The SIR Model Applied to Covid-19 Case Data from Brazil

Daniel de Castro

December 16, 2021

Abstract

The Susceptible-Infected-Recovered (SIR) model is a model of disease transmission that has been used to study many different disease outbreaks in the past, and which has found itself particularly relevant in the discussion of the Covid-19 pandemic. In this paper, we discuss the fundamental properties of the SIR model and conduct two simulations of the Covid-19 outbreak in Brazil using two different sets of parameters. We then use the results of this simulation to demonstrate the theoretical properties of the SIR model. Finally, we compare the two simulations against one another and make a judgement as to which set of parameters is a better fit for the actual Covid-19 case data. In the end, we find that neither model is truly a good fit for the actual data, and provide potential reasons for these shortcomings based on the limitations of the SIR model and its underlying assumptions.

1 Introduction

The Susceptible-Infected-Recovered (SIR) model is a model of disease transmission that has been used in the study of the H7H9 flu, whopping cough, and MERS-CoV, as well as other infectious diseases. The SIR model has also found itself particularly relevant in discussion of the Covid-19 outbreak, especially with respect to its basic and effective reproductive numbers (R_0 and R_e), which were discussed frequently in the early days of the ongoing global pandemic. In this paper, we will discuss the properties of the SIR model, perform two separate simulations of Covid-19 spread in Brazil using this model and two different sets of parameters, and assess the performance of these models in reflecting the actual data from Brazil.

The SIR model is a compartmental model in that it separates a fixed population of size N into the three categories of variable size from which it draws its name. The sizes of each of these segments of the population at time t are given by $S(t)$, $I(t)$, and $R(t)$. The model as a whole is described by the following three coupled non-linear ordinary differential equations,

$$\frac{dS}{dt} = -\beta SI \tag{1}$$

$$\frac{dI}{dt} = \beta SI - \nu I \tag{2}$$

$$\frac{dS}{dt} = \nu I \tag{3}$$

where $\beta > 0$ and $\nu > 0$, and the duration of infection can be defined as $D = \frac{1}{\nu}$. Additional parameters of the model are κ, which represents the number of contacts an infected individual has per unit time, independent of population size; and τ, the *transmissibility* of the disease, or the fraction of these contacts that result in transmission of the disease. Each infected individual thus infects $\kappa\tau\frac{S}{N}$ susceptible individuals per unit time; from this result, we find that $\beta = \frac{b}{N}$, where $b = \kappa\tau$. In this way, the term βSI represents the total number of susceptible individuals infected in a single unit of time [1].

1.1 Underlying Assumptions of the SIR Model

Before we can analyze the SIR model in further detail, it will be necessary to account for the assumptions on which the model is built. These assumptions are as follows:

1. The population of size N is large and closed.

2. No natural births or deaths occur in the population.
3. An individual is infectious as soon as they have contracted the disease (the infection has a zero latency period).
4. Once recovered from the infection, an individual has lifetime immunity from the infection.
5. The individuals of the population perform *mass action mixing*, where the rate of encounter between susceptible and infected individuals is proportional to the product of the population sizes. This also requires that the members of the population are homogeneously distributed in space and that the every person will encounter every other person with equal probability per unit time.

These five key assumptions can of course be relaxed to improve upon the SIR model in various ways. Weiss notes that there has been much discussion as to how reasonable is Assumption 5 (that of mass action mixing) [1].

1.2 Properties of the SIR Model

In this section, we will discuss the properties of the SIR model that are key to understanding it in full. We will then show how these properties hold true in a simulation that we performed with the SIR model.

1.2.1 Constant Total Population

Though it may not be immediately obvious, in the SIR model, the total population N is constant. We can show this by showing that the rate of change of the population, dN/dt, is zero:

$$N = S + I + R \tag{4}$$

$$\frac{dN}{dt} = \frac{d}{dt}(S + I + R)$$

$$\frac{dN}{dt} = \frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt}$$

$$\frac{dN}{dt} = -\beta SI + \beta SI - \nu I + \nu I$$

$$\frac{dN}{dt} = 0 \tag{5}$$

1.2.2 Will there be an epidemic? The Epidemic Threshold Theorem

In the SIR model, there are two paths to the end of disease spread. In the first, the disease dies out from the start, with $I(t)$, the number of infected individuals, decreasing monotonically to zero. In the second, an epidemic of the disease breaks out, and the number of infected increases first to a maximum before decreasing to zero (we will formalize this definition below).

