclear fission, and $C_i(t)$ represents the delayed neutrons, i.e, neutrons generated from the radioactive decay of fission products. They are also called precur neutrons.

Table 2 shows nuclear parameters that make us able to understand the factors that influence the operation of a nuclear reactor and its monitoring. These factors are reactivity, delayed neutron fraction, and radioactive decay constant.

Table 2: List of nuclear parameters of point kinetics and their respective descriptions.

Nuclear parameters	Description
$\boldsymbol{\rho(t)}$	Reactivity = parameter related to the multiplicative factor of the chain reaction, that is, how much the concentration of neutrons increases, decreases, or remains constant.
$\boldsymbol{\beta(t)}$	Delayed neutron fraction, considering all groups of precursors, in relation to total neutrons.
$\boldsymbol{i(t)}$	Delayed neutron fraction of each group of precursors in relation to total neutrons.
$\boldsymbol{\lambda_i}$	Radioactive decay Constant for each group of precursors.

1.3. Mathematical model foundation for the SARS-Cov-2 pandemic

Knowing that the mathematical model for a pandemic must have some similarities to the nuclear model, but it must also have its peculiarities, it is important to stress some premises that are necessary for model preparation:

a) The Susceptible group consists of the entire population of the country that has not yet been infected.

b) All Infected (Symptomatic or Asymptomatic) can transmit the virus to susceptible individuals.

c) All Symptomatic Infected are tested and make up the official data released by the WHO in the country chosen to validate the method.

d) None of the Asymptomatic Infected is tested and make up the official data released by the WHO in the country chosen to validate the method.

e) The person recovered from SARS-Cov-2 can no longer infect other individuals.

f) The person Recovered from SARS-Cov-2 can no longer be reinfected during the term of this study.

g) All Symptomatic Infected people who die have their cause of death determined by SARS-Cov-2.

h) All Asymptomatic Infected who die have their cause of death determined by causes unrelated to SARS-Cov-2. The mortality rate of this group is the same as the rest of the population that is not infected.

i)The Recovered ones are divided into two groups, namely: Recovered from the Symptomatic (who were Symptomatic Infected) and Recovered from the Asymptomatic (who were Asymptomatic Infected)

j) The total population of the country is considered to be variable during the pandemic.

From the fundamentals of the model, we present in Table 3 its variables and parameters.

Table 3: List of variables and parameters of the nuclear model for the SARS-Cov-2 pandemic.

Model´s variables	Model´s coefficients or parameters
$S(t) =$ Susceptible	$\rho(t) =$ Specific parameter of infection
$I_S(t) =$ Symptomatic Infected	$\alpha(t) =$ Birth rate of the country
$I_A(t) =$ Asymptomatic Infected	$\mu(t)$ = Mortality rate of the country
$R_S(t) =$ Recovered from the Symptomatic	$\beta(t)$ = Death fraction per minute due to SARS-Cov-2
$R_A(t) =$ Recovered from the Asymptomatic	$\Upsilon(t)$ = Immigration rate of the country
Total population = $N(t)$	$\lambda_S(t) =$ Fraction of recovered ones from the symptomatic group
----------	$\lambda_A(t)$ = Fraction of recovered ones from the Asymptomatic group

The model´s equations are:

$$\frac{dN(t)}{dt} = \alpha \cdot N(t) + \gamma \cdot N(t) - \mu \cdot [I_A(t) + R_A(t) + R_S(t) + S(t)] - \beta I_S(t) \quad \text{(eq.7)}$$

$$\frac{dS(t)}{dt} = \alpha \cdot N(t) + \gamma \cdot N(t) - \rho(t)\frac{I_S(t)}{N(t)}S(t) - \rho(t)\frac{I_A(t)}{N(t)}S(t) - \mu S(t) \quad \text{(eq.8)}$$

$$\frac{dI_S(t)}{dt} = \rho(t)\frac{I_S(t)}{N(t)}S(t) - \lambda_S R_S(t) - \beta I_S(t) \quad \text{(eq.9)}$$

$$\frac{dI_A(t)}{dt} = \rho(t)\frac{I_A(t)}{N(t)}S(t) - \lambda_A R_A(t) - \mu I_A(t) \quad \text{(eq.10)}$$

$$\frac{dR_S(t)}{dt} = \lambda_S R_S(t) - \mu R_S(t) \quad \text{(eq.11)}$$

$$\frac{dR_A(t)}{dt} = \lambda_A R_S(t) - \mu R_A(t) \quad \text{(eq.12)}$$

$$N(t) = S(t) + I_S(t) + I_A(t) + R_A(t) + R_S(t) \quad \text{(eq.13)}$$

2. METHODOLOGY

Although it is necessary to choose a region or country with relatively reliable data to perform the tests in order to validate the model, no country will have its data on the SARS-Cov-2 pandemic without any deviation or inaccuracy. Besides that, there is a natural underreporting stemming from the fact that a large percentage of those infected develop the disease of milder form, almost without symptoms, or even with a complete absence of symptoms. They are called patients asymptomatic. They will rarely discover that they have contracted the virus because without symptoms they will not take the test and will be out of the statistic. Even so, these individuals may be virus transmitting agents (MALLAPATY, 2020) [17].

The collection of large amounts of data from trustworthy sources is an important factor for the success of the mathematical model developed in this project. It is paramount to expand the number of countries investigated so that those political factors exposed can be considered negligible in the model´s final result.

Two parameters are key in this analysis of the reliability level of each country's official data. The first one is the number of tests carried out for every hundred thousand inhabitants of that territory. The lower this number, the less official data is recognized. The second parameter is the lethality rate of SARS-Cov-2 in that land (DONSIMONI, 2020) [23].

