Epidemiological Models: SARS-CoV-2 in Portugal

Maria Beatriz Silva Santiago
mbeatrizsantiago@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

December 2021

Abstract

This analysis includes the computation of the level of under report, positivity rate, lethality and the (basic) reproduction number accompanied by a new formula with a non-constant viral load. At last, there was fitting of epidemiological models to the second and third wave of the pandemic in Portugal. To do so, we resorted to the SIR and SEIRD models, where the latter had significantly better results. The error metrics used were squared errors and mean absolute percentage error. For a more complete analysis, this process was also made per region and gender. Health care systems endured a test like no other as COVID-19 patients filled up hospitals leading to increasingly crowded hospitals. Overcrowded hospitals have severe consequences and for that reason, the proposal of a model that takes into account the number of patients in the infirmary and in ICUs is the main contribution of this work.

Keywords: SARS-CoV-2, SIR Model, SEIRD Model, Reproduction Number

1. Introduction

Even though epidemics has been around for thousands of years, its mathematical study is a relatively new field of research. The first statistical study [7] developed in this area was in 1662 by John Graunt. His small book consisted of statistical analysis of weekly records of diseases and casualties. It was not until the 19^{th} and early 20^{th} century that remarkable breakthroughs concerning causes and prevention of diseases were accomplished, paving the way for mathematical modelling of infectious diseases. Major progress was made from this point on.

The most influential contributions to the area are from Kermack and McKendrick in 1927. Their famous trilogy [11, 12, 13] captures diseases that established themselves and persist in population. Their work considers a deterministic epidemic model that takes into account susceptible, infected and removed individuals.

Around the same time, the mathematical study of differential equations was of the utmost importance, creating a bridge between epidemics and its modelling. Regarding ordinary differential equations, we follow the perspective of [8] as a basic text in qualitative theory. The mathematician Herbert Hethcote published groundbreaking work, such as [3, 4, 5, 6], that were crucial to the fast development and analysis of compartment models.

In 2015, the mathematical biologist Maia Martcheva published her first book [14], an introduction to mathematical modelling and analysis of infectious diseases. This book covers ordinary differential equation models, which is the foundation of this work.

With the global phenomenon of COVID-19, the study of epidemiological diseases becomes of the utmost importance. Suddenly, global economies had to come to an alt and people's livelihood were in danger. Mathematical models are of great importance to get a better understanding of a given system, providing us an opportunity to seek optimal performances, intervention strategies and predictions about its behaviour, all of which can be life saving.

The necessity to do further research on epidemic models and make a thorough analysis of the evolution of COVID-19 in Portugal throughout the past year becomes clear. That is the main goal of this work. Hopefully, this work will pave the way for how to approach a future pandemic. To fit epidemiological models, we will resort to square errors metrics and mean absolute percentage error to choose the best model to the given data. A similar type of work has been made before in [1].

As the number of cases of COVID-19 increased, hospitals were getting more crowded by the day. The overall panorama was so worrying that some feared the collapse of national health care systems. The main contribution of this work is the proposal of a model that accounts for the number of patients in infirmaries and intensive care units. In terms of model fitting, a similar process will occur. The only difference relies in choosing the best model with the lowest mean error metric of the three curves.

The present work is organised as follows. Section 2 covers some basics regarding infections and its transmission. This section also provides some mathematical background for different types of models and computation of important parameters. On Section 3 a brief overview of the evolution of SARS-CoV-2 is provided, followed by a careful analysis of the virus in Portugal. In this section, models discussed in Section 2 will be fitted into the data. Finally, on Section 4 a new model is introduced, accounting for individuals in infirmaries and in intensive care units.

2. Background

2.1. Basics: Infections, Transmission and Models

This subsection provides an introduction to the definitions of infections, transmission, mathematical models and key concepts [14, 21] in infectious disease epidemiology.

All species carry infections of a wide variety. Many are harmless, some are beneficial, but some, the pathogens, harm their hosts and lead to diseases. This article will focus on the latter. When a transmission of the infection occurs between two individuals it is usually called *effective contact*. There are several ways to characterise the transmission of a disease, including vertically/horizontally, direct/indirect contact, airborne infection, droplet infection, vector-borne and fecal-oral route. Some viruses are spread using multiple mechanisms.

Once the pathogen establishes itself in the host, typically it takes a certain period of time for the infectious agent to replicate before being able to infect other individuals. To understand the behaviour of infectious diseases and its dynamics, three essential time periods must be distinguished. The *latent period* is defined as the time interval between the infection of an host by a pathogen and when this host becomes infectious, i.e. capable of transmitting pathogens to susceptible individuals. The *incubation period* refers to the time period between exposure to an infectious agent and the onset of symptoms of the disease. As for the *infectious period* it refers to the period where an host can transmit the pathogen to other individuals.

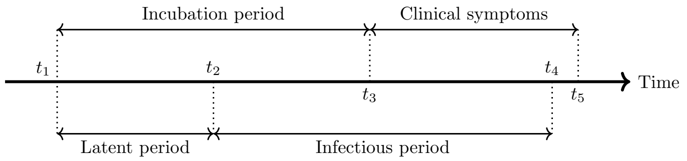


Figure 1: Summary of the relevant time periods.

Note that there might be individuals that make a potentially disease transmitting contact, becoming *exposed*, and may or may not develop the disease (typically are non infectious). However, mathematical models often assume that all exposed individuals develop the disease. Therefore, for the purpose of this work individuals in the latent period will be considered as *exposed individuals*.

There are different outcomes after an infection. Some may be mild, causing little to no illness to their host, while there are some more extreme that may be fatal. As far as the duration of the infectiousness goes, it is usually determined by the ability of the host to create an immune response, or vaccine induced immunity.

Mathematical epidemiological models are developed to help explain a system, study the effects of each component and to help make predictions about their behaviour. Mathematical models consist of parameters and variables that are somehow connected. These variables represent a part of the system that can be quantified/measured.

In this work, ordinary differential equation models will be used to model the distribution of infectious diseases in populations. These models are nonlinear, dynamic, continuous and deterministic.

2.2. Epidemiological Models

This subsection was written using [8, 14] which explore two types of epidemic models, SIR and SEIRD model.

2.2.1 SIR Model

This model dates back to 1927 from the famous article [10] of Kermack and McKendrick, and takes into account susceptible (S), infected (I) and recovered (R) individuals. This model supports itself on several assumptions: there are neither births nor deaths in the population; the population is closed, there are no entries or exits to/from the population; infected individuals are considered infectious; recovered individuals are considered to have full immunity and cannot be reinfected. Sometimes deceased individuals are also included in this class.

The movement of individuals is unidirectional, i.e. an individual cannot return to a previous class. Susceptible individuals get infected according to β, the *transmission rate constant*. Individuals leave the infected class at a per capita probability per time unit ρ, known as *recovery rate*.