Which path a given disease will take can be determined with the help of two constants, the basic reproductive number R_0 and the effective reproductive number R_e, both of were the subject of much discussion even in popular media at the beginning of the COVID-19 pandemic. These constants are defined as follows:

$$R_0 = \frac{b}{\nu} = \frac{\kappa\tau}{\nu} \tag{6}$$

$$R_e = \frac{S(0)}{N} \cdot \frac{b}{\nu} = \frac{S(0) \cdot \beta}{\nu} = \frac{S(0)}{N} \cdot R_0 \tag{7}$$

Note that if the disease begins with only one patient — i.e. $S(0) = N - 1$, $I(0) = 1$ — then $R_e = (N - 1/N)R_0 \approx R_0$. The *Epidemic Threshold Theorem* states that whether or not an epidemic breaks out depends on the effective reproductive number R_e.

Theorem 1 (Epidemic Threshold [1])**.** *If $R_e \leq 1$, then $I(t)$ decreases monotonically to zero as $t \to \infty$. If $R_e > 1$, then $I(t)$ increases, reaches a maximum, then decreases to zero as $t \to \infty$. Note that we define an* epidemic *by this second pattern of disease spread.*

1.2.3 End Behavior of the SIR Model

Note that in both cases fleshed out in Theorem 1, the end behavior of the model is such that $I(t) \to 0$ as $t \to \infty$. Because individuals cannot be reinfected with the disease, the disease will always eventually die out in the SIR model.

Interestingly enough, while it is easy to assume that the disease dies out in the SIR model because of a lack of susceptible individuals, Weiss 2013 noted that the disease eventually dies out because of a lack of newly infected individuals [1]. When $S(t)$ drops below $\frac{\nu}{\beta}$, the rate at which new individuals are infected, $\frac{dI}{dt}$, drops below the rate at which infected individuals recover, $\frac{dR}{dt}$. Because of this, $S(\infty)$ actually has a lower bound, as Weiss 2013 showed through the following derivation. We begin by dividing equation (1) by equation (3).

$$\frac{dS}{dR} = \frac{-\beta SI}{\nu I} = \frac{-\beta S}{\nu}$$

$$\int_0^t S^{-1} dS = \int_0^t \frac{-\beta}{\nu} dR$$

$$\ln S(t) - \ln S(0) = \frac{-\beta}{\nu}(R(t) - R(0))$$

$$\frac{S(t)}{S(0)} = \exp\left(\frac{-\beta}{\nu}(R(t) - R(0))\right)$$

$$S(t) = S(0) \exp\left(\frac{-R_0}{N}(R(t) - R(0))\right) \tag{8}$$

Now, since $0 \leq R(t) - R(0) \leq N$,

$$S(t) \geq S(0)e^{-R_0}$$

Thus,

$$S(\infty) \geq S(0)e^{-R_0} \tag{9}$$

In this way, the basic reproductive R_0 also plays a role in determining how many susceptible individuals will remain at the end of the outbreak of the disease.

1.2.4 The Maximum Number of Infections

Weiss 2013 shows that a formula can be found for I_{max}, the maximum number of infections, without solving the system of ODEs [1]. One can begin by dividing equation (1) by equation (2):

$$\frac{dS}{dI} = \frac{-\beta SI}{\beta SI - \nu I}$$

$$\int \frac{\beta S - \nu}{\beta S} dS = -\int dI$$

$$\int 1 - \frac{\nu}{\beta S} dS = -I + C$$

$$S - \frac{\nu}{\beta} \int S^{-1} dS = -I + C$$

$$S - \frac{\nu}{\beta} \ln S = -I + C$$

$$S + I - \frac{\nu}{\beta} \ln S = C \tag{10}$$

Thus for any time t, the sum $S(t) + I(t) - \frac{\nu}{\beta} \ln S(t)$ is equal to the some constant C. This allows us to assert that

$$S(t) + I(t) - \frac{\nu}{\beta} \ln S(t) = S(0) + I(0) - \frac{\nu}{\beta} \ln S(0)$$