As a rule, the lethality rate should be the same in countries that are similar when it comes to data on the demographic distribution of risk groups and relevant environmental situations. If there is a similarity between populations, the lethality rate is expected to be alike. If one country has a higher lethality rate compared to the other analogous one, it may indicate that the number of infected people is greater than what the official data indicate.

It is important to point out that the overload and quality of the health system of any country will also influence its lethality rate, either by SARS-Cov-2 or because of any other disease. All things mentioned will determine an extremely judicious scientific methodology for the choice of countries, or country, that will be used to validate the mathematical model that is going to be developed.

Discrepancies between the mathematical model´s results and reliable official data will determine adjustments in the model itself. Accordingly, Germany was chosen for the validation of the nuclear model for the SARS-Cov-2 pandemic. The criteria for such a choice are in Table 4. The German study on the pestilence was developed by Donsimoni (2020) [23] and Barbarossa (2020) [24].

Table 4: Criteria that resulted in the choice of Germany as the place for the model´s validation

Country Choice Criteria for Model Validation
1) A large number of SARS-Cov-2 tests taken per million inhabitants.
2) Low mortality rate due to SARS-Cov-2
3) Great percentage of recovered
4) Long period of evolution of the pandemic in the country, coming close to normality.
5) High population density.
6) Large number of infected.
7) Occupancy rate of ICU beds below 95% throughout the period.

In Table 5 we have a sample of the data collected for Germany: (considering that the complete table has approximately 100 lines). All this comprehensive information on the pest in Germany came from the WHO (2020) [25], John Hopkins (2020) [26], and CountryMeter (2020) [27].

Table 5: Sampling of official data for Germany.

Date (2020)	Day	Confirmed cases	Infected	Deaths	Recovered individuals	Mortality rate
March 10	50	1112	1095	2	15	0,180%
March 20	60	8198	8001	20	177	0,244%
March 21	61	14138	13887	45	206	0,318%
April 05	76	85778	55739	1342	28697	1,565%
April 07	78	95391	57706	1607	36078	1,685%
April 15	86	127584	51733	3254	72597	2,550%
April 16	87	130450	49884	3569	76997	2,736%
April 28	99	156337	33027	5913	117397	3,782%
April 29	100	157641	31129	6115	120397	3,879%

From the equations of the nuclear model for SARS-Cov-2 and the complete set of collected data, it was possible to obtain the following function for the Symptomatic Infected:

$$\boldsymbol{I_S(t) = A \cdot e^{at^2+bt+c} + B \cdot sen(w_1 t) + C} \qquad \text{(eq.14)}$$

In which:

$$\boldsymbol{A = 0.8}, \quad \boldsymbol{B = 6}, \quad \boldsymbol{C = 4},$$
$$\boldsymbol{a = -8.87058 \times 10^{-6}}, \quad \boldsymbol{b = 0.03504}, \quad \boldsymbol{c = -23.37399} \quad \& \quad \boldsymbol{w_1 = 0.8}.$$

To obtain the above equation we used several approaches and basic techniques for solving differential equations, as it was found out that this function has no analytical solution. Notice that the above solution is a particular solution for equation 14 (eq.14) of the $\boldsymbol{I_S(t)}$ derivative, taking into account the initial and boundary conditions established by the results of table 5.

These same conditions let us obtain other parameters and variables. Therefore, all variables are obtained in accordance with the model. It is worth mentioning that all mathematical models, no matter which area of knowledge they might be applied, presuppose the use of simplifying hypotheses so that it is possible to identify the most relevant variables and parameters and find a solution for the problem. This solution is called analytic when we discover a function for each variable that satisfies the system of differential equations of the model.

If it is impossible or difficult to obtain an analytical solution, a numerical solution can be obtained using computational methods. The numerical solution reaches values that support the construction of tables and graphics that fully describe the variable´s growth, but without necessarily having a corresponding analytical function.

The numerical method used to determine the model´s coefficients was the finite difference method. A test of the data evolution per minute and hour was executed. As the difference between both results was less than 5%, for the sake of speed, we decided to consider the time interval between each iteration ($\boldsymbol{\Delta t}$) equal to one hour. This hypothesis is reasonable since our sample universe is approximately 100 days.

Consequently, the finite differences method was used to calculate the parameter of specific infection, the fraction of deaths per minute of COVID-19, the fraction of recovered from symptomatic and asymptomatic.

Calculations of Germany's mortality, population growth, immigration, and birth rates were obtained directly from Germany's real-time demographic data available on the Country Matters website (2020) [27].

As the amount of information collected and the data production by interpolation for this work was huge, it is necessary to have a specific methodology for the data´s treatment. Pursuant to Trafimow (2016) [28], the most common is for a researcher to calculate the group's averages and use the null hypothesis significance test procedure to draw conclusions about the populations from which the groups were taken from. In our project, it is possible to ensure that the constituent groups of the country´s populations under study are dynamic and their state is variable over time. It also admits the use of methodologies to eliminate or reduce sources of uncertainties in the process (SIMONSOHN, 2020) [29].

Vergura (2009) [30] describes a methodology for monitoring a photovoltaic plant in Italy through statistical inference. The quantity of real variables involved in this process resembles the evolution of the new coronavirus pandemic, which would allow a genuine statistical analysis in our work. On account of the difficulties in the current pandemic situation and the basic need for social isolation, this process of statistical analysis has been greatly reduced, as will be seen in the next section.

3. RESULTS

From the foregoing in previous sections and the complete data on the SARS-Cov-2 pandemic in Germany, we present the graphs in Figures (1) until (5).

