Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE

Corso di Laurea in Informatica

A compartmental model for the analysis and prediction of COVID 19 spread

Relatore:
Chiar.ma Prof.ssa.
Elena Loli Piccolomini

Presentata da:
Pietro Miotti

Anno accademico 2019/2020
II Sessione

"Nay, be a Columbus to whole new continents and worlds within you, opening new channels, not of trade, but of thought".

Henry David Thoreau

Contents

1 Modeling Infectious Disease **9**

1.1 What are models and why modeling? 9
1.2 What is a good model? 10
1.3 A modeling dichotomy 10
1.4 SIR Model 11
1.5 Ordinary Differential Equations 12
1.6 The Threshold Phenomenon 14
1.7 R_0 Index 15
1.8 Introducing Demography 16
1.9 Other Models 17

2 Time forced and SEIRD model **19**

2.1 Time forced models 20
2.2 SEIRD time forced model 21

3 Numerical Results **25**

3.1 Epidemic Data 25
3.2 Cauchy Problem initialization 26
3.3 Interval Variations 26
3.4 Parameter Estimation and Prevision 29
3.5 Computational Strategies 29
3.6 Results 30
3.7 An alternative approach for Exposed Population 36

4 Conclusions **39**

List of Figures

1.1 Agent Based Approach 11
1.2 Compartment based model (SIR) 11
1.3 SIRD model 17
1.4 SIS model 17
1.5 SICA model 18

2.1 Influenza cases in US per week 19
2.2 Influenza cases in US per year [5] 20
2.3 Time Line infection 22
2.4 SEIRD model 23

3.1 Iteration scheme of the optimization process 26
3.2 Large Error estimation without splitting the integration period 27
3.3 Updated Iteration Scheme with Variational Interval approach 28
3.4 Definitive Schema of the optimization process used 30
3.5 Estimation and Prevision of Infected people in Emilia Romagna using time-forced beta 3.2 31
3.6 Estimation and Prevision of Infected people in Emilia Romagna using time-forced beta 3.1 31
3.7 Estimation and Prevision of Infected people in Lombardia using time-forced beta 3.2 31
3.8 Estimation and Prevision of Infected people in Lombardia using time-forced beta 3.1 32
3.9 Estimation and Prevision of Infected people in Emilia Romagna using time-forced beta 3.2 starting from 22 June 33
3.10 Estimation and Prevision of Infected people in Emilia Romagna using time-forced beta 3.1 starting from 22 June 33
3.11 Estimation and Prevision of Infected people in Lombardia using time-forced beta 3.2 starting from 22 June 34

3.12 Estimation and Prevision of Infected people in Lombardia using time-forced beta 3.1 starting from 22 June 34
3.13 Emilia Romagna prediction with ρ of the penalty factor set to 1 37
3.14 Emilia Romagna prediction with ρ of the penalty factor set to 0 37

Introduction

I am quite confident to say that the topic of this thesis doesn't need a proper presentation. We all know what SARS-CoV-2 / Covid-19 is, since for all the past year it plays a main role in our lives: it impacted and changed our habits, our relationships, our hobbies and more general our every day life. It is quite strange to say that Covid became a constant in our lives and I feel that little by little we are getting used to it: we are getting used to wear masks and to avoid crowded places, we are getting used to always new restrictions and we are getting used to online lessons.
By the way, In this thesis I will try to explain what SARS-CoV-2 is from a more mathematical perspective and treat it as an infectious disease.
To get a full understanding we firstly have to start from definitions and clarify what we exactly mean by disease. As main reference we take in consideration the definition given by the Oxford English Dictionary: a disease is "a condition of the body, or some part or organ of the body, in which its functions are disturbed or deranged, a morbid physical condition, a departure from the state of health, especially when caused by structural change". As we expected this definition is quite general and hence leads to a wide range of different categories of diseases, but for modelling purpose we will only distinguish two types of diseases: infectious and noninfectious ones. An infectious disease can be passed between individuals, whereas noninfectious ones develop over an individual's lifespan.
In literature there are many more other sub-classes of infectious diseases (such as microparasite vs macroparasite), but for a our purpose - in order to keep our model more comprehensive - we are not interested in further classification based in biological details and we will just analyze the main dynamics and behaviours that are characteristics of an infectious disease.
We know from medical studies and diagnosis that SARS-CoV-2 belongs to the Coronaviruses family of viruses - known to cause diseases ranging from the common cold to more serious diseases such as the Middle East Respiratory Syndrome (MERS) and the Severe Acute Respiratory Syndrome (SARS) - and it is demonstrated that this virus can be spread from an infected person in several ways. Accordingly to the definition stated above, we can say that SARS-CoV-2 is an infectious disease and hence we can apply to it

all the mathematical modelling theories that were developed in this field over the past decades in order to better understand it and change our actions accordingly to minimize its damage in our lives.