The model is given by the ODEs

$$\begin{cases} S'(t) = -\beta S(t)I(t) \\ I'(t) = \beta S(t)I(t) - \rho I(t) \\ R'(t) = \rho I(t), \end{cases} \tag{1}$$

with initial conditions $S(0)$, $I(0)$ and $R(0)$. The population remains constant. Adding all equations from (1), results in $N'(t) = S'(t) + I'(t) + R'(t) = 0$ yielding $N(t) = N$, $\forall t$.

For a better understanding of the behaviour of this model, a few computations are made with respect to the numbers of susceptible and recovered individuals. By dividing $S'(t)$ for $R'(t)$ and then, integrating and rearranging both sides, it follows

$$S(t) = S(0)e^{\frac{\beta}{\rho}R(t)}. \tag{2}$$

The number of recovered individuals is monotone and bounded by N, and consequently $S(t) > 0$. Therefore, the epidemics does not end. Some individuals always escape the disease.

In order to solve the differential equations, let us divide both equations

$$\frac{I'}{S'} = \frac{\beta SI - \rho I}{-\beta SI} \Leftrightarrow I' = \Big(-1 + \frac{\rho}{\beta S}\Big) S', \tag{3}$$

where one can inspect the behaviour of $I(t)$. The infected population arises at first reaching an all time high and then declines.

A simple integration on both sides allow us to explicitly attain $I(t)$ as a function of $S(t)$. Since $S(t)$ is a monotonically decreasing function, the maximum number of infections can be computed

$$I' = 0 \Leftrightarrow S(t) = \frac{\rho}{\beta}, \tag{4}$$

which is an extremely important information when fighting this type of disease. An important threshold is hidden in the previous equation. The *effective reproduction number* is

$$\mathcal{R}_t = \frac{\beta}{\rho} S(t). \tag{5}$$

2.2.2 SEIRD Model

Many infectious diseases have a latency period. Giving its importance to how the spread of an infection can occur, this particular model considers a new class, *exposed individuals* (E). In addition, a class for the *deceased individuals* (D) is also taken into consideration.

For an easier understanding of the model, Table 1 presents a summary of the notation of the model.

Table 1: Summary of notation - SEIRD model.

β	Transmission rate
$1/\sigma$	Average latent period
$1/\rho$	Average infectious period
γ	Fraction of recovered individuals

The ODEs that define this model are

$$\begin{cases} S'(t) = -\beta S(t)I(t) \\ E'(t) = \beta S(t)I(t) - \sigma E(t) \\ I'(t) = \sigma E(t) - \rho I(t) \\ R'(t) = \gamma \rho I(t) \\ D'(t) = (1-\gamma)\rho I(t), \end{cases} \tag{6}$$

with initial conditions $S(0)$, $E(0)$, $I(0)$, $R(0)$ and $D(0)$. Once again, this model is closed and consequently the population size remains constant, which is easily proven by adding all equations from (6).

Linearization

Since the last two variables, R and D, can be obtained using simple quadratures, the first three equations will provide full information on the behaviour of this model. With the respective field, one can linearize the system near its equilibrium point $(1, 0, 0)$, resulting in

$$\begin{cases} S'(t) = -\beta I(t) \\ E'(t) = -\sigma E(t) + \beta I(t) \\ I'(t) = \sigma E(t) - \rho I(t). \end{cases} \tag{7}$$

With the previous system, one can attain the formula for the number of infected individuals: $I(t) = \frac{e^{\lambda t}}{N}$, where λ is the dominant eigenvalue. If one considers the daily growth rate a, where $I(t+1) = aI(t)$, the result $\lambda = \log a$ follows. On the other hand, the eigenvalue can be computed explicitly from the previous system. The transmission rate β can be written as

$$\beta = \frac{\rho\sigma + \rho \log a + \sigma \log a + \log^2 a}{\sigma}. \tag{8}$$

Consequently, the *basic reproduction number*, $\mathcal{R}_0$, can be computed as

$$\mathcal{R}_0 = \frac{\beta}{\rho} = 1 + \frac{\log a}{\rho} + \frac{\log a}{\sigma} + \frac{\log^2 a}{\rho\sigma}. \tag{9}$$

2.3. Herd Immunity

If the fraction of susceptible individuals is sufficiently low, then the pathogen will not be able to successfully spread. The reduction of susceptible individuals in a population is achieved by individuals acquiring immunity, either through natural infection or through vaccination.

Herd immunity defines itself as the indirect protection from an infectious disease when a significant fraction of the population is immune to the virus [19]. Herd immunity is of the utmost importance since it allows immunocompromised and younger people to remain unvaccinated.

The value of the reproduction number is required to compute the percentage threshold of the population that must be immune to block sustained transmission, i.e. the herd immunity threshold

$$\mathcal{H} = 1 - \frac{1}{\mathcal{R}_0}. \tag{10}$$

The herd immunity formula relies itself on a few assumptions. There must be an homogeneous mixing of individuals within a population and all individuals must develop immunity that provides a lifelong protection against the virus. In real-world cases, population density differs immensely from region to region and vaccines may not confer full immunity, specially in new viruses where new variants can emerge. With new variants, vaccines lose some of its efficacy resulting in a need to adjust the herd immunity value

$$\mathcal{H}_{adj} = \left(1 - \frac{1}{\mathcal{R}_0}\right)\frac{1}{V_e}, \tag{11}$$

where V_e represents the vaccine effectiveness (the immunity to the virus that the vaccine confers to an individual).

3. Analysis of SARS-CoV-2 in Portugal

Before entering into the analysis of any pathogen in a given population, it is necessary to fully comprehend how the virus spreads and its characteristics. SARS-CoV-2 is a respiratory infectious disease and has the ability to spread rapidly across all continents in our globalised world. This virus can be spread through multiple ways: contact with an infected person, touching a contaminated surface, by droplet transmission of respiratory particles that contain the virus and lastly by airborne transmission droplets/particles suspended in the air for longer periods of time and distance. All ages are susceptible to infection, however clinical manifestations differ with age. As far as the length of the latency and incubation period, recent studies [17, 22] point to a shorter latency period.

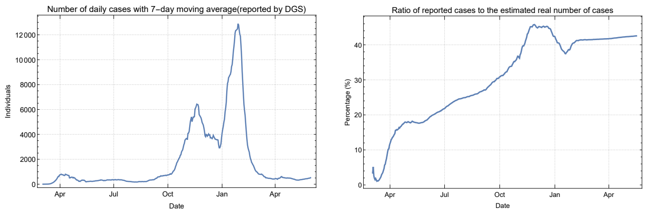
3.1. Preliminary Analysis

The data here used is of public access and provided daily in [20]. All the results here presented were obtained resorting to the software *Wolfram Mathematica*. There are several information of Portugal and its regions, including number of confirmed cases (by age and gender), deaths, recovered individuals, patients in the infirmary and ICU, active cases, among others.