$I(t)$ should have a maximum when $\frac{dI}{dt} = 0$. According to equation (2), this should happen when $S(t) = \frac{\nu}{\beta}$. Thus,

$$I_{max} + \frac{\nu}{\beta} - \frac{\nu}{\beta} \ln \frac{\nu}{\beta} = S(0) + I(0) - \frac{\nu}{\beta} \ln S(0)$$

$$I_{max} = S(0) + I(0) - \frac{\nu}{\beta} \left(\ln S(0) + 1 - \ln \frac{\nu}{\beta} \right) \tag{11}$$

1.2.5 The Total Number of Infections

The total number of infections can be expressed by $1 - S(\infty)$. As Weiss 2013 did [1], we can begin with equation (8), letting $R(0) = 0$ and setting $t = \infty$:

$$S(\infty) = S(0) \exp \left(\frac{-R_0}{N} R(\infty) \right)$$

Additionally, if we consider that at the end of the epidemic $I(\infty) = 0$,

$$S(\infty) = S(0) \exp \left(-R_0 \left(1 - \frac{S(\infty)}{N} \right) \right)$$

$$\frac{S(\infty)}{N} = \frac{S(0)}{N} \exp \left(-R_0 \left(1 - \frac{S(\infty)}{N} \right) \right)$$

If we assume that $S(0)$ is asymptotically close to N,

$$\frac{S(\infty)}{N} = \exp \left(-R_0 \left(1 - \frac{S(\infty)}{N} \right) \right)$$

$$\ln \frac{S(\infty)}{N} = R_0 \left(\frac{S(\infty)}{N} - 1 \right) \tag{12}$$

Equation (12) can be solved numerically for $S(\infty)$, as we do with our concrete example in Section 3.1.5.

2 Methods

In this section, we discuss our application of the SIR model to Covid-19 Case Data from Brazil. Case data was drawn from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [2]. The CSSE data was given in terms of cumulative confirmed Covid-19 cases, cumulative total deaths, and cumulative total recovered individuals for each day from 22 January 2020 to 2 December 2021; the last day of complete data, however, was 5 August 2021. The first Covid-19 infection in Brazil was reported on 26 February 2020, so we begin all of our analyses with this date ($t = 0$). We were thus left with 526 days of complete data ($t \in [0, 525]$).

Because the CSSE data was not given in terms of current susceptible ($S(t)$), infected ($I(t)$), and recovered ($R(t)$) individuals on each day, we had to manipulate the data set in order to fit the constraints of the SIR model. First, we calculated $R(t)$ for each day by adding the cumulative number of Covid-19 deaths and the cumulative number of individuals recovered from Covid-19 for each day in the data set. $I(t)$ was calculated by subtracting $R(t)$ from the cumulative number of Covid-19 infections for each day in the data set. Lastly, $S(t)$ was calculated by subtracting both $I(t)$ and $R(t)$ from the initial number of susceptible individuals, $S(0)$. For our calculations, $S(0) = 212,559,408$, that is, one less than the 2020 population of Brazil as reported by the World Bank [3]; $I(0) = 1$; and $R(0) = 0$. Figure 1 is a plot of this data.