The aim of this thesis was trying to understand better COVID-19 disease using a compartment based model, more specifically a SEIRD model, in order to make predictions and get a deeper comprehension on its dynamic and behaviour. For our analyses we used the data provided by the Italian Government, we defined our model and then performed parameter estimations using an optimization process to set the values of its coefficients that better describe the spreading. In this thesis we have reported all the choices that we made both for modelling and implementation parts.

The first chapter describes the fundamentals that underlie the activity of modelling: it presents the definition of modelling and some of the theoretical background that are needed for a better understanding such as ordinary differential equations. Moreover it reports some examples on how it is possible to model different aspects of the reality through the language of mathematics and some important model examples too.

The second chapter presents what is time-forcing, why it is necessary and what it wants to model. This chapter describes also in details the SEIRD model that we used for the analysis and all the motivations that lead to this choice.

In the third chapter are described the strategies that we used for the numerical analysis and the results that we obtained. It is presented in detail the variational approach that is used to split the integral period and moreover are discussed the outcomes given by the choices of different time-forced beta for the italian regions Lombardia and Emilia Romagna. We presented also some possible ideas for further studies.

Chapter 1

Modeling Infectious Disease

1.1 What are models and why modeling?

In general, mathematical models are considered as concrete attempts to describe particular phenomenas (that can be seen also as a system of agents) in quantitative terms, this means describing the behaviour and dynamics of that phenomena using formulas and equations (language of mathematics). The purpose of modeling some phenomena is in general to better understand it and hence, possibly, making predictions.
In epidemiology, models allow us to predict the evolution of the epidemic in the whole population knowing some individual-level informations such as epidemiological factors (e.g the average infectious period, the incubation rate etc): in other words, models allow us to generalize our knowledge from the individual to the population.
In the present situation, with the spread of the SARS-CoV-2, it's clear how models can be crucial in the management of the pandemic: for governments for example is essential knowing what is the best thing to do and when taking action in order to minimize the number of death people and the impact on the economy.
Unfortunately, as will be discussed in the further paragraphs, every model is "wrong" in the sense that it is not possible to build a fully accurate model for our phenomena, this is due to the fact that modelling is a constant trade-off process between accuracy and generalization: every model will make some simplifying assumptions.
Historically mathematical models were heavly used for the past pandemics (Spanish Flu: 1918-1920, AIDS pandemic and epidemic: 1981-present day, H1N1 Swine Flu pandemic: 2009-2010, West African Ebola epidemic: 2014-2016) and the last years have seen an increasing interest in this field as evidenced by the increasing number of papers published in these topics. Hopefully new more accurate models will be discovered that will help us for the present and future pandemics.

1.2 What is a good model?

As mentioned above every model is "wrong", then what is a good model?
A good model is, very simply, a model that is usefull, in the sense that describes - in more or less - general terms some phenomena and can be used for practical applications. In order to find the best model there are several features that can be used for making comparisons. In particular, each model is judged by its:

- **Accuracy**: the ability to reproduce the observed data and reliably predict future dynamics
- **Transparency**: the ability to understand how model components influence the dynamics and iteract
- **Flexibility**: measures the ease with which the model can be adapted to new situations

The real job of the mathematical modeller is the very creative process that consists on creating, adding or removing new components to the model in the pursue of the best trade-off between these features.

1.3 A modeling dichotomy

There are two main different approaches that are used in modelling infectious diseases:

- **Agent Based**: an agent-based model is the approach that is focused on the single interactions between autonomous agents (single individuals or groups) in an attempt to re-create and predict the appearance of complex phenomena, it is also called *microscale* model.
- **Compartment Based**: in this approach the population is divided in dynamic compartments - in the sense that people can shift from one compartment to another - that represent the *state* of people with reference to the evolution of the disease. Eg. the SIR model which is well described in the following paragraph.

Each approach has its own pros and cons: in general the agent based is more fined-grained but usually leads to large dimension systems with the consequently explosion of the number of parameters. On the other hand the compartment based is less grained and hence tends to make more modelling approximations but typically reduce the dimension of the system and hence the number of parameters. However the agent based approach is beyond the scope of this thesis, we will focus our attention on the compartment based approach.

