Epidemiological Models: SARS-CoV-2 in Portugal

Maria Beatriz Silva Santiago

Thesis to obtain the Master of Science Degree in

Mathematics and Applications

Supervisor: Prof. Henrique Manuel dos Santos Silveira de Oliveira

Examination Committee

Chairperson:	Prof. António Manuel Pacheco Pires
Supervisor:	Prof. Henrique Manuel dos Santos Silveira de Oliveira
Members of the Committee:	Prof. José Rui De Matos Figueira

December 2021

To my late grandparents.

Acknowledgements

First and foremost, I would like to thank my supervisor, Professor Henrique Oliveira, whose expertise was invaluable for the accomplishments presented in this thesis. Your help, motivation, guidance and tireless commitment are the main reason this work is still standing. It was a privilege and an honour to have worked with someone not only with exceptional scientific knowledge but human qualities.

I would also like to thank my family for all of the unconditional support throughout my academical path. In particular, I would like to express my deepest heartfelt gratitude to my (step)parents for always believing in me and without whom none of this would have been possible. To my siblings, Mariana and Afonso, I wish to thank for all the memories, laughs, arguments and so much more. I will cherish them forever. To my grandparents, Mariana and José, thank you for the love, generosity and for always making sure I never leave your house hungry. Last but not least, to my late grandparents, Alzira and António, whom I hope to have made proud.

I am deeply grateful to have so many people in my life who wish me well and helped in my life. To all my friends, I would like to thank from the bottom of my heart for always having my back and being there for me when I needed the most. To my *flores*, André Brito and Alice Carvalho, thank you for being there from day one and inspiring me by following your dreams. To Tiago Santos for being a shoulder I can always depend on. To Andreia Bastos for your amazing advices and for being the best unexpected friendship one could have asked for. To *alentejano* Bruno Melgão for bringing truckloads of joy and happiness in my life. I am deeply thankful for all the moments the academic life provided and for my second family, Tomás Freire and Rodrigo Girão. To my uncle Francisco Silva and bestie Miguel Barata, I have to thank for all of your patience and time spent helping me study.

To my friends from back home, I am deeply sorry for being less present during these years and please know I am deeply thankful for having all of you in my life. Nonetheless, I have to mention Pedro Bigodinho, Maria Duarte and Carlos Fresco for understanding me like no other. Finally, I would like to thank my high school teacher, Ana Antunes, for making me fall in love with mathematics during those six years.

To all of those not here mentioned but that had an impact on my life, positive or negative, thank you regardless for you have helped me shape into the person I am today.

Abstract

Mathematical models are key tools to guide public health measures, and the field of epidemiological modelling must play an intrinsic part when deciding how to tackle a pandemic. The complexity of epidemiological models varies. The approach chosen by epidemiologists deeply depends on how much is known about the disease, purpose of the study and the amount/quality of the data available. Despite unavoidable uncertainties and errors associated with the data, models can reveal extremely important information such as predictions on the number of infected and casualties, how vaccination efforts will impact the disease's development, and how one can limit the spread of the disease.

Over the last decades remarkable developments have been made in the field and with the global phenomenon of COVID-19, the opportunity to study a pandemic in the modern days arose. In this work a thorough analysis of the evolution of SARS-CoV-2 took place. 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 intensive care units is the main contribution of this work.

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

Resumo

Modelos matemáticos são ferramentas chave para gerir medidas de saúde pública e como tal, a área de modelação epidemiológica deverá ser uma parte intrínseca no processo de decisão de como enfrentar uma pandemia. A complexidade dos modelos epidemiológicos varia e consequentemente, a abordagem escolhida pelos epidemiologistas irá depender de diversos fatores. Os mais importantes incluem o quão vasto é o conhecimento sobre a doença, o objetivo do estudo e a quantidade/qualidade dos dados disponíveis. Apesar de incertezas inevitáveis e erros associados, os modelos podem revelar informações extremamente importantes, como previsão sobre o número de infetados e mortes, quão eficazes serão os esforços de vacinação e como se pode limitar a propagação do vírus.

Durante as últimas décadas, avanços notáveis foram feitos na área e com o fenómeno global da doença COVID-19, surgiu a oportunidade de estudar uma pandeia nos tempos modernos. Neste trabalho foi feita uma análise minuciosa à evolução do vírus SARS-CoV-2 em Portugal. Esta análise inclui o cálculo da percentagem de casos por reportar, da taxa de positividade, da letalidade e do índice de transmissibilidade acompanhado por uma nova fórmula que considera uma carga viral não constante ao longo do tempo.

Por fim, houve ajuste de modelos epidemiológicos à segunda e terceira onda da pandemia em Porugal. Para tal recorreu-se aos modelos SIR e SEIRD, em que o último obteve resultados significativamente melhores. Quanto às métricas de erro foram utilizados erros quadrados e erro médio percentual absoluto. Para uma mais completa análise, este processo foi também realizado por região e por sexo.

Durante a pandemia, o número de pacientes admitidos em hospitais aumentava de forma perigosa, levando os sistemas nacionais de saúde a enfrentar um teste nunca antes visto. Hospitais sobrelotados têm consequências gravíssimas. Tudo isto culmina na introdução de um novo modelo que considera o número de indivíduos nas enfermermarias e cuidados intensivos, que é a principal contribuição deste trabalho.

Palavras-Chave: SARS-CoV-2, Modelo SIR, Modelo SEIRD, Índice de Transmissibilidade

Contents

List of Figures

List of Tables

Glossary

SUSCEPTIBLE An individual who has not yet been infected and is at risk of infection.

EFFECTIVE CONTACT A contact that is sufficient to lead to transmission. Occurs between an infectious and susceptible individual.

EXPOSED A susceptible individual that has a potentially-transmitting contact. Exposed individual may or may not develop the disease. Typically, these individuals are non infectious.

INFECTED Referring to an exposed individual in which the pathogen has established itself.

INFECTIOUS Infected individuals who can transmit the disease to others.

ASYMPTOMATIC An infected person who does not have any recognised signs of the disease.

LATENT PERIOD Time from infection to when the host can transmit the agent to others.

INCUBATION PERIOD The time period between infection and onset of clinical symptoms of the disease.

INCIDENCE Number of new infections in a specified interval of time. Occasionally, incidence is the number of individuals who contract the disease per unit time and per unit person.

PREVALENCE Number of individuals that have the disease at a specific time. It can also be written as the number of infected people per unit person.

FORCE OF INFECTION The rate at which susceptible individuals become infected per time unit.

REPRODUCTION NUMBER Number of secondary cases produced by a single infectious individual in a population consisting only of susceptible individuals.

HERD IMMUNITY THRESHOLD The indirect protection from infection conferred to susceptible individuals when a sufficiently large proportion of immune individuals exist in a population.

EFFICACY (VACCINE) The direct protection provided by vaccination against infection or other outcome of interest. It excludes any indirect herd effect.

IMMUNE A person who has complete protection to an infection, which results from either vaccination or previous infection. Individuals are said to be *partially immune* if they are not fully protected.

OUTBREAK A time when something suddenly begins, especially a disease or something else dangerous or unpleasant.

Acronyms

SI	susceptible-infectious model
SIS	susceptible-infectious-susceptible model
SIR	susceptible-infectious-recovered model
SEIRS	susceptible-exposed-infectious-recovered-suscpetible model
SEIRD	susceptible-exposed-infectious-recovered-deceased model
HIV	human immunodeficiency virus
SARS-CoV-2	severe acute respiratory syndrome coronavirus 2
COVID-19	coronavirus disease 2019
ICU	intensive care unit
IFR	infection fatality ratio
CFR	case fatality ratio
LoS	length of stay
ODE	ordinary differential equation
SSE	sum of squared errors
MSE	mean squared error
RMSE	root mean squared error
MAPE	mean absolute percentage error

List of Symbols

N Population size

S Susceptible individuals

E Exposed individuals

I Infected individuals

H Hospitalised (infirmary) individuals

C Critical (ICU) individuals

R Recovered individuals

D Deceased individuals

$\mathcal{R}_0$ Basic reproduction number

$\mathcal{R}_t$ Effective reproduction number

$\mathcal{H}$ Herd immunity threshold

Chapter 1

Introduction

1.1 Motivation

Since its beginnings, mathematics aims to observe and formalise real world issues. And now, more than ever, with the global phenomenon of COVID-19, the study of epidemiological diseases is of the utmost importance.

Epidemiology studies the patterns and causes/risk factors of health-related states and events in specified populations, tacitly the spread of infectious viruses and its consequences in human populations. The top global causes of deaths are associated with cardiovascular malfunctions such as heart disease and stroke [42], which positions diseases that do not transmit from one person to another the main concern of epidemiology. Lower respiratory infections (tuberculosis or pneumonia, for example) and HIV are among the deadliest infectious diseases.

Throughout the years humanity has faced several infectious diseases, which are caused by the presence of a pathogenic microbial agent. This agent can be bacterial, viral, fungal, parasitic or toxic proteins (commonly named as prions).

The latest respiratory infectious disease caught everyone off guard. Suddenly, global economies had to come to an alt and people's livelihood were in danger. The governments had to turn to epidemiologists to properly respond to such an unusual situation.

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.

Having said this, 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, specifically certain indicators that one should be aware of.

1.2 Historical Remarks on Infectious Diseases

The first plague ever described by historians was the Plague of Athens ($430-426$ BCE), which occurred during the Peloponnesian War and ravaged the city of Athens, killing nearly two thirds of the Athenians' population. The scientific historian Thucydides ($460-400$ BCE), who himself suffered the disease, vividly describes in his book [56] (*The History of the Peloponnesian War*) the symptoms, the evolution and the number of deaths caused by such plague, stating that "...a pestilence of such extent and mortality was nowhere remembered.". Thucydides' precise description remained influential for centuries and it is considered a landmark in epidemic history.

The Black Death ($1347-1352$) was one of the most significant and devastating events occurred in the late medieval ages, and it is considered by some historians as a turning point in European history. Overall mortality estimates are extremely difficult to obtain, taking into account that such areas did not kept consistent/accurate census nor burial reports. Nevertheless, continuous efforts to approximate those estimates from historians working on multiple locations yield a mortality rate ranging from 45 to 60 percent across the affected areas. Over the following centuries, Europeans suffered more outbreaks from this bubonic plague.

A devastating pandemic in history was the $1918-1919$ flu pandemic, commonly known as the Spanish flu. With its highly morbidity rate, it was responsible for over 40 million casualties. This was the first time due to a virus that citizens were ordered to use masks to protect themselves and others.

Up to the present time there are a few other significant pandemics worth mentioning. The HIV/AIDS in the 1980's and early 1990's, and with no vaccine to eradicate the disease, it is still a major global public health issue, having claimed over 33 million lives so far, and according to the World Health Organization an estimate of 38 million people were still living with HIV by the end of 2019 [41]. The 2003 severe acute respiratory syndrome (SARS) affected 8.000 people worldwide and killing close to 800, leading to the discovery of a new strain of coronavirus called SARS-CoV. More recently, another coronavirus outbreak took place in 2012 in the Middle East, naming the virus as Middle East respiratory syndrome (MERS-CoV). There was also the H1N1 swine flu pandemic (2009), a subtype of Influenza A virus.

Please note that the previous information was extracted from [6, 7, 9].

1.3 State of the Art

Even though epidemics has been around for thousands of years, its mathematical study is a relatively new field of research. The first statistical study developed in this area was in 1662 by a respected businessman in London, John Graunt. His small book concerned public health statistics with the analysis of *bills* - weekly records of diseases and casualties made public. The 5^{th} edition [24] of his study was published in 1899.

What is now considered the first epidemiological model would only be published over a century later in 1766 by one of the greatest scientist of the 18^{th} century, Daniel Bernoulli. The main question of his work [5] was whether inoculation (the voluntary introduction of a small amount of the less virulent smallpox in the body to protect it against later infections) should be encouraged in order to reduce the death rate and thereby increase the population, even if the inoculation itself is at times a deadly operation.

There were other valuable contributions to the understanding of infectious diseases before there was knowledge about the transmission process of such diseases, namely [53] the study of the temporal and spatial pattern of cholera cases in London in 1855 by John Snow, who was able to pinpoint the source of infection to the Broad Street water pump. In 1873, with [8] William Budd also achieved a similar understanding concerning the spread of the typhoid. Another relevant study to mention was a letter [14] submitted by William Farr to the *Annual Report of the Registrar General of Births, Deaths and Marriages in England* in 1840, that was focused on the laws that underlie the rise and fall of epidemics.

In the 19^{th} and early 20^{th} century remarkable breakthroughs concerning causes and prevention of diseases were accomplished, which provided a direct support to the germ theory of disease. Louis Pasteur reduced mortality from puerperal fever and created the first vaccines for rabies and anthrax. Robert Koch, a German physician and microbiologist, not only identified the specific causative agents of tuberculosis, cholera and anthrax but also provided experimental support for the concept of infectious disease. He is also notable for the development of Koch's postulates. At this point in time, science could finally comprehend the mechanism of how an individual became ill, paving the way for mathematical modelling of infectious diseases. Major progress was made in this area from this point on.

In 1906, William H. Hamer proposed [17] where he stated that the spread of infection should depend on the number of susceptible and infected individuals. Even though Hamer was the first to suggest a mass action law for the rate of new infections, the father of modern mathematical epidemiology is Sir Ronald Ross. He won the Nobel Prize in Medicine with his pioneer work on the life cycle of the malaria parasite, proposing a simple compartmental model in which he showed that the reduction of the mosquito population below a critical level (known nowadays as basic reproduction number) would be sufficient to control malaria [50]. In 1929, the work of William Hamer was followed up by Herbert E. Soper in [54], providing a treatment of the periodicity of measles epidemics only departing in detail from Hamer's method.

Rising the level of mathematical modelling of infectious diseases and likely the most influential contribution to this area is the paper published by Kermack and McKendrick in 1927. Their joint paper [29] is cited in a great number of papers and is the first to consider a deterministic epidemic model that takes into account susceptible, infected and removed individuals. This model disregards natural births and deaths and for this particular reason it only models outbreaks. In order to capture diseases that established themselves and persist in population, Part II and Part III of *A contribution to the mathematical theory of epidemics* were published in 1932 and 1933, respectively. Due to their groundbreaking work in mathematical epidemiology, the Kermack-McKendrick trilogy was reprinted in 1991 [30, 31, 32].

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 [25] published originally in the 70's as a basic text in qualitative theory.

In the 80's and early 90's, the study of modelling mathematical diseases increased rapidly with the outbreak of HIV/AIDS across the world, though the disease originated decades earlier.

Herbert Hethcote published scientific papers, such as [20, 21, 22, 23], that were crucial to the rapidly development and analysis of compartment models. Hethcote did several research on several types of models besides the common at the time SI, SIS and SIR models. He did groundbreaking work on multiple fronts of mathematical epidemiology, including age-structured models that included passively immune newborns, asymptomatic infectives, recovered individuals, casualties originated from all classes, vaccination.

In 2015, the mathematical biologist Maia Martcheva published her first book [37] consisting of an introduction to mathematical modelling and analysis of infectious diseases. This book covers very important topics including ordinary differential equation models, which is the foundation of this thesis.

1.4 Claim of Contributions

With this work, we aim to properly analyse the evolution of SARS-CoV-2 in Portugal, which will allow a better understanding on what went wrong and what were the signs one should have paid more attention to in order to avoid not only countless deaths but lock downs that had severe consequences on multiple fronts. Other important computations such as the (basic) reproduction number were made, including a new formula that considers a non-constant viral load.

Furthermore, we will use mathematical models previously discussed to fit into the number of infected individuals. A similar type of work has been made before in [4]. The metrics used to choose the best model were the square errors: Sum of Square Errors (SSE), Mean Sum of Square Errors (MSE) and Residual Mean Square Errors (RMSE). It was also used the Mean Absolute Percentage Error (MAPE) that has the big advantage of being invariant to scales, and hence comparisons between different curves can be made. One quick example is with the number of infected individuals in North versus Alentejo.

As the number of cases of COVID-19 increased, so did the number of individuals to require medical assistance in hospital. As a direct consequence, 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 takes into consideration 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.

1.5 Overview of the Thesis

This section concludes with a brief overview of this dissertation. Chapter 2 covers some basics regarding infections and its transmission. This chapter also provides some mathematical background for different types of models and computation of important parameters. On Chapter 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 chapter, models discussed in Chapter 2 will be fitted into the data. Finally, on Chapter 4 a new model is introduced, accounting for individuals in infirmaries and in intensive care units.

Chapter 2

Background

2.1 Basics: Infections, Transmission and Models

This section provides an introduction to the definitions of infections, transmission, mathematical models and key concepts in infectious disease epidemiology. To write this section, we took advantage of [37, 58].

Infection is a frequent phenomenon in nature. All species carry infections of a wide variety. Many of them are harmless, some are actually beneficial, but some, the pathogens, harm their hosts and lead to diseases. This thesis will focus on the latter. When a transmission of the infection occurs between two individuals it is usually called *effective contact*. The transmission of a disease can happen one of two ways: vertical or horizontal transmission. Vertical transmission occurs when pathogens are transmitted from parent to offspring during pregnancy, birth, or breastfeeding. On the other hand, horizontal diseases are transmitted from one individual to another in the same generation. There are several other ways to categorise transmissions of infections:

- **Airborne Infection:** refers to infectious agents that are spread via droplet nuclei. These organisms can survive outside the body and remain suspended in the air for long periods of time. (examples: tuberculosis, chickenpox, measles)
- **Droplet Infection:** refers to the transmission of a pathogen to a new host over distances of one meter or less, when an individual coughs or sneezes, small droplets of mucus, ejecting the pathogen. (examples: influenza virus, rubella, corona viruses)
- **Direct Contact:** occurs through skin-to-skin contact, kissing and sexual intercourse. (examples: athlete's foot, syphilis, conjunctivitis, Ebola)
- **Indirect Contact:** involves transmission through contamination of inanimate objects, also known as *vehicle-borne transmission*. (examples: hepatitis A, cholera, salmonella)
- **Vector-borne Transmission:** a vector is an organism that does not cause the disease itself, but transmits the infection from one host to another. (examples: malaria, viral encephalitis, Lyme disease, bubonic plague)

- **Fecal-oral Route:** refers to pathogens in the fecal particles that pass from one individual to the mouth of another. (examples: cholera, hepatitis A, polio)

Note: 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. For the exact purpose of understanding the behaviour of infectious diseases and its dynamics, it is of the utmost importance to distinguish three essential time periods. The *latent period*, occasionally called *pre-infectious 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.

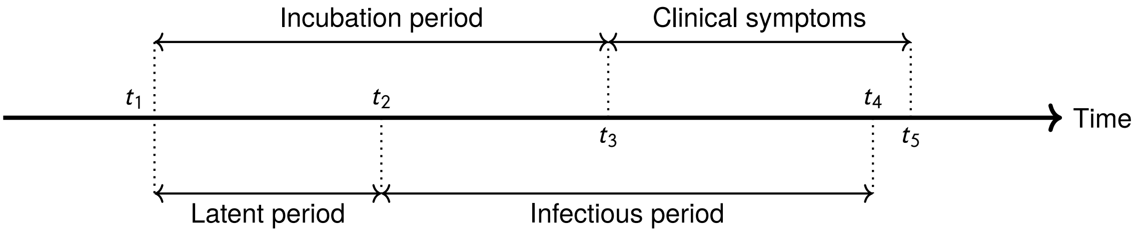


Figure 2.1: Summary of the relevant time periods.

As it can be observed above, the incubation period does not necessarily coincide with the latent period. In fact, for some diseases individuals become infectious before starting to exhibit symptoms (influenza, for example). For others, symptoms develop before the infectious period, allowing people to recognise themselves as ill and prevent the spread of the virus (Ebola fits this description). A similar situation occurs with the infectious period and the phase with clinical symptoms, depending on the disease considered we have one of three scenarios: $t_4 < t_5$, $t_4 = t_5$ or $t_4 > t_5$.

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. Disregarding the latter, infected individuals may become solidly immune (measles for example), may rid themselves of the infections but still remain susceptible to a greater or lesser degree of the disease (occurs with malaria), or develop small or no immunity whatsoever and hence, remain infectious for life (Herpes simplex). There are many other variations regarding the nature of immunity developed to infections, but are well beyond the scope of this thesis.

A mathematical epidemiological model is a simplified representation of complex phenomenons. These 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. As mentioned in [37], there are several ways to classify a model:

- **Linear vs. nonlinear:** If there is nonlinear dependence on variables (for example, product of variables), the model is nonlinear.
- **Static vs. dynamic:** A static model describes a system in equilibrium (time-invariant), while a dynamic model represents the time-dependent aspects of a system. The latter are typically represented by differential equations.
- **Discrete vs. continuous:** Discrete models treat time/states as discrete (distinct and separate values). Continuous models treat time/states as continuous (any value within a finite or infinite interval).
- **Deterministic vs. stochastic:** In deterministic models, the output of the model is fully determined by the parameters in the model initial conditions. Stochastic models possess some inherent randomness, where variable states are described by probability distributions.

In this thesis, ordinary differential equation models will be used to model the distribution of infectious diseases in populations. These models are nonlinear, dynamic, continuous and deterministic. Nevertheless, stochastic epidemic models have been developed and used in literature in works like [1].

2.2 Epidemic Models

Throughout time epidemic models have evolved and been enhanced. This section was written taking into account [25, 37] which explore different types of epidemic models.

2.2.1 SI Model

We start by considering one of the simplest epidemiological models, in which the population splits into two non-intersecting classes: susceptible (S) and infected (I) individuals. In all of the models here presented, the total population size N remains the same throughout time, and it is the sum of all classes:

$$N = S(t) + I(t). \tag{2.1}$$

This epidemic model considers a non existent latency period, hence all infected individuals are infectious. When an infectious individual comes in contact with a susceptible the latter becomes infected (and infectious) with a certain probability moving from the susceptible class to the infected class. The number of individuals who become infected per unit of time is called *incidence*. These types of models are called compartment models. In these models, an individual can reside in a single compartment and move to another according to movement arrows. In Figure 2.2, a flowchart for the SI model is presented, where β is the *transmission rate constant*.

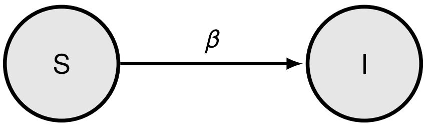


Figure 2.2: Compartment flowchart for SI model.

For a better understanding of this model, let us define some variables:

- c is the contact rate per capita (regarding the total population size N);
- p_t is the probability that a contact of a susceptible individual with an infectious results in transmission;

With these definitions, the explicit computation of the incidence is achieved quite easily. First and foremost, $p_t cS$ is the number of susceptible individuals who are infected per unit time per infectious individual and consequently, βSI is the number of new infected individuals per unit time (incidence), with the coefficient $\beta = p_t c$. Another important parameter to the study of infectious diseases is the *force of infection* which represents the rate at which susceptible acquire the infection, that is defined as $\lambda(t) = \beta I$.

Due to the simplicity of this model, it is simple to write the system of ODEs that define the model considering that those who leave the susceptible class must go to the infected class.

$$\begin{cases} S'(t) = -\beta S(t)I(t) \\ I'(t) = \beta S(t)I(t). \end{cases} \tag{2.2}$$

The associated initial conditions are $S(0)$ and $I(0)$. By adding the equations from system (2.2), results in $N'(t) = S'(t) + I'(t) = 0$ for all t. Thus, the population size is constant and equal to its initial value, $N(t) = N$.

Table 2.1: Summary of notation for the SI model.

β	Transmission rate

A very important matter that we should concern are the units of the quantities involved. Both sides of the equation must have the same units. Undoubtedly, the units of the derivatives are number of people per time unit. Let us take a closer look at the right hand side of the equations. S and I are expressed in number of people. Parameter β has units $1/(\text{time unit} \times \text{number of people})$. Consequently, the right hand side of the equations has the desired units of number of people per time unit. In order to solve the system (2.2) of ODEs, S can be replaced by $N - I$ since the population size is constant, which leads to the first-order, autonomous, nonlinear differential equation

$$\frac{dI}{dt} = \beta SI = \beta(N - I)I = \beta N(1 - \frac{I}{N})I. \tag{2.3}$$

The previous equation is a logistic differential equation [25] and its solution is easily found by integrating on both sides, yielding

$$\int \frac{1}{N(1-\frac{I}{N})I}\,dI = \int \beta\,dt \Leftrightarrow \int \frac{1}{(N-I)I}\,dI = \int \beta\,dt. \tag{2.4}$$

Taking a closer look at the left hand side of the equation, and taking advantage of the method of the partial fractions decomposition, the integral can be simplified as follows

$$\int \frac{1}{(N-I)I}\,dI = \int \frac{1}{I} + \frac{1}{(N-I)}\,dI. \tag{2.5}$$

Combining equations (2.4) and (2.5), and solving for $I(t)$ we have

$$I(t) = \frac{Nc_1e^{\beta t}}{1+c_1e^{\beta t}}, \tag{2.6}$$

where c_1 is the arbitrary integration constant. By evaluating the expression at $t = 0$, it follows that $c_1 = I(0)/(N - I(0))$ and replacing in (2.6) the final expression is

$$I(t) = \frac{I(0)N}{I(0) + e^{-\beta t}(N - I(0))}. \tag{2.7}$$

Regardless of the initial condition $I(0) > 0$, $I(t)$ and $S(t)$ are monotone and positive. On one hand, the number of infected individuals approaches N monotonically and on the other hand, the number of susceptible individuals approaches 0 monotonically.

$$\lim_{t\to\infty} I(t) = N \tag{2.8a}$$

$$\lim_{t\to\infty} S(t) = N - \lim_{t\to\infty} I(t) = 0. \tag{2.8b}$$

From equations (2.8a) and (2.8b) it becomes clear that in the SI model, all individuals will be infected regardless of the transmission rate constant β.