The first reported case was on the 2nd of March, however the dataset starts on the 26th of February with 25 individuals already in vigilance. All mentioned information was reported daily, and the last day of the pandemic here considered is May 31st, 2021, which results in a total of 461 days.

One of the biggest problems when modelling this pandemic was the level of unreliability and inaccuracy of the data. There were limitations regarding the number of tests performed and most laboratories only being open during the weekdays. Thus, to smooth the data, a 7-day moving average was applied. There are three epidemiological waves during this period, which are very clear on Figure 2(a), where the number of daily cases is presented.

In order to fit epidemiological models into the data at our disposal, we must take advantage of the number of infected individuals. However, since the data provided by *"Direcção Geral de Saúde"*, the Portuguese health authority (DGS) is unreliable, one must turn to other alternatives. For the purpose of achieving a good approximation of the real number of infected individuals, one can take advantage of the number of casualties. A recent article [2] used age specific COVID-19 data from multiple countries to investigate the consistency of infection and fatality patterns. The infection fatality ratio estimated for Portugal was 0.86% with a 95% C.I. of 0.75 − 0.99%. With this information, and considering fourteen days from onset symptoms of COVID-19 to death, we are able to compute the total number of cases. The computation of the level of under report of cases follows (see Figure 2(b)).



(a) Daily cases

(b) Level of reported cases

Figure 2: New daily cases vs. ratio between the number of cases reported and the number of estimated cases.

Also, if we know how much time an individual spends infected, the number of active cases per day can be computed. Consequently, the desired real number of infected individuals for each point in time can be obtained. Two data sets were created considering 7 and 14 days to recover. These two curves and the original provided by DGS (presented in Figure 3) will be used to fit epidemiological models in the next section.

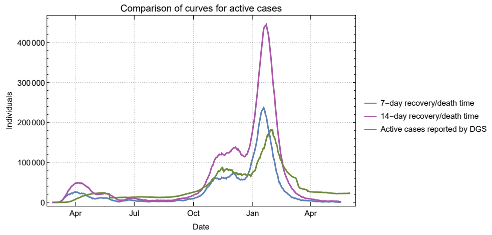
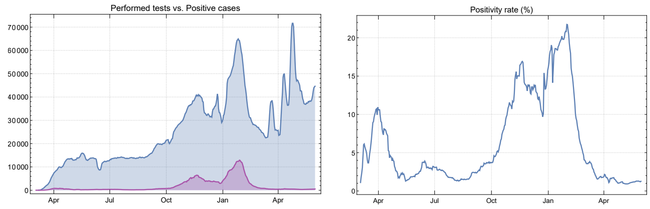


Figure 3: Curves for the number of active cases.

There are several indicators that can help us have a better understanding of how the pandemic is evolving. An important indicator that a new wave may be emerging is the positivity rate, the daily percentage of performed tests that are actually positive. If the percentage of positive tests starts increasing, it is a good indicator that the level of under reporting is increasing and consequently, more tests should be performed. During the weekend the number of performed tests is reduced given that some laboratories are only open on weekdays, and for this particular reason a 7-day moving average is performed to stabilise these weekly irregularities.

As it can be easily perceived on Figure 4(b), there were spikes in the positivity rate representing all three waves. This actively demonstrates that there was an higher number of unreported cases during these periods.

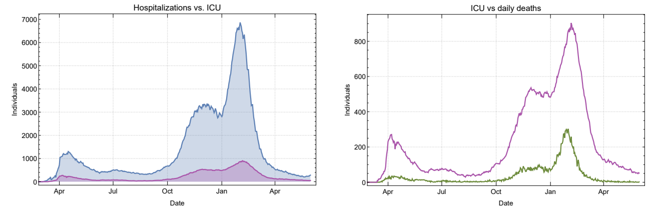
Another indicator that help us evaluate the pandemic is the number of individuals hospitalised. As we know, the number of resources available is limited, whether it is in terms of personnel or ventilators. Throughout time, the number of beds available in the ICU for COVID-19 patients has been adjusted according to the needs, but nonetheless there is a ceiling for the maximum capacity of critical beds. This value was reported to be 900 during the third wave of the pandemic, even though an extra 4 critical beds above this threshold were occupied by February 5^{th}.



(a) Performed tests and positive cases

(b) Positivity rate

Figure 4: On the left is the comparison between the number of performed tests (blue line) and the number of positive tests (purple line). On the right, the positivity rate (%) is presented.



(a) Hospitalisations vs. ICU

(b) ICU vs. daily deaths

Figure 5: The number of hospitalisations (infirmary and ICU) is presented by a blue line. The purple curve represents the number of individuals in intensive care units. The green line stands for the number of daily deaths.

Not only the population density is relevant when fighting a virus, but also how elderly the population is on different regions. Alentejo is the region with the most elderly population, followed by Centre of Portugal, as stated in [18]. The report [16] published in 2020 by the Organisation for Economic Co-operation and Development (OECD) declares Alentejo as the region with the lowest ratio of hospital beds per 1000 inhabitants, followed in order by Algarve, Centre of Portugal, North and Lisbon and Tagus Valley. It is expected that regions with worse scores on these factors to be more likely to have an higher case fatality rate (CFR).

To compute the daily CFR, we can consider a 14-day gap from the moment an individual tests positive for SARS-CoV-2 to death. With the goal of smoothing the data, a 7-day moving average is performed. For the sake of having more reliable computations, we will solely compute this ratio during the time period where the number of deaths were more significant, i.e. during the second and third wave. Our suspicions are confirmed when observing Figure 6, where the orange curve representing Alentejo stands out for the worst reasons. For the most part, Alentejo has the highest CFR of all regions achieving an extremely high value of

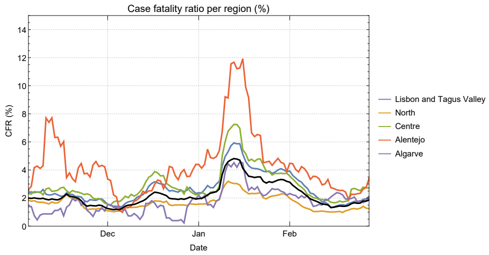


Figure 6: Case fatality ratio per region. The black curve represents the overall CFR in Portugal.

11.9213% on January 16th, 2021. The other regions have a CFR that does not differ as much, even though there is some relevance in pointing out that from all the curves, the Centre of Portugal scores higher as expected from the information in the previous paragraph.

3.2. Reproduction Number Computation

The reproduction number is a very important indicator in order to have a better understanding on how fast a pathogen is spreading. Hence, we must compute how many individuals does in fact an infected person transmits the virus.