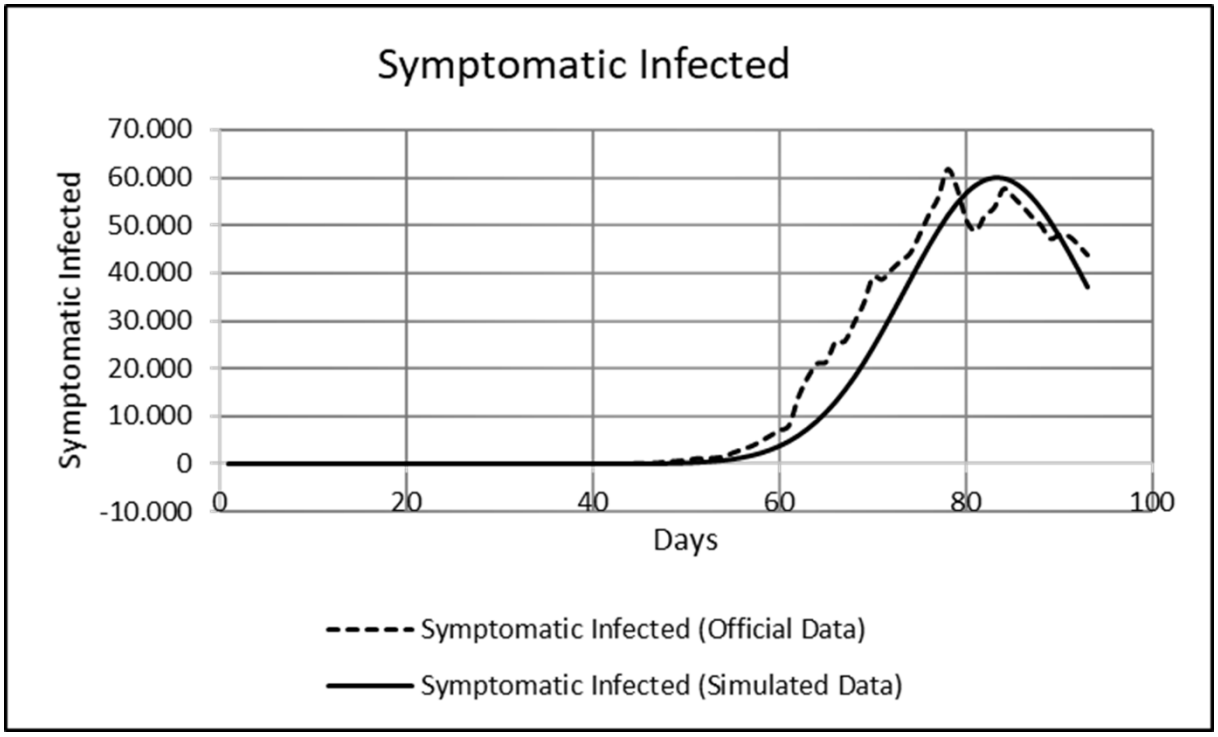

Figure 1. *Symptomatic infected. Official data in relation to the results obtained in the equation (eq. 14)*