2.2.2 SIS Model

Similarly to the previous model, the SIS model only has two non-overlapping classes: susceptible and infected, where infected are considered infectious. Opposite to the SI model, the relation between the two classes flows in both directions. By disregarding full immunity, infected individuals become susceptible at a given *recovery rate* α.

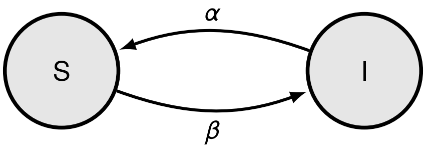


Figure 2.3: Compartment flowchart for SIS model.

The ordinary differential equations have the form

$$\begin{cases} S'(t) = -\beta S(t)I(t) + \alpha I(t) \\ I'(t) = \beta S(t)I(t) - \alpha I(t), \end{cases} \tag{2.9}$$

with the initial conditions $S(0)$ and $I(0)$. Once again, condition (2.1) holds, translating in a constant population size. This can be proven by summing all equations, resulting in $N'(t) = S'(t) + I'(t) = 0$ for all t and hence $N(t)$ is constant throughout time.

Table 2.2: Summary of notation for the SIS model.

β	Transmission rate
α	Recovery rate

In order to solve the ODE (2.9), let us express S as $S = N - I$. The resulting equation can be expressed as a logistic equation

$$\frac{dI}{dt} = \beta SI - \alpha I = \beta(N - I)I - \alpha I = (N\beta - \alpha)I\left(1 - \frac{I}{N - \frac{\alpha}{\beta}}\right) = rI\left(1 - \frac{I}{r\beta^{-1}}\right) \tag{2.10}$$

where $r = N\beta - \alpha$ is often called the *growth rate*. Following a very similar process as in the SI model, the solution of (2.9) is

$$I(t) = \frac{I(0)r\beta^{-1}}{I(0) + e^{-rt}(r\beta^{-1} - I(0))}. \tag{2.11}$$

With the explicit number of infected individuals over time, the system of ODEs is automatically defined since $S(t) = N - I(t)$. Now, there are three possible scenarios:

- $r = 0$: the obvious case, where the number of infected is constant.
- $r < 0$: If the growth rate is negative, then the number of infected individuals will tend to 0 as $t \to +\infty$ as the whole class becomes susceptible

$$\lim_{t\to+\infty} I(t) = \frac{I(0)r\beta^{-1}}{+\infty} \to 0 \tag{2.12}$$

$$\lim_{t\to\infty} S(t) = N - \lim_{t\to\infty} I(t) = N. \tag{2.13}$$

- $r > 0$: With a positive growth rate, $I(t)$ will converge to an equilibrium point

$$\lim_{t\to+\infty} I(t) = \frac{I(0)r\beta^{-1}}{I(0) + 0} = r\beta^{-1} \tag{2.14}$$

$$\lim_{t\to\infty} S(t) = N - \lim_{t\to\infty} I(t) = N - r\beta^{-1}, \tag{2.15}$$

where the disease remains in the population for an indefinite amount of time.

The threshold condition $r > 0$ is extremely important and can be rewritten as

$$r > 0 \Leftrightarrow N\beta - \alpha > 0 \Leftrightarrow \frac{N\beta}{\alpha} > 1 \Leftrightarrow \mathcal{R}_0 > 1, \tag{2.16}$$

where $\mathcal{R}_0 = \frac{N\beta}{\alpha}$ is the *basic reproduction number*. This value plays a very important role in modelling epidemics, representing the number of secondary infections produced by a single infectious individual in a population consisting only of susceptible individuals. This number allows one to understand if the disease will remain in the population indefinitely or if it gradually dies out, disappearing from the population.

2.2.3 SIR Model

This model dates back to 1927 from the famous article [29] of Kermack and McKendrick, and takes into account susceptible (S), infected (I) and recovered (R) individuals. This model supports itself on several assumptions:

1. There are neither births nor deaths in the population;
2. The population is closed, there are no entries or exits to/from the population. The following equation is verified for all time t

$$N = S(t) + I(t) + R(t); \tag{2.17}$$

3. Infected individuals are considered infectious;
4. Recovered individuals are considered to have full immunity and cannot be reinfected. Sometimes deceased individuals are also included in this class.

Analogous to the SI model, the movement of individuals is unidirectional and hence if an individual changes classes, he cannot return to the previous. Individuals, that either recover or die, leave the infected class at a per capita probability per time unit ρ, known as the *recovery rate*.

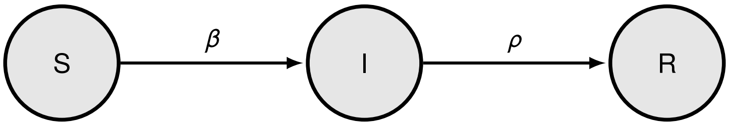


Figure 2.4: Compartment flowchart for SIR model.

The model is given by the ordinary differential equations

$$\begin{cases} S'(t) = -\beta S(t)I(t) \\ I'(t) = \beta S(t)I(t) - \rho I(t) \\ R'(t) = \rho I(t), \end{cases} \tag{2.18}$$

with initial conditions $S(0)$, $I(0)$ and $R(0)$. Similar to the previous model, we can verify that the population remains constant by adding all equations from (2.18), resulting $N'(t) = S'(t) + I'(t) + R'(t) = 0$ yielding again $N(t) = N$ for all 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 $R'(t)$, it yields

$$\frac{S'}{R'} = -\frac{\beta}{\rho} S(t). \tag{2.19}$$

Table 2.3: Summary of notation for the SIR model.

β	Transmission rate
ρ	Recovery rate

By rearranging and integrating the previous equation, it follows

$$\begin{gathered} S'(t) = -\frac{\beta}{\rho} S(t) R'(t) \Leftrightarrow \int S'(t)\, dt = -\int \frac{\beta}{\rho} S(t) R'(t)\, dt \Leftrightarrow \\ \Leftrightarrow S(t) = S(0) e^{\frac{\beta}{\rho} R(t)}. \end{gathered} \tag{2.20}$$

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.

Taking advantage of (2.17), the number of recovered individuals, $R(t)$, one can express it explicitly as a function of number of susceptible and infected individuals, $R(t) = N - S(t) - I(t)$. Thereby, the system (2.18) can be solved solely by considering the first two equations.

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

$$\frac{I'}{S'} = \frac{\beta S I - \rho I}{-\beta S I} \Leftrightarrow \frac{I'}{S'} = -1 + \frac{\rho}{\beta S} \Leftrightarrow I' = \Big(-1 + \frac{\rho}{\beta S} \Big) S'. \tag{2.21}$$

By a simple integration on both sides, the following result is obtained

$$\int I'\, dt = \int \Big(-1 + \frac{\rho}{\beta S} \Big) S'\, dt \Leftrightarrow I(t) = -S(t) + \frac{\rho}{\beta} \log S(t) + c_2, \tag{2.22}$$

where c_2 is an integration constant. Using the initial conditions, the arbitrary constant c_2 can be computed

$$c_2 = I(0) + S(0) - \frac{\rho}{\beta} \log S(0). \tag{2.23}$$

Bearing in mind that $S(t)$ is a monotonically decreasing function and by analysing equation (2.21), the maximum number of infections can be easily computed. This number expresses the maximum severity of the disease and occurs when

$$I' = 0 \Leftrightarrow S(t) = \frac{\rho}{\beta}. \tag{2.24}$$

Hence, the maximum number of infected individuals can be estimated for newly infectious diseases, which is an extremely important information when fighting this type of diseases. Combining and rearranging equations (2.22), (2.23) and (2.24), the desired value is achieved

$$I_{max} = -\frac{\rho}{\beta} + \frac{\rho}{\beta} \log \frac{\rho}{\beta} + I(0) + S(0) - \frac{\rho}{\beta} \log S(0). \tag{2.25}$$

In equation (2.24), an important threshold quantity is hidden. First and foremost, with a simple analysis of $I'(t)$, we know that the infected population arises at first reaching an all time high and then declines. As mentioned previously this maximum value occurs for $S(t) = \frac{\rho}{\beta}$, which allows us to compute the *effective* *reproduction number*

$$S(t) = \frac{\rho}{\beta} \Leftrightarrow \frac{\beta}{\rho} S(t) = 1 \Rightarrow \mathcal{R}_t = \frac{\beta}{\rho} S(t), \tag{2.26}$$

where for $t = 0$, we are considering the basic reproduction number $\mathcal{R}_0$. In this particular case, $S(0) \approx N$ returning the same value computed for the SIS model in (2.16). The effective reproduction number is of great significance for occurring infectious diseases since it allows a daily study of the pandemic's evolution.

Linearization

Another approach to solve this problem is resorting to linearization. First of all, to simplify computations let us consider the number of individuals per class as a percentage of the whole population N. With this modification, we have

- Incidence: $\lambda(t) = N\beta S(t)I(t)$;
- Force of infection: $\lambda(t) = N\beta I(t)$;
- Basic reproduction number: $\mathcal{R}_0 = \frac{\beta}{\rho}$;

Bearing in mind that $R(t)$ is easily obtained with simple quadratures from the variable $I(t)$, it suffices to consider a two dimensional system, disregarding the number of recovered individuals. Let us consider the field

$$\begin{aligned} f_1(S,I) &= -\beta SI \\ f_2(S,I) &= \beta SI - \rho I \end{aligned} \tag{2.27}$$

and its Jacobian matrix

$$\mathbf{J}_{(S,I)} = \begin{bmatrix} -\beta I & -\beta S \\ \beta I & \beta S - \rho \end{bmatrix}. \tag{2.28}$$

The disease-free equilibrium happens at $(1, 0)$, representing the scenario where all are susceptible and none is infected. Linearizing near this point leads to the following Jacobian matrix

$$\mathbf{J}_{(1,0)} = \begin{bmatrix} 0 & -\beta \\ 0 & \beta - \rho \end{bmatrix}. \tag{2.29}$$

With the previous matrix, important information such as the basic reproduction number can be withdrawn. There are two eigenvalues: 0 and $\beta - \rho$. The latter is negative if $\frac{\rho}{\beta} > 1$ and positive if $\frac{\rho}{\beta} < 1$. Hence the infected population increases while the susceptible decreases monotonically and $\mathcal{R}_0 = \frac{\rho}{\beta}$, as proved earlier.

The linearized equation is given by

$$\begin{cases} S'(t) = -\beta I(t) \\ I'(t) = (\beta - \rho)I(t), \end{cases} \tag{2.30}$$

which is very easy to integrate and only depends on $I(t)$. Integrating the last equation on both sides results in

$$I(t) = I(0)e^{(\beta-\rho)t}, \tag{2.31}$$

where $I(0)$ is the initial number of infected individuals. Taking into account that we study the population on a daily basis, the next relation can be established

$$I(t+n) = a^n I(t), \tag{2.32}$$

where a represents the daily growth of infected individuals per day. Considering that only one individual out of all the population is infected at $t = 0$, it is possible to obtain an interesting expression

$$I(t+1) = aI(t) \Leftrightarrow \frac{e^{(\beta-\rho)(t+1)}}{N} = a\frac{e^{(\beta-\rho)t}}{N} \Leftrightarrow \beta = \rho + \log a. \tag{2.33}$$

This provides us the value of β from official data from on going pandemic. Finally, the value of $\mathcal{R}_0$ is

$$\mathcal{R}_0 = \frac{\beta}{\rho} = 1 + \frac{\log a}{\rho}. \tag{2.34}$$

2.2.4 SIRD Model

With a small modification on the model discussed in the previous section 2.2.3, the number of deceased individuals is now included, representing a fraction of those that are retired from the infected class. Thus, we are facing a model with four non-overlapping classes: susceptible (S), infected (I), recovered (R) and deceased (D). An individual can never return to a previous state.

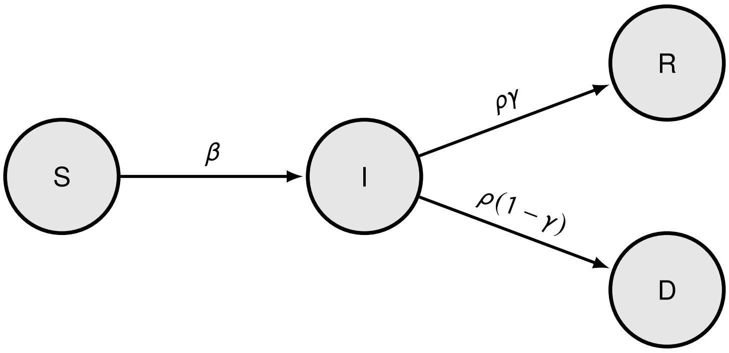


Figure 2.5: Compartment flowchart for SIRD model.

Once again, this is a closed system. Neither exits nor entries are possible and, consequently, the total population size will not suffer any alterations whatsoever. This can be verified by performing a similar process as stated in previous sections

$$N = S(t) + I(t) + R(t) + D(t). \tag{2.35}$$

Thus, the parameters mentioned in Figure 2.5 will be used to shape the model. The summary of its notation is presented in Table 2.4. The value of $1/\rho$ is the average period of infection. The fraction of recovered and deceased individuals from the infected class is given by γ and $1-\gamma$, respectively.

Table 2.4: Summary of notation for the SIRD model.

β	Transmission rate
$1/\rho$	Infection period
γ	Fraction of recovered individuals

The SIRD model is a system of four ordinary differential equations

$$\begin{cases} S'(t) = -\beta S(t)I(t) \\ I'(t) = \beta S(t)I(t) - \rho I(t) \\ R'(t) = \gamma \rho I(t) \\ D'(t) = (1-\gamma)\rho I(t), \end{cases} \tag{2.36}$$

with initial conditions $S(0)$, $I(0)$, $R(0)$ and $D(0)$. Due to its similarity to the Kermack-McKendrick model, with the first two equations equal to (2.18), most of the outcomes will be the same and it is unnecessary to explicitly show, yet again, the results.

By dividing the third and fourth equation from (2.36), a simple result follows

$$\frac{D'}{R'} = \frac{(1-\gamma)\rho I}{\gamma \rho I} \Leftrightarrow D' = \frac{(1-\gamma)}{\gamma} R' \Leftrightarrow D(t) = \frac{(1-\gamma)}{\gamma} R(t). \tag{2.37}$$

Concerning the evolution of susceptible and infected, the results obtained for the SIR model still apply. The only difference between the models is that here individuals that leave the infected class can either be considered as recovered or as deceased.

2.2.5 SEIRD Model

As stated before, many infectious diseases have a period after the transmission of the pathogen where individuals carry the disease but are not yet contagious (latent period). Giving its importance to how the spread of an infection can occur, this particular model considers a new class, *exposed individuals* (E).

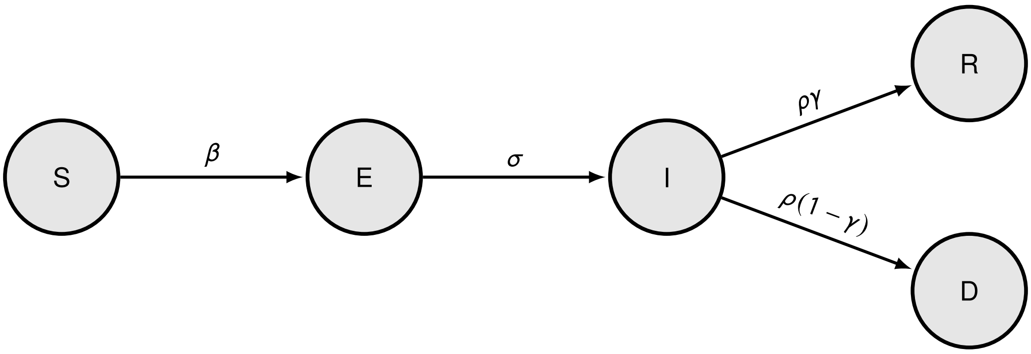


Figure 2.6: Compartment flowchart for SEIRD model.

In Figure 2.6 a flowchart represents how the movement of individuals between classes occurs in this model and in Table 2.5 is presented a summary of the notation here used.

The new parameter σ is the rate of conversion of exposed individuals to contagious individuals, where the inverse is the average number of days an individual spends as non-contagious after being exposed to the disease. The parameter γ is the rate at which individuals leave the infected class and $1/\gamma$ can be interpreted as the time one remains contagious.

Table 2.5: Summary of notation for the SEIRD model.

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

The ordinary differential equations that define this model are

$$\begin{cases} S'(t) = -\beta S(t)I(t) \\ E'(t) = \beta S(t)I(t) - \sigma E(t) \\ I'(t) = \sigma E(t) - \rho I(t) \\ R'(t) = \gamma\rho I(t) \\ D'(t) = (1-\gamma)\rho I(t), \end{cases} \tag{2.38}$$

with initial conditions $S(0)$, $E(0)$, $I(0)$, $R(0)$ and $D(0)$. As previously mentioned, all models are closed with no entries nor exits. Thus, again, the population size remains constant which can be easily proven.

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. To obtain approximate solutions of these equations, we will be resorting to linearization once again. Considering the field

$$\begin{aligned} f_1(S,E,I) &= -\beta SI \\ f_2(S,E,I) &= \beta SI - \sigma E \\ f_3(S,E,I) &= \sigma E - \rho I, \end{aligned} \tag{2.39}$$

its Jacobian matrix is as follows

$$\mathbf{J}_{(S,E,I)} = \begin{bmatrix} -\beta I & 0 & -\beta S \\ \beta I & -\sigma & \beta S \\ 0 & \sigma & -\rho \end{bmatrix}. \tag{2.40}$$

Aiming to linearize the system near the disease-free equilibrium point $(1,0,0)$, i.e. when everyone is susceptible and there are neither exposed nor infected individuals, the Jacobian matrix yields

$$\mathbf{J}_{(1,0,0)} = \begin{bmatrix} 0 & 0 & -\beta \\ 0 & -\sigma & \beta \\ 0 & \sigma & -\rho \end{bmatrix}. \tag{2.41}$$

The linearized equation is now

$$\begin{cases} S'(t) = -\beta I(t) \\ E'(t) = -\sigma E(t) + \beta I(t) \\ I'(t) = \sigma E(t) - \rho I(t). \end{cases} \tag{2.42}$$

Looking at the last two equations, the system can be written as

$$X'(t) = AX(t) \Leftrightarrow \begin{bmatrix} E'(t) \\ I'(t) \end{bmatrix} = \begin{bmatrix} -\sigma & \beta \\ \sigma & -\rho \end{bmatrix} \begin{bmatrix} E(t) \\ I(t) \end{bmatrix}. \tag{2.43}$$

The previous system has a simple solution and considering that at time $t = 0$ there are no exposed cases and only one individual carries the disease ($\frac{1}{N}$), the result follows

$$I(t) = I(0)e^{\lambda t} = \frac{e^{\lambda t}}{N}, \tag{2.44}$$

where λ is the dominant eigenvalue, which can be obtained one of two ways. On one hand, taking equation (2.32) with $n = 1$, we have

$$I(t+1) = aI(t) \Leftrightarrow \frac{e^{\lambda(t+1)}}{N} = a\frac{e^{\lambda t}}{N} \Leftrightarrow \lambda = \log a. \tag{2.45}$$

On the other hand, the dominant eigenvalue can be computed directly from matrix A

$$\det(A - \lambda I) = 0 \Leftrightarrow \begin{vmatrix} -\sigma - \lambda & \beta \\ \sigma & -\rho - \lambda \end{vmatrix} = 0 \Leftrightarrow \lambda = \frac{1}{2}\left(-\rho - \sigma - \sqrt{(\rho - \sigma)^2 + 4\beta\rho} \right). \tag{2.46}$$

With the relations provided in (2.45) and (2.46), the value for the transmission rate can be derived

$$\beta = \frac{\rho\sigma + \rho \log a + \sigma \log a + \log^2 a}{\sigma}. \tag{2.47}$$

This relation is very interesting, in the sense that provides us the value of β from real time data, and by having information regarding the average days of latency and infectious period of the disease. Additionally, the value for the basic reproduction number $\mathcal{R}_0$ is

$$\mathcal{R}_0 = \frac{\beta}{\rho} = \frac{\rho\sigma + \rho \log a + \sigma \log a + \log^2 a}{\rho\sigma} = 1 + \frac{\log a}{\rho} + \frac{\log a}{\sigma} + \frac{\log^2 a}{\rho\sigma}. \tag{2.48}$$

2.3 Herd Immunity Threshold

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

Herd immunity arises from the effects of individual immunity in a whole population, and defines itself as the indirect protection from an infectious disease when a significant fraction of the population is immune to the virus [47]. Herd immunity is of the utmost importance since it allows immunocompromised and younger people to remain unvaccinated.

In a population full of non-bearers of the pathogen, the virus will spread in an unchecked manner, whereas if a portion of the population is immune, the likelihood of an effective contact between a susceptible and an infected individual decreases.

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

$$\mathcal{H} = 1 - \frac{1}{\mathcal{R}_0}. \tag{2.49}$$

Thus, the greater the number of secondary infections caused by a single person, the bigger the fraction of immunised individuals needed for the purpose of stopping the virus from spreading. Nevertheless, the herd immunity formula relies itself on a few assumptions, that most of the times are not met in real scenarios. 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:

$$\mathcal{H}_{adj} = \left(1 - \frac{1}{\mathcal{R}_0}\right)\frac{1}{V_e}, \tag{2.50}$$

where V_e represents the vaccine effectiveness, i.e. the immunity to the virus that the vaccine confers to an individual.

Chapter 3

Analysis of SARS-CoV-2 in Portugal

Bearing in mind everything that has been stated so far, an extensive and thorough analysis of SARS-CoV-2 in Portugal will take place in this chapter.

3.1 SARS-CoV-2: An Overview

Before entering into the analysis of any pathogen in a given population, it is of the utmost importance to fully comprehend how the virus spreads and its characteristics, gathering as much information as one can. The real world context in which a mathematical problem settles is crucial to its analysis.

In December 2019, multiple cases of pneumonia of unknown origin were reported in Wuhan (Hubei Province, China). On January 9^{th}, China announced a novel coronavirus as the agent of the outbreak. Within a short span of time, human-to-human transmission was confirmed, Wuhan went under confinement, all provinces across China reported cases of coronavirus, the WHO declared the outbreak a Public Health Emergency of International Concern (PHEIC), and by February 11^{th} the International Committee on Taxonomy of Viruses (ICTV) named the virus as SARS-CoV-2, short for severe acute respiratory syndrome coronavirus 2 [26]. The WHO designated the disease resulting from the infection as coronavirus disease (COVID-19).

With the goal of containing COVID-19, unprecedented strict public health measures were implemented in China. Not only the uncertainties around the disease but also the abundance of international travel enabled a rapid worldwide spread. By the end of February, large clusters of infections were reported over several countries and the world was on the verge of facing a never before seen global event. On March 11^{th}, the World Health Organization officially characterised COVID-19 as a pandemic [43]. Soon enough a majority of countries were taking severe unparalleled public health and social measures.

Even though the scientific community cannot pinpoint the specific circumstances in which the virus was developed and transmitted to humans, careful and thorough research on this pathogen allowed a better understanding on the subject. Coronaviruses can be grouped in four genera, and this particular virus is very similar to betacoronaviruses (β-CoV) found in bats [16, 62], albeit distinct from the 2003 SARS-CoV and the 2012 MERS-CoV. There was more proximity to the SARS-CoV in terms of genome, and hence the name SARS-CoV-2. Nevertheless, bats are the natural reservoir for a wide variety of CoVs and thus it can be assumed that the origin of the virus was indeed the bats and has been transmitted at some point in time to other animal hosts and eventually to humans. Articles like [3, 26, 62] provide a lot more information on a biological level, however the biological theory is out of scope of this work and will not be here presented.

Respiratory infectious diseases spread through multiple ways: contact with an infected person or touching a surface that is contaminated, by droplet transmission of respiratory particles that contain the virus (occurs near an infected person) and lastly by airborne transmission droplets/particles suspended in the air for longer periods of time and distance. There were reports of other routes of transmission, but those seem to be isolated cases and do not amount to a significant slice of the worldwide transmissions, and consequently can be disregarded. In early stages of the pandemic, there were greater concerns regarding surface transmission, however recent studies [11, 38] consider as unlikely this type of transmission to be a major route for infections, even though the virus can persist for days on inanimate surfaces.

All ages are susceptible to infection, however clinical manifestations differ with age. Most young people or children have only mild diseases (non-pneumonia or mild pneumonia) or are asymptomatic whereas elderly (> 65 years of age) with co-morbidities are more susceptible to develop severe disease. The underlying health conditions that increase susceptibility are hypertension, diabetes, chronic obstructive pulmonary disease and cardiovascular disease. The most common symptoms include fever and cough. Other less common clinical manifestation include fatigue, sputum production, shortness of breath, sore throat, headache and diarrhea (uncommon). As far as more severe complications go, there is acute respiratory distress syndrome (ARDS), pneumonia, septic shock, liver problems, coagulation dysfunction, multiple organ failure and even death. This information can be obtained in articles such as [15, 16, 26, 52].

As far as the length of the latency and incubation period, today persists a lot of uncertainty. It varies greatly from one individual to another. Nevertheless, recent studies [44, 61] point to a shorter latency period, meaning that an individual becomes infectious before developing onset symptoms.