To find β and ν, we first followed the method outlined by Lounis and Bagal (referred to as Model 1 from here on out) [4]. Using this method, $\beta = 2.965 \times 10^{-10}$ and $\nu = 0.05168$; however, we found that our model

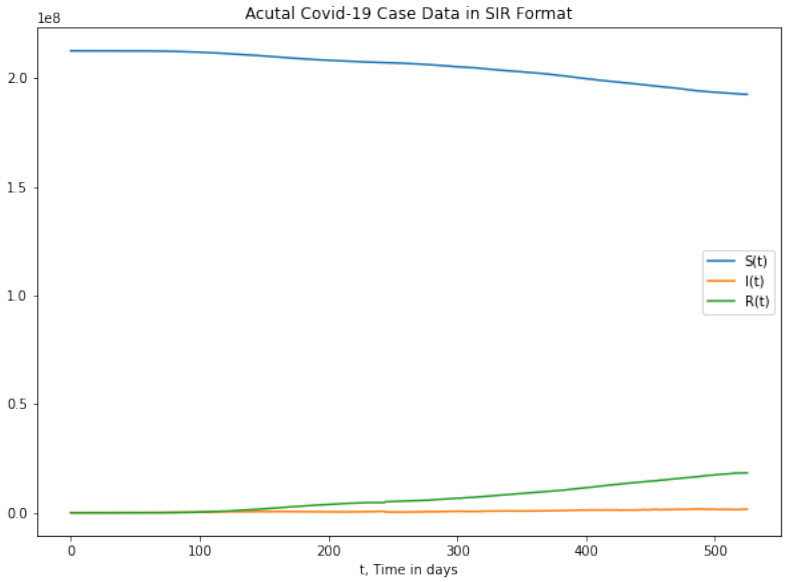

Figure 1: Covid-19 case data in SIR format, as calculated from the data given by the CSSE at Johns Hopkins University [2].

was quite erroneous. In particular, the value for β appeared to be so low that while the actual data saw a growth in cases in the interval $t \in [0, 1000]$, our model saw cases remain near zero. We thus decided to calculate β and ν using a different method (referred to as Model 2 from here on out). To do so, we used the center difference method to estimate $\frac{dS}{dt}$ and $\frac{dR}{dt}$ at every point (except for the end points), used this value to calculate β and ν at these points, and then averaged these β and ν values. The results for this method were $\beta = 3.957 \times 10^{-10}$ and $\nu = 0.05277$.

In both cases, we used scipy.integrate.solve_ivp to find the solutions to our system of differential equations. Both implicit and explicit methods were tested — including RK45, backwards differentiation, and an order 5 Radau method — but the solutions given by each method differed only slightly from one another. To be safe, all of the following results are those we found when working with the LSODA method, an Adams/backwards differentiation method drawn from ODEPACK (c.f. [5]) that employs automatic stiffness detection and switching [6].

3 Results

Figure 2 shows the resulting solutions to our system of differential equations for Model 1, while Figure 3 shows the same for Model 2. Clearly, there is some error inherent in our both versions of our model. As a fraction of the total population, we calculated the average absolute error for $S(t)$, $I(t)$, and $R(t)$ in Model 1 to be 3.5%, .33%, and 3.3%. The same error figures for Model 2 were 2.6%, 0.39%, and 2.6%, respectively. In both cases, the error is rather high.

Nevertheless, we can still demonstrate that the properties of the SIR model discussed in section 1.2 hold with our model, as we do in the following section. Useful to this end will be the values of R_0 and R_e for our models. For Model 1, $R_0 \approx R_e \approx 1.220$; for Model 2, $R_0 \approx R_e \approx 1.597$.

3.1 Demonstration of the Properties of the SIR Model

In this subsection, we will discuss the properties of the SIR model laid out in Section 1.2 in the context of our simulation.