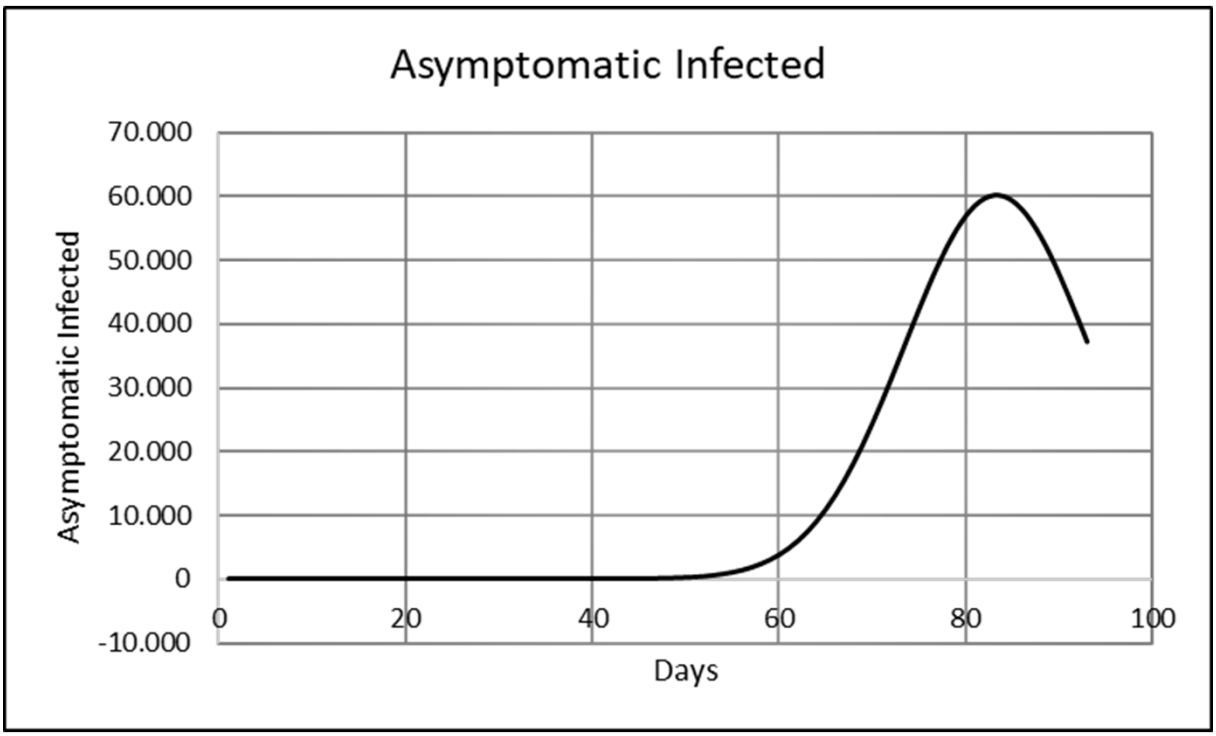

Figure 2. *Asymptomatic infected obtained through the model.*

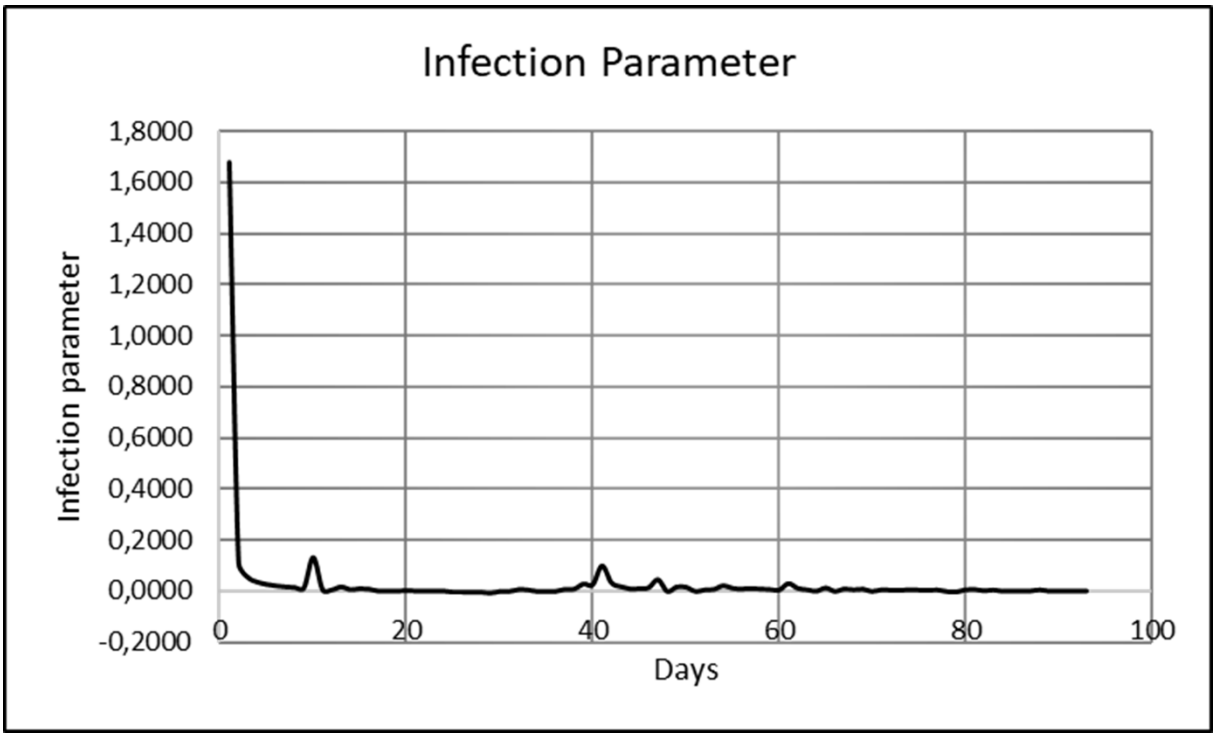

Figure 3. *Infection parameter Rho obtained through the model.*

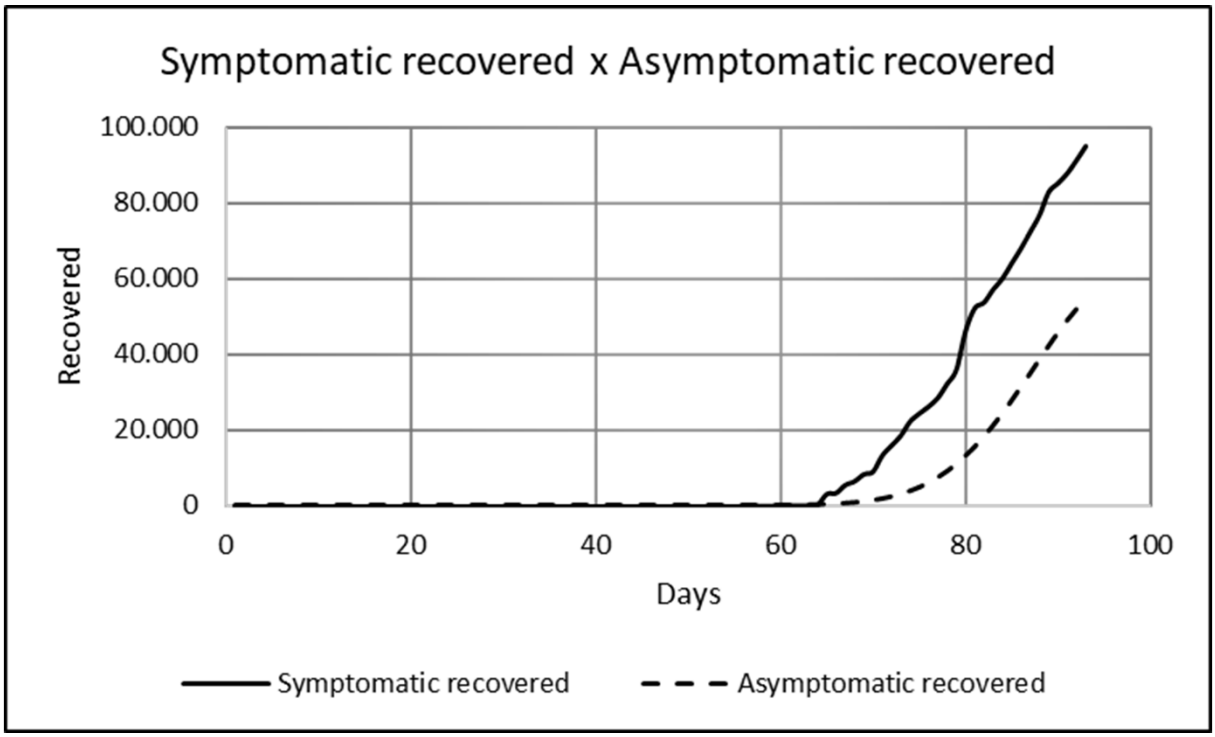

Figure 4. *Symptomatic and Asymptomatic Recovered, obtained, respectively, by official data and model equations.*