3.2 Preliminary Analysis

The data here used is of public access and provided daily by the Directorate General for Health (DGS) in [51] and by the Environmental Systems Research Institute (ESRI) [36]. In this work, we will take advantage of the data available in [49] which was built from the last two references. All the results here presented were obtained resorting to the software *Wolfram Mathematica*.

A large amount of information was collected of Portugal and its regions (North, Centre, Lisbon and Tagus Valley, Alentejo, Algarve, Madeira and Azores):

- total number of confirmed cases by gender and age;
- new daily cases;
- total number of deaths;

- new daily deaths;

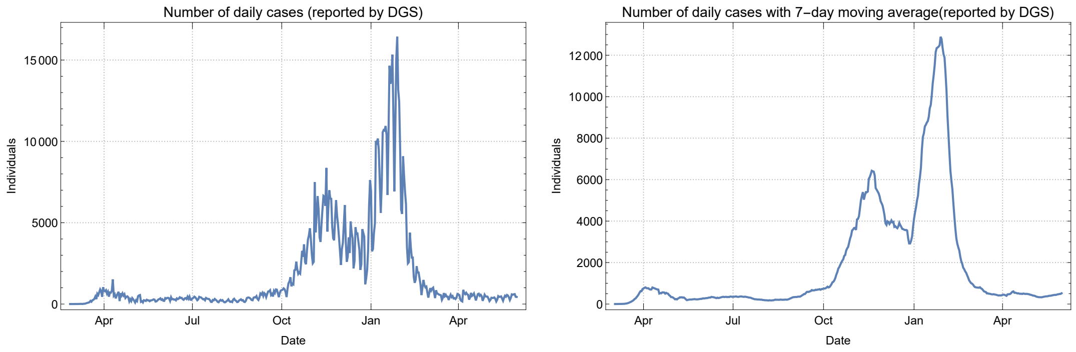
- total number of recovered individuals;

There is also overall information regarding the number of confirmed, recovered and casualties cases by gender and age (sprawled in ten year classes), number of infirmary and ICU patients, active cases and number of individuals in vigilance. Data regarding vaccination is also available since January 2021.

On a couple of days, October 5th, 2020 and April 25th, 2021, the number of casualties and daily cases by age and gender were not reported. To avoid problems with computations, these values were interpolated. In order to do so, we resorted to spline interpolation with third degree piecewise polynomials. By using cubic polynomials, we guarantee that the first and second derivative are continuous everywhere, i.e. each successive polynomial will have identical values in the knots in terms of first and second derivative.

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.

During this time period, the pandemic in Portugal consisted of three waves, where the first one started on March 2nd, 2020 and ended around mid May, with a great level of under reporting of cases. In the first few months, there was a lot of uncertainty around the disease and how to properly attack it on all fronts. There was a deficit on the number of tests performed and not enough control on individuals that had been in contact with infected ones for example. Portugal spent the summer with a rather controlled number of cases, and the second wave only began mid September, around the time schools opened and classes were given on site. On the 13th of September, the daily incidence surpassed the threshold of 500, a worrying indication of the approach of a new wave. This second wave hit its peak early December. This wave was apparently on a slow downward trend when the celebration of two major events (Christmas and New Year) caused a major outbreak on all regions and the deadliest wave of the pandemic in Portugal.



(a) Without 7-day moving average

(b) With 7-day moving average

Figure 3.1: Number of daily new cases reported by DGS without and with a 7-day moving average from left to right.

One of the biggest problems when modelling this pandemic was the level of unreliability of the data published by *"Direcção Geral de Saúde"*, the Portuguese health authority (DGS). Due to asymptomatic individuals, limitations regarding the number of tests performed and most laboratories only being open during the weekdays, the daily reported number of new cases was always inaccurate. Hence, on Figure 3.1 the number of daily cases are presented with and without a 7-day moving average, this computation will help stabilise some irregularities. Nonetheless, the three waves that hit Portugal during this period are still very clear.

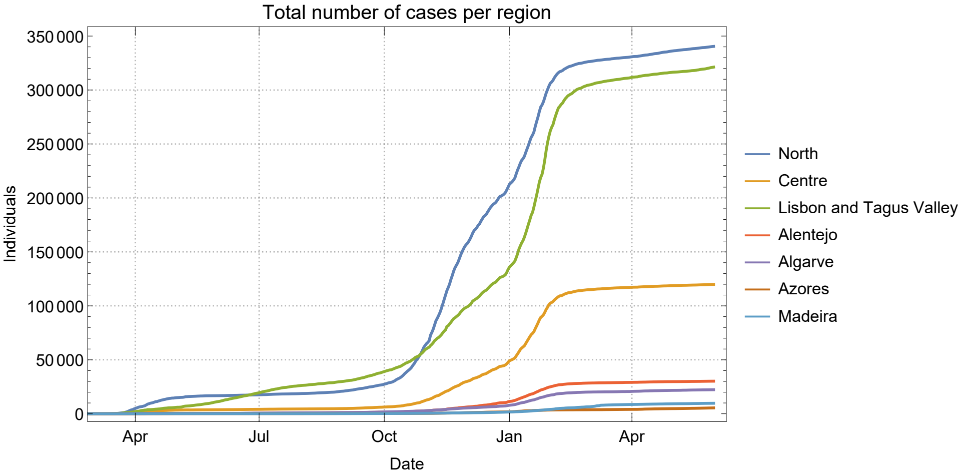


Figure 3.2: Total number of cases per region.

On Figure 3.2, the number of cumulative cases per region is presented. Three significant rises in the number of reported cases are very clear, indicating the existence of each wave. The population density is a crucial factor for the evolution of a pandemic. For example, Alentejo is Portugal's biggest region but it also has the lowest ratio of people per square kilometre. With relatively small cities, when compared to cities like Lisbon or Oporto, the number of people each individual contacts with is significantly lower and thus, the transmission rate will decrease accordingly.

For the purpose of achieving a good approximation of the real number of infected individuals in Portugal, one can take advantage of the number of casualties. A recent article [12] 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% confidence interval of $0.75 - 0.99\%$. With this information, and considering x days from onset symptoms of COVID-19 to death, we are able to compute the total number of cases. Also, if we know how much time an individual spends infected, the number of active cases per day can be computed. By tuning these two parameters, we can obtain the real number of infected individuals for each point in time, which is of the utmost importance to fit models discussed in Chapter 2.

The average time interval from onset symptoms to death can be extracted from the data at our disposal by computing the difference between peaks of two curves: active cases and daily deaths. The curves should have a similar shape with a slight delay for the deceased curve. However, we face an unexpected outcome when analysing this difference for the most protuberant wave of the pandemic. The peak of deaths occurs a day prior to the maximum number of active cases, when it was to be expected a delay in the number of deaths (see Figure 3.3). This actively demonstrates that there was an higher level of under-reporting in the days prior to the peak of infected individuals. Unfortunately, this entails the necessity to solely resort to scientific papers in order to obtain the best value for the referred time interval.

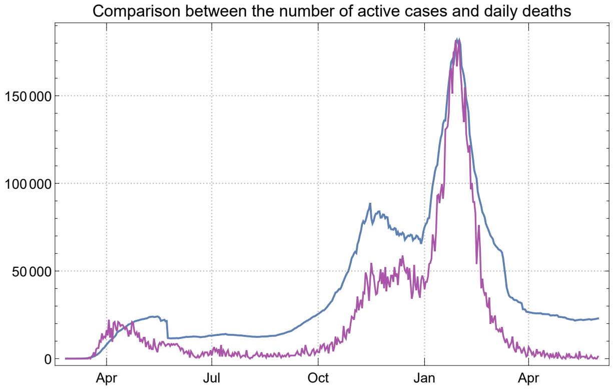


Figure 3.3: Comparison between the number of active cases (blue line) and six hundred times the number of daily casualties provided by DGS (purple line), exhibiting an unexpected overlap of the peaks of both curves.

The Scientific Advisory Group for Emergencies of the United Kingdom wrote an unpublished paper [18] where they questioned themselves on the time from symptom onset to death and if it differs by age, sex and first/second wave. Median of time from symptom onset to death was shorter in the second wave (7 days) compared to the first wave (13 days). The scientific article [33] synthesises results obtained from several studies, where three studies in China report an overall 15.93 ± 2.86 days from symptoms to casualty. Another study [59] in Wuhan reports a mean 17.4 ± 8.4 days to death. These articles have wide confidence intervals, making it extremely difficult to establish an accurate value for this number since it depends greatly on each individual and health care conditions of each country. We must also take into account that once an individual starts developing symptoms, there is still a small time gap until a person is officially considered positive. It might be a day or two until a person is fully aware of their symptoms in order to call DGS's line for medical information and advice on COVID-19, to then schedule a date to finally be tested.

Bearing everything that has been stated in mind, we will consider a two-week period (14 days) for an individual to die after symptoms start to manifest, and it is now possible to calculate the estimated number

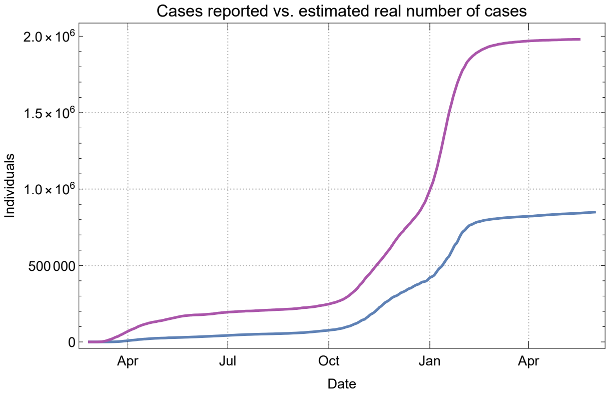


Figure 3.4: Comparison between the number of cases reported by DGS (blue line) and the estimated real number of cases computed with the number of deaths (purple line), taking advantage of the infection fatality ratio 0.086% and a 14 days from onset symptoms to death.

of overall infections (check Figure 3.4). With a 14-day delay, a surprising total of 1.97965×10^6 people have been infected by May 17^{th} even though by the same date only 842.381 cases have been reported, representing only 42.55% of the cases.

With an approximation for the real cumulative number of infected individuals, an estimation for the level of under-reporting in Portugal can be computed. It is expected an higher lever of undetected cases at the beginning of the pandemic. As it progresses the government starts to have a better understanding of the virus, allowing the increase of number of tests performed and enhancement of methods to control the spread of the virus. Eventually, a more stable level of under-reported cases should be achieved, representing individuals that are asymptomatic and never had a reason to get tested. By observing Figure 3.5, we can infer that, from November 2020 forward, the government was only able to assess $40-50\%$ of the estimated real number of cases.

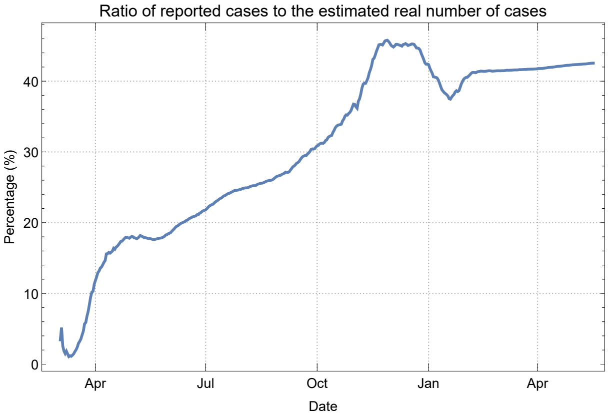


Figure 3.5: Ratio between the number of cases reported and the number of estimated cases.

As far as the number of days it takes for a person to recover, let us consider 7, 14 or 21 days. We take these values on account of that for individuals with none to mild symptoms it takes a week or two to recover, while patients with severe complications caused by the virus may spend over a month in the hospital. To best choose the dataset we wish to use, we will take advantage of the third wave of the pandemic. First and foremost, by analysing the values for each recovery time, the maximum number of active cases can be computed.

Table 3.1: Maximum number of active cases with different recovery times.

	7 days	14 days	21 days
Peak of active cases	236.744	444.186	619.302

The highest number of active cases reported by DGS was 181.811, which is lower than all values presented in Table 3.1. There will always be missing cases, generally asymptomatic individuals. A recovery time of 21 days represents a report of approximately 30% of the real number of cases, and even if this number is correct, after two weeks after onset of symptoms if a person presents itself with a positive test for COVID-19, its viral load will most likely be too low to effectively infect another individual, which is not relevant for the purposes of mathematical epidemiology. For that reason, a 21-day recovery time will be disregarded.

The recovery time that more accurately represents the level of under-reporting is a 14-day period, which can be verified by checking Figure 3.5 and performing simple computations. Notwithstanding the relevance of a 7-day period, since it may be a better representation of real-world scenarios. Once an individual becomes aware of bearing such virus, he will most likely confine to prevent the spread of the virus to others, and despite the fact that this person is still infectious, in practical terms he will no longer be responsible for other infections on his behalf.

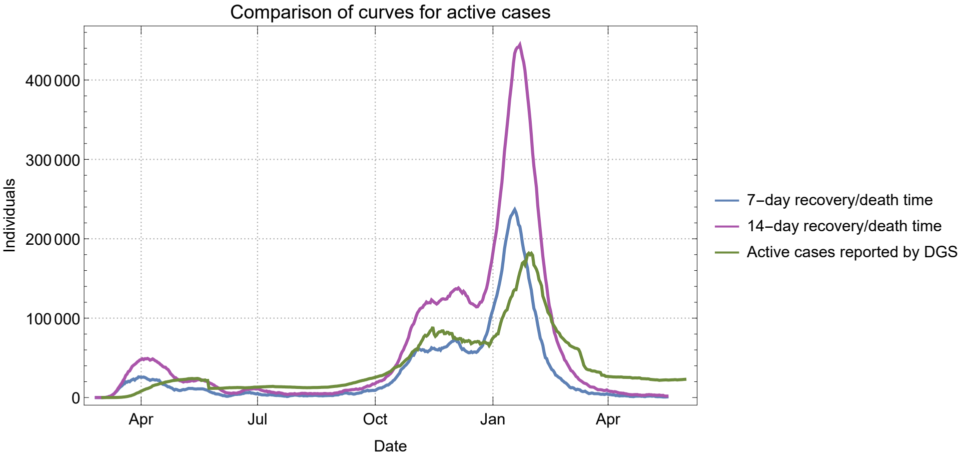
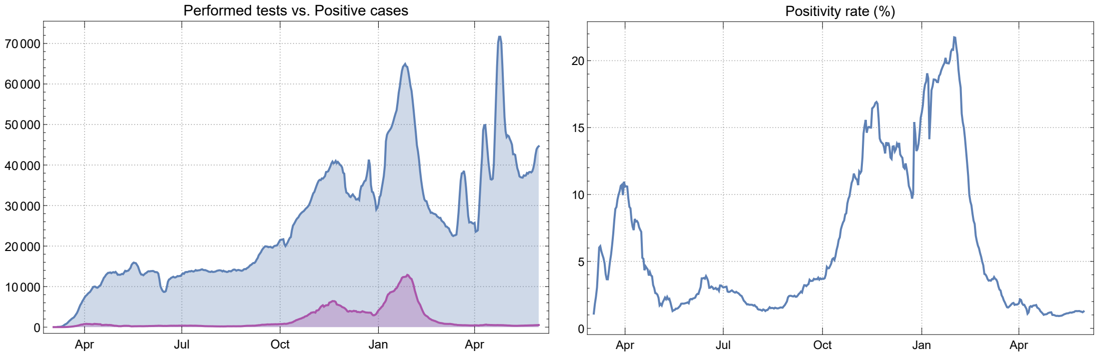


Figure 3.6: Comparison of different curves for the number of active cases. The blue and purple lines represent the number of active cases considering 7 and 14 days, respectively, to leave the infectious class either by recovery or death. The green one stands for the number of active cases reported by DGS.

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.



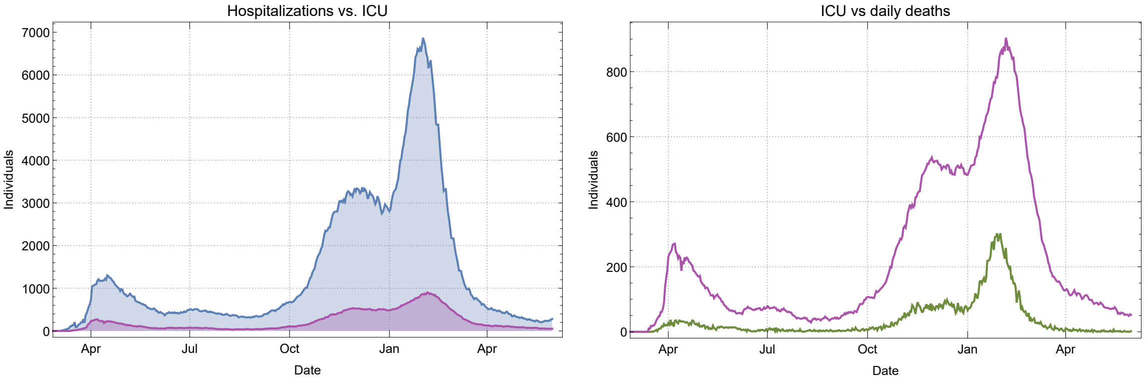
(a) Performed tests and positive cases

(b) Positivity rate

Figure 3.7: On the left is the comparison between the number of performed tests (represented by a blue line) and the number of positive tests (represented by a purple line), both with a 7-day moving average. On the right, the positivity rate (%) is presented.

As it can be easily perceived on Figure 3.7(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 the 5^{th}.



(a) Hospitalisations vs. ICU

(b) ICU vs. daily deaths

Figure 3.8: On the left, the number of patients in the hospital (infirmary and ICU) is presented by a blue line. In both figures, the purple curve represents the number of individuals in intensive care units. The green line stands for the number of daily deaths.

It is common for individuals in the hospital to move from the infirmary to the ICU or vice versa according to one's condition. Consequently, it was expected both curves in Figure 3.8(a) to have such similar trends. At first, one would also anticipate similar trends for the daily casualties but with a slight delay in comparison to the other two curves. However, if we focus on the third wave of the pandemic, we notice that the peak of daily deaths (January 28th) occurs earlier than the peak of ICU individuals (February 5^{th}). This can be explained by the sudden increase of hospitalisations, catching the government and health care professionals by surprise. Rapid adjustments had to be made, which included new wings for intensive care patients, reallocation of non COVID-19 patients, new and more exhausting schedules for professionals, among others. We must also bear in mind that different bodies react differently to the disease and it is possible that an higher number of individuals with a worse reaction to the virus may have been infected in an earlier stage of the wave.

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 [46]. The report [40] 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. Since the number of deaths in the beginning of the second wave were still low, the computations will start on November 4^{th}, 2020, the day in which the threshold of fifty deaths was exceeded. Our suspicions are confirmed when observing Figure 3.9, where the orange curve representing Alentejo stands out for the worst reasons. For the most part, throughout the time window presented, Alentejo has the highest CFR of all regions achieving an extremely high value of 11.9213% on January 16^{th}, 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.

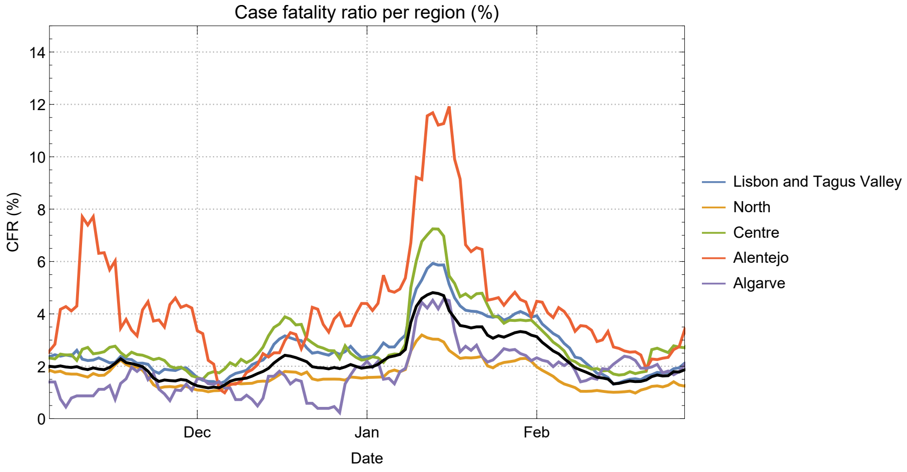


Figure 3.9: Case fatality ratio per region. The black curve represents the overall CFR in Portugal.

Out of curiosity, it is particularly interesting to check how the number of deaths increased during this pandemic in comparison to previous years. Even though the number of casualties due to COVID-19 were extremely high, the fact that people were quarantining and following protection measures against COVID-19 such as regularly using masks and social distancing helped in reducing the number of contagion

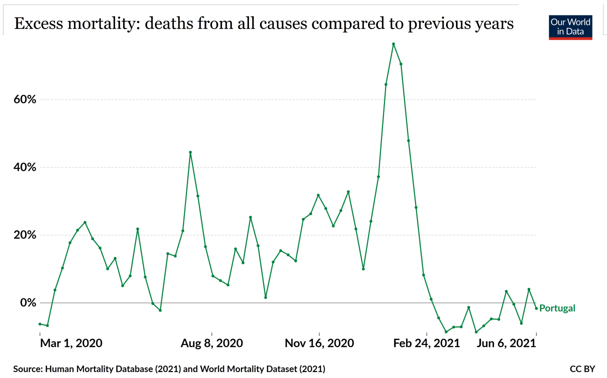


Figure 3.10: Excess mortality (%) in Portugal during the pandemic in comparison to the years 2015-2019.

for others diseases, resulting in zero flu deaths [57]. Obviously, this leads to a lower immunity, increasing the risk of a more serious disease during the winter of 2021-2022. Resorting to [48], it is shown in Figure 3.10 the number of weekly or monthly deaths in 2020–2021 differs as a percentage from the average number of deaths in the same period over the years 2015–2019. To be noticed that there might some error in these values due to incomplete coverage or delays in death reporting. The fact that even with a decrease of cases for other infectious diseases, the excess mortality is almost always positive shows how deadly COVID-19 was. The highest percentage recorded was unexpectedly during the third wave, from January 17^{th} to January 24^{th}, with an astonishing 76.45% more deaths when compared to same period during 2015-2019.

3.3 Reproduction Number Computation

As discussed in the previous chapter, the reproduction number is a very important indicator in order to have a better understanding on how fast a pathogen is spreading. With the goal of empirically compute the reproduction number, it must be taken into account its definition discussed in a prior chapter. Hence, we want to compute how many individuals does in fact an infected person transmits the virus.

3.3.1 Robert Koch Institute Formula

The Robert Koch Institute (RKI) published a report [27] 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

$$\mathcal{R}_{t,\tau} = \frac{\sum_{i=t-\tau+1}^{t} E_i}{\sum_{i=t-\tau+1}^{t} E_{i-\nu}}, \tag{3.1}$$

where / 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 3.11 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.

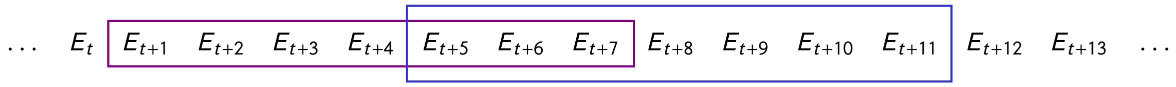
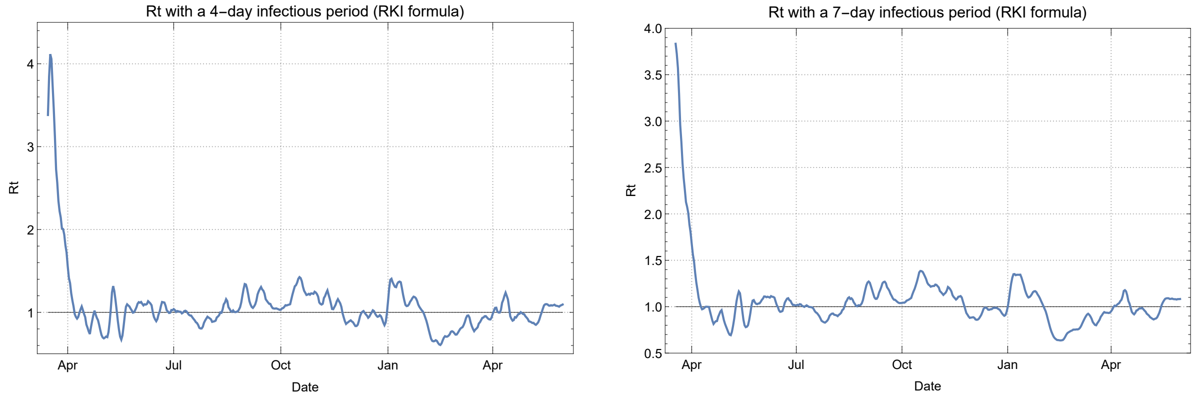


Figure 3.11: Representation of $\mathcal{R}_{t+11}$ by RKI with $\tau = 7$ and $\tau = 4$.

As mentioned previously, the data provided by the Portuguese government was affected by the reduction of tests performed on the weekends and by laboratories that are only open during week days. These factors led to major irregularities on the number of cases that were daily reported. It was observed a decrease of cases on the weekend followed by a big increase on Monday and Tuesday to compensate the previous days. For that reason, a 7-day moving average on the daily new cases was performed before applying the formula (3.1). A big advantage of this formula is that allows the computation of the reproduction number of the current day.