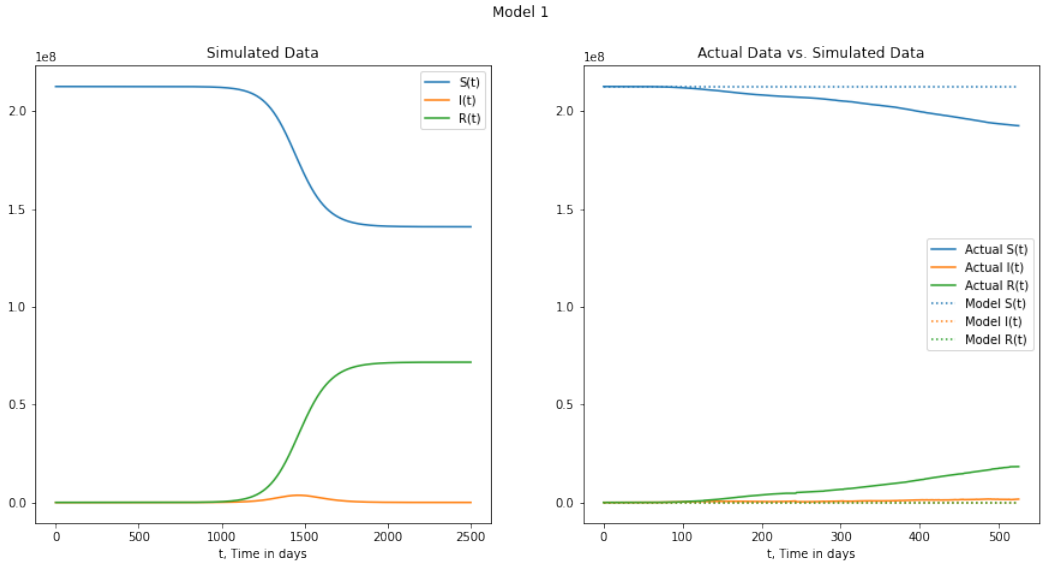

Figure 2: The solutions to Model 1.

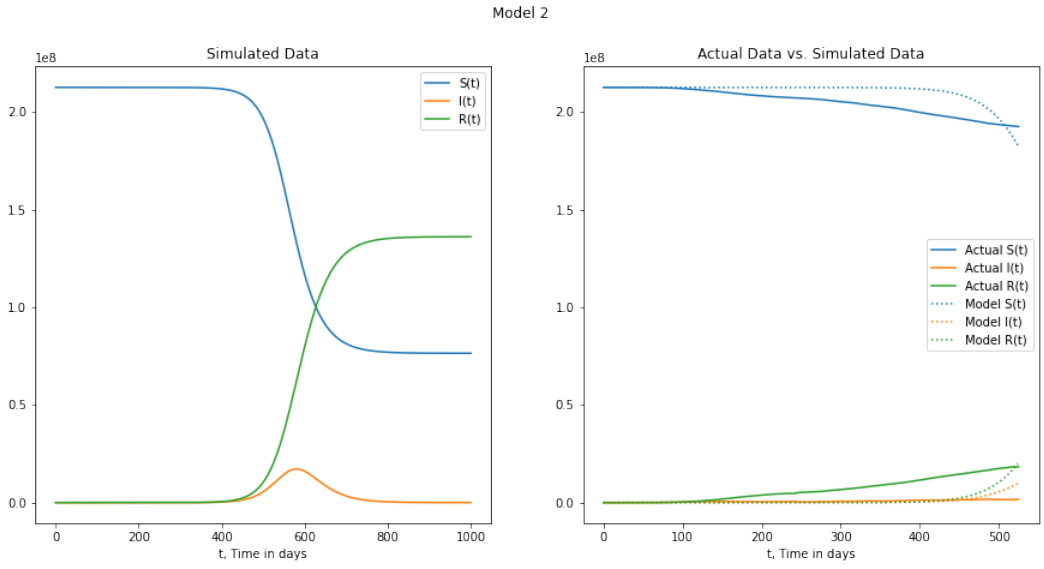

Figure 3: The solutions to Model 2.