3.2.1 Robert Koch Institute Formula

The Robert Koch Institute (RKI) published a report [9] with an empirical formula for the $\mathcal{R}_t$. This formula consists on studying the number of new daily cases, by considering a moving given window of time to check the number of new cases that emerge from the wave of infected from that previous time. The formula is as follows

$$\mathcal{R}_{t,\tau} = \frac{\sum_{i=t-\tau+1}^{t} E_i}{\sum_{i=t-\tau+1}^{t} E_{i-\nu}}, \tag{12}$$

where l the time lag corresponding to the latent period and τ represents the time an individual spends as infectious. The RKI assumes a time lag $\nu = 4$ and considers two possible scenarios, $\tau = 4$ or a more stable 7-day $\mathcal{R}_t$ value ($\tau = 7$). On Figure 7 is a visual representation of the formula, where for the computation of $\mathcal{R}_{t+11}$ it is considered that the new cases in the blue box originated from contacts with individuals that were initially infected during the purple period.

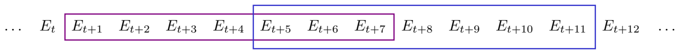
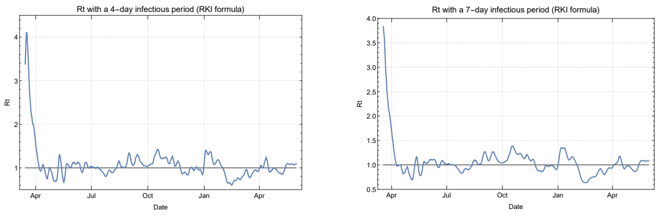


Figure 7: Representation of $\mathcal{R}_{t+11}$ with $\tau = 7$ and $\tau = 4$.

A big advantage of this formula is that allows the computation of the reproduction number of the current day.

On Figure 8 the reproduction number for each day is represented for an infectious period of $\tau = 4$ and $\tau = 7$, from left to right. As expected, a 7-day period of infectiousness creates a more stable and smoother $\mathcal{R}_t$. It is also visible when the pandemic waves occurred.



(a) $\mathcal{R}_t$ with $\tau = 4$

(b) $\mathcal{R}_t$ with $\tau = 7$

Figure 8: Reproduction number in Portugal using the RKI formula. The black line represents the important threshold of $\mathcal{R}_t = 1$.

3.2.2 Non-constant Viral Load Formula

Throughout this pandemic, we also resorted to a new empirical formula to compute the reproduction number. In a very similar manner, we took advantage of the new daily cases and a 7-day infectious period. The major difference with this formula relies on the fact that here we consider that SARS-CoV-2 viral load is not constant during the infectious period, and therefore an additional weight must be added. The viral load will follow a Gaussian distribution with its peak in the middle of the infectious period. The formula is

$$\mathcal{R}_t = \frac{\sum_{i=t-\omega}^{t+\omega} w_{i-t+4} E_i}{\sum_{i=t-\omega}^{t+\omega} w_{i-t+4} E_{i-\nu}}, \tag{13}$$

where E_i is the number of new cases on day i, ν is the number of days in the latency period, w_j with $j \in \{1, \ldots, \tau\}$ is the weight associated with j^{th} day of infection, and $\omega = \lfloor \frac{\tau}{2} \rfloor$, with τ an odd number (of days spent as infectious). A disadvantage when comparing equations (12) and (13) is that the latter will always compute the reproduction number with a $\lfloor \frac{\tau}{2} \rfloor$ days delay.

Similar to the RKI, a latency period of 4 days and an infectious period of 7 days will be considered. As for the percentage of viral load on each day, it will be considered a Gaussian distribution with mean value $\mu = 0$ and standard deviation $\sigma = 2$, leading to the weight vector $w \approx (0.325, 0.607, 0.882, 1, 0.882, 0.607, 0.325)$.

One major disadvantage of this formula is that it computes the reproduction number with a delay of 3 days, which can be an important factor when fighting a pandemic on real time. Since there are several variables that might influence the value of the reproduction number in real-world scenarios, a 10% error margin can be considered (see Figure 9).

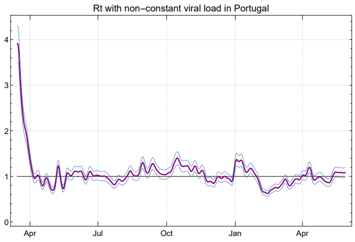


Figure 9: Reproduction number using the number of active cases and the viral load during a 7-day infectious period follows $\mathcal{N}(0, 4)$.

3.3. Model Adjustment

Here we will use the SIR and SEIRD model to fit into the data at our disposal. To properly measure which epidemiological model is the best fit, some error measurements [15] were considered (mean absolute percentage error (MAPE) and square errors metrics).

With the goal of adjusting the models discussed in Section 2, the class that we must try to replicate with the epidemiological models is the infected class. The models were fitted into the second (from September 13th to December 26th, 2020) and third wave (from December 26th, 2020 to February 28th, 2021) of the pandemic.

The functions were implemented on *Mathematica*, one for each model. At each time step, with a different set of parameters, we resorted to command *NDSolve* to find a numerical solution to the ordinary differential equations. The method used by this command is automatic, meaning that accordingly to the set of ODEs given, it will be chosen the method that best fits the problem. This step is followed by computing the error metrics (SSE, MSE, RMSE and MAPE) between the numerical solution obtained via *Mathematica* and the number of infected individuals in Portugal.

Finally, the function returns the models that had the best scores in terms of square errors and MAPE.

Last but not least, the SEIRD model requires the knowledge of how many individuals are already bearing the virus but as non-contagious (i.e. in the exposed class) for the initial condition of the ordinary differential equation. To solve this problem, we will consider that the number of individuals in the exposed class is a ratio of how many are on the infectious class. Thus, a variable p_e is going to represent the ratio between the number of individuals on the exposed class and the ones on the infectious class. This ratio will be tuned alongside the parameters of the model.

3.3.1 Second Wave

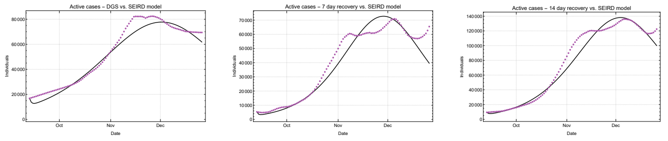
Focusing ourselves in the second wave, it lasts 105 days and accounts for a total of 4696 casualties during this time gap. For all three data sets, the curve does not have the common shape known of epidemiological models. It has a steady increase during a month and a half, but then it stabilises in a dangerous area, leaving open the possibility of major outbreak if a significant event were to disturb the sensitive system, which unfortunately did occur with the Christmas' celebrations.

Table 2: Parameters and error measurements for the models that best fit each data for the second wave.