On Figure 3.12 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 3.12: The blue line stands for the reproduction number in Portugal using the RKI formula. On the left, an infectious period of 4 days is considered while on the right there is a 7-day infectious period. The black line represents the important threshold of $\mathcal{R}_t = 1$.

Another interesting subject worth exploring is to compute this number for each gender and for different age classes, that we present in Figure A.1 and A.2 in Appendix A, respectively. All plots presented were computed using a 7-day infectious period and a time lag $\nu = 4$. In terms of age, each class has a ten year range and do not overlap. As expected, the plots are extremely similar and do not indicate significantly different spread rates according to the gender. The government only made available information by gender on March 19^{th}, and consequently the values for the $\mathcal{R}_t$ start a bit latter and preclude the computation of the basic reproduction number $\mathcal{R}_0$ specific to gender.

3.3.2 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

$$\mathcal{R}_t = \frac{\sum_{i=t-\omega}^{t+\omega} w_{i-t+4} E_i}{\sum_{i=t-\omega}^{t+\omega} w_{i-t+4} E_{i-\nu}}, \tag{3.2}$$

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 (3.1) and (3.2) is that the latter will always compute the reproduction number with a $\lfloor \frac{\tau}{2} \rfloor$ days delay.

Similar to the computations made by the German RKI, we will use a latency period of 4 days and an infectious period of 7 days. 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 following weight vector

$$w = (0.324652, 0.606531, 0.882497, 1, 0.882497, 0.606531, 0.324652). \tag{3.3}$$

As expected the probability of transmitting the disease is not constant ($w_i \neq 1, \forall i$), instead it is lower on the first and last day of the infectious period ($\approx 32.5\%$), increasing symmetrically, hitting its peak on the 4^{th} day with 100%.

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 3.13).

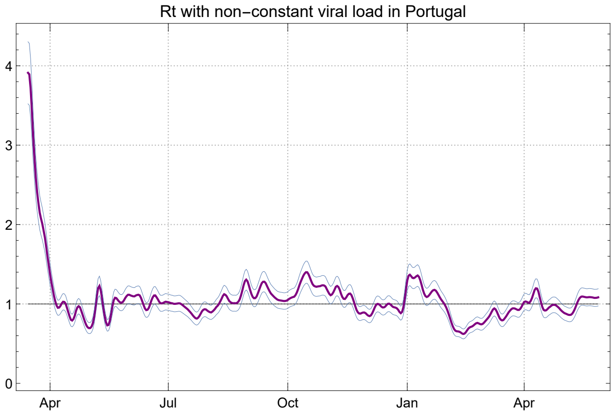
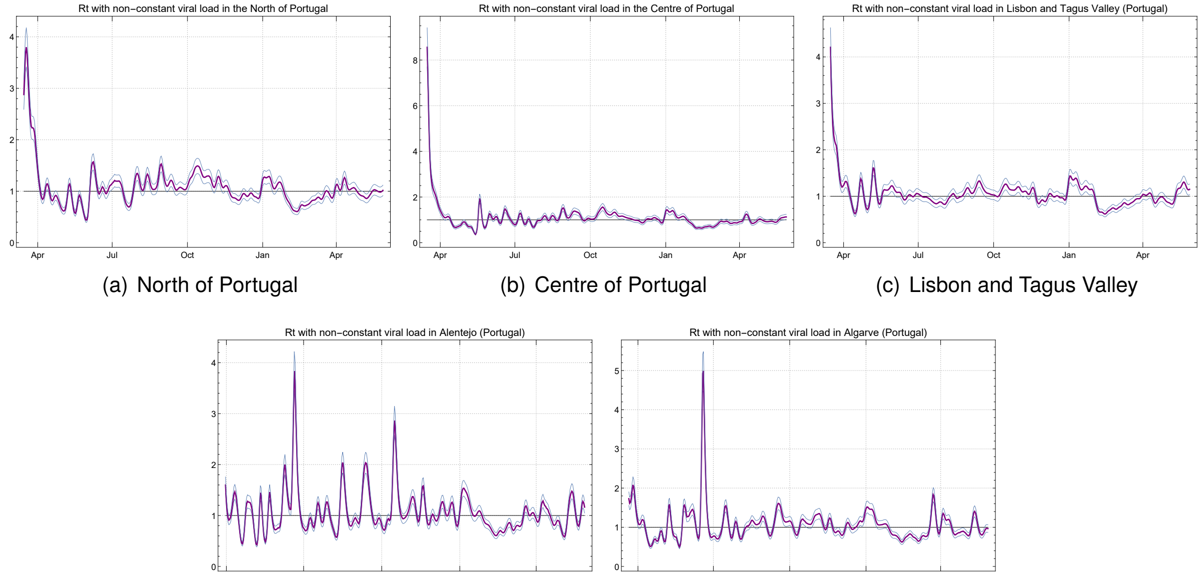


Figure 3.13: Reproduction number represented by the purple curve, using the number of active cases and the viral load during a 7-day infectious period follows $\mathcal{N}(0, 4)$. The blue lines describe a 10% error margin and $\mathcal{R}_t = 1$ is illustrated by the black line.

As mentioned in the previous section, the population density is of the utmost importance in regards of fast a virus can spread, and thus on Figure 3.14 are presented plots for the reproduction number for Portugal's regions. Açores and Madeira were left out in view of the low number of cases that were daily presented. Alentejo and Algarve are proof of how a lower number of cases can impact the reproduction number plot. Both present quite a spiky plot (Figures 3.14(d) and 3.14(e), respectively) that only gets stabler around the second and third wave.

An estimation of the $\mathcal{R}_t$ can also be performed resorting to number of deceased individuals. To perform this computation a time gap of fourteen days from infection to death is, once again, considered. This very reason leads to its first major disadvantage: computing the $\mathcal{R}_t$ with a delay of 14 days can reveal itself deadly when fighting a real pandemic.

The other disadvantage lies with the fact that during periods with lower deaths, the reproduction number becomes more unstable. During the second and specially, the third wave of the pandemic the number of casualties were much higher which allows better estimates. As for the other intervals of time, the curve tends to be rougher and very often with spikes which do not represent the situation at the time



(a) North of Portugal (b) Centre of Portugal (c) Lisbon and Tagus Valley

(d) Alentejo (e) Algarve

Figure 3.14: Reproduction number using equation (3.2) (with $\tau = 7$ and $\nu = 4$) for regions of Portugal, described by the purple line with blue lines illustrating a 10% error margin. The black line represents the important threshold of $\mathcal{R}_t = 1$.

of the pandemic. For this reason, it is not sound to take advantage of this basic reproduction number to compute the herd immunity.

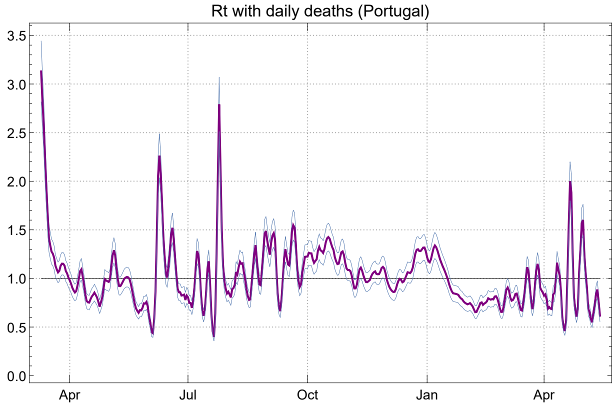


Figure 3.15: Reproduction number represented by the purple curve, using the number of deceased individuals instead of the new daily cases in equation (3.2) with a 7-day infectious period following a $\mathcal{N}(0, 4)$. The blue lines describe a 10% error margin and $\mathcal{R}_t = 1$ is illustrated by the black line.

Similar to the previous subsection, the plots of the reproduction number by gender and age class are presented in appendix A in Figures A.3 and A.4, respectively. The plots obtained from the non-constant viral load formula are very similar to the corresponding plot using the RKI formula, only a bit smoother. Regarding the values achieved for the age classes it is relevant to mention the basic reproduction number, $\mathcal{R}_0$. Up until the age of nineteen, this value is quite low in comparison to the other plots that indicate a $\mathcal{R}_0$ above two. The age class 60-69 stands out with an higher basic reproduction number of 5.24732.

3.4 Third Wave Analysis

It is particularly interesting to analyse how the variables are intertwined. To give a better perception on how the emergence of a new wave can be predicted and consequently, contained, we will focus ourselves on the third wave of the pandemic (from 26^{th} December, 2020 to 28^{th} February, 2021).

The first alarm signals one should be aware of is the number of daily new cases (Figure 3.1) and the positivity rate (Figure 3.7(b)). As the second wave was on its apparent downward trend, the number of daily cases started to rise. On New Year's Eve hits what it was at the time the second highest value ever recorded in Portugal, 7.627 cases, which occurs a week after Christmas. This was the first warning that the number of cases would start increasing in the following days if no actions were taken. The positivity rate accompanies the same movement and on January 6^{th} hits a new all time high with 19.0528% of tests positive. On the same day the threshold of 10.000 daily cases is crossed, with an alarming 10.027 cases. On this day, one could predict that the worst was yet to come as the effects of irresponsible celebrations of Christmas and New Year started to reveal itself catastrophic. The virus was already spread widely across the country and severe measures had to be taken immediately.

Unfortunately, the government was not duly prepared and no major decisions were made before the meeting at INFARMED (a Government agency accountable to the Health Ministry, that evaluates, authorises, regulates and controls human medicines as well as health products, namely, medical devices and cosmetics for the protection of Public Health) that took place on January 12^{th}, 2021. The decisions made on this day were only made effective on the 15^{th}, representing an eight day delay that allowed a further spread of the disease. The reproduction number computed in Section 3.3 is also another confirmation of how serious the situation was. A fast increase in this value occurred, with the threshold of $\mathcal{R}_t = 1$ being surpassed in a matter of days. On December 29^{th}, 2020, the reproduction number was 0.901646 and on January 7^{th}, 2021, this value was at an impressive 1.3514 (these values were obtained from the RKI formula with $\tau = 7$ presented in Figure 3.12(b)).

The lack of a fast response to a serious problem had tremendous consequences. As it can be observed in Figure 3.8, the number of hospitalised individuals increased immensely, including the number of critically ill patients in the ICUs. Unfortunately, the number of people that lost their lives too soon due to COVID-19 increased accordingly, leading to an higher case fatality ratio represented in Figure 3.9.

Finally, after the country went through a strict lockdown, only complete on January 22^{nd}, 2021, it was possible to contain a great amount of virus' outbreaks throughout the country and as time went by all indicators started its downward trend, indicating that this wave was finally ending.

3.5 Comparison with other Countries

Even if one believes that the previous analysis solely on Portugal is not enough to accurately analyse and predict the pandemic, a quick comparison with other countries will reveal how critical the situation in Portugal was. Taking advantage of a scientific online publication, Our World in Data (OWID), that uses interactive charts and maps to illustrate research findings on large global problems, we present Figure

3.16 and 3.17 attained from [48]. On such figures is a comparison of the daily new confirmed COVID-19 cases/deaths per million people in Portugal versus other countries and regions, respectively. To be noted that all plots have a 7-day moving average and in both cases there is some inaccuracy involved, mainly due to limited testing and also challenges to attribute the cause of death.

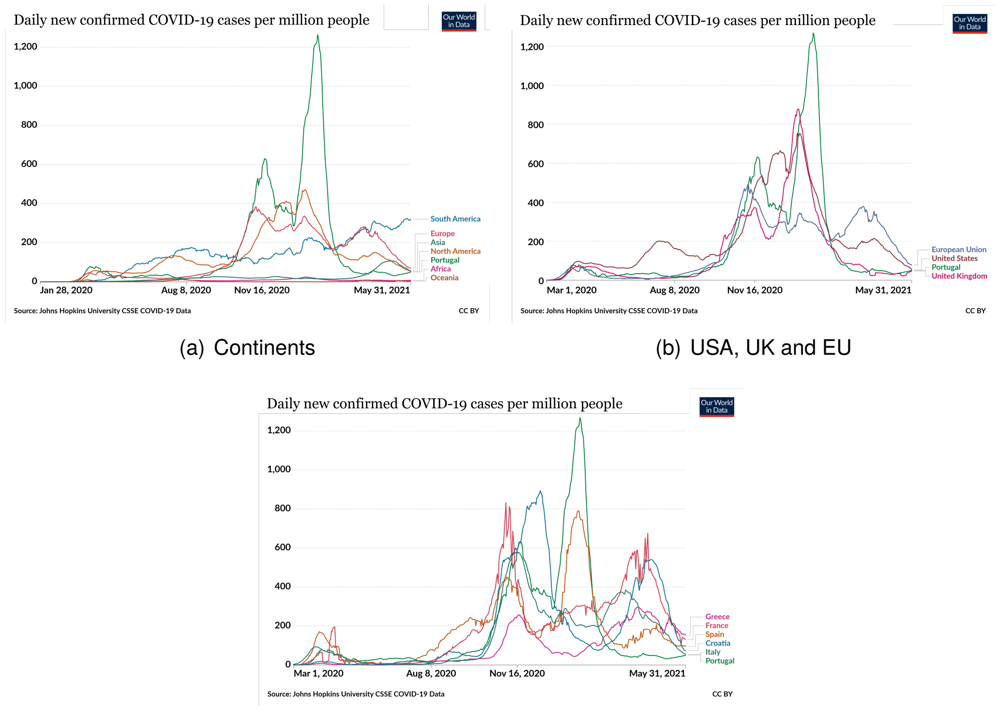
First of all, when comparing Portugal's numbers with the ones for each continent (Figures 3.16(a) and 3.17(a)), it is noticeable two very flat curves that correspond to Africa and Oceania. Each has very distinct reasons to justify such behaviour, mentioned in [60] and [19, 28], respectively.

Besides the weather, where higher temperatures have been linked to lower cases of COVID-19 in works like [10], the large majority of Africa is constituted by developing countries (with a low Human Development Index), which implies several characteristics associated with such countries. With a very youthful population, it is expected that most COVID-19 cases are either asymptomatic or just mild symptoms, resulting in a high rate of undetected cases. Combining high levels of pollution and poverty with low levels of access to safe drinking water, sanitation and hygiene leads to an high proportion of people with tropical and infectious diseases. The previous sentence makes us talk about the elephant in the room: with so many other diseases and problems these countries face, there is most likely an inaccurate counting of the number of cases and casualties and a lower CFR due to age structure.

In Oceania, countries like Australia and New Zealand tackled the virus very efficiently. The major key factor for such successful containments and management of all SARS-CoV-2 variants is a quick response based on the new daily data. As soon as a single new case emerged, severe measures were taken immediately to contain that chain of transmission as soon as possible, which contrasts with how European governments dealt with similar situations. This success proves that a strong public health response enforced by a government focused on regular testing, tracing and quarantine is the key to fighting a pandemic in its initial stages.

As far as the other continents go, its numbers of cases and deaths are quite low in comparison to Portugal which was expected since we are discussing extremely wide regions that might have extremely different numbers according to how each country managed the pandemic and consequently, it is not possible to withdraw significant conclusions on its regard.

At this point, it makes sense to shorten the spectrum to countries/regions that are more similar to Portugal. In order to do so, we resorted to the Country Similarity Index (CSI) [34] that is built based on 5 pillars: Demographics, Culture, Politics, Infrastructure and Geography. This index will allow a comparison not solely on sanitation, medical infrastructures and how the government responds to emergencies but also in terms of culture (traditions, religion, recreation and behaviour). This is of the utmost importance when one intends to inspect how a population will react in terms of following the measures/guidelines/restrictions even if it means breaking some habits/traditions. As expected, the first twenty positions are occupied by countries that are members/candidates of the EU. Even though countries like the United States of America (USA) and United Kingdom (UK) do not have an index with Portugal as high as other countries, both have reported extremely high numbers of cases/deaths throughout the pandemic. For this particular reason, these countries are represented on Figures 3.16(b) and 3.17(b) along with Portugal and the EU. The EU's number of cases/deaths was considerably low up until early October, when new waves started

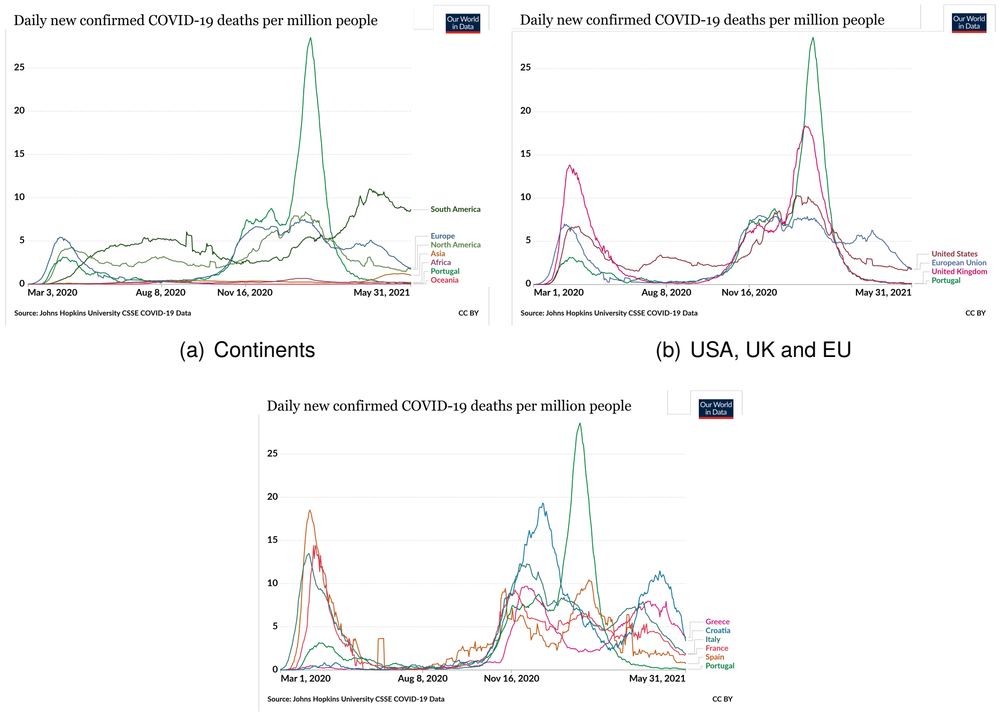


(a) Continents

(b) USA, UK and EU

(c) EU member countries

Figure 3.16: Comparison between the daily new confirmed COVID-19 cases per million people in Portugal and throughout the continents, United Kingdom (UK), United States of America (USA), European Union (EU) and some of its member countries.



(a) Continents

(b) USA, UK and EU

(c) EU member countries

Figure 3.17: Comparison between the daily new confirmed COVID-19 casualties per million people in Portugal and throughout the continents, United Kingdom (UK), United States of America (USA), European Union (EU) and some of its member countries.

emerging throughout Europe. Similar to what happened to the continents' curves, it is worth mentioning that EU's curves do not have a similar shape to the ones from epidemiological models (remaining more or less steady until the end of April). This can be explained by the fact that not all waves occurred during the same time period on all countries. Looking in particular to the evolution of COVID-19 in the USA and the UK, we notice curves that resemble what occurred in Portugal, especially the UK where the curves are extremely alike. Nonetheless, neither hit values per million people as high as Portugal did during the third wave.

At last, on Figures 3.16(c) and 3.17(c) is the comparison with the countries that score the highest on the CSI: Spain, Italy, France, Greece and Croatia, from highest to lowest. With these figures one can indeed confirm that the waves hit the European countries in different time periods, which is expected from infectious disease. Once again no country since the beginning of the pandemic had the amount of cases and deaths per million people as Portugal did in January, 2021. It is particularly interesting to notice that our neighbour Spain had the second and third wave during similar time periods as Portugal. Even if not reaching the high values of Portugal, Croatia had a very worrying second wave with almost 900 cases of COVID-19 per million and close to 20 deaths per million.

To sum up everything that has been stated so far, one thing becomes clear: the third wave of the pandemic in Portugal could and should have been better handled. By the 10^{th} of January, from the mentioned regions only the UK had more cases per million people than Portugal and not to mention that Portugal was considered worldwide to be the country with the highest daily incidence per million people from January 18^{th} to February 5^{th}, 2021. To be noted that at that time a further study on the circumstances of how each country was handling the disease should have been made before jumping to conclusions, such as the positivity rate and restrictions imposed by the government. Nonetheless, these results allow us to conclude that the third wave in Portugal was very worrying and if the government had decided to delay a few more days to take the severe measures required, the consequences would have been even more catastrophic: the national health care system could have easily collapsed, a bigger impact on heath care workers' physical and mental health, a larger number of deaths would have been registered, among others.

3.6 Model Adjustment

In order to adjust epidemiological models to the data in hands, research is deeply valued, given that it will allow for a better tuning of parameters. To properly measure which epidemiological model is the best fit, some error measurements were considered. The simpler and most well-known technique [39] to measure the amount of variance in a data set is the sum of square errors (SSE), occasionally known as residual sum of squares (RSS)

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{3.4}$$

Since we will be working on data with sufficiently large values, the SSE can very easily take extremely high values, and for that reason measures like the mean squared error (MSE) and residual mean squared error (RMSE) provide simple transformations of the SSE that can be more easily interpreted

$$MSE = \frac{\sum_{i=1}^{n}(y - \hat{y}_i)^2}{n} = \frac{SSE}{n}, \tag{3.5}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y - \hat{y}_i)^2}{n}} = \sqrt{MSE}. \tag{3.6}$$

It is relevant to point out that all the three previous measures rely on the sum of square errors, and subsequently if a set of parameters of a given model produces the lowest value for one of the previous error measurements, it will be the lowest for the remaining two measures as well.

Nonetheless, at times it will be necessary to compare curves that will be on different scales, and therefore the previous error measurements will not be able to accurately compare the results obtained for each curve. The solution for this problem is the mean absolute percentage error

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i}. \tag{3.7}$$

With the goal of adjusting the models discussed in Chapter 2, the class that we must try to replicate with the epidemiological models is the infected class. To do so, two functions were created, one for each model, SIR and SEIRD. Both functions receive an initial and final date corresponding to the time interval we aim to adjust, and a set of possible values for each parameter we are intending to tune. For the SIR model, only the transmission rate and the recovery rate are needed, whereas for the SEIRD model an extra parameter, responsible for controlling the latency period, must be tuned. 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.

The code used to implement both functions is available on Listings 1 and 2 in Appendix B. At each time step, with a different set of parameters, we resorted to *Mathematica*'s 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 mean absolute percentage error.

First and foremost, the time period of each wave must be defined. The second wave will be assumed to start on September 13^{th}, 2020, the day where the daily incidence of five hundred cases was surpassed. As for its end we cannot consider the same criteria since the emerge of a deadlier wave is the cause for its early and abrupt end. The 26^{th} of December, the day with least new daily cases, will mark both the end and the start of the second and third wave, respectively. Maintaining the same criteria of a threshold of five hundred cases per day, the 28^{th} of February, 2021, is the last day of the biggest wave Portugal has faced.

As previously stated, there are three data sets that account for the number of active cases on each day: the reports provided by DGS and the data obtained from the number of deaths with 7 and 14 days to recover/die. For each dataset, the two models, SIR and SEIRD, will be adjusted with the set of parameters that minimise the error measurements described above.

Some thresholds for the parameters were used when computing the functions. When adjusting the SEIRD model, the time one spends in the latency period is of at least two days and hence $\sigma \leq 0.5$. The ratio p_e was analysed by observing the infected curve. If within a small interval of time the number of individuals in the infectious class increased a given percentage, then these new arrivals came from the exposed class, creating an initial value to begin with. As for the SIR model, there was not a need to reduce as much the possibilities for the values of β and ρ since there are only two parameters and there is no problem in regards of computational cost. With both β and ρ ranging from 0 to 1 with 0.01 interval jumps, all $101 \times 101 = 10.201$ combinations of parameters will be considered for this model. Occasionally, a second tuning will occur considering a parameter error smaller than 10^{-3}.

3.6.1 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.

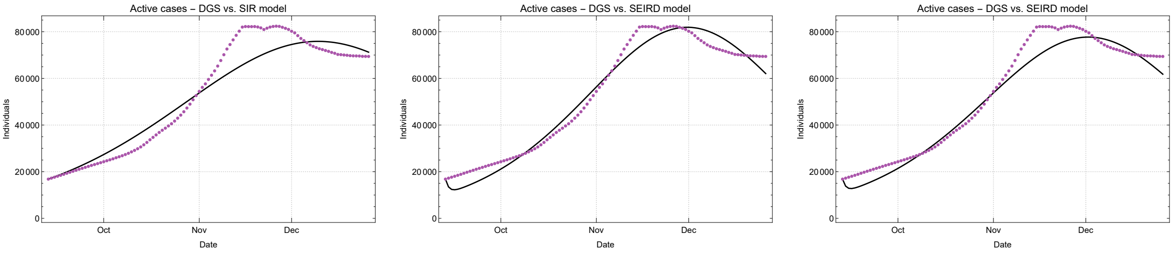
On Table 3.2 are presented the models that better adjust the number of active cases provided by the Portuguese government. The SIR model presented obtained the lowest values on all error measurements, with an infected period of $1/\rho \approx 4.17$ days.

Table 3.2: Error measurements for the models that best fit the data provided by DGS for the second wave.