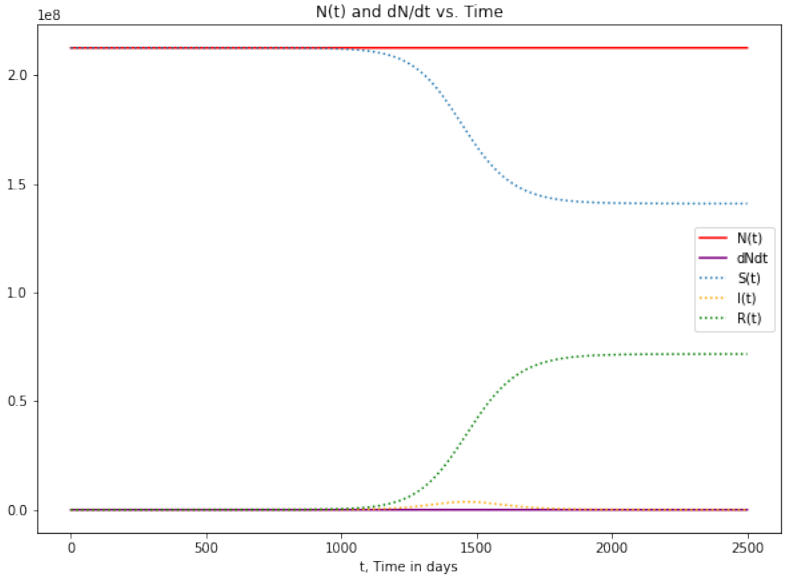

Figure 4: Plotting $N(t)$ and dN/dt shows that both remain constant, with $dN/dt = 0$, as $t \to \infty$. This plot uses the solutions to Model 1; the same holds true for Model 2.

3.1.1 Constant Total Population

As per equation (4), the total population of our model, N, can, at any given time t, be expressed as the sum of the individuals that are susceptible to, are infected by, or have recovered from the virus. As Figure 4 clearly shows, $N(t)$ is constant with a value of 212,559,409, and $\frac{dN}{dt}$ is constant with a value of 0 as $t \to \infty$.

3.1.2 The Epidemic Threshold Theorem

Does the Epidemic Threshold Theorem hold true for out simulation? Given that $R_e > 1$ in both models, we would expect our model to predict an epidemic of Covid-19 in Brazil. By inspection, the behavior of $I(t)$ in both Figures 2 and 3 does indeed match the behavior of an epidemic as described by the Epidemic Threshold Theorem.

3.1.3 End Behavior of the SIR Model

By inspection of Figures 2 and 3, we can clearly see that $I(t) \to 0$ as $t \to \infty$, as we asserted would be the case in Section 1.2.3. Figures 2 and 3 also demonstrate that an epidemic does not die out because of a lack of susceptible individuals: When $t = 10,000$, $I = 0$ (with rounding error) — meaning the epidemic has effectively died out — but $S \approx 140,881,982$ in Model 1 and $S \approx 76,453,053$ in Model 2 (rounded to nearest whole number). We also know $S(10,000)$ to be the final number of susceptible individuals, since $dS/dt = -\beta SI = 0$ from this point forwards, because $I = 0$.

We were also able to calculate the lower bound for $S(\infty)$ for both models, which was $62,772,376$ for Model 1 and $43,045,250$ for Model 2. These lower bounds are accurate, since both are less than the corresponding $S(10,000)$ values.

3.1.4 The Maximum Number of Infections

Equation (11) gives the theoretical value of I_{max}, the maximum number of infections. We calculate this value to be $3,676,461$ for Model 1 and $17,151,726$ for Model 2 (rounded to nearest integer).

We can also find I_{max} by employing numerical methods to find the critical point of $I(t)$ and evaluating the function at this point. In this method, we find a Cubic Hermite Spline interpolation polynomial for each of the three sets of data points given by the scipy IVP solver. Cubic Hermite Spline interpolation polynomials were used in order to force the first derivatives of these polynomial to match those given by equations (1), (2), and (3). We then write a function to calculate dI/dt using equation (2), and find the root of this function using scipy.optimize.root_scalar and the bisection method. This method gave I_{max} to be $3,680,392$ for Model 1 and $17,169,329$ for Model 2, which have an error of 0.11% and 0.10% from the theoretical value given by equation (11), respectively. We thus find equation (11) to be consistent with the true results given by both models.

3.1.5 The Total Number of Infections

Using scipy.optimize.root_scalar and the secant method, we are able to solve the following modified form of equation (12):

$$\ln \frac{S(\infty)}{N} - R_0 \left(\frac{S(\infty)}{N} - 1 \right) = 0$$

The result given by this method is $140,882,163$ for Model 1 and $76,447,789$ for Model 2 (rounded to the nearest integer). If we take $S(10,000)$ to be the true value of $S(\infty)$, the result given by equation (12) has negligible error (less than 10^{-4}) for both Model 1 and Model 2.