		Parameters				Error Measurements	
		β	σ	ρ	p_e	RMSE	MAPE (%)
DGS	SIR	0.27	-	0.24	-	6446.64	9.56513
	SEIRD	0.53	0.42	0.44	0.3	3995.62	8.55783
		0.50	0.45	0.42	0.3	4736.89	8.29674
7-day	SIR	0.46	-	0.4	-	5181.58	17.5323
		0.39	-	0.34	-	8469.04	17.1954
	SEIRD	0.87	0.5	0.7	0.26	6776.08	12.0944
14-day	SIR	0.34	-	0.28	-	11205.5	21.2055
		0.29	-	0.24	-	12432.1	14.938
	SEIRD	0.52	0.47	0.4	0.25	8868.81	9.05434
		0.49	0.5	0.38	0.25	8984.05	8.5526

The first thing that stands out in Table 2 is how the SEIRD models scored better. This was expected since one of characteristics of SARS-CoV-2 is the existence of a latency period that varies from $2 - 5$ days. All these models indicate that the number of days before an individual becomes infectious $(1/\sigma)$ range from 2 to 2.38. As far as the infectious period goes in the SEIRD models, there is the estimation of values from 2.27 to 2.63 days for the DGS and 14-day data sets. The 7-day data set indicates a sightlier smaller time period of $1/\rho \approx 1.43$ days. The SIR model only has a period for the infected class, and consequently it was expected values for this time period at least equal to the ones attained for the infectious class in the model referred previously. This sentence can be corroborated by the values on Table 2, where the values for this time period range from 2.5 to 4.17 days.

One should also be aware when using the RMSE to compare different data sets. A downside of this error metric is the inability to adjust to different scales, resulting in a biased estimator that would give more importance to bigger values. Thus, when comparing the three data sets we must look at the MAPE scores. On Figure 10 are presented each data set and the model that obtained the best MAPE score.



(a) DGS (b) 7-day (c) 14-day

Figure 10: Visual representation (by a black line) of the best models presented in Table 2 for each data set in terms of MAPE. The data sets are represented by the purple points. The remaining models will be represented with the same color scheme.

With Figure 10, it becomes clear the dangerous period where the pandemic could start its downward trend or a new wave.

3.3.2 Third Wave

The third wave is shorter, but much deadlier. It lasts 65 days, and is responsible for around 9761 deaths, disregarding that some in the beginning still belong to the second wave and others occur after we consider this wave as over. In consequence of being the deadlier wave, the data sets that were generated from the daily deaths are a more accurate representation of a epidemiological curve of an outbreak, so better results are expected.

Similar to the previous wave, all the SEIRD models point to latency periods that range from 2 to 2.5 days. For the 7-day data set, the model seems to have a bit of trouble fitting since the values obtained for the remaining parameters differ quite a bit in comparison to the other data sets. In fact, this is the only scenario where the SIR model scored better in terms of MAPE. As for the infectious period, even though all point to relatively small time periods, there is a bit of discrepancies when comparing data sets. The SEIRD model in the DGS data set, states that a person remains infectious for an average of 1.82 days, whereas in the 7-day data set there is a slightly smaller infectious period of 1.43 and 1.47 days, depending on which error metric is used.

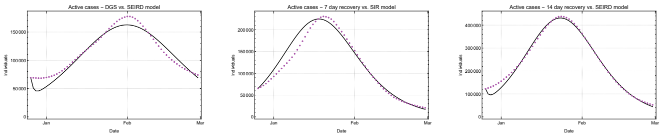
The model that scored the best out of the three data sets was the one were a 14-day recovery period was considered. This model states an infectious period of $1/\rho \approx 2.60$ and 2.56 days for the RMSE and MAPE best models, respectively. The only reason this model did not have a better score can be perceived on Figure 11(c). All SEIRD models had a hard time fitting the initial days of the third wave of the pandemic. The level of under report was not as high in the beginning of the wave as one expected and hence, the parameter p_e was of the utmost importance to regulate the initial number of infected individuals.

Even though the SEIRD model for the DGS data had quite a low score, one can notice that fitting the shape of the wave was not ideal. This did not happen with the other two data sets, corroborating once again the unreliability of the data provided by DGS.

A common occurrence on all results was the unexpected values of the low time an individual spends as infectious. The models return average infectious periods that vary from

Table 3: Parameters and error measurements for the models that best fit each data for the third wave.

		Parameters				Error Measurements	
		β	σ	ρ	p_e	RMSE	MAPE (%)
DGS	SIR	0.37	-	0.31	-	18169.5	13.9784
		0.34	-	0.29	-	20615.1	13.3031
	SEIRD	0.75	0.42	0.55	0.25	8394.9	6.87465
		0.75	0.4	0.55	0.3	8692.67	6.82885
7-day	SIR	0.54	-	0.41	-	12710.5	10.787
		0.56	-	0.42	-	14111.8	8.5582
	SEIRD	1.1	0.49	0.7	0.4	9607.74	11.4101
		1.05	0.5	0.68	0.4	10085.1	9.58651
14-day	SIR	0.366	-	0.256	-	32499.4	16.4994
		0.388	-	0.27	-	38439.2	12.8535
	SEIRD	0.66	0.415	0.385	0.18	11635.3	4.86516
		0.67	0.41	0.39	0.18	11724	4.64901



(a) DGS (b) 7-day (c) 14-day

Figure 11: Visual representation of the best models presented in Table 3 for each data set in terms of MAPE.

one to three days. From scientific research on the virus, this value should be higher, around fourteen days in most cases even though the viral load might not be sufficiently high during this whole period to make an effective infection. This phenomenon can be explained by people's behaviour. In real-world scenarios, with a disease as dangerous as COVID-19, once a person discovers that it is infected, the normal course of action is to isolate himself in order to avoid infecting others that could potentially lead to severe diseases/complications and eventually death. Consequently, this individual will only be at risk of infecting others for the first three days, until he realises that he carries the virus.

Nevertheless, the results here achieved, particularly the last scenario, are remarkable taking into consideration that the curves we were adjusting our models into have a lot of observational error associated.

3.4. Herd Immunity Threshold

Hopefully, vaccines for infectious diseases are able to be fabricated in order to save a great number of lives. As discussed in Subsection 2.3, the percentage threshold of the population that must be immune, either by infection or vaccination, allows a better understanding of the disease and consequently, governments can have the most educated decisions in regards of fighting the pandemic.

As mentioned previously, to compute the herd immunity threshold one can use of the basic reproduction numbers computed in Subsection 3.2, or one may use equation (9) if the values σ and ρ are known. For the latter and similar to Subsection 3.2, we will consider a latency period of 4 days and an infectious period of 4 and 7 days. We will also take into account the best results obtained in the previous Subsection for each wave.

Table 4: Basic reproduction number and herd immunity values.