	Parameters				Error Measurements			
	β	σ	ρ	p_e	SSE	MSE	RMSE	MAPE (%)
SIR	0.27	-	0.24	-	4.36372×10^9	4.15592×10^7	6446.64	9.56513
SEIRD	0.53	0.42	0.44	0.3	1.67633×10^9	1.5965×10^7	3995.62	8.55783
	0.50	0.45	0.42	0.3	2.35601×10^9	2.24381×10^7	4736.89	8.29674

As for the SEIRD model, two set of parameters are presented in view of the fact that one obtain better results in terms of square errors and another has a slightly better mean absolute percentage error. Nonetheless, the values obtained were quite similar. For the first, there is a latency period of $1/\sigma \approx 2.38$ days whereas the second indicates an approximate 2.22 days. This difference is not significant in practical terms and a similar situation occurs for the remaining parameters, where is relevant to point out that for both SEIRD models, there is an indication of very short infectious periods, approximately 2.27 and 2.38 days, respectively.

By observing Figure 3.18, we can state than even if the mean absolute percentage error is relatively low, no model has a similar shape to the data provided by DGS, as expected since there are some irregularities associated. To be noted that it was applied a 7-day moving average to try and smooth the data, and even so the curve is not as clean as one would like.



(a) Best SIR model (b) Best SEIRD model for square errors (c) Best SEIRD model for MAPE

Figure 3.18: Visual representation of the models presented in Table 3.2 for the second wave. The purple points correspond to the number of daily active cases provided by DGS. On the left, the black line represents the SIR model that performed better on all error measurements. On the remaining figures, the two best SEIRD models are presented by a black line. Figure 3.18(b) had the best results for square errors metrics and Figure 3.18(c) had the lowest value for the MAPE.

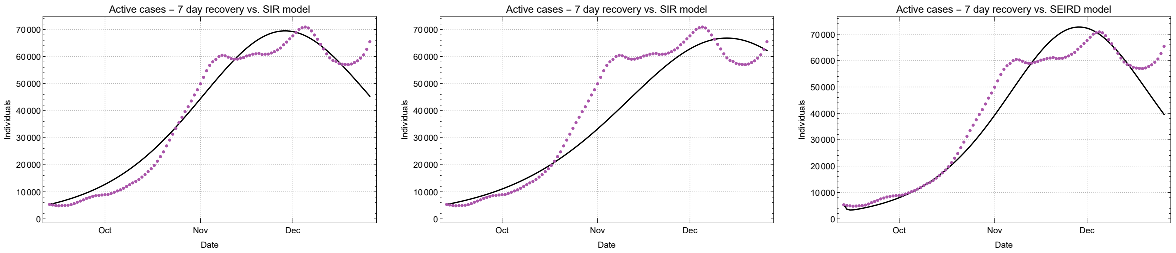
Next, the same process was repeated, but now taking advantage of the dataset built from the number of casualties and then a 7 day period for the individuals to leave the class, either by recovery or death. Before analysing the results, it is relevant to emphasise that this dataset is not very smooth during the second wave. The number of casualties increased but eventually reached a zone where there were small variations but never surpassing a given interval. Consequently, it is only to be expected a period of instability where the mathematical models will have a hard time adjusting.

Table 3.3: Error measurements for the models that best fit the data built from the number of deceased and with a 7-day period in the infected class for the second wave.

	Parameters				Error Measurements			
	β	σ	ρ	p_e	SSE	MSE	RMSE	MAPE (%)
SIR	0.46	-	0.4	-	2.81912×10^9	2.68488×10^7	5181.58	17.5323
	0.39	-	0.34	-	7.53108×10^9	7.17246×10^7	8469.04	17.1954
SEIRD	0.87	0.5	0.7	0.26	4.8211×10^9	4.59153×10^7	6776.08	12.0944

On Table 3.3 the best results in terms of metrics are presented. The more complex model has a significant difference for the MAPE when comparing to the one that attained the best value for this metric within the SIR model. This model has a 2-day latency period and $1/\rho \approx 1.43$ days as infectious.

The time spend in the infected class is, approximately, 2.5 and 2.94 days for the two best SIR models with respect to the square errors metrics and the MAPE, respectively.



(a) Best SIR model for square errors

(b) Best SIR model for MAPE

(c) Best SEIRD model

Figure 3.19: Visual representation of the models presented in Table 3.3 for the second wave. The purple points correspond to the number of daily active cases obtained from the data created from the number of casualties and with a 7-day period. On the left and middle, the black line represents the SIR model that performed better on the square error measures and on the MAPE, respectively. Figure 3.19(c) illustrates the SEIRD model with the best overall results with a black line.

At last, we fit SIR and SEIRD models to a dataset that was built in a similar manner to the previous, but with a 14-day period to leave the active class. As anticipated, the values for SSE, MSE and RMSE increase with this data. This is due to an higher level of infected individuals, justifying why these metrics lose reliability if we aim to compare data sets with very distinct values for the number of active cases.

In this case, it is presented on Table 3.4 two distinct set of parameters for each type of model. For a better reading, the first set will always be the one with the best values for the square errors metrics and the second set will be the one that scored the lowest in the MAPE.

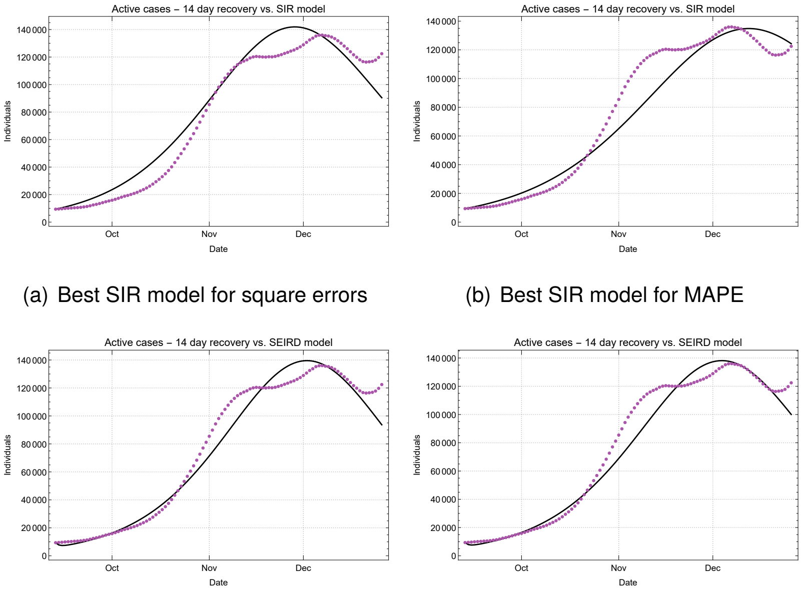
Table 3.4: Error measurements for the models that best fit the data built from the number of deceased and with a 14-day period in the infected class for the second wave.

	Parameters				Error Measurements			
	β	σ	ρ	p_e	SSE	MSE	RMSE	MAPE (%)
SIR	0.34	-	0.28	-	1.31841×10^{10}	1.25563×10^{8}	11205.5	21.2055
	0.29	-	0.24	-	1.62285×10^{10}	1.54557×10^{8}	12432.1	14.938
SEIRD	0.52	0.47	0.4	0.25	8.25886×10^{9}	7.86558×10^{7}	8868.81	9.05434
	0.49	0.5	0.38	0.25	8.47488×10^{9}	8.07131×10^{7}	8984.05	8.5526

In this data set, it becomes extremely clear that a model that incorporates a latency period adjusts better to the data for all error measurements. Both SIR models provide similar values for the infected period, 3.57 and 4.17 days, correspondingly.

As for the SEIRD model, both set of arguments are quite alike. The first model has a latency period of 2.13 days and an infectious period of 2.5 days, whereas the MAPE's best scorer has a 2-day period as non-contagious and considers that an individual spends 2.63 days as infectious.

Even with a 7-day moving average on the data set that considers a 14-day recovery period, the curve is not perfect to adjust mathematical models since, from mid-November until the 26^{th} of December, 2020, the number of active cases oscillates from 110.000 to 130.000, making it difficult to reduce the error measurements presented in Table 3.4.



(a) Best SIR model for square errors

(b) Best SIR model for MAPE

(c) Best SEIRD model for square errors

(d) Best SEIRD model for MAPE

Figure 3.20: Visual representation of the models presented in Table 3.4 for the second wave. The purple points correspond to the number of daily active cases obtained from the data created from the number of casualties and with a 14-day period. The black line represents the SIR model for the figures on top and the SEIRD model for the figures on the bottom. On the left are the graphs that had the best score for the square errors metrics. On the right are the graphs with the best MAPE values.

3.6.2 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 it is expected better results than those attained so far.

National Results

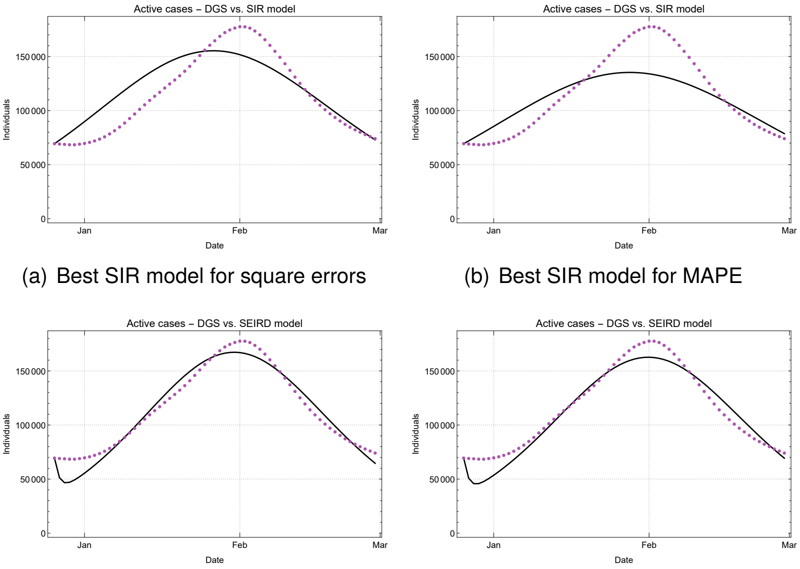
A big discrepancy regarding the error measurements is observed between the models that consider the host of the virus to pass through a latency period and those who do not (check Table 3.5). These are the lowest values obtained so far for the mean absolute percentage error, 6.87% and 6.83%. As usual the values of the parameters do not differ that much for the SEIRD models. The first has a latency of $1/\sigma = 2.5$ days, whilst the second has a 2.38 days period. Both models share an approximate value of 1.82 days as infectious.

The SIR model had a poor performance with this data, as it can be seen in Figure 3.21, where it has an hard time getting a similar shape to the curve of actives cases by DGS. Nonetheless, the time an individual spends as infectious is, according to these models, 3.23 and 3.45 days.

Unsurprisingly, when inspecting Figures 3.21(c) and 3.21(d), the mathematical models seem to adjust quite better to the third wave, and it is expected the improving of such values when using the two more reliable data sets.

Table 3.5: Error measurements for the models that best fit the data provided by DGS for the third wave.

	Parameters				Error Measurements			
	β	σ	ρ	p_e	SSE	MSE	RMSE	MAPE (%)
SIR	0.37	-	0.31	-	2.14586×10^{10}	3.30132×10^{8}	18169.5	13.9784
	0.34	-	0.29	-	2.7624×10^{10}	4.24984×10^{8}	20615.1	13.3031
SEIRD	0.75	0.42	0.55	0.25	4.58083×10^{9}	7.04743×10^{7}	8394.9	6.87465
	0.75	0.4	0.55	0.3	4.91156×10^{9}	7.55625×10^{7}	8692.67	6.82885



(a) Best SIR model for square errors

(b) Best SIR model for MAPE

(c) Best SEIRD model for square errors

(d) Best SEIRD model for MAPE

Figure 3.21: Visual representation of the models presented in Table 3.5 for the third wave. The purple points correspond to the number of daily active cases provided by DGS. The black line represents the SIR model for the figures on top and the SEIRD model for the figures on the bottom. On the left are the graphs that had the best score for the square errors metrics. On the right are the graphs with the best MAPE values.

Now, the models will be fitted into the most reliable data sets. Starting with the one that considers a 7-day period to leave the infected, the results are displayed in Table 3.6. In terms of square errors metrics, both SEIRD models achieved, once again, better results than the SIR models. As for the MAPE, the best model is, surprisingly, the one that does not have an exposed class.

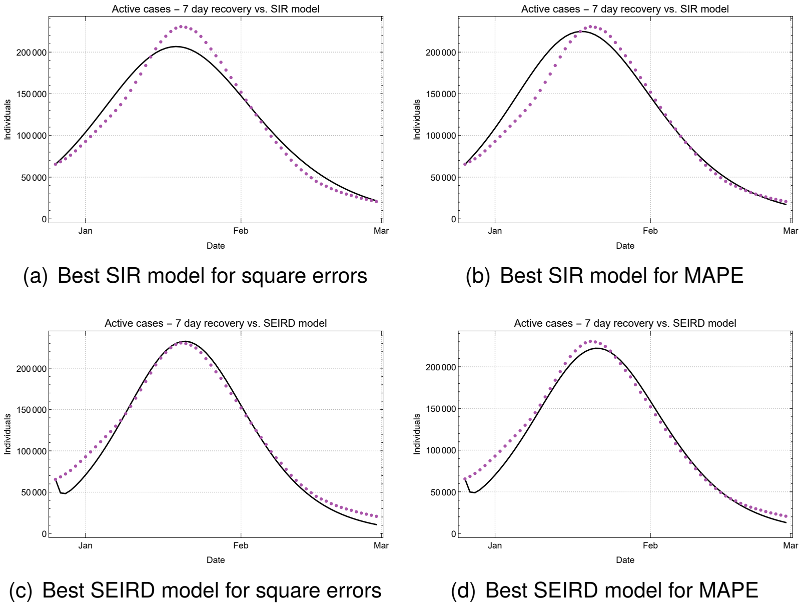
For the SIR models, the length of the infected period are very similar, 2.44 and 2.38 days. As for the SEIRD models, individuals spend approximately 2 days as non-contagious and around 1.45 days as infectious.

Even though the SEIRD models did not score as well as expected, Figure 3.22 explains why. The models are able to capture quite well the real number of active cases including its peak, however the first

Table 3.6: Error measurements for the models that best fit the data built from the number of deceased and with a 7-day period in the infected class for the third wave.

	Parameters				Error Measurements			
	β	σ	ρ	p_e	SSE	MSE	RMSE	MAPE (%)
SIR	0.54	-	0.41	-	1.05013×10^{10}	1.61558×10^{8}	12710.5	10.787
	0.56	-	0.42	-	1.29442×10^{10}	1.99142×10^{7}	14111.8	8.5582
SEIRD	1.1	0.49	0.7	0.4	6.00007×10^{9}	9.23087×10^{7}	9607.74	11.4101
	1.05	0.5	0.68	0.4	6.61105×10^{9}	1.01708×10^{7}	10085.1	9.58651

days underestimate slightly. This actively demonstrates that the level of under reporting might have been lower on these first few days than one anticipated.



(a) Best SIR model for square errors

(b) Best SIR model for MAPE

(c) Best SEIRD model for square errors

(d) Best SEIRD model for MAPE

Figure 3.22: Visual representation of the models presented in Table 3.6 for the third wave. The purple points correspond to the number of daily active cases obtained from the data created from the number of casualties and with a 7-day period. The black line represents the SIR model for the figures on top and the SEIRD model for the figures on the bottom. On the left are the graphs that had the best score for the square errors metrics. On the right are the graphs with the best MAPE values.

It is relevant to mention how the ratio p_e can have a great impact on these computations. This parameter is allowing the models to adjust the initial number of infected individuals and consequently, a possible improvement on the remaining days of the data set.

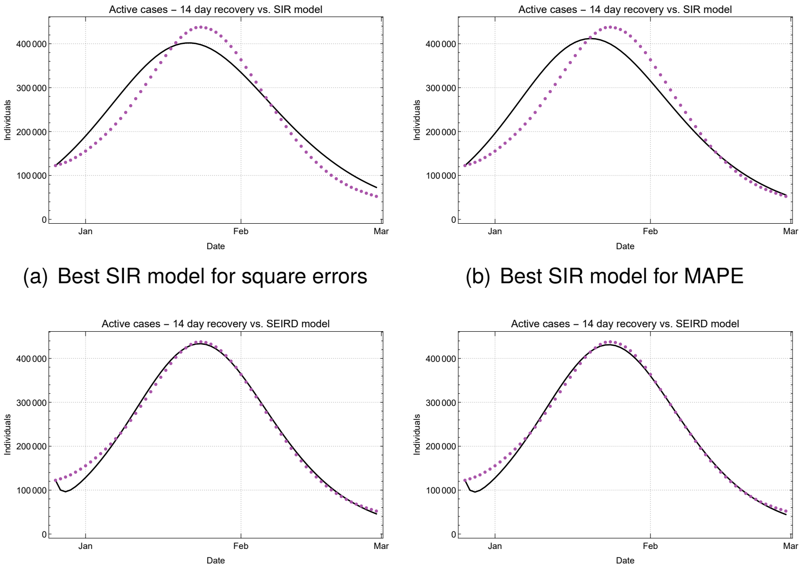
Last but not least, we will fit the epidemiological models to what is probably the most stable data. The previous statement is confirmed by the mean absolute percentage errors obtained by the SEIRD models (see Table 3.7), where the best score has only a 4.65% error. In spite of an high level of infected individuals (over 400.000), the errors for the square errors metrics are quite low.

The MAPE's best scorer states that the latency period is $1/\sigma \approx 2.44$ days and the infectious period is $1/\rho \approx 2.56$ days. The difference between SIR and SEIRD models is evident when checking Figure 3.23,

Table 3.7: Error measurements for the models that best fit the data built from the number of deceased and with a 14-day period in the infected class for the third wave.

	Parameters				Error Measurements			
	β	σ	ρ	p_e	SSE	MSE	RMSE	MAPE (%)
SIR	0.366	-	0.256	-	6.86537×10^{10}	1.05621×10^{9}	32499.4	16.4994
	0.388	-	0.27	-	9.60423×10^{10}	1.47757×10^{8}	38439.2	12.8535
SEIRD	0.66	0.415	0.385	0.18	8.79979×10^{9}	1.35381×10^{7}	11635.3	4.86516
	0.67	0.41	0.39	0.18	8.93433×10^{9}	1.37451×10^{7}	11724	4.64901

which was a result we intended on proving on this work. The SARS-CoV-2 has an average latency period of $2-5$ days, and thereby the models should corroborate this fact.



(a) Best SIR model for square errors

(b) Best SIR model for MAPE

(c) Best SEIRD model for square errors

(d) Best SEIRD model for MAPE

Figure 3.23: Visual representation of the models presented in Table 3.7 for the third wave. The purple points correspond to the number of daily active cases obtained from the data created from the number of casualties and with a 14-day period. The black line represents the SIR model for the figures on top and the SEIRD model for the figures on the bottom. On the left are the graphs on the left that had the best score for the square errors metrics. On the right are the graphs with the best MAPE values.

Once again, it seems that there was a bit of over estimation in the first few days of the data set built from the number of deaths. Regardless, the number of infected individuals predicted in the first few days of the third wave are still higher than what was reported by DGS.

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

As for the latency period obtained in the SEIRD models, all give an indication for a very short period, 2 to 2.5 days.

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.

Results by Gender

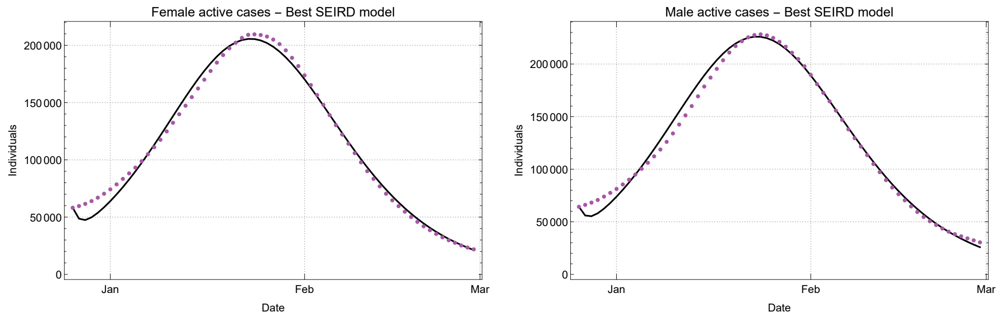
Since the results were better for the data with a 14-day time period from infection to death/recovery, we will now fit models for each gender during the third wave. Hence, it will be considered fourteen days from infection to death and/or recovery according to each model. In terms of gender, Portugal tends very slightly to a more female population accounting for 52.8% of the population, whereas the male population is 47.2% [45].

On table 3.8 is clear once again how much better the SEIRD model adjusts to this infectious disease. The female population seems to have slighter shorter latency and infectious period, $1/\sigma \approx 2.44$ and $1/\rho \approx 2.38$ days, than the male population, $1/\sigma \approx 2.36$ and $1/\rho \approx 2.94$ days, respectively. Of course, these small differences are not relevant in practical terms. Nonetheless, it seems that a bigger percentage of the number of females were in the latency period at the beginning of the wave, 20.635 versus 16.231 male individuals.

Table 3.8: Error measurements for the models that best fit each gender during the third wave.

		Parameters				Error Measurements			
		β	σ	ρ	p_e	SSE	MSE	RMSE	MAPE (%)
Female	SIR	0.384	-	0.275	-	1.62214×10^{10}	2.4956×10^{8}	15797.5	17.8803
		0.41	-	0.294	-	2.46836×10^{10}	3.79748×10^{8}	19487.1	13.4098
	SEIRD	0.7	0.41	0.42	0.32	1.98224×10^{9}	3.0496×10^{7}	5522.32	4.74318
		0.7	0.41	0.42	0.38	2.37107×10^{9}	3.64781×10^{7}	6039.71	4.29528
Male	SIR	0.349	-	0.238	-	1.83127×10^{10}	2.81734×10^{8}	16784.9	15.6141
		0.37	-	0.252	-	2.5768×10^{10}	3.96431×10^{8}	19910.6	12.3034
	SEIRD	0.59	0.42	0.34	0.26	1.99808×10^{9}	3.07398×10^{7}	5544.34	4.85753
		0.59	0.43	0.34	0.27	2.40264×10^{9}	3.69636×10^{7}	6079.77	4.45436

The best fit for each gender is presented on Figure 3.24 confirming how well the models fit the data.



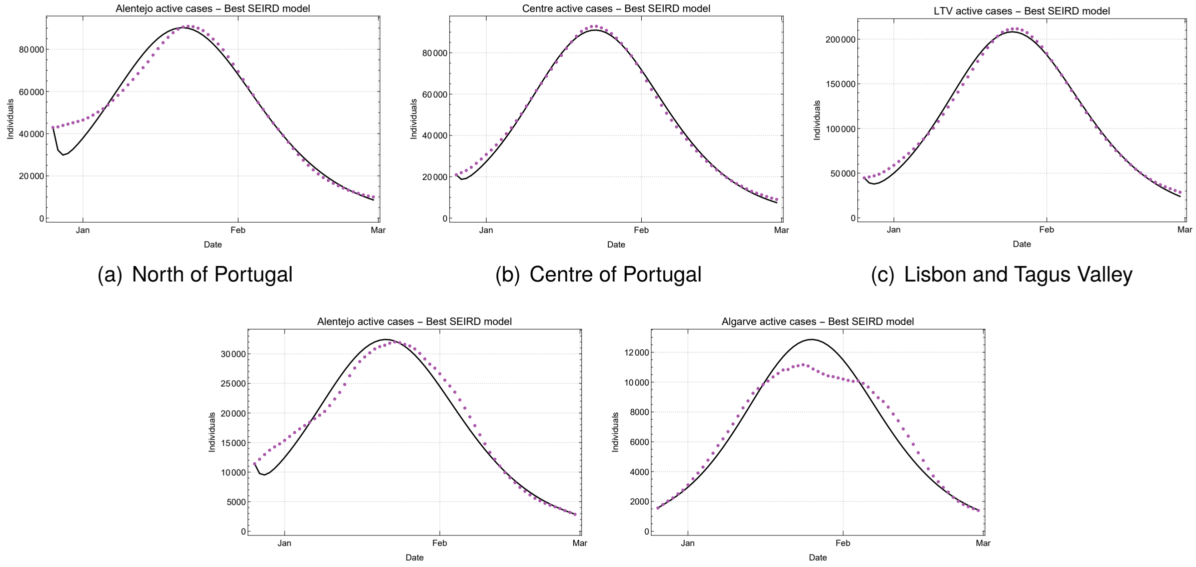
(a) Best female SEIRD model (b) Best male SEIRD model

Figure 3.24: Visual representation of the best SEIRD model in terms of MAPE for the female and male population during the third wave. The purple points correspond to the daily number of female/male active cases. The black line represents the curve for the infected class obtained from each model.

Results by Region

Taking into account, once more, how the population density can affect the spread of a given virus, the process made previously for each gender will be repeated for each region.

It is particular relevant to notice on Table 3.9 how misleading it is to compare regions based on square error measures. As it can be observed, regions with smaller populations (Alentejo and Algarve) have extremely good results for the RMSE, however this does not necessarily mean we are facing a good fit. For example, Lisbon and Tagus Valley has a poorer result for the square errors but in terms of MAPE score, it has a way better result.



(a) North of Portugal (b) Centre of Portugal (c) Lisbon and Tagus Valley

(d) Alentejo (e) Algarve

Figure 3.25: Visual representation of the best SEIRD model in terms of MAPE for each region during the third wave. The purple points correspond to the daily number of active cases. The black line represents the curve for the infected class obtained from each model.