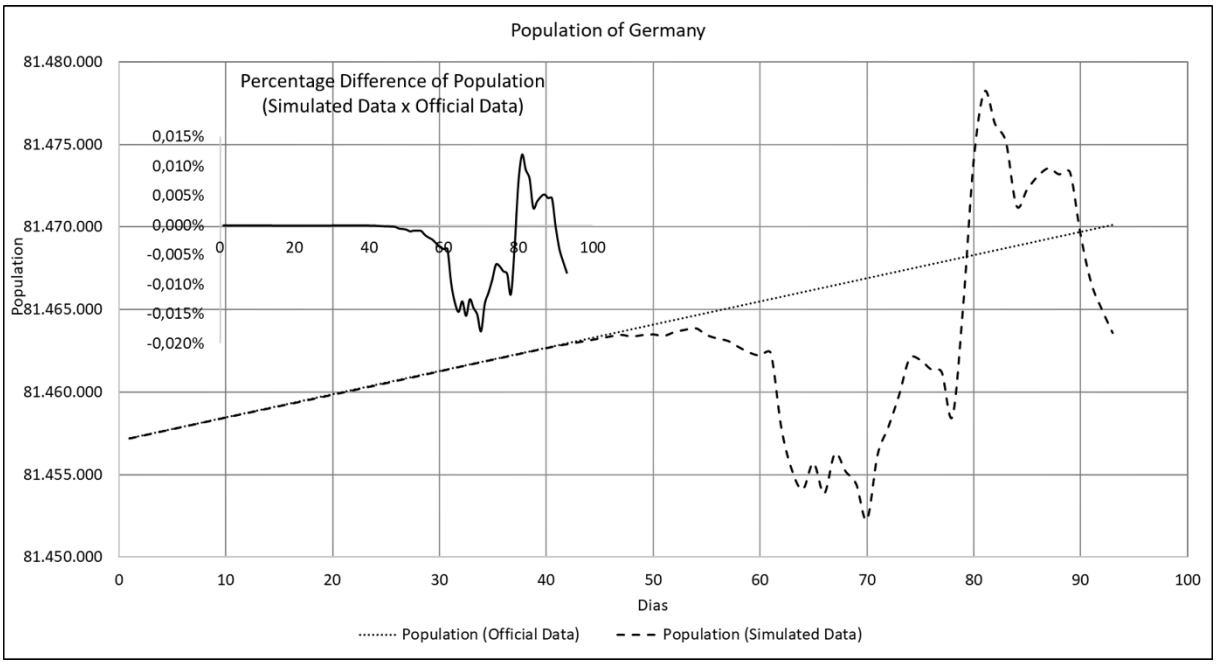

Figure 5. *Population of Germany, in comparison, the data obtained by the model and the official data. Also, the percentage difference between the data in the graphs.*

A cut of the tables with a sample of the respective results is coming up next. The original tables, produced in spreadsheets, couldn´t be put in their entirety in this treatise since there was a lot of measured data, with approximately 2,400 lines for the calculation per hour and 144,000 lines for the calculation per minute.

It is notable that a data interpolation process was carried out so as to make a gradual variation of the data, varying from hour to hour, until we complete a whole day. The official data were released daily, but we interpolated it to write the increment of each hour. This procedure was done for all variables of the mathematical model: infected, recovered, population, etc. In consequence, we are able to expand the volume of data based on the hypothesis of a linear growth within that day, considering that in each hour there was the same growth. This likewise can be applied when there is a reduction of the parameter, for example, when the number of infected decreases overnight.

We executed this increment per minute too, but the discrepancy in the result obtained per hour was almost nil, which determined the decision to use increment per hour. In a single day, we have minutes. As the analysis was made in 90 days, we are going to have 129,600 data for each parameter. We executed this increment per minute too, but the discrepancy in the result obtained per hour was almost nil, which determined the decision to use increment per hour. We have minutes in a single day. As the analysis was made in 90 days, we are going to have 129,600 data for each parameter. This makes the application operation much more laborious, for practically no result greater than interpolation made per hour. In this interpolation produced per hour, we will have over the 90 days of observation a total of 2160 data per model variable, and it has proven to be really easier to operate.

Table 6: Sampling of data obtained through the model

Date (2020)	Day	Asymptomatic Infected $I_A(t)$	Parameter of infection $\rho(t)$	Recovered from the Symptomatic $R_S(t)$	Recovered from the Asymptomatic $R_A(t)$
March 10	50	293	0.0008	21	9
March 20	60	4754	0.0310	177	71
March 21	61	5937	0.0125	206	103
April 05	76	49092	0.0062	28697	7403
April 07	78	54701	-0.0020	36078	10965
April 15	86	56028	0.0023	72597	35130
April 16	87	53669	0.0065	76997	38811

Table 7: Comparison between the official data and the model during and beyond the sample collected.

Data (2020)	Day	Symptomatic Infected from Official Data	Symptomatic Infected from model	Percentage Discrepancy
March 20	60	8001	4754	40.3%
April 05	76	55739	49092	11.9%
April 07	78	57706	54701	5.2%
May 20	121	13361	7335	45.0%
June 10	142	6966	6890	1.1%
July 30	192	8432	5989	29.0%

4. DISCUSSION

From the graphic shown in Figure 1, we can observe that the implemented mathematical model is compatible with the information collected in the period from January 21, 2020 to April 29, 2020. This fact by itself is not sufficient to ensure that the proposed nuclear model for the SARS-Cov-2 pandemic is valid in the period after that of the data sample. But the verification is made from the analysis statistics presented in Table 7.