Input Parameters		Results		
σ	ρ	$\mathcal{R}_0$	$\mathcal{H}$	$\mathcal{H}_{adj}$
1/4	1/7	10.7097	0.906627	1
1/4	1/4	7.27586	0.862559	0.980181
0.45	0.42	3.90608	0.743989	0.845442
0.41	0.39	4.24918	0.764661	0.868933
-	-	3.38217	0.704331	0.800377
-	-	3.83298	0.739106	0.839894
-	-	3.91142	0.744338	0.845839

All computations indicate that at least 70% of the population must be immune to the virus if there were no variants besides the original one. With the Delta variant, the previous threshold rises to at least 80%, requiring an extra effort on behalf of national health organizations to achieve such goal.

By the end of May, a total of 809.135 individuals have been reported to recover from the disease and at this point, 1.987.389 people have been fully vaccinated. This represents approximately 27% of individuals immune to the virus, which is nowhere near enough to the herd immunity threshold and to allow the population to fully return to the lifestyle one had before COVID-19.

Note: After we closed our study, a number of nearly 85% individuals were vaccinated which is very close to the number we believe to be safe to return to our almost regular pre pandemic lives.

4. Proposal of a New Model

An alarming number of new cases occurred on a daily basis during the fall and winter of 2020/21, leading to increasingly crowded hospitals. The consequences of overcrowded hospitals include shortage of medical beds, delays in laboratory tests, shortage of healthcare professionals, increased waiting times, higher mortality, emotional/physical exhaustion of health care professionals and not to mention economic costs. The creation of a model that considers hospitalised patients was necessary.

4.1. SEIHCRD Model

In order to create the desired model, two new classes were added to the preexisting model SEIRD: hospitalised (H) and critical (C) individuals. Please note that even though patients in critical conditions (ICU) are indeed hospitalised, here we consider the class of hospitalised formed only by those in the infirmary. In contrast to the previously mentioned models, more variables must be considered to correctly represent this new model (see Table 5). The compartment flowchart for the model is displayed in Figure 12, where the scenario with overcrowded hospitals, specifically saturation of ICUs, is taken into account by modifying the flow between classes. This model will consider infected individuals to move directly to the critical class, and then

Table 5: Summary of notation for the SEIHCRD model.

β	Transmission rate
$1/\sigma$	Average latent period
$1/\rho$	Average infectious period outside the hospital
$1/\tau$	Average period spent in the infirmary
$1/\gamma$	Average period spent in the ICU
T_h	Rate of individuals in the hospital (infirmary and ICU)
T_c	Rate of ICU patients from the ones in the hospital
L_h	Lethality in the infirmary
$L_{\overline{h}}$	Lethality without attending the hospital
L_c	Lethality in ICU

move either to the infirmary or the deceased class. Susceptible individuals are automatically considered as recovered if they are vaccine-immune to the virus (transitioning from class S to class R). The rate at which the vaccination occurs is explained by $f(t, S(t))$.

Similarly to models that take into account the latency period, an extra parameter hides beneath the system when fitting it to real data. Generally, during the latency period it is unbeknownst to the individual the bearing of such virus, making it hard for the scientific community to accurately predict the number of exposed individuals. Even though this model is a better representation of the evolution of a virus in the modern day society, it will have other troublesome problems of its own, including an high number of parameters to tune leading to computationally expensive algorithms.

The model is modelled by the system of equations

$$\begin{cases} S'(t) = -\beta S(t)I(t) - f(t, S(t)) \\ E'(t) = \beta S(t)I(t) - \sigma E(t) \\ I'(t) = \sigma E(t) - \rho I(t) \\ H'(t) = T_h(1 - T_c)\rho I(t) + \\ \qquad\qquad \gamma(1 - L_c)C(t) - \tau H(t) \\ C'(t) = (1 - Incr(t)Sat(t))[T_h T_c \rho I(t) - \gamma C(t)] \\ R'(t) = (1 - T_h)(1 - L_{\overline{h}})\rho I(t) + \\ \qquad\qquad (1 - L_h)\tau H(t) + f(S(t), t) \\ D'(t) = (1 - T_h)L_{\overline{h}}\rho I(t) + \\ \qquad\qquad L_h \tau H(t) + \gamma L_c C(t) + \\ \qquad\qquad Incr(t)Sat(t)[T_h T_c \rho I(t) - \gamma C(t)], \end{cases}$$

where N_{cb} is the number of beds available in the ICU. Function $Sat(t)$ will be responsible for monitoring if there is saturation in the intensive care units. If all beds destined to COVID-19 patients (N_{cb}) are occupied, the function will return 1, otherwise $Sat(t) = 0$. As for function $Incr(t)$, it will indicate whether the number of ICU patients is in an downward trend ($Incr(t) = 0$) or in a upward trend ($Incr(t) = 1$). The previous functions are represented as sigmoids to avoid discontinuity points

$$Sat(t) = \frac{1}{1 + e^{-10(NC(t) - N_{cb})}} \tag{14}$$

$$Incr(t) = \frac{1}{1 + e^{-10^{10}(T_h T_c \rho I(t) - \gamma C(t))}}. \tag{15}$$

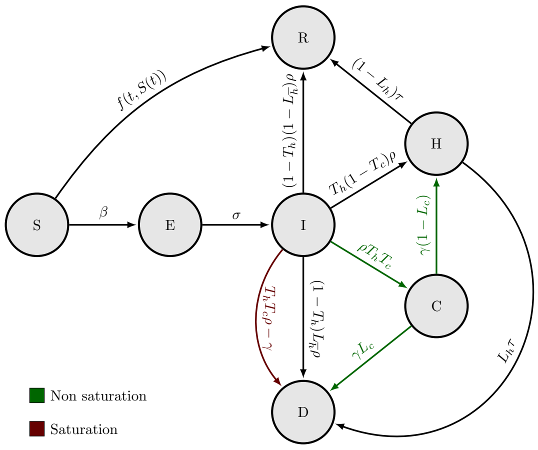


Figure 12: Compartment flowchart for SEIHCRD model.

Finally, to have a mathematically well defined system of ODEs, initial conditions $S(0)$, $E(0)$, $I(0)$, $H(0)$, $C(0)$, $R(0)$ and $D(0)$ must be established. The total population size N remains constant for all time t, and can be obtained by adding all classes.

4.2. Model Adjustment

Before starting with model fitting and its results, it must be pointed out that only the third wave will be taken into consideration here since it is the one with higher values of hospitalisations, ICU admissions and deaths. During this time period, some vaccines had already been administrated, specially to health care workers, however it was a small portion of individuals and consequently, the function $f(t, S(t))$ will be disregarded. Once again, we will resort to the three data sets of infected individuals at our disposal.

Regarding error measurements, there is a need to use a metric that incorporates the fit of all three curves simultaneously. This is where the square errors metrics lose relevance on account of disregarding the scale of each curve. Therefore, the only observational error here used will be the mean absolute percentage error (MAPE). We will compute this value for all curves with a given set of parameters and then proceed to calculate the mean of these three values. The chosen set of parameters shall be the one with the lowest mean MAPE value.