To give a better perception on how each model fits the data on each region, plots for the best SEIRD model concerning the MAPE score are represented on Figure 3.25. Both regions with a smaller number of individuals have a more unstable curve, which explains why the models were never able to reach as lower MAPE scores as the other regions. It is remarkable how low the percentage error ($\leq 5\%$) is for the Centre of Portugal and Lisbon and Tagus Valley, specially taking into account the amount of errors that are associated with the data.

Table 3.9: Error measurements for the models that best fit each region during the third wave.

		Parameters				Error Measurements			
		β	σ	ρ	p_e	SSE	MSE	RMSE	MAPE (%)
North	SIR	0.399	-	0.3	-	5.24171×10^9	8.06418×10^7	8980.08	21.5112
		0.443	-	0.33	-	7.86073×10^9	1.20934×10^8	10997	15.1552
	SEIRD	0.76	0.45	0.49	0.3	1.10596×10^9	1.70147×10^7	4124.89	8.08957
		0.85	0.41	0.53	0.24	1.28722×10^9	1.98033×10^7	4450.09	6.21033
Centre	SIR	0.408	-	0.291	-	1.32212×10^9	2.03403×10^7	4510.02	11.3012
		0.424	-	0.302	-	1.87227×10^9	2.88041×10^7	5366.94	7.8828
	SEIRD	0.64	0.55	0.4	0.33	1.52967×10^8	2.35334×10^6	1534.06	4.26066
		0.64	0.54	0.4	0.33	1.84782×10^8	2.84279×10^6	1686.06	4.24997
LTV	SIR	0.309	-	0.187	-	1.60374×10^{10}	2.4673×10^9	15707.6	16.9956
		0.326	-	0.2	-	2.40606×10^{10}	3.70164×10^8	19239.6	13.5241
	SEIRD	0.63	0.25	0.28	0.31	1.25514×10^9	1.93098×10^7	4394.3	4.60571
		0.63	0.25	0.28	0.33	1.28047×10^9	1.96995×10^7	4438.41	4.52945
Alentejo	SIR	0.329	-	0.241	-	4.28573×10^8	6.59343×10^6	2567.77	20.6542
		0.365	-	0.267	-	8.75067×10^8	1.34626×10^7	3669.14	14.1445
	SEIRD	0.62	0.35	0.37	0.31	1.5406×10^8	2.37015×10^6	1539.53	9.91095
		0.61	0.39	0.37	0.31	1.85166×10^8	2.8487×10^6	1687.81	7.34761
Algarve	SIR	0.498	-	0.384	-	3.19065×10^7	4.90870×10^5	700.621	10.6266
		0.512	-	0.393	-	4.15505×10^7	6.39238×10^5	799.524	9.26353
	SEIRD	0.64	0.8	0.46	0.8	3.49047×10^7	5.36995×10^5	732.8	10.5888
		0.68	0.78	0.48	0.8	5.64738×10^7	8.68827×10^5	932.109	9.16083

It would also have been interesting to fit the data per age class, however we did not find information online regarding the total number of individuals with the same time gaps as the ones provided in the dataset at our disposal. The lack of the total number of people in a given age class forbids us to know how many susceptible individuals are in the population at all times and, consequently the initial condition is incomplete preventing the fit of any of the epidemic models discussed in this work.

3.7 Threshold of Herd Immunity

Hopefully, vaccines for infectious diseases are able to be fabricated in order to save a great number of lives. As discussed in section 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.

Nonetheless, genetic variants of SARS-CoV-2 have been rapidly emerging for which the vaccines are not as effective and natural infection no longer provide full immunity. Thereby, a bigger percentage of the population will require vaccination to achieve herd immunity. To this date, the Delta variant is believed to be the most transmissible yet, although more research is needed. The study [55] conducted by the Public Health England reports an 88% vaccine effectiveness for the Pfizer-BioNTech COVID-19, which is the one with the higher score in regards of V_e.

As explored in Subsection 2.2.5, by resorting to linearization, the explicit formula (2.48) for the basic reproduction number is attained. In order to stabilise the daily growth, a, of the number of infected individuals, a 7-day geometric mean is performed. The initial growth rate is $a_0 = 1.52859$. To finally obtain the number $\mathcal{R}_0$, we must know how long is the latency and infectious period.

On Table 3.10 is presented the basic reproduction number and both herd immunity values obtained from each pair of initial parameters chosen. In the first two lines, similar to what the RKI in Germany states, a 4-day latency period is considered and as far as the infectious period goes, it will be considered a 4 and 7 day period. The last two entries of the table have the best pair of input parameters, in terms of MAPE, obtained in the last section for the second and third wave, respectively.

Table 3.10: Basic reproduction number and herd immunity values attained from equations (2.48), (2.49) and (2.50). The input parameters on the first two lines are values used by the RKI. The last two are the best values obtained in the previous section for the second and third wave, respectively.

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

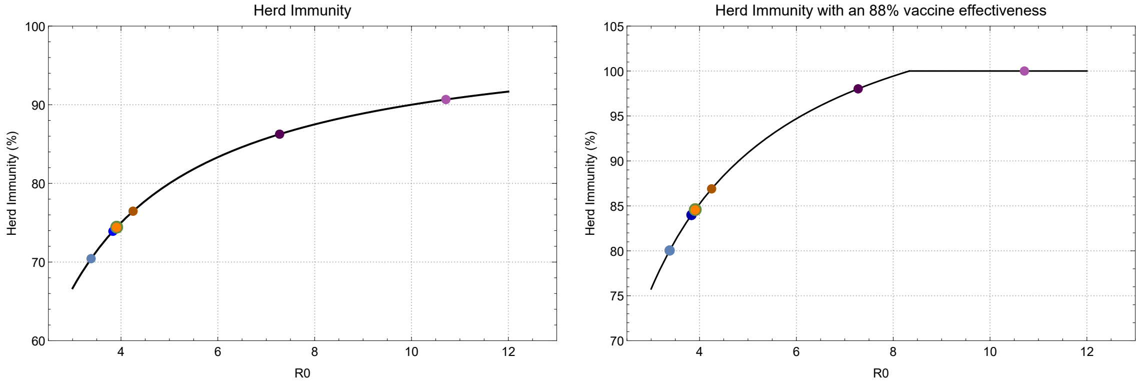
Another way to approach the problem is to take advantage of Section 3.3 to get the basic reproduction number, $\mathcal{R}_0$. Using the initial values of Figures 3.12(a), 3.12(b) and 3.13, the computation for the herd immunity easily follows, which are presented on Table 3.11. As one can observe, these results fall close to the case where the values of σ and ρ used were computed from model fitting of the second and third wave. Once again this suggests that we must consider a shorter period for the time an individuals spends as infectious since people isolate themselves once they discover to be carrying the virus. A major conclusion that can be withdrawn from such results, specifically on the first entry of the table, is the real estimation of basic reproduction number if people did not take precautions/measures regarding the virus, $\mathcal{R}_0 = 10.7097$.

The initial growth rate is computed with the initial values of the pandemic where the most simple measures such as wearing a mask in closed spaces did not exist and then it is considered a 7-day infectious period which even though is believed to be close to its real value, a person that follows the government measures regarding COVID-19 will not infect individuals during this whole period. With such a high value for the reproduction number, the herd immunity thresholds are extremely high and for the value that considers the SARS-CoV-2 variants there is indication that the whole population should be vaccinated.

Table 3.11: Basic reproduction number obtained from Figures 3.12(a), 3.12(b) and 3.13 in this particular order, and its herd immunity thresholds.

$\mathcal{R}_0$	$\mathcal{H}$	$\mathcal{H}_{adj}$
3.38217	0.704331	0.800377
3.83298	0.739106	0.839894
3.91142	0.744338	0.845839

For a better visualization, on Figure 3.26 are presented the results obtained above for the herd immunity values. 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.



(a) Herd Immunity

(b) Adjusted Herd Immunity

Figure 3.26: Visual representation of the herd immunity thresholds presented in Tables 3.10 and 3.11. The light and dark purple points were attained from the 4 and 7-day infectious period, respectively, presented on Table 3.10. The light and dark orange correspond to the best values obtained from model fitting for the second and third wave, respectively. The light and dark blue were derived from the values obtained using the $\mathcal{R}_t$ formula from RKI for $\tau = 4$ and $\tau = 7$. At last, the green point derived from the non-constant viral load formula of the reproduction number.

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.

Chapter 4

Proposal of a New Model

As mentioned in the previous chapter, 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 4.1).

Table 4.1: 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

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.

In Figure 4.1, the compartment flowchart for the model is displayed, where the scenario with overcrowded hospitals, specifically saturation of ICUs, is taken into account by modifying the flow between classes. The most common path for an infected individual in an hospital is to be admitted in the infirmary, and if its condition worsens he will be moved to the ICU. At this point, the patient either improves, returning to the infirmary, or the disease gets the best out of him, resulting in a casualty. However, it is not always this straightforward, there are infected individuals who are immediately admitted to the ICU, occasionally patients go back and forth from the ICU and the infirmary and some can die while still in the infirmary. After careful consideration, this model will not represent the exact dynamics described on the grounds that each body reacts to the virus differently and it is difficult to pinpoint the time spent in the infirmary before moving to the ICU. Bearing this in mind, this model will consider infected individuals to move directly to the critical class, and then move either to the infirmary or the deceased class.

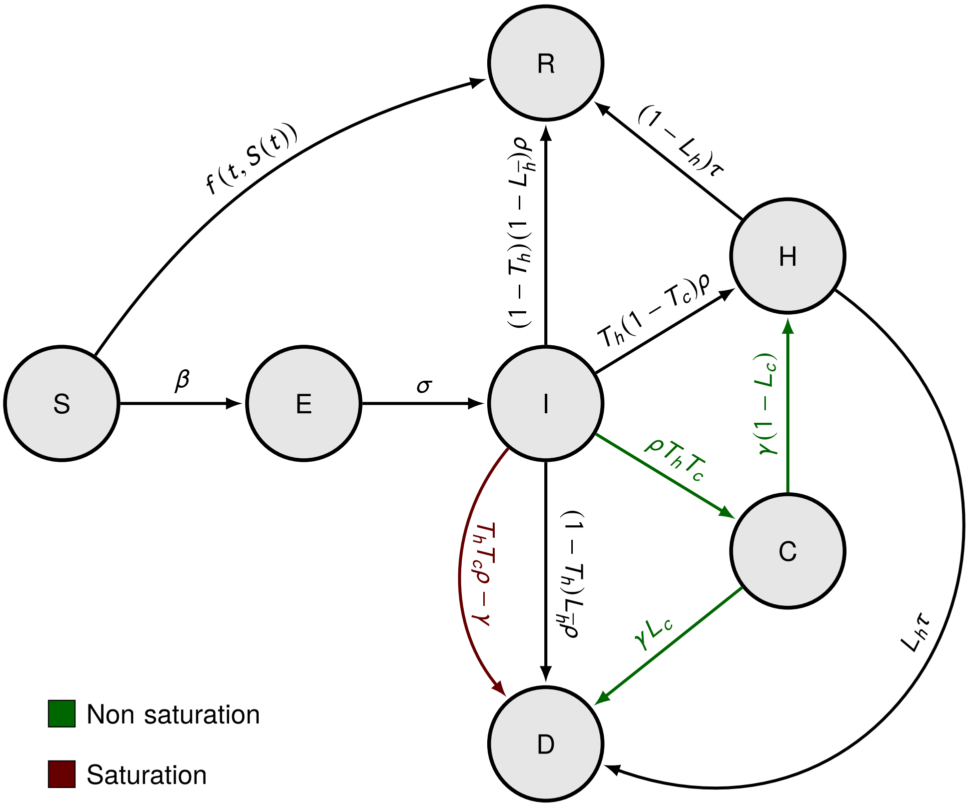


Figure 4.1: Compartment flowchart for SEIHCRD model.

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 $f(t, S(t))$ presented in Figure 4.1. For new viruses, not only the development of a vaccine takes time but also the large production necessary to meet the world's needs. Thereby, it is expected for an initially low value that increases as time goes by, explained by the arrival of an higher number of vaccines as the producing intensifies and several types of vaccines become available. Eventually, a threshold is achieved for which no more vaccines can be administrated daily due several factors, such as shortage of staff and lack of appropriate locations/equipment. During this period of time $f(t, S(t))$ is constant, until the number of susceptible begins approaching zero leading to a fast decrease on the number of daily vaccines administrated.

The model is modelled by the system of equations

$$\begin{cases} S'(t) = -\beta S(t)I(t) - f(t, S(t)) \\ E'(t) = \beta S(t)I(t) - \sigma E(t) \\ I'(t) = \sigma E(t) - \rho I(t) \\ H'(t) = T_h(1 - T_c)\rho I(t) + \gamma(1 - L_c)C(t) - \tau H(t) \\ C'(t) = (1 - Incr(t)Sat(t))[T_hT_c\rho I(t) - \gamma C(t)] \\ R'(t) = (1 - T_h)(1 - L_{\overline{h}})\rho I(t) + (1 - L_h)\tau H(t) + f(S(t), t) \\ D'(t) = (1 - T_h)L_{\overline{h}}\rho I(t) + L_h\tau H(t) + \gamma L_c C(t) + Incr(t)Sat(t)[T_hT_c\rho I(t) - \gamma C(t)], \end{cases} \tag{4.1}$$

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

$$Sat(t) = \frac{1}{1 + e^{-10(NC(t) - N_{cb})}} \tag{4.2}$$

$$Incr(t) = \frac{1}{1 + e^{-10^{10}(T_hT_c\rho I(t) - \gamma C(t))}}. \tag{4.3}$$

By taking advantage of sigmoid functions, there will be a rapid increase from 0 to 1 in specific points in time. The function regarding the trend of critical patients requires a steepest increase from 0 to 1 than the function that controls the saturation of critical beds, which are controlled by the factors 10 and 10^{10} respectively. The reason behind the previous statement relies with the fact that $Incr(t)$ must be more sensitive to sign changes in $T_hT_c\rho I(t) - \gamma C(t)$, since it represents the trend of the critical individuals, whereas if $NC(t) - N_{cb} < 0.5$ it can be considered that no more beds are available in the intensive care units. Consequently, if the ICUs are at its full capacity and there are still more patients who need specialised medical attention than those leaving the critical class, these will move to the deceased class.

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

$$N = S(t) + E(t) + I(t) + H(t) + C(t) + R(t) + D(t). \tag{4.4}$$

To prove the previous statement, all equations from (4.1) can be added resulting in the desired result $N'(t) = 0$ for all t

$$\begin{aligned}
N'(t) &= S'(t) + E'(t) + I'(t) + H'(t) + C'(t) + R'(t) + D'(t) = \\
&= -\beta S(t)I(t) - f(t, S(t)) + \beta S(t)I(t) - \sigma E(t) + \sigma E(t) - \rho I(t) + T_h(1 - T_c)\rho I(t) + \\
&\quad + \gamma(1 - L_c)C(t) - \tau H(t) + (1 - Incr(t)Sat(t))[T_h T_c \rho I(t) - \gamma C(t)] + \\
&\quad + (1 - T_h)(1 - L_{\bar{h}})\rho I(t) + (1 - L_h)\tau H(t) + f(S(t), t) + (1 - T_h)L_{\bar{h}}\rho I(t) + \\
&\quad + L_h \tau H(t) + \gamma L_c C(t) + Incr(t)Sat(t)[T_h T_c \rho I(t) - \gamma C(t)] = \\
&= -\rho I(t) + T_h \rho I(t) - T_h T_c \rho I(t) + \gamma C(t) - \gamma L_c C(t) - \tau H(t) + T_h T_c \rho I(t) - \gamma C(t) + \\
&\quad + \rho I(t) - L_{\bar{h}}\rho I(t) - T_h \rho I(t) + T_h L_{\bar{h}}\rho I(t) + \tau H(t) - L_h \tau H(t) + L_{\bar{h}}\rho I(t) - \\
&\quad - T_h L_{\bar{h}}\rho I(t) + L_h \tau H(t) + \gamma L_c C(t) = \\
&= 0.
\end{aligned} \tag{4.5}$$

4.2 Model Adjustment

In the previous chapter, computational programs were created to adjust models to the real number of infected individuals. However, in this case our main goal is to predict the number of intensive care patients. Consequently, it was necessary the creation of a function that simultaneously adjusted three curves (hospitalised, critical and deceased individuals), returning the best set of parameters.

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, resulting in a biased estimator that would give more importance to data with bigger values. 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.

The code used to implement the fit of SEIHCRD can be observed on listing 3 in Appendix B. Similar to the code used for the previous chapter, we take advantage of the command *NDSolve* to solve the ODEs for each set of parameters. As mentioned in the previous paragraph, the computation of the MAPE for the model obtained is the average of the errors for three curves: hospitalised, critical and deceased. The function returns not only the set of parameters that scored the lowest mean MAPE but also the ones with the best score for each curve. These last results show the importance of tuning all three curves at once. The cost of trying to have the best fit on only one curve is extremely high, resulting in poor results on the other two curves. Consequently, the overall panorama of the pandemic is not accurate and the model loses its relevance.

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 previous Chapter 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.

From papers like [2], we know that the mortality in the ICU must lie with the interval $35 - 50\%$. The article [35] was published early 2021, and it gathered 33 studies with 13.000 patients with a COVID-19 diagnosis and analysed the fatality rates of the disease. It was here stated that the mortality rate in hospitals, aside from critical care patients, is 11.5% with a 95% confidence interval of $7.7 - 16.9\%$. It is worth mentioning that this article reports a 40.5% mortality with a 95% C.I. of $31.2 - 50.6$ among critical ill patients, which supports the statement made in the beginning of the paragraph from [2].

The scientific article [13] published on October 2020 studies a group of over 14.000 patients from different hospitals in Belgium regarding the time from onset symptoms of COVID-19 to hospitalisation as well as the length of stay (LoS) in the hospital. The results for both time intervals are very similar, ranging from 3 to 10.4 days depending on the age of the patient. This work also has information on the LoS in ICUs, reporting an interval that varies from 3.8 to 15.7 days.

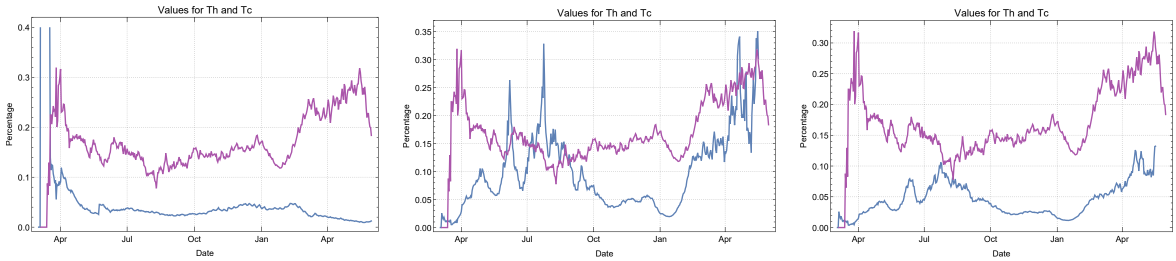
Two particular parameter need some extra attention. The way our model was created, there is not a constant to regulate how much it takes an individual to leave the infected class in case he is moving to the infirmary or to the ICU. The parameter ρ only controls how much time it takes to leave class I, 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.

With the information from the last paragraphs, we have some guidance for all parameters except one. There is still few information on the lethality of COVID-19 in individuals that never attended an hospital, nevertheless the health care conditions existent in a developed country, such as Portugal, are quite good, and thus this value should be low.

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.

The values for T_h' and T_c' can be estimated by computing the ratio of the total number of individuals that are in the hospital (infirmary and ICU) out of the total of infected and how many of the former are in the ICU. These values can be computed on a daily basis for each of the data.

The previous computations will allow us to focus solely on the time each individual stays on each class.



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

Figure 4.2: Visual representation of the rate of T_h' and T_c' by a blue and purple line, respectively. The three figures illustrate these values for each data set of infected individuals for the whole period of the pandemic.

To simplify we will take the mean of the previous ratios for the period of the 3^{rd} wave, 26^{th} December, 2020 to 28^{th} February, 2021. These values can be consulted on Table 4.2.

Table 4.2: Mean value for 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

Without further ado, we present the results. Starting with the data provided by DGS and taking advantage of the SEIRD model that had the best parameters with this data, the results for the SEIHCRD parameters are presented on the table below.

Table 4.3: Parameters of the best SEIHCRD model for the DGS data set, where the first four values are from the best MAPE performance of Table 3.5 and did not require tuning.

β	σ	ρ	p_e	τ	γ	T_h	T_c	L_h	$L_{\overline{h}}$	L_c
0.75	0.4	0.55	0.3	0.285714	0.203008	0.0256451	0.108343	0.077	0.0005	0.4

Now, it is necessary to analyse the previous parameters to check if they adjust well to reality. The average period spent in the infirmary is $1/\tau \approx 3.5$ days, and the LoS in the ICU is $1/\gamma \approx 4.92591$ days. Both values lie within the expected range mentioned earlier on this section. As for the values of T_h, we have

$$T_h = T_h' t_h \frac{1}{\rho} \Leftrightarrow t_h = \frac{\rho T_h}{T_h'} \Leftrightarrow t_h = 0.358025. \tag{4.6}$$

This value represents an approximate of 2.7931 days for an individual to require medical care in an hospital, a very close value to the mentioned range. As for the time until one is admitted to the ICU, it can be easily computed by solving

$$T_c = T_c' t_c \frac{1}{t_h} \Leftrightarrow t_c = \frac{T_c t_h}{T_c'} \Leftrightarrow t_c = 0.250617, \tag{4.7}$$

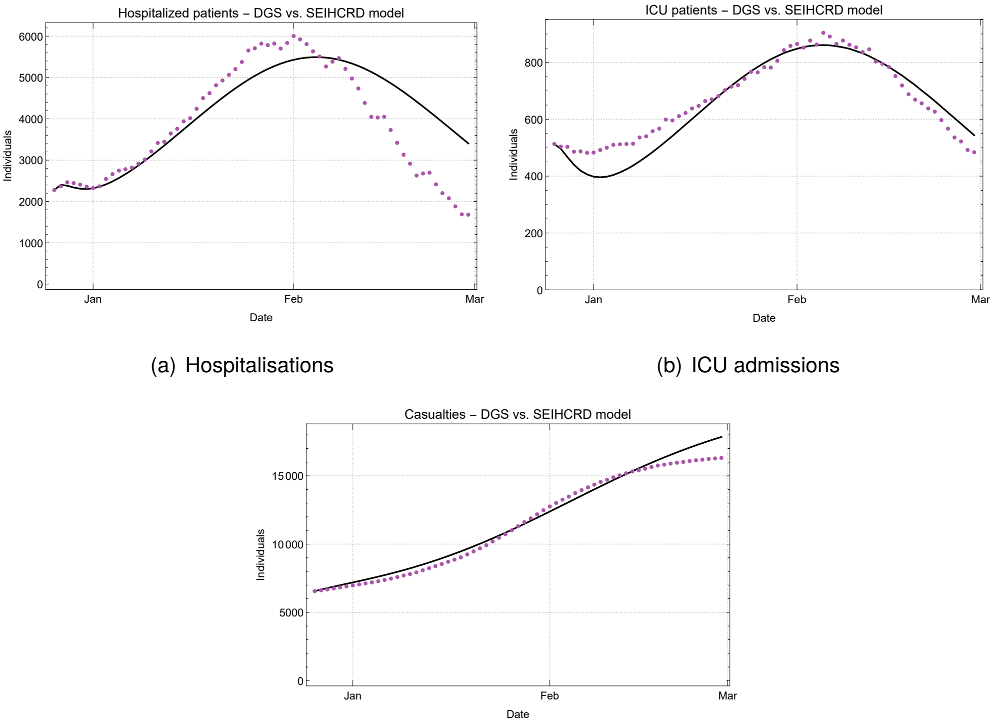
resulting in 3.99015 days. The remaining parameters do not need computations to be interpreted.

The following results confirm what has been previously commented: the number of active cases provided by DGS are not as reliable as one would like. The error measurements obtained in this model are quite good, except the hospitalised curve that scored an 19% error. Taking a closer look on Figure 4.3, where the curves of relevance are represented, 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 4.4: Mean absolute percentage error for the three relevant classes and its mean value, with the parameters from Table 4.3.

	Hospitalised (H)	Critical (C)	Deceased (D)	Mean value
MAPE (%)	19.1644	7.04538	3.18098	9.79691

It must be also pointed out that even though the MAPE for the deceased curve is very satisfactory, it can be perceived 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.



(a) Hospitalisations

(b) ICU admissions

(c) Deaths

Figure 4.3: Visual representation of the model presented in Table 4.3 for the three curves of interest. Hospitalised patients on the top left, ICU patients on the top right and number of deaths on the bottom. The red points represent the data we aim to fit the models into, which are represent by black curves.

For the data set that considers a 7-day infected period, the best model presents itself with the parameters shown on Table 4.5.

A similar process must be made once again. The average period spent in the infirmary is $1/\tau = 10.9091$

Table 4.5: Parameters of the best SEIHCRD model for the 7-day data set, where the first four values are from the best MAPE performance of Table 3.6 and do not needed tuning.