A percentage discrepancy ranging from 1% to 45% in the period after the data sampling shows that it is necessary to refine the process and further improve the quality of the proposed model. The fluctuation of the inconsistency percentage is very similar to the period of data collection, i.e., the period from January to April.

An oscillatory component of the periodic evolution curve of pandemic variables in the mathematical model, when compared to the actual data, highlights the fact that the oscillation period is not regular for the real data. Consequently, the cycles of pandemic parameters variation are not equal, resulting in a discrepancy up to 45% at most. A method to reduce this disparity is using the data to predict a period of a few days ahead. In fact, it is exceedingly difficult to make a 90-day prediction because there are many variables involved in a mathematical model. Based on the results obtained, it is possible to point out that the data collected at the beginning of the pandemic in a given location are sufficient to predict its behavior until about three months after the end of the date of the initial data collection, with percentage discrepancies decreasing as the number of days to be predicted is less, that is, if we use the model to predict the next 10 days the results will be much better than if we were to make a forecast for the next 90 days.

It is expected, in the current state of development of the nuclear model for the SARS-Cov-2 pandemic, to have a discrepancy of a maximum percentage of 45%. In order to make a comparison, in an article published by Donsimoni (2020) a percentage discrepancy of up to 50% was found using a Markov time model continuous.

This project presents its results similar to the article by Bärwolff (2020) [31]. The behavior of graphs of figures 1 through 4 find a similarity with those presented in what concerns the evolution of the new coronavirus pandemic in Germany. Note that this implies greater data reliability obtained from the nuclear model for epidemiology presented here.

In the works of Donsimoni (2020) [23] and Barbarossa (2020) [24] we also found a reasonable similarity with the results, although the mathematical methods adopted by them are quite different from that developed here in this work. The predictability capability of the nuclear model for the new coronavirus pandemic is confirmed also from a comparison with these articles that present results for Germany.

CONCLUSION

From the discussion about the study and the results obtained, taking into consideration Germany´s epidemiological situation, the nuclear model for the SARS-Cov-2 pandemic was considered valid. Despite this, the precision of the model can still be improved, since this work is part of an institutional research project in IFRJ and is in its early stages.

One possible way to verify the efficiency of the method proposed here is the analysis made and presented in the figure´s graph 5. It is viable to observe the calculation of the population from the demographic data in comparison to the calculation from our model, that we use symptomatic infected, asymptomatic infected, symptomatic recovered, asymptomatic recovered, and susceptible to obtain the population from within the model. Therefore, note that the discrepancy of this result of the population is minimal, which definitively proves the consistency of the model.

Along these lines, we can consider that the results for validation of the model here are preliminary. In spite of this, the mathematical model presented is very promising. This model can and will be revisited with the objective of being distributed in the future to predict the evolution of the pandemic in Brazil, regionally and nationally.

LIST OF ABBREVIATIONS

COVID-19 = Corona Vírus Disease (2019).

IFRJ = Federal Institute of Education, Science and Technology of Rio de Janeiro

Sars-Cov-2 = Severe Acute Respiratory Syndrome Virus (New Coronavirus).

SIR = Susceptible, Infected and Removed in the mathematical model in epidemiology.

U 235 = Uranium 235.

USA = United States of America

WHO = World Health Organization

CONSENT FOR PUBLICATION

We at this moment declare that the present paper " VALIDATION METHOD OF THE MATHEMATICAL MODEL FOR SARS-Cov-2 PANDEMIC FROM DATA MINING AND STATISTICAL ANALYSIS" is our original work and has not been previously considered, either in whole or in part, for publication elsewhere. Besides, we warrant the authors will not submit this paper for publication in any other journal. We also guarantee that this article is free of plagiarism and that any accusation of plagiarism will be the authors' sole responsibility. The undersigned transfer all copyrights to the present paper (including without limitation the right to publish the work in any and all forms) to BJEDIS, understanding that neglecting this agreement will submit the violator to undertake the legal actions provided in the Law on Copyright and Neighboring Rights (No. 9610 of February 19, 1998). Also, we, the authors, declare no conflict of interest. Finally, all funders were cited in the acknowledgments section.

CONFLICT OF INTEREST

This work is the result of the research project "Construction and improvement of descriptive Mathematical Models of the SARS-Cov-2 Pandemic", which was approved in the Integrated Teaching, Research, Innovation and Extension nº 01 e 02/2020 of the Rio de Janeiro´s Federal Institute of Education, Science and Technology (IFRJ). This research project was contemplated with a scientific initiation scholarship, approved by the general direction of the Duque de Caxias Campus of IFRJ and is valid from August (2020) to July (2021).

ACKNOWLEDGEMENTS

We thank the Physics Team - (IFRJ Cduc Physics) for their contribution with ideias and for the use of the Physics Laboratory Nilton de Souza Medeiros as a space for guidance, work and research meetings.

We would also like to thank the Coordination of Research and Innovation (Copi), the Initiation Scholarships Program Scientific - PIBIC/IFRJ, and the Dean of Graduate Studies, Research and Innovation - PROPPI, for the approval of the project and donation of scientific initiation grant.

A special thanks to the Federal Institute of Rio de Janeiro - Campus Duque de Caxias, for the opportunity to act as teachers and researchers at this institution.

Last but not least, we would like to express an exceptional thanks to the entire organizing team of the 1st Brazilian Conference on Experimental Planning and Data Analysis (ConBraPa), for the opportunity to present the initial results of our research project, in an event of such relevance.