Considering the large amount of parameters that can be tuned, we will take advantage of the best set of parameters obtained for the SEIRD model in the Subsection 3.3 for each dataset. Nonetheless, that still leaves us with seven variables to adjust. At this point, information from scientific articles is extremely relevant, allowing us to start with a more or less accurate interval for where the parameter value must belong.

Two particular parameter need some extra attention. The way our model was created, there is not a constant to regulate how much it takes an individual to leave the infected class in case he is moving to the infirmary or to the ICU. The parameter ρ only controls how much time it takes to leave class I, disregarding where the person is heading to. One solution to overcome this obstacle is to write $T_h = T'_h t_h \frac{1}{\rho}$, where T'_h represents the percentage of infected individuals that will go to the hospital (infirmary and ICU) and $\frac{1}{t_h}$ is the time it takes for an individual to leave the infected class if moving to the infirmary. A similar process will occur for the value of T_c, except in this case it will help us separate the infirmary from the intensive care units and we will have $T_c = T'_c t_c \frac{1}{t_h}$. The nomenclature follows the same pattern: $\frac{1}{t_c}$ is the time it takes to leave the infected class to go to the ICU and T'_c is a fraction of T'_h representing the number of hospitalised individuals that will go to ICUs. The values, T'_h and T'_c, for each data set are presented on Table 6. These values were obtained by taking the mean of each ratio during the third wave.

Table 6: Mean of T'_h and T'_c during the third wave period.

	DGS data	7-day data	14-day data
T'_h	0.0393962	0.0614479	0.0254821
T'_c	0.154776	0.154776	0.154776

Excluding the parameters obtained from the best fittings of SEIRD models in Table 3, the best set of parameters for each dataset is presented on Table 7. The error metrics obtained for each model are presented on Table 8.

Table 7: Parameters of the best SEIHCRD model for each data set.

	τ	γ	T_h	T_c	L_h	$L_{\overline{h}}$	L_c
DGS	0.285714	0.203008	0.0256451	0.108343	0.077	0.0005	0.4
7-day	0.0916667	0.04	0.00451823	0.0742926	0.14	0.0011	0.48
14-day	0.11	0.05	0.00458367	0.0773881	0.08	0.0011	0.49

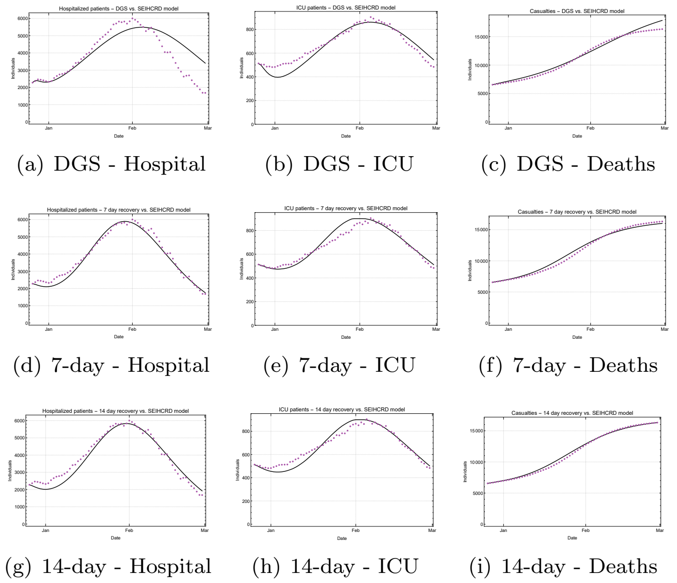
The error measurements obtained for the DGS data are quite good, except the hospitalised curve that scored an 19% error. Taking a closer look on Figure 13(a), the slightly poor adjustment on the hospitalisations' curve is confirmed. This curve has a particular hard time fitting into the real data regardless of the imputed parameters. As discussed previously, the number of active cases provided by DGS are severely affected by under reporting, especially in the beginning of the third wave, leading to an uneven curve with different levels of under report throughout time. For this particular reason, the parameters have trouble tuning since there are a lot of irregularities.

Table 8: Mean absolute percentage error for the three relevant classes and its mean value, with the parameters from Table 7.

	Hospitalised	Critical	Deceased	Mean value
DGS	19.1644	7.04538	3.18098	9.79691
7-day	5.2134	4.11551	2.86027	4.06306
14-day	8.00091	4.61049	1.80115	4.80419

It must be also pointed out that even though the MAPE for the deceased curve in the DGS data set is very satisfactory, it can be perceived (see Figure 13(c)) that the model is increasing faster in the last few days than the real number of deaths, a direct consequence of the poor adjustment of the infected curve.

The data sets built from the number of deaths achieve a very low score ($\leq 5\%$) on the MAPE considering the amount of bias in the data (for example irregularities on the number of tests performed and uneven population density). The good fit of such models is confirmed when inspecting Figures 13(d) to 13(i).



(a) DGS - Hospital (b) DGS - ICU (c) DGS - Deaths

(d) 7-day - Hospital (e) 7-day - ICU (f) 7-day - Deaths

(g) 14-day - Hospital (h) 14-day - ICU (i) 14-day - Deaths

Figure 13: Visual representation each model presented in Table 7 for the three curves of interest.

The admission time in infirmary and ICU is overestimated. This is a problem of this method. Nevertheless, the second and third modelation in this section are very good to estimate the actual numbers observed and can be used in the future to predict pandemic waves.

All three fittings have mortality rates (in the infirmary and ICU) within the expected range. The model attained from the DGS data points to 3.5 days spent in the infirmary and close to 5 days in the ICU, whereas the other two data sets indicate a longer stay in the infirmary and ICU, around 10 and $20 - 25$ days, respectively.

One of the biggest struggles in this section was the amount of parameters that needed tuning and the different levels of sensitiveness. While in the previous section, one had few parameters to tune and it was possible to try a great amount of combinations with a given error, here it was not possible to do so in a reasonable amount of time.

5. Conclusions and Future Work

The computation of reproduction number, $\mathcal{R}_t$, was extremely important to understand on a daily basis how the pandemic was evolving. Later on, these estimations were helpful to compute the herd immunity threshold. All values attained for $\mathcal{H}_{adj}$ were very close to the 85% threshold. The Portuguese government established, after we closed our study, this threshold to be the goal of vaccines administrated in order to start returning to life pre COVID-19.

From the two models fitted into the data, SIR and SEIRD model, the latter was the one with the best results. The major difference from these models is that the SEIRD model accounts for individuals in the latency period, which is a characteristic of SARS-CoV-2, explaining why it performed better. On Section 3, we were able to obtain models with a mean absolute percentage error lower than 5% which is remarkable considering the amount of bias in the data. Not only the irregularities in the number of tests performed (due to closed laboratories and lack of testing) took a toll on the reliability of the data, but also factors as uneven population density and more importantly, how unpredictable people's behaviour can be.

Regarding the latency period, all SEIRD models indicate a period of 1.5 to 4 days, which are results that confirm the literature on the disease.