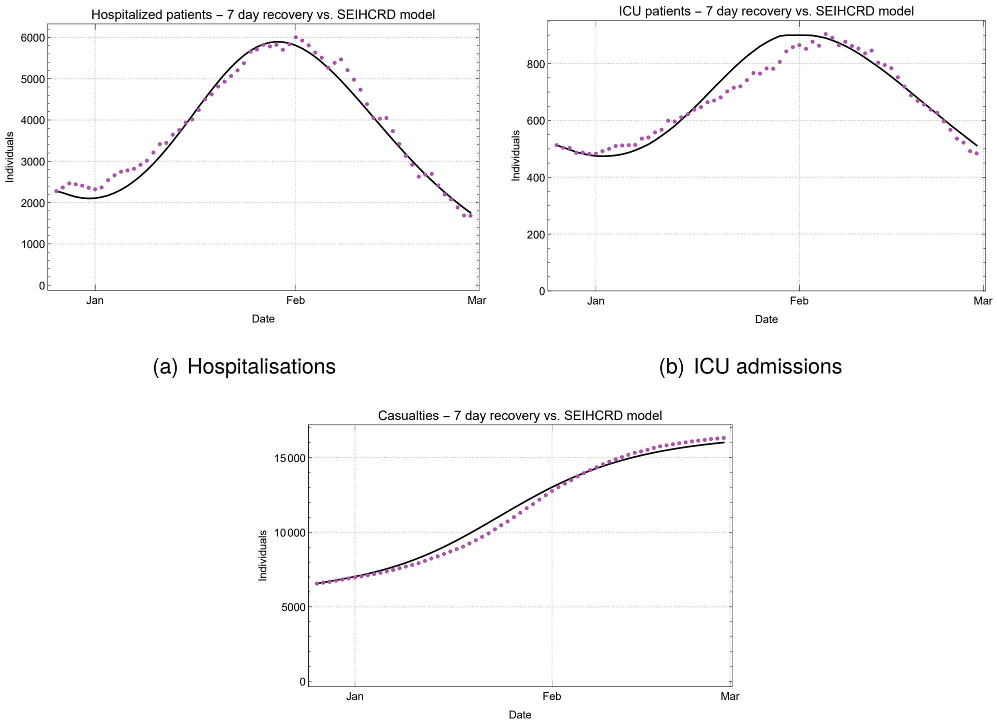
β	σ	ρ	p_e	τ	γ	T_h	T_c	L_h	$L_{\bar{h}}$	L_c
1.05	0.5	0.68	0.4	0.0916667	0.04	0.00451823	0.0742926	0.14	0.0011	0.48

days, and the LoS in the ICU is $1/\gamma = 25$ days. The first value lies right outside the edge of the interval mentioned earlier on this section. The LoS in the ICU is significantly higher than the expected maximum value of 15.7 days. As for T_h and T_c, the results are not satisfactory. The time it takes for one to be admitted the infirmary or ICU are too high and do not make sense in practical terms.

It is necessary to mention that when tuning the parameters at first, we took advantage of intervals obtained from other studies and scientific articles. Nonetheless, by analysing the graphs it was very straight forward which values could be changed in order to improve the results and that enhancement occurred instantly.

Table 4.6: Mean absolute percentage error for the three relevant classes and its mean value, with the parameters from Table 4.5.

	Hospitalised (H)	Critical (C)	Deceased (D)	Mean value
MAPE (%)	5.2134	4.11551	2.86027	4.06306



(a) Hospitalisations

(b) ICU admissions

(c) Deaths

Figure 4.4: Visual representation of the model presented in Table 4.5 for the three curves of interest. Hospitalised patients on the top left, ICU patients on the top right and number of deaths on the bottom. The red points represent the data we aim to fit the models into, which are represent by black curves.

This model is extremely well-fitted to the data (check Figure 4.4) with a mean MAPE of 4% approximately.

To be noticed that on Figure 4.4(b) the number of patients in the ICU hits the threshold of 900 for four days from January 30^{th} to February 2^{nd}, 2021.

At last, the SEIHCRD will be fitted into the 14-day recovery data. The best performance on the previous chapter was with this data, and the parameters obtained for this new model are presented on Table 4.7.

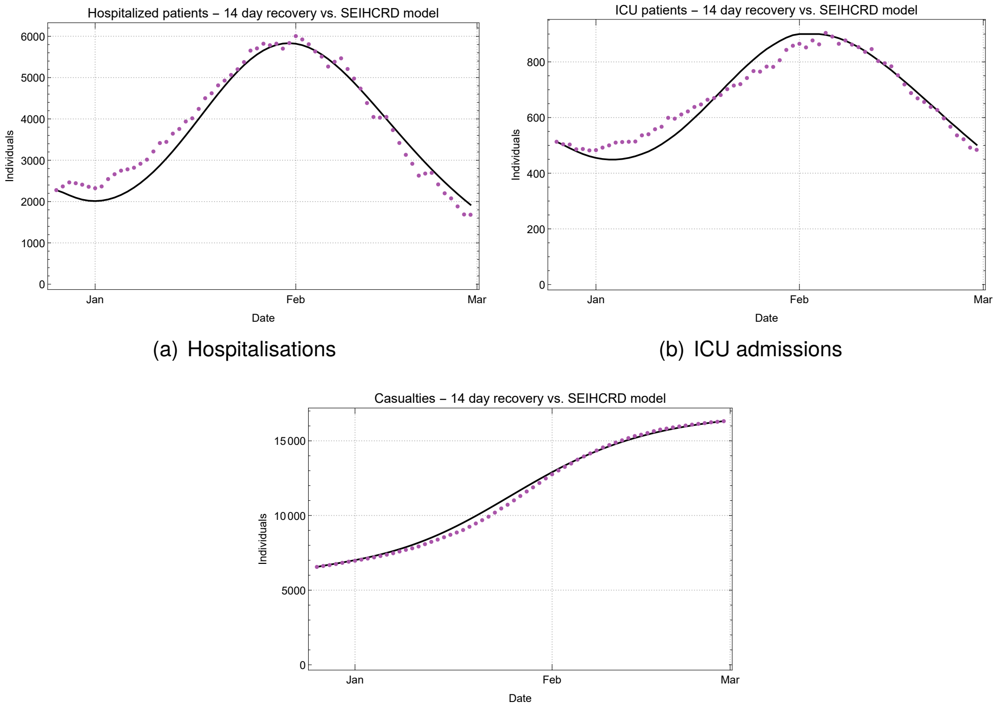
Table 4.7: Parameters of the best SEIHCRD model for the 14-day data set, where the first four values are from the best MAPE performance of Table 3.7 and do not needed tuning.

β	σ	ρ	p_e	τ	γ	T_h	T_c	L_h	$L_{\bar{h}}$	L_c
0.67	0.41	0.39	0.18	0.11	0.05	0.00458367	0.0773881	0.08	0.0011	0.49

For this model, an individual spends an average of $1/\tau = 9.09$ days in the infirmary and around $1/\gamma = 20$ days in the ICU. The first value lies within the expected range, whereas the second has a bit over four days to the expected top limit range. With the values of T_h and T_c, we can derive that it takes around 14.2547 days for a patient to be admitted in the infirmary and approximately 28.5094 days to arrive to the ICU. These values are a bit higher than expected, where the first value considers an extra four days than our initial expected range. As for the time it takes to get to the ICU, we still observed an unexpected high

Table 4.8: Mean absolute percentage error for the three relevant classes and its mean value, with the parameters from Table 4.7.

	Hospitalised (H)	Critical (C)	Deceased (D)	Mean value
MAPE (%)	8.00091	4.61049	1.80115	4.80419



(a) Hospitalisations

(b) ICU admissions

(c) Deaths

Figure 4.5: Visual representation of the model presented in Table 4.7 for the three curves of interest. Hospitalised patients on the top left, ICU patients on the top right and number of deaths on the bottom. The red points represent the data we aim to fit the models into, which are represent by black curves.

value, even if is not as high as the one attained with the previous data set.

This model has great scores for all three curves as it can be easily viewed with Table 4.8 with a mean error of 4.80%. The values presented in the table can be corroborated by observing Figure 4.5, where the curves have a great fit to the data.

All three fittings have mortality rates (in the infirmary and ICU) within the expected range.

One of the biggest struggles in this chapter was the amount of parameters that needed tuning and the different levels of sensitiveness. While in the previous chapter, 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. For example, for the T_h, we were tuning the third decimal place and on wards whereas for the lethality in the intensive care units L_c it was only by percentage point at most.

Our way to overcome this obstacle was to tune taking into account the parameter at hands. For example, for the rate of individuals in the hospital (infirmary and ICU), T_h, we aimed to know the value of t_h (number of days to get to the infirmary), thus we decomposed T_h so that the possible tuning parameters implied $1/t_h$ to be integer.

There are several observational errors that occurred until this point associated with population density, lack of resources at some point in time, under-reporting, inadequate control on the number of active cases, and so on. Bearing all of this in mind, an error below 5% is very impressive since the data itself has a significant error associated to disturb computations.

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

Chapter 5

Conclusions

5.1 General Overview

If there is something we can learn from this past year is that infectious diseases can have profound impact on the most diverse sectors and on society at large. Its implications are broad and complex, and as of today its depths are still unknown.

The amount of recent articles on this subject was both a bless and a curse. There was never lack of information regarding any subject, but there was the need to filter such huge amounts of information. Countries behave differently according to several factors (government, economic situation, religion, traditions, weather and so on), which require a careful analysis on what the circumstances of each study were. Furthermore, there was also a constant flow of articles emerging and new information about the virus that made it harder to keep up to date, and for that reason some of the bibliography might not be as correct and recent as one would like.

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 Chapter 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. The only exception is Algarve, with a latency period of 1.28 days, approximately. As mentioned previously, this region's curve does not have the shape one would like for different reasons, explaining the poorer and more unreliable results.

An important and interesting result was how the models accurately represented smaller infectious periods that are directly related to the isolation of individuals once they discovered to be bearers of the virus. No model pointed out an infectious period greater than six days, and the majority does not exceed three days.

As far as the model proposed in Chapter 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 Section 3.6, 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 government must keep a close connection to mathematicians and epidemiologists so that action can be taken fast and effectively and hence, severe consequences are avoided.

5.2 Future Directions of Work

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 chapter. The enhancement of these functions should be made.

As mentioned in Section 3.5, the comparison made between countries can be very important. This allows the assessment of the evolution of the pandemic in other countries that may be already experiencing a new wave. Consequently, countries can better prepare themselves. Nonetheless, a further and more thorough study should be made. The fit of epidemiological models in more populated countries could be really helpful considering that in Portugal it became harder to study the pandemic when the number of cases reduced significantly.

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: events, religion, weather, medical infrastructure, sanitation, how the government will respond to emergencies, among others.

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. Moreover, it would be very useful to have information regarding time spent in the hospital to possibly make adjustments to the model developed in Chapter 4.

With this type of information, the door to the field of survival analysis would also open. This branch of statistics could allow one to answer important questions such as: what is the proportion of a population which will survive past a certain time? Can multiple causes of death or failure be taken into account? How do particular circumstances or characteristics increase or decrease the probability of survival?

With the need to develop this field of study, several articles on the subject are being published and without a doubt remarkable advances in the area will be made.

Bibliography

[1] L. J. S. Allen. 'An Introduction to Stochastic Epidemic Models'. In: vol. 1945. Apr. 2008, pp. 81–130. ISBN: 978-3-540-78910-9. DOI: 10.1007/978-3-540-78911-6_3.

[2] R. A. Armstrong et al. 'Mortality in patients admitted to intensive care with COVID-19: an updated systematic review and meta-analysis of observational studies'. In: *Anaesthesia* 76.4 (2021), pp. 537–548. DOI: https://doi.org/10.1111/anae.15425. URL: https://associationofanaesthetists-publications.onlinelibrary.wiley.com/doi/abs/10.1111/anae.15425.

[3] I. Astuti and Ysrafil. 'Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2): An overview of viral structure and host response'. In: *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 14 (Apr. 2020). DOI: 10.1016/j.dsx.2020.04.020.

[4] 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.

[5] D. Bernoulli and S. Blower. 'An attempt at a new analysis of the mortality caused by smallpox and of the advantages of inoculation to prevent it'. In: *Reviews in Medical Virology* 14.5 (2004), pp. 275–288. DOI: 10.1002/rmv.443.

[6] F. Brauer. 'Mathematical epidemiology: Past, present, and future'. In: *Infectious Disease Modelling* 2 (Feb. 2017). DOI: 10.1016/j.idm.2017.02.001.

[7] F. Brauer and C. Castillo-Chavez. *Mathematical Models in Population Biology and Epidemiology*. 2nd ed. Texts in Applied Mathematics 40. Springer-Verlag New York, 2012. ISBN: 978-1-4614-1686-9.

[8] W. Budd. 'Typhoid Fever Its Nature, Mode Of Spreading, And Prevention'. In: *American Journal of Public Health* 8.8 (1918), pp. 610–612. DOI: 10.2105/ajph.8.8.610.

[9] J. P. Byrne. *Encyclopedia of Pestilence, Pandemics, and Plagues*. Encyclopedia of Pestilence, Pandemics, and Plagues v. 1. Greenwood Press, 2008. ISBN: 9780313341014.

[10] S. Chen et al. 'Climate and the spread of COVID-19'. In: *Scientific reports* 11 (Apr. 2021), p. 9042. DOI: 10.1038/s41598-021-87692-z.

[11] Atlanta (GA): Centers for Disease Control and Prevention (US). 'Science Brief: SARS-CoV-2 and Surface (Fomite) Transmission for Indoor Community Environments'. In: *National Center for Immunization and Respiratory Diseases (NCIRD), Division of Viral Diseases* (Apr. 2021).

[12] 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.

[13] C. Faes et al. ‘Time between Symptom Onset, Hospitalisation and Recovery or Death: Statistical Analysis of Belgian COVID-19 Patients’. In: *International Journal of Environmental Research and Public Health* 17.20 (2020). ISSN: 1660-4601. DOI: 10.3390/ijerph17207560. URL: https://www.mdpi.com/1660-4601/17/20/7560.

[14] W. Farr. ‘Second Annual Report of the Registrar-General on Births, Deaths, and Marriages in England, in 1840’. In: (1840), pp. 69–98.

[15] Wei-jie Guan et al. ‘Clinical Characteristics of Coronavirus Disease 2019 in China’. In: *New England Journal of Medicine* 382.18 (2020), pp. 1708–1720. DOI: 10.1056/NEJMoa2002032. eprint: https://doi.org/10.1056/NEJMoa2002032. URL: https://doi.org/10.1056/NEJMoa2002032.

[16] Y. R. Guo et al. ‘The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak- A n update on the status’. In: *Military Medical Research* 7 (Dec. 2020). DOI: 10.1186/s40779-020-00240-0.

[17] W. H. Hamer. ‘The Milroy Lectures on Epidemic Disease in England — The Evidence of Variability and of Persistency of Type.’ In: *The Lancet* 167.4305 (1906). Originally published as Volume 1, Issue 4305, pp. 569–574. ISSN: 0140-6736. DOI: https://doi.org/10.1016/S0140-6736(01)80187-2.

[18] E. M. Harrison, A. Docherty and Calum S. ‘COVID-19: time from symptom onset until death in UK hospitalised patients’. https://www.gov.uk/government/publications/co-cin-covid-19-time-from-symptom-onset-until-death-in-uk-hospitalised-patients-7-october-2020.

[19] W. A. Haseltine. ‘What Can We Learn From Australia’s Covid-19 Response?’ In: *Forbes* (24th Mar. 2021). (Accessed on 26/08/2021). URL: https://www.forbes.com/sites/williamhaseltine/2021/03/24/what-can-we-learn-from-australias-covid-19-response/?sh=435cec23a01a.

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

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.

[26] B. Hu, H. Guo and P. Zhou. 'Characteristics of SARS-CoV-2 and COVID-19'. In: *Nature Reviews Microbiology* 19 (Oct. 2020), pp. 1–14. DOI: 10.1038/s41579-020-00459-7.

[27] 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.

[28] A. Jones. 'How did New Zealand become Covid-19 free?' In: *BBC News* (10th July 2021). (Accessed on 26/08/2021). URL: https://www.bbc.com/news/world-asia-53274085.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] M. Khalili et al. 'Epidemiological Characteristics of COVID-19: A Systematic Review and Meta-Analysis'. In: *Epidemiology and Infection* 148 (June 2020), pp. 1–39. DOI: 10.1017/S0950268820001430.

[34] Objective Lists. *Which Countries are Most Similar to Portugal? 2.0*. https://objectivelists.com/2021/07/20/which-countries-are-most-similar-to-portugal/. (Accessed on 28/08/2021). July 2021.

[35] A. Macedo, N. Gonçalves and C. Febra. 'COVID-19 fatality rates in hospitalized patients: systematic review and meta-analysis'. In: *Annals of Epidemiology* 57 (May 2021), pp. 14–21. DOI: 10.1016/j.annepidem.2021.02.012.

[36] A. Marques. *COVID19Portugal_view*. https://esriportugal.maps.arcgis.com/home/item.html?id=803d4c90bbb04c03999e65e5ce411cf8#data. 2020.

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

[38] M. Mondelli et al. 'Low risk of SARS-CoV-2 transmission by fomites in real-life conditions'. In: *The Lancet Infectious Diseases* 21 (May 2021), e112. DOI: 10.1016/S1473-3099(20)30678-2.

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

[40] 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.

[41] World Health Organization. *HIV/AIDS*. https://www.who.int/news-room/fact-sheets/detail/hiv-aids. (Accessed on 20/03/2021). Nov. 2020.

[42] World Health Organization. *The top 10 causes of death*. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. (Accessed on 19/03/2021). Dec. 2020.

[43] World Health Organization. *WHO Director-General's opening remarks at the media briefing on COVID-19*. https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020. (Accessed on 18/05/2021). Mar. 2020.

[44] 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.

[45] PORDATA. *População residente, média anual: total e por sexo*. https://www.pordata.pt/Portugal/Popula%C3%A7%C3%A3o+residente++m%C3%A9dia+anual+total+e+por+sexo-6. (Accessed on 23/08/2021). June 2021.

[46] 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.

[47] 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.

[48] H. Ritchie et al. 'Coronavirus Pandemic (COVID-19)'. In: *Our World in Data* (2020). URL: https://ourworldindata.org/coronavirus.

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

[50] R. Ross. *The Prevention of Malaria*. London: J. Murray, 1911.

[51] Direção-Geral de Saúde. *COVID-19 - Relatório de Situação*. https://covid19.min-saude.pt/relatorio-de-situacao/. (Accessed on 31/05/2021). Mar. 2020.

[52] Yu Shi et al. 'An overview of COVID-19'. English. In: *Journal of Zhejiang University: Science B* 21.5 (May 2020). DOI: 10.1631/jzus.B2000083.

[53] J. Snow. 'The Mode Of Propagation Of Cholera.' In: *The Lancet* 67.1694 (1856), p. 184. DOI: 10.1016/s0140-6736(02)67846-8.

[54] H. E. Soper. 'The Interpretation of Periodicity in Disease Prevalence'. In: *Journal of the Royal Statistical Society* 92.1 (1929), pp. 34–73. DOI: 10.2307/2341437.

[55] J. Stowe, N. Andrews and C. et al. Gower. 'Effectiveness of COVID-19 vaccines against hospital admission with the Delta (B.1.617.2) variant'. https://khub.net/web/phe-national/public-library/-/document_library/v2WsRK3ZlEig/view/479607266.

[56] Thucydide, R. Warner and M. I. Finley. *History of the Peloponnesian War*. Penguin Books, 1972.

[57] Unknown. 'Porque não houve mortes por gripe em Portugal no último ano?' In: *SIC Notícias* (26th July 2021). (Accessed on 26/08/2021). URL: https://sicnoticias.pt/saude-e-bem-estar/2021-07-26-Porque-nao-houve-mortes-por-gripe-em-Portugal-no-ultimo-ano--f3701052.

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

[59] K. Wang et al. 'Analysis of the clinical characteristics of 77 COVID-19 deaths'. In: *Scientific reports* 10 (Oct. 2020), p. 16384. DOI: 10.1038/s41598-020-73136-7.

[60] P. Whiteside, P. Oakley and C. Aguilar Garcia. 'COVID-19 in Africa: Why is the official death rate so low?' In: *Sky News* (27th Mar. 2021). (Accessed on 26/08/2021). URL: https://news.sky.com/story/covid-19-in-africa-why-is-the-death-rate-so-low-12236347.

[61] 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.

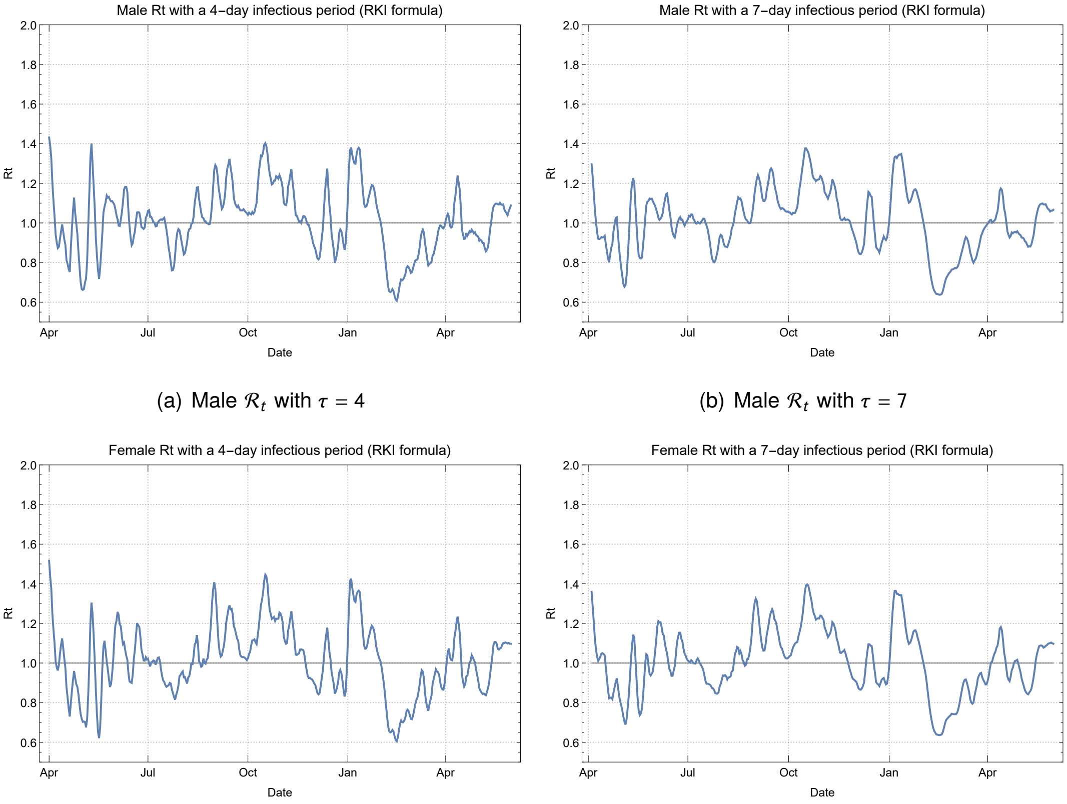
[62] N. Zhu et al. 'A Novel Coronavirus from Patients with Pneumonia in China, 2019'. In: *New England Journal of Medicine* 382.8 (2020). PMID: 31978945, pp. 727–733. DOI: 10.1056/NEJMoa2001017. URL: https://doi.org/10.1056/NEJMoa2001017.

Appendix A

Complementary Reproduction Number Plots

Complementary plots for the reproduction number according to the gender and age class.

A.1 Robert Koch Institute Formula



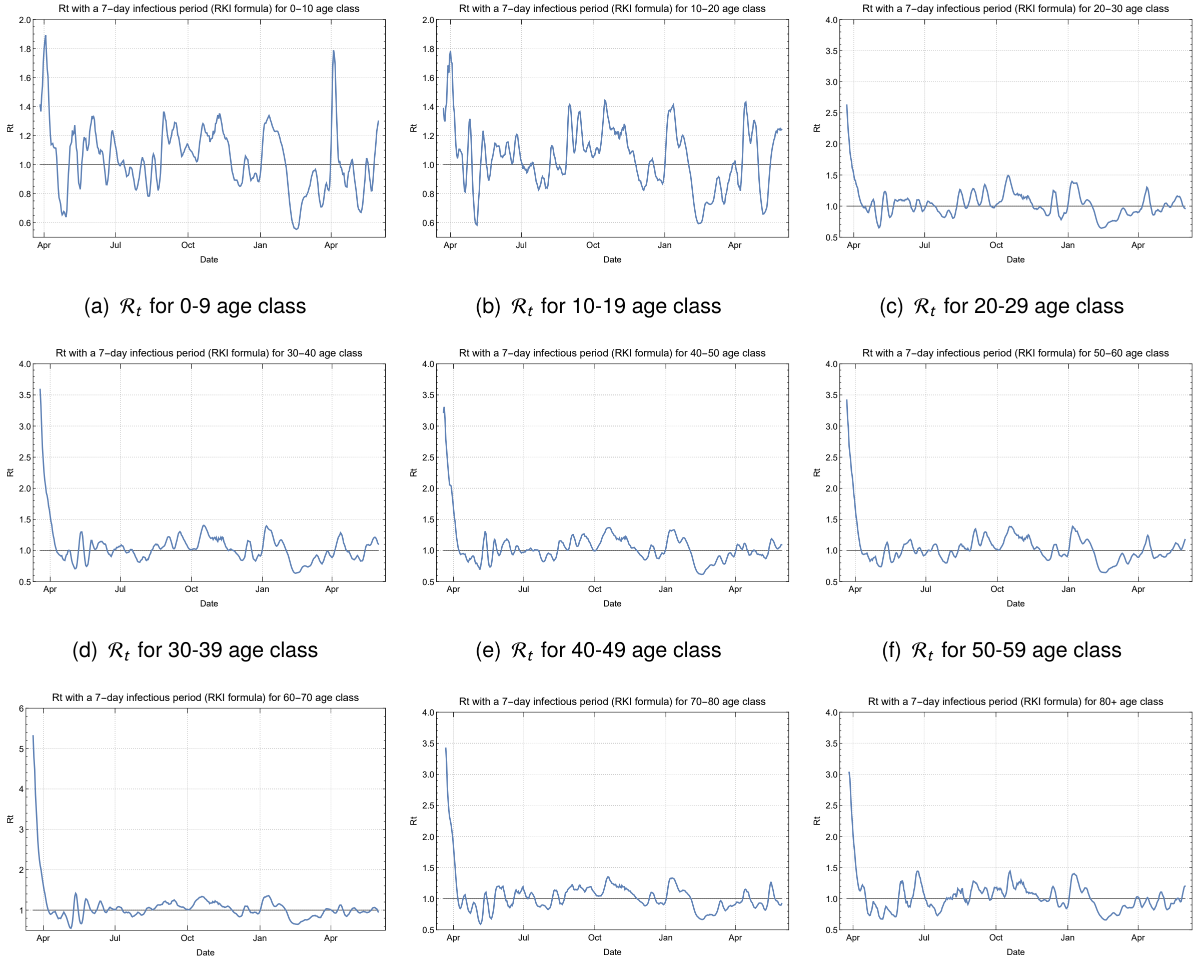
(a) Male $\mathcal{R}_t$ with $\tau = 4$

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

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

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

Figure A.1: Reproduction number (by RKI formula) for each gender and for an infectious period of $\tau = 4$ and $\tau = 7$, represented by the blue curve. The black line is threshold for $\mathcal{R}_t = 1$.