SUPPORTIVE/SUPPLEMENTARY MATERIAL

As supplementary material, we will send figures from 1 to 6, tables 1 to 6 and we are going to add the complete tables 5 and 6, namely:

- Figure 1. Symptomatic infected.
- Figure 2. Asymptomatic infected.
- Figure 3. Rho infection parameter.
- Figure 4. Symptomatic and Asymptomatic Recovered.
- Figure 5. Population of Germany, by comparison, model data and official data.
- Table 1. Comparative table between point kinetic model and model for SARS-Cov-2 pandemic.
- Table 2. List of nuclear parameters of point kinetics and their respective descriptions.

- Table 3. List of variables and parameters of the nuclear model for the SARS-Cov-2 pandemic.
- Table 4. Criteria that resulted in the choice of Germany as the place for the model´s validation.
- Table 5. Sampling of official data for Germany.
- Table 6. Sampling of data obtained through the model.
- Table 7. Comparison between the official data and the model during and beyond the sample collected.
- Table. Infected and deaths daily
- Table. Infected and deaths per hour

Sample CRediT author statement

Rafael Pereira Santana is coordinator and advisor of this research project, responsible for writing the article and reviewing the text, tables and graphs**. Anderson Lupo Nunes** is the collaborator and co-advisor of this project, responsible for writing the article, interpolation of data, and adjustment of curves. **Pedro Maia Salomone** is a scholarship student of the project's scientific initiation. Responsible for the interpolation of data, preparation of tables, and adjustment of curves.

REFERENCES

1. CHAVES, T. S. S.; BELLEI, N. SARS-COV-2: The new coronavirus: a reflection about “One Health” and the importance of travel medicine when new pathogens emerge. **Revista de Medicina**, São Paulo, v. 1, p. 99, i-iv, jan-fev 2020. Available in: <https://www.revistas.usp.br/revistadc/article/view/167173>. Access in: March 06, 2021.
2. GOULART, A. D. C. Revisiting the Spanish flu: the 1918 influenza pandemic in Rio de Janeiro. **História, Ciência e aúde - Manguinhos**, Rio de Janeiro, v. 12, n. 1, p. 101-142, Jan/April 2005. Available in: <https://www.nescon.medicina.ufmg.br/biblioteca/imagem/0915.pdf>. Access in: March 05, 2021.
3. WHO. COVID-19 Weekly Epidemiological Update in 29 Dez. 2020. **Worlds Health Organization**, 2020. Available in: <https://www.who.int/publications/m/item/weekly-epidemiological-update---29-december-2020>. Access in: Feb 22, 2021.
4. LUIZ, M. H. R. **Modelos Matemáticos em Epidemiologia**. Rio Claro - SP: UNESP, 2012. Available in: <https://repositorio.unesp.br/bitstream/handle/11449/94348/luiz_mhr_me_rcla.pdf;jsessionid=3A9005292828646A6A77FB5BF8C73F43?sequence=1>. Access in: March 05, 2021.
5. COSTA, S. D. S. Pandemia e desemprego no Brasil. **Revista de Administração Pública**, Rio de Janeiro, v. 54, n. 4, p. 969-978, Aug 2020. Available in: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-76122020000400969&tlng=pt>. Access in: March 17, 2021.
6. MACEDO, L. D. D.; MACEDO, J. R. D. D. A pandemia de Covid-19: aspectos do seu impacto na sociedade globalizada do século XXI. **Caderno de Ciências Sociais Aplicadas**, Vitória da Conquista - BA, v. 17, n. 30, p. 40-43, jul 2020. Available in: <https://periodicos2.uesb.br/index.php/ccsa/article/view/7315>. Access in: March 17, 2021.
7. CRISTÓVÃO, R. B. **Modelo SIR:** Uma Aplicação à Hepatite A. São Paulo: USP, 2015. Available in: <https://www.ime.usp.br/~map/tcc/2015/Rafael%20Belmiro.pdf>. Access in: March 05, 2021.
8. OKHUESE, V. A. Mathematical Predictions for COVID-19 as a Global Pandemic. **MedRXiv: The Preprint Server for Health Sciences**, p. 01-16, mar 2020. Available in: <https://www.medrxiv.org/content/10.1101/2020.03.19.20038794v1>. Access in: March 08, 2021.

9. COSTA, G. S.; COTA, W.; FERREIRA, S. C. Metapopulation modeling of COVID-19 advancing into the countryside: an analysis of mitigation strategies for Brazil. **MedRXiv: The Preprint Server for Health Sciences**, p. 01-13, may 2020. Available in: <https://www.medrxiv.org/content/10.1101/2020.05.06.20093492v2>. Access in: March 08, 2021.

10. CIUFOLINI, I.; PAOLOZZI, A. Mathematical prediction of the time evolution of the COVID-19 pandemic in Italy by a Gauss error function and Monte Carlo simulations. **The European Physical Journal Plus**, v. 135, n. 355, p. 1-8, april 2020. Available in: <https://link.springer.com/article/10.1140%2Fepjp%2Fs13360-020-00383-y>. Access in: March 08, 2021.

11. FOKAS, A. S.; DIKAIOS, N.; KASTIS, G. A. Mathematical models and deep learning for predicting the number of individuals reported to be infected with SARS-CoV-2. **Journal of The Royal Society Interface**, v. 17, p. 01-12, jul 2020. Available in: <https://royalsocietypublishing.org/doi/10.1098/rsif.2020.0494>. Access in: March 08, 2021.

12. KERMACK, W. O.; MCKENDRICK, A. G. A contribution to the mathematical theory of epidemics. **Proceedings of the Royal Societ a Mathematical, Physical and Engineerings Sciences**, p. 700-721, 01 Aug 1927. Available in: <https://royalsocietypublishing.org/doi/10.1098/rspa.1927.0118>. Access in: March 05, 2021.

13. OLIVEIRA, M. H. D. Análise do modelo SIR: Comportamento da curva de infectados em relação à inclusão de novas semanas epidemiológicas. **Tabalho de Conclusão de Curso - Instituto de Matemática e Estatística - USP**, São Paulo, 2018.