As far as the model proposed in Section 4, the results are quite satisfactory. The mean MAPE of the three curves of interest was particularly low (smaller than 5%) for the two data sets built from the number of deaths. Since the fit of the number of infected individuals for these models came from Subsection 3.3, we are able to obtain models with extremely good fits on four curves (infected, hospitalised, critical and deaths). Nonetheless, unexpected results arose when computing the time it takes to arrive to the hospital and LoS in the ICUs. We were not able to fully grasp why these values were so high, but one should always keep in mind the amount of errors associated in the process, namely under report of cases, oscillation in the number of daily tests performed and population density.

The major result that should be captured from this work when a similar pandemic falls upon us is the need for a daily study on the evolution of the virus on all indicators.

The fitting of the models was made on a trial and error, i.e. with a given possibilities for a each parameter, the function would try all scenarios and choose the one with the best score. This type of programming is extremely expensive, which explains the problems we had on the last section. The enhancement of these functions should be made.

Another direction of work relies with the use of dynamic models, where the transmission rate β varies throughout time. This type of dynamism could be very useful to take advantage on real world scenarios since the rate at which a disease spreads can depend on numerous factors.

If there was further information on each infected individual, a wide range of possibilities to analyse the data would emerge. With information regarding symptoms, time spent in the infirmary and ICU, age, preexisting medical conditions, and so on, one could take advantage of machine learning algorithms to check, for example, which variables have an higher relevance when predicting which individuals will attend the infirmary, ICU or even death. With this type of information, the door to the field of survival analysis would also open.

References

[1] M. J. Beira and P. J. Sebastião. "A differential equations model-fitting analysis of COVID-19 epidemiological data to explain multi-wave dynamics". In: *Scientific Reports* 11.1 (2021). DOI: 10.1038/s41598-021-95494-6.

[2] M. Driscoll et al. "Age-specific mortality and immunity patterns of SARS-CoV-2 infection in 45 countries". In: 590 (Nov. 2020), pp. 140–145. DOI: https://doi.org/10.1038/s41586-020-2918-0. URL: https://www.nature.com/articles/s41586-020-2918-0.

[3] H. W. Hethcote. *A Thousand and One Epidemic Models.* Ed. by Simon A. Levin. Berlin, Heidelberg: Springer Berlin Heidelberg, 1994, pp. 504–515.

[4] H. W. Hethcote. "An age-structured model for pertussis transmission". In: *Mathematical Biosciences* 145.2 (1997), pp. 89–136. ISSN: 0025-5564. DOI: https://doi.org/10.1016/S0025-5564(97)00014-X. URL: https://www.sciencedirect.com/science/article/pii/S002555649700014X.

[5] H. W. Hethcote. "The mathematics of infectious diseases". In: *SIAM Review* 42 (2000), pp. 599–653.

[6] H. W. Hethcote. *Three Basic Epidemiological Models.* Ed. by Simon A. Levin, Thomas G. Hallam, and Louis J. Gross. Berlin, Heidelberg: Springer Berlin Heidelberg, 1989, pp. 119–144. ISBN: 978-3-642-61317-3. DOI: 10.1007/978-3-642-61317-3_5.

[7] H. Higgs and C. H. Hull. "The Economic Writings of Sir William Petty, Together with the Observations upon the Bills of Mortality, more Probably by Captain John Graunt." In: *The Economic Journal* 9.36 (1899), p. 564. DOI: 10.2307/2956578.

[8] M. W. Hirsch, S. Smale, and R. L. Devaney. *Differential Equations, Dynamical Systems, and an Introduction to Chaos.* English. 3rd. Elsevier Inc., 2013. ISBN: 9780123820105.

[9] Robert Kock Institute. *Nowcasting and R estimate: Estimation of the current development of the SARS-CoV-2 epidemic in Germany.* https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Projekte_RKI/Nowcasting.html. (Last accessed on 03/06/2021). May 2020.

[10] W. Kermack and A. McKendrick. "A Contribution to the Mathematical Theory of Epidemics". In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 115.772 (1927), pp. 700–721. ISSN: 09501207.

[11] W. Kermack and A. McKendrick. "Contributions to the mathematical theory of epidemics—I". In: *Bulletin of Mathematical Biology* 53.1-2 (1991), pp. 33–55. DOI: 10.1016/s0092-8240(05)80040-0.

[12] W. Kermack and A. McKendrick. "Contributions to the mathematical theory of epidemics—II. the problem of endemicity". In: *Bulletin of Mathematical Biology* 53.1-2 (1991), pp. 57–87. DOI: 10.1016/s0092-8240(05)80041-2.

[13] W. Kermack and A. McKendrick. "Contributions to the mathematical theory of epidemics—III. Further studies of the problem of endemicity". In: *Bulletin of Mathematical Biology* 53.1-2 (1991), pp. 89–118. DOI: 10.1016/s0092-8240(05)80042-4.

[14] M. Martcheva. *An Introduction to Mathematical Epidemiology.* 1st. Springer Publishing Company, Incorporated, 2015. ISBN: 1489976116.

[15] J. Neter et al. *Applied Linear Statistical Models.* Irwin series in statistics. Irwin, 1996. ISBN: 9780256117363. URL: https://books.google.pt/books?id=q2sPAQAAMAAJ.

[16] OECD. *OECD Regions and Cities at a Glance 2020 - Country Note: Portugal.* 2020. DOI: https://doi.org/10.1787/959d5ba0-en. URL: https://www.oecd.org/cfe/oecd-regions-and-cities-at-a-glance-26173212.htm.

[17] M. Peirlinck et al. "Outbreak dynamics of COVID-19 in China and the United States". English. In: *Biomechanics and modeling in mechanobiology* 19.6 (Dec. 2020). DOI: 10.1007/s10237-020-01332-5.

[18] PORDATA. *População residente: total e por grandes grupos etários (%).* https://www.pordata.pt/Municipios/Popula%C3%A7%C3%A3o+residente+total+e+por+grandes+grupos+et%C3%A1rios+(percentagem)-726. (Accessed on 07/08/2021). June 2021.

[19] H. E. Randolph and L. B. Barreiro. "Herd Immunity: Understanding COVID-19". In: *Immunity* 52.5 (2020), pp. 737–741. DOI: 10.1016/j.immuni.2020.04.012.

[20] B. D. Rodrigues. *Dados relativos à pandemia COVID-19 em Portugal.* https://github.com/dssg-pt/covid19pt-data. 2020.

[21] E. Vynnycky and R. G. White. *An introduction to infectious disease modelling.* Oxford Univ. Press, 2011.

[22] S. Zhao. "Estimating the time interval between transmission generations when negative values occur in the serial interval data: Using COVID-19 as an example". English. In: *Mathematical Biosciences and Engineering* 17.4 (May 2020). DOI: 10.3934/MBE.2020198.