(a) $\mathcal{R}_t$ for 0-9 age class

(b) $\mathcal{R}_t$ for 10-19 age class

(c) $\mathcal{R}_t$ for 20-29 age class

(d) $\mathcal{R}_t$ for 30-39 age class

(e) $\mathcal{R}_t$ for 40-49 age class

(f) $\mathcal{R}_t$ for 50-59 age class

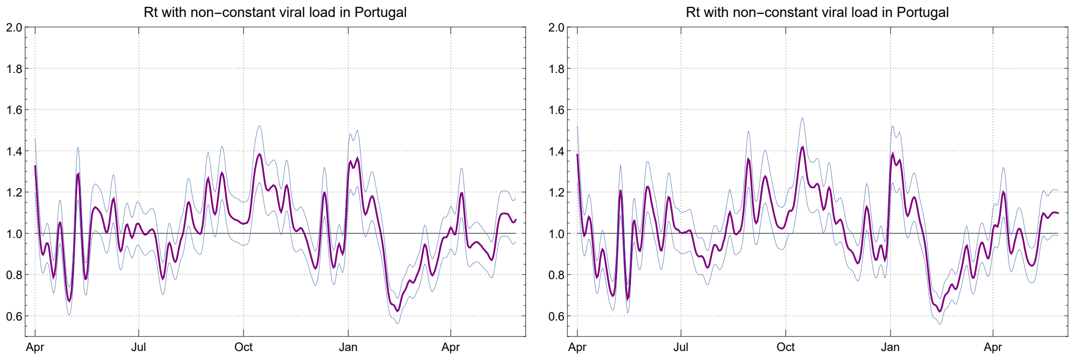
(g) $\mathcal{R}_t$ for 60-69 age class

(h) $\mathcal{R}_t$ for 70-79 age class

(i) $\mathcal{R}_t$ for 80+ age class

Figure A.2: Reproduction number (by RKI formula) for each age class of ten years each and for an infectious period of $\tau = 7$ and $\nu = 4$, represented by the blue curve. The black line is threshold for $\mathcal{R}_t = 1$.

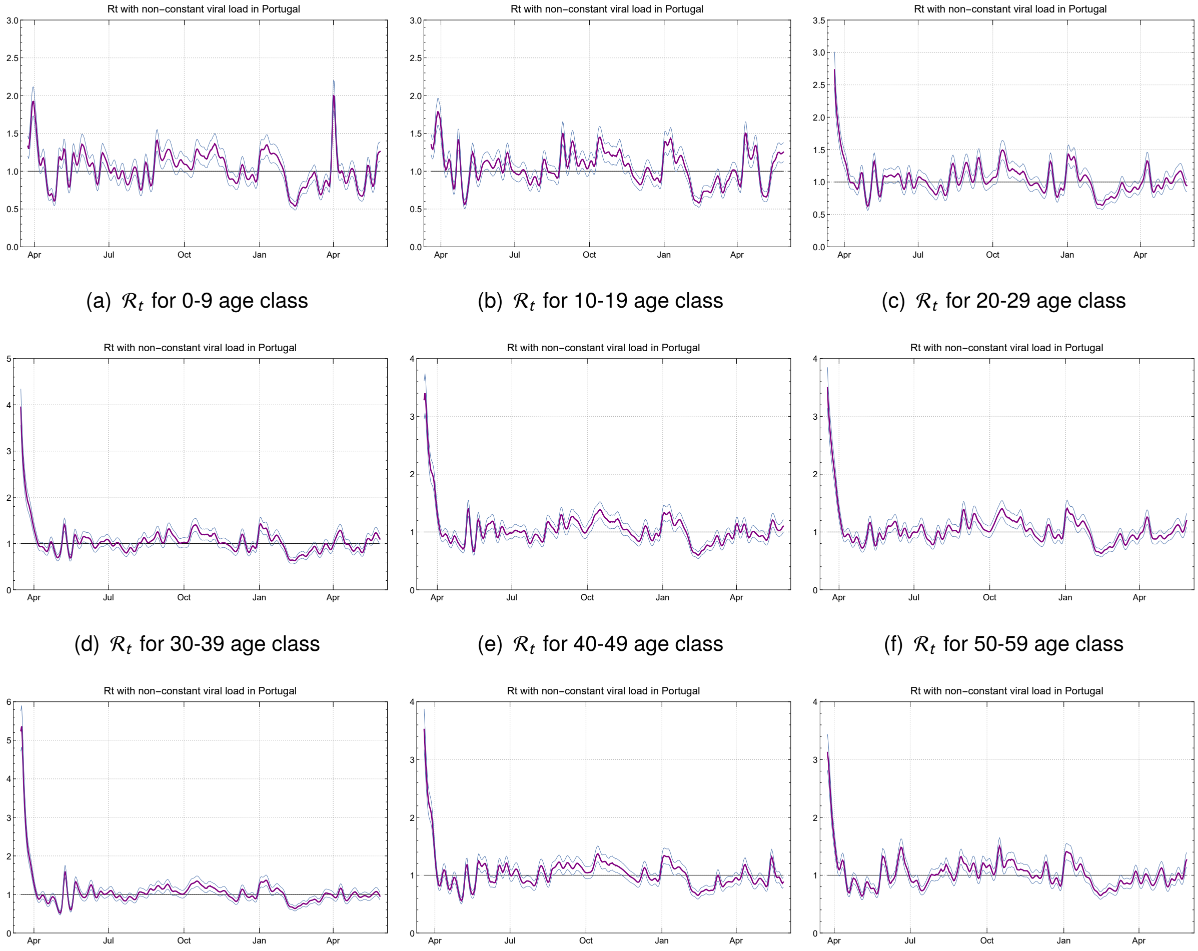
A.2 Non-constant Viral Load Formula



(a) Male $\mathcal{R}_t$

(b) Female $\mathcal{R}_t$

Figure A.3: Reproduction number (with non-constant viral load formula) for each gender and for $\tau = 7$ and $\nu = 4$, represented by the blue curve. The black line is threshold for $\mathcal{R}_t = 1$.



(a) $\mathcal{R}_t$ for 0-9 age class

(b) $\mathcal{R}_t$ for 10-19 age class

(c) $\mathcal{R}_t$ for 20-29 age class

(d) $\mathcal{R}_t$ for 30-39 age class

(e) $\mathcal{R}_t$ for 40-49 age class

(f) $\mathcal{R}_t$ for 50-59 age class

(g) $\mathcal{R}_t$ for 60-69 age class

(h) $\mathcal{R}_t$ for 70-79 age class

(i) $\mathcal{R}_t$ for 80+ age class

Figure A.4: Reproduction number (with non-constant viral load formula) for each age class of ten years each and for an infectious period of $\tau = 7$ and $\nu = 4$, represented by the blue curve. The black line is threshold for $\mathcal{R}_t = 1$.

Appendix B

Implementations in Wolfram Mathematica

Listing 1: SIR function.

```
(* Function that fits the best SIR model into the infected curve *)
(* The function receives:
        - the datasets (active cases and total number of cases) as a vector;
        - all possible values for each parameter of the model (\[Beta], \[Sigma]) as vectors;
        - first and last day of the data to fit, written in the format {year, month, day};
        - the population number (npop);
        *)
(* Function that fits the best SIR model into the infected curve *)

bestmodelSIR[betas_,rhos_,initialday_,finalday_,dataActive_,dataTotal_,npop_]:=Module[
        {start,end,point,realData,rmse,mse,sse,mape,infected,k,\[Beta],\[Rho],sir,
                    s,i,r,results,finalpairSSE,finalpairMAPE,posSSE,posMAPE,combinations},
        (* all the data in Portugal starts on the 26^{th} of February, 2020 *)
        start=DayCount[DateObject[{2020,2,26}],DateObject[initialday]]+1;
        Print[Style["Initial index of the dataset: ",Bold],start];
        end=DayCount[DateObject[{2020,2,26}],DateObject[finalday]]+1;
        Print[Style["Final index of the dataset: ",Bold],end];
        (* 7-day moving average to smooth the number of active cases *)
        realData=MovingAverage[dataActive[[start-6;;end]],7];
        (* Initial conditions *)
        point={(npop-dataTotal[[start]])/npop,realData[[1]]/npop,(dataTotal[[start]]-realData[[1]])/npop};
        Print[Style["Initial conditions: ",Bold],point];
        results={};
        combinations=Outer[List,betas,rhos]//Flatten[#,1]&; (* All possible combinations *)
        k=1;
        While[k<=Length[combinations],
                \[Beta]=combinations[[k,1]];
                \[Rho]=combinations[[k,2]];
                sir=NDSolve[ (* Solving the ODE *)
                    {s'[t]==-\[Beta]* i[t]*s[t], (* Susceptible class *)
                    i'[t]==\[Beta]*i[t]*s[t]-\[Rho]*i[t],  (* Infected class *)
                    r'[t]== \[Rho]*i[t], (* Recovered class *)
                    Thread[{s[0],i[0],r[0]}==point]},{s,i,r},{t,0,end-start}
                    ];
                (* Extracting the values for the infected for each day and computing the scores *)
                infected=Table[Floor[Extract[npop*i[j]/.sir,1]],{j,0,end-start}];
                mape=MAPE[realData,infected];
                rmse=RMSE[realData,infected];
                mse=MSE[realData,infected];
                sse=SSE[realData,infected];
                AppendTo[results,{{\[Beta],\[Rho]},sse,mse,rmse,mape}];
                k=k+1;
                ];
        (* Checking the best square errors and MAPE score *)
        posSSE=Position[results[[All,2]],Min[results[[All,2]]]][[1,1]];
        posMAPE=Position[results[[All,5]],Min[results[[All,5]]]][[1,1]];
        finalpairSSE=results[[posSSE]];
        finalpairMAPE=results[[posMAPE]];
        Print[Style["Best fit for square error metrics:",Bold]];
        Print["    ",Grid[Prepend[{Flatten[finalpairSSE]},{"\[Beta]","\[Rho]","SSE","MSE","RMSE","MAPE"}],Frame->All]];
        Print[Style["Best fit for MAPE:",Bold]];
        Print["    ",Grid[Prepend[{Flatten[finalpairMAPE]},{"\[Beta]","\[Rho]","SSE","MSE","RMSE","MAPE"}],Frame->All]];
        Return[{finalpairSSE,finalpairMAPE,point}]
        ];
```

Listing 2: SEIRD function.

```
(* Function that fits the best SEIRD model into the infected curve *)
(* The function receives:
        - the datasets (active cases, total number of cases, number of deceased individuals) as a vector;
        - all possible values for each parameter of the model (\[Beta], \[Sigma], \[Rho], factor) as vectors;
        - first and last day of the data to fit, writeen in the format {year, month, day};
        - the population number (npop);
        *)
bestmodelSEIRD[betas_,sigmas_,rhos_,initialday_,finalday_,dataActive_,dataTotal_,factor_,npop_,deathsRegion_]:=Module[
        {start,end,point,realData,sse,mse,rmse,mape,res2,infected,k,\[Beta],\[Sigma],\[Rho],
                seird,s,e,i,r,d,results,posSSE,posMAPE,finalpairSSE,finalpairMAPE,combinations,fator},
        (* all the data in Portugal starts on the 26^{th} of February, 2020 *)
        start=DayCount[DateObject[{2020,2,26}],DateObject[initialday]]+1;
        Print[Style["Initial index of the dataset: ",Bold],start];
        end=DayCount[DateObject[{2020,2,26}],DateObject[finalday]]+1;
        Print[Style["Final index of the dataset: ",Bold],end];
        (* 7-day moving average to smooth the number of active cases *)
        realData=MovingAverage[dataActive[[start-6;;end]],7];
        results={};
        combinations=Outer[List,betas,sigmas,rhos,factor]//Flatten[#,3]&; (* All possible combinations *)
        k=1;
        While[k<=Length[combinations],
                \[Beta]=combinations[[k,1]];
                \[Sigma]=combinations[[k,2]];
                \[Rho]=combinations[[k,3]];
                fator=combinations[[k,4]];
                (* Initial conditions *)
                point={(npop-dataTotal[[start]]-fator*realData[[1]])/npop,(fator*realData[[1]])/npop,realData[[1]]/npop,
                                (dataTotal[[start]]-realData[[1]]-deathsRegion[[start]])/npop,deathsRegion[[start]]/npop};
                seird=NDSolve[ (* Solving the ODE *)
                        {s'[t]==-\[Beta]* i[t]*s[t], (* Susceptible class *)
                        e'[t]==\[Beta]*i[t]*s[t]-\[Sigma] e[t], (* Exposed class *)
                        i'[t]==\[Sigma] e[t]-\[Rho] i[t], (* Infected class *)
                        r'[t]== \[Gamma]*\[Rho]*i[t], (* Recovered class *)
                        d'[t]== (1-\[Gamma])*\[Rho]*i[t], (* Deceased class *)
                        Thread[{s[0],e[0],i[0],r[0],d[0]}==point]},{s,e,i,r,d},{t,0,end-start}
                        ];
                (* Extracting the values for the infected for each day and computing the scores *)
                infected=Table[Floor[Extract[npop*i[j]/.seird,1]],{j,0,end-start}];
                sse=SSE[realData,infected];
                mse=MSE[realData,infected];
                rmse=RMSE[realData,infected];
                mape=MAPE[realData,infected];
                AppendTo[results,{{\[Beta],\[Sigma],\[Rho]},fator,sse,mse,rmse,mape,point}];
                k=k+1;
                ];
        (* Checking the best square errors and MAPE score *)
        posSSE=Position[results[[All,3]],Min[results[[All,3]]]][[1,1]];
        posMAPE=Position[results[[All,6]],Min[results[[All,6]]]][[1,1]];
        finalpairSSE=results[[posSSE]];
        finalpairMAPE=results[[posMAPE]];
        (* Printing the results *)
        Print[Style["Best fit for square error metrics:",Bold]];
        Print["    ","Initial conditions: ",finalpairSSE[[-1]]];
        Print["    ",Grid[Prepend[{Flatten[finalpairSSE[[;;-2]]]},
                            {"\[Beta]","\[Sigma]","\[Rho]","factor","SSE","MSE","RMSE","MAPE"}],Frame->All]];
        Print[Style["Best fit for MAPE:",Bold]];
        Print["    ","Initial conditions: ",finalpairMAPE[[-1]]];
        Print["    ",Grid[Prepend[{Flatten[finalpairMAPE[[;;-2]]]},
                            {"\[Beta]","\[Sigma]","\[Rho]","factor","SSE","MSE","RMSE","MAPE"}],Frame->All]];
        Return[{finalpairSSE,finalpairMAPE}]
        ];
```

Listing 3: SEIHCRD function.

```
(* Function that fits the best SEIHCRD model considering three curves: hospitalised, critical and deceased individuals *)
(* The function receives:
        - the datasets (active cases, total number of cases, number of hospitalised, critical and deceased individuals) as a
    vector;
        - all possible values for each parameter of the model (\[Beta], \[Sigma], \[Rho], \[Tau], \[Gamma], TH, TC, LH, LSH,
        LCs, factor) as vectors;
        - first and last day of the data to fit, written in the format {year, month, day};
        - the population number (npop);
        *)
bestmodelSEIHCRD[betas_,sigmas_,rhos_,taus_,gammas_,THs_,TCs_,LHs_,LSHs_,LCs_,initialday_,finalday_,dataActive_,
                dataHosp_,dataCrit_,dataDeath_,dataTotal_,factor_,npop_]:=Module[
        {start,end,point,realData,sseH,mseH,rmseH,mapeH,sseC,mseC,rmseC,mapeC,sseD,mseD,rmseD,mapeD,res2,
            hospitalised,critical,deaths,k,\[Beta],\[Sigma],\[Rho],\[Tau],\[Gamma],Th,Tc,Lh,Lsh,Lc,seihcrd,s,e,i,h,c,r,d,
            results,posH,posC,posD,pos,finalpairH,finalpairC,finalpairD,finalpair,combinations,fator,sat,incr},
        (* all the data in Portugal starts on the 26^{th} of February, 2020 *)
        start=DayCount[DateObject[{2020,2,26}],DateObject[initialday]]+1;
        Print[Style["Initial index of the dataset: ",Bold],start];
        end=DayCount[DateObject[{2020,2,26}],DateObject[finalday]]+1;
        Print[Style["Final index of the dataset: ",Bold],end];
        (* 7-day moving average to smooth the number of active cases *)
        realData=MovingAverage[dataActive[[start-6;;end]],7];
        results={};
        (* all possible parameter combinations *)
        combinations=Outer[List,betas,sigmas,rhos,taus,gammas,THs,TCs,LHs,LSHs,LCs,factor]//Flatten[#,10]&;
        k=1;
        While[k<=Length[combinations],
            {\[Beta],\[Sigma],\[Rho],\[Tau],\[Gamma],Th,Tc,Lh,Lsh,Lc,fator}=combinations[[k]];
            (* Initial conditions *)
            point={(npop-dataTotal[[start]]-fator*realData[[1]])/npop,(fator*realData[[1]])/npop,realData[[1]]/npop,
                    (dataHosp[[start]]-dataCrit[[start]])/npop,dataCrit[[start]]/npop,
                    (dataTotal[[start]]-realData[[1]]-dataHosp[[start]])/npop,dataDeath[[start]]/npop};
            seihcrd=NDSolve[ (* Solving the ODE *)
                    {sat[t_]:=1/(1+Exp[-10( Npop*c[t] -Ncb)]);(* if no more beds are available sat = 1 *)
                    (* if c(t) is an upward trend (risk of saturation) then incr(t) = 1 *)
                    incr[t_]:=1/(1+Exp[-10^10( Th*Tc*\[Rho]*i[t]-\[Gamma]*c[t])]);
                    s'[t]==-\[Beta]*i[t]*s[t], (* Susceptible class *)
                    e'[t]==\[Beta]*i[t]*s[t]-\[Sigma]*e[t], (* Exposed class *)
                    i'[t]==\[Sigma]*e[t]-\[Rho]*i[t], (* Infected class *)
                    h'[t]==Th*(1-Tc)*\[Rho]*i[t]+\[Gamma]*(1-Lc)*c[t]-\[Tau]*h[t], (* Hospitalised class *)
                    c'[t]==(1-sat[t]*incr[t])*(Th*Tc*\[Rho]*i[t]-\[Gamma]*c[t]), (* Critical class *)
                    r'[t]==(1-Th)*(1-Lsh)*\[Rho]*i[t]+(1-Lh)*\[Tau]*h[t], (* Recovered class *)
                    d'[t]==(1-Th)*Lsh*\[Rho]*i[t]+Lh*\[Tau]*h[t]+\[Gamma]*Lc*c[t]+
                            sat[t]*incr[t]*(Th*Tc*\[Rho]*i[t]-\[Gamma]*c[t]), (* Deceased class *)
                    Thread[{s[0],e[0],i[0],h[0],c[0],r[0],d[0]}==point]},{s,e,i,h,c,r,d},{t,0,end-start}
                    ];
            (* Extracting all the values for the three curves of interest and computing the MAPE score *)
            hospitalised=Table[Floor[Extract[npop*h[j]/.seihcrd,1]],{j,0,end-start}];
            mapeH=MAPE[dataHosp[[start;;end]]-dataCrit[[start;;end]],hospitalised];
            critical=Table[Floor[Extract[npop*c[j]/.seihcrd,1]],{j,0,end-start}];
            mapeC=MAPE[dataCrit[[start;;end]],critical];
            deaths=Table[Floor[Extract[npop*d[j]/.seihcrd,1]],{j,0,end-start}];
            mapeD=MAPE[dataDeath[[start;;end]],deaths];
            AppendTo[results,{{\[Beta],\[Sigma],\[Rho],\[Tau],\[Gamma],Th,Tc,Lh,Lsh,Lc},mapeH,mapeC,mapeD,point,fator}];
            k=k+1;
            ];
        (* Checking the best MAPE score for each curve *)
        posH=Position[results[[All,2]],Min[results[[All,2]]]][[1,1]];
        finalpairH=results[[posH]];
        posC=Position[results[[All,3]],Min[results[[All,3]]]][[1,1]];
        finalpairC=results[[posC]];
        posD=Position[results[[All,4]],Min[results[[All,4]]]][[1,1]];
        finalpairD=results[[posD]];
        (* Best overall MAPE score *)
        pos=Position[results[[All,2]]+results[[All,3]]+results[[All,4]],
                           Min[results[[All,2]]+results[[All,3]]+results[[All,4]]]][[1,1]];
        finalpair=results[[pos]];
        (* Printing the results *)
        Print[Style["Best fit for the Hospitalised curve:",Bold]];
        Print["    Initial conditions: ",finalpairH[[5]],";"];
        Print["    Parameters: \n    ",Grid[Prepend[
            {finalpairH[[1]]},{"\[Beta]","\[Sigma]","\[Rho]","\[Tau]","\[Gamma]","Th","Tc","Lh","Lsh","Lc"}],Frame->All]];
        Print["    MAPE Scores (\%): \n    ",Grid[Prepend[{Flatten[
            {finalpairH[[2;;4]],(Plus@@finalpairH[[2;;4]])/3}]},{"Hospitalised","Critical","Deceased","Mean"}],Frame->All]];
        Print["    Factor: ",finalpairH[[6]]];
        Print[Style["Best fit for the Critical curve:",Bold]];
        Print["    Initial conditions: ",finalpairC[[5]],";"];
        Print["    Parameters: \n    ",Grid[Prepend[
            {finalpairC[[1]]},{"\[Beta]","\[Sigma]","\[Rho]","\[Tau]","\[Gamma]","Th","Tc","Lh","Lsh","Lc"}],Frame->All]];
        Print["    MAPE Scores (\%): \n    ",Grid[Prepend[{Flatten[
            {finalpairC[[2;;4]],(Plus@@finalpairC[[2;;4]])/3}]},{"Hospitalised","Critical","Deceased","Mean"}],Frame->All]];
        Print["    Factor: ",finalpairC[[6]]];
        Print[Style["Best fit for the Deceased curve:",Bold]];
        Print["    Initial conditions: ",finalpairD[[5]],";"];
        Print["    Parameters: \n    ",Grid[Prepend[
            {finalpairD[[1]]},{"\[Beta]","\[Sigma]","\[Rho]","\[Tau]","\[Gamma]","Th","Tc","Lh","Lsh","Lc"}],Frame->All]];
        Print["    MAPE Scores (\%): \n    ",Grid[Prepend[{Flatten[
            {finalpairD[[2;;4]],(Plus@@finalpairD[[2;;4]])/3}]},{"Hospitalised","Critical","Deceased","Mean"}],Frame->All]];
        Print["    Factor: ",finalpairD[[6]]];
        Print[Style["Best fit for the Hospitalised + Critical + Deceased curve:",Bold]];
        Print["    Initial conditions: ",finalpair[[5]],";"];
        Print["    Parameters: \n    ",Grid[Prepend[
            {finalpair[[1]]},{"\[Beta]","\[Sigma]","\[Rho]","\[Tau]","\[Gamma]","Th","Tc","Lh","Lsh","Lc"}],Frame->All]];
        Print["    MAPE Scores (\%): \n    ",Grid[Prepend[{Flatten[
            {finalpair[[2;;4]],(Plus@@finalpair[[2;;4]])/3}]},{"Hospitalised","Critical","Deceased","Mean"}],Frame->All]];
        Print["    Factor: ",finalpair[[6]]];
        Return[{finalpairH,finalpairC,finalpairD,finalpair}]
        ];
```

TÉCNICO
LISBOA

Epidemiological Models: SARS-CoV-2 in Portugal

Maria Beatriz Silva Santiago