14. CRUZ, P. A. D.; CREMA-CRUZ, L. C.; CAMPOS, F. S. Modeling transmission dynamics of severe acute respiratory syndrome coronavirus 2 in São Paulo, Brazil. **Revista da Sociedade Brasileira de Medicina Tropical**, Uberaba, v. 54, p. 1-8, jan 2021. Available in: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0037-86822021000100304&tlng=en>. Access in: March 08, 2021.

15. NETO, O. P. et al. Mathematical model of COVID-19 intervention scenarios for São Paulo - Brazil. **Nature Communications**, v. 12, p. 418, jan 2021. Available in: <https://www.nature.com/articles/s41467-020-20687-y>. Access in: March 08, 2021.

16. KAMRUJJAMAN, M.; JUBYREA, J.; ISLAM, M. S. Data analysis and mathematical model: control measures and prediction to prevent COVID-19 outbreak. **Arabian Journal of Medical Sciences**, v. 3, n. 2, p. 5-9, May 2020. Available in: <https://www.researchgate.net/publication/341448410>. Access in: March 08, 2021.

17. MALLAPATY, S. The Mathematical Strategy that could transform coronavirus testing. **Nature**, v. 583, p. 504, jul 2020. Available in: <https://www.nature.com/articles/d41586-020-02053-6>. Access in: March 08, 2021.

18. DUDERSTADT, J. J. **Nuclear Reactor Analysis**. New York: Wiley & Sons, 1987. Available in: <https://documents.pub/document/nuclear-reactor-analysis-56075ae4de6bc.html>. Access in: March 08, 2021.

19. HETRICK, D. L. **Dynamics of nuclear reactors**. [S.l.]: The University of Chicago Press Ltda, 1971. Available in: <https://openlibrary.org/books/OL1425869M/Dynamics_of_nuclear_reactors>. Access in: March 08, 2021.

20. AKCASU, Z.; LELLOUCHE, G. S.; SHOTKIN, L. M. **Mathematical Methods in Nuclear Reactor Dynamics**. [S.l.]: Academic Press, 1971. Available in: <https://www.sciencedirect.com/book/9780120471508/mathematical-methods-in-nuclear-reactor-dynamics>. Access in: March 17, 2021.

21. NUNES, A. L. et al. A New Formulation to the Point Kinetics Equations Considering the Time Variation of the Neutron Currents. **World Journal of Nuclear Science and Technology**, v. 5, p. 57-71, jan 2015. Available in: <https://www.scirp.org/journal/paperinformation.aspx?paperid=53644>. Access in: March 08, 2021.

22. PALMA, D. A. P.; LUPONUNES, A.; MARTINEZ, A. S. Effect of the time variation of the neutron current density in the calculation of the reactivity. **Annals of Nuclear Energy**, v. 96, p. 204-211, 2 Oct 2016. Available in: <https://www.sciencedirect.com/science/article/pii/S0306454916303292>. Access in: March 08, 2021.

23. DONSIMONI, J. R. et al. Projecting the spread of COVID-19 for Germany. **De Gruyer**, v. 21, n. 2, p. 181-216, 2020. Available in: <https://www.degruyter.com/document/doi/10.1515/ger-2020-0031/html>. Access in: March 08, 2021.

24. BARBAROSSA, M. V. et al. Modeling the spread of COVID-19 in Germany: Early assessment and possible scenarios. **Journal Plos One**, v. 15, n. 9, p. 01-22, sep 2020. Available in: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0238559>. Access in: March 08, 2021.

25. WHO. Coronavirus Disease (COVID-19) Situation Report in 21 Set. 2020. **Worlds Health Organization**, 2020. Available in: <https://www.who.int/emergencies/diseases/novel-coronavirus-2019>. Access in: March 10, 2021.

26. JOHNS HOPKINS UNIVERSITY. Mortality in the most affected countries. **MORTALITY ANALYSES**, Johns Hopkins, 29 dec 2020. Available in: <https://coronavirus.jhu.edu/data/mortality>. Access in: March 08, 2021.

27. COUNTRYMETERS. Relógio de Mortalidade da População Mundial, 2020. Available in: <https://countrymeters.info/pt>. Access in: March 08, 2021.

28. TRAFIMOW, D.; MACDONALD, J. A. Performing Inferential Statistics Prior to Data Collection. **Educational and Psychological Measurement**, v. 77, n. 2, jul 2016. Available in: <https://journals.sagepub.com/doi/10.1177/0013164416659745>. Access in: March 08, 2021.

29. SIMONSOHN, U.; SIMMONS, J. P.; NELSON, L. D. SUPPLEMENTARY MATERIALS FOR: Specification Curve: Descriptive and Inferential Statistics On All Reasonable Specifications. **ResearchBOX: Open Research Made Easy**, Pennsylvania, p. 01-19, Oct 2019. Available in: <http://urisohn.com/sohn_files/wp/wordpress/wp-content/uploads/Supplement-Specification-Curve-2019-10-29.pdf>. Access in: March 08, 2021.

30. VERGURA, S. et al. Descriptive and Inferential Statistics for Supervising and Monitoring the Operation of PV Plants. **IEEE Transactions on Industrial Electronics**, v. 56, n. 11, p. 4456 - 4464, nov 2009. Available in: <https://ieeexplore.ieee.org/abstract/document/4555650>. Access in: March 08, 2021.

31. BÄRWOLFF, G. Mathematical Modeling and Simulation of the COVID-19 Pandemic. **Life in the Time of a Pandemic: Social, Economic, Health and Environmental Impacts of COVID-19 - Systems Approach Study**, v. 8, n. 24, p. 01-12, jul 2020. Available in: <https://www.mdpi.com/2079-8954/8/3/24>. Access in: March 10, 2021.