4 Discussion

As the error calculations in Section 3 and Figures 2 and 3 clearly demonstrate, both simulations of the Covid-19 data in Brazil using the SIR model are flawed. As it turns out, Model 1 is a better fit for the data, though for the most part through the fault of Model 2. Indeed, Model 2 predicts a sharp uptick in cases around day 500 that is not present in the actual data. The steepness of all three curves at this point suggests an earlier peak in the pandemic than is reflected in the actual data, which at this point does not appear to be approaching its peak (see Figures 2 and 3). Thus, despite having a higher average error, Model 1 seems to better mirror the actual behavior of the Covid-19 pandemic in Brazil.

There are many reasons why both of our SIR models presented such high error in their reflections of the actual Covid-19 case data. The most likely of these reasons stem from the limitations of the SIR model, which Tolles and Luong [7] outlined as follows:

1. The SIR model does not account for the latency period of the SARS-Cov-2 virus, that is, the time after infection and before the virus has replicated itself enough in an individual to reach levels of transmission. The SEIR model, where E represents the number of indivuals currently exposed but not contagious, is an improvement upon the SIR model that takes this into account.

2. The SIR model assumes that there are no births, no deaths, and no migration into/out of the country.

3. The SIR model assumes that there is homogeneous mixing in the population, while in reality most contacts occur within limited social networks.

4. The parameters of the SIR models are point estimates, which do not allow researches to take into account the uncertainty in their values. This is the primary reason why we had to perform two separate simulations in this paper. Clancy and O'Neill [8] show how statistical distributions for model parameters can be used, as well as how parameters can be estimated from incoming data.

Given all these sources of error, is the SIR model worth our time? The answer is still yes. The SIR model is useful because of the ease and speed with which simulations can be performed. It also serves as a good starting point for other, more complicated models, which can be achieved by adapting some of the assumptions of the SIR model. As already discussed, the SEIR model is an example of this method of creating more advanced, more accurate models of disease transmission. Weiss also discusses in detail how an analysis of the SIR model has revealed to public health officials the Epidemic Threshold Theorem, the existence of herd immunity, and the fact that an epidemic ends because of a lack of new infected individuals [1].

References

[1] Howard Weiss, *The SIR model and the Foundations of Public Health*, MATerials MATemàtics (2013), available at `https://mat.uab.cat/web/matmat/wp-content/uploads/sites/23/2020/05/v2013n03.pdf`.

[2] *COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University*, `https://github.com/CSSEGISandData/COVID-19`. Accessed 1 Dec. 2021.

[3] *DataBank: World Development Indicators*, `https://databank.worldbank.org/reports.aspx?source=world-development-indicators`. Accessed 1 Dec. 2021.

[4] Mohamed Lounis and Dilip Kumar Bagal, *Estimation of SIR model's parameters of COVID-19 in Algeria*, Bull Natl Res Cent. (2020), available at `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7570398/`.

[5] *ODEPACK: Fortran ODE Solvers*, `https://computing.llnl.gov/projects/odepack`. Accessed 15 Dec. 2021.

[6] *SciPy API Reference: scipy.integrate.solve_ivp* (2021), `https://docs.scipy.org/doc/scipy/reference/generated/scipy.integrate.solve\_ivp.html`. Accessed 15 Dec. 2021.

[7] Juliana Tolles and ThaiBinh Luong, *Modeling Epidemics With Compartmental Models*, JAMA Guide to Statistics and Materials (May 27, 2020), available at `https://jamanetwork.com/journals/jama/fullarticle/2766672`.

[8] Damian Clancy and Philip D. O'Neill, *Bayesian estimation of the basic reproduction number in stochastic epidemic models*, Bayesian Analysis (2008), available at `file:///Users/danieldecastro/Downloads/08-BA328.pdf`.