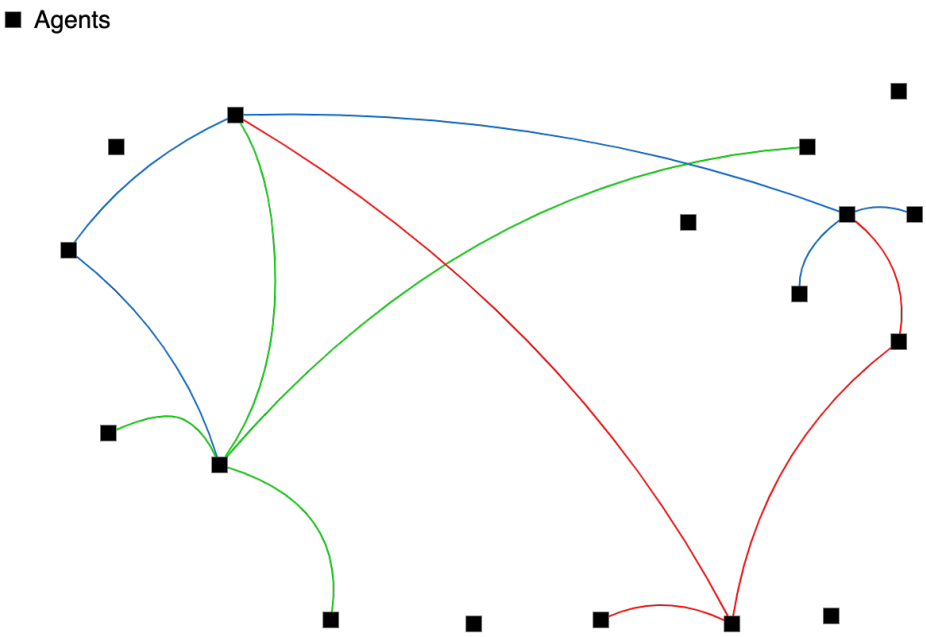

Figure 1.1: Agent Based Approach

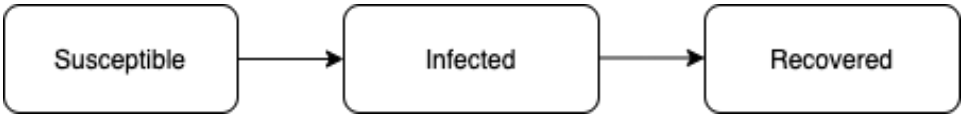

Figure 1.2: Compartment based model (SIR)

1.4 SIR Model

To get the idea of what is a mathematical model in epidemiology let's introduce the **SIR** compartment based model, which despite its simplicity it summaries very well the essence and main principles that are used in this field. The core idea is to divide a closed population (which means that new born, deaths or migrations are not taken into account) with N individuals in three categories:

- **Susceptible**: people that can be infected.
- **Infected**: people that can spread virus and infect others (susceptible).
- **Recovered**: people that have been infected but they recovered

where $S + I + R = N$
A virus has been spread in this population (and hence there has to be at least one infected person) and the aim of modeling is to find which are the relations that describe the dynamics between the three categories mentioned above: what happen when an infectious person enter in contact with a susceptible one? What can be done to slow down the

velocity of growing infectious people?
Since this is a continuous dynamic phenomena that change through time, the mathematical tools used to describe and model the movement of people from one compartment to another are differential equations and more in particular ordinary differential equations since the evolution is analyzed only with respect to one dimension: time
Since there are three different categories, the phenomena is described by the following system of ODEs:

$$\begin{aligned} \frac{\mathrm{d}S}{\mathrm{d}t} &= -\beta SI \\ \frac{\mathrm{d}I}{\mathrm{d}t} &= \beta SI - \gamma I \\ \frac{\mathrm{d}R}{\mathrm{d}t} &= \gamma I \end{aligned} \tag{1.1}$$

where:

- γ is called the removal or recovery rate, or the coefficient that tells how many days an infected person takes to heal and hence become a recovered.
- β is the infection rate: the coefficient that describe how strong is the virus load and in this sense bigger it is certainly faster is the spread.

The solution of this system are three continuous function that provide the numbers of people in each compartment during time.
The closed population assumption is now modelled in a more rigorous way. It's easy to prove that:

$$\frac{\mathrm{d}S}{\mathrm{d}t} + \frac{\mathrm{d}I}{\mathrm{d}t} + \frac{\mathrm{d}R}{\mathrm{d}t} = 0 \tag{1.2}$$

and hence the population is constant through time.
Just analyzing this simple model it is possible to understand some important feature of the disease and its behaviour. A key example is reported in the following paragraph.
Considering the system from a theoretical point of view, it can be classified as:

- Non linear
- Time invariant, since time is not explicitly present in the equations.
- 1st order, since no higher order derivatives are present
- normal form, that is the left hand side only contains the derivatives

1.5 Ordinary Differential Equations

A first order ODE [9] is an equation in which the unknown is a function itself $y(t)$ that appears in the equation with its first derivative $y'(t)$. In general these equations are

presented in this form:

$$\phi(t, y(t), y'(t)) = 0$$

where ϕ is a function defined in an open set $A \in \mathbb{R}$ x $\mathbb{R}$ x $\mathbb{R}$.
If Ω is an open set of $\mathbb{R}^2$, $A = \Omega$x$\mathbb{R}$ and $\phi(t, y(t), y'(t)) = y'(t) - f(t, y(t))$, the equation becomes:

$$y' = f(t, y(t)) \tag{1.3}$$

and it is called *normal formed* equation because the higher order of derivative appears on the left hand side.

As already mentioned a solution of this equation is a function itself $v \in C^1(I, \mathbb{R}^n)$ with $I \subset \mathbb{R}$ and (t, v(t)) $\in \Omega$ such that:

$$v'(t) = f(t, v(t)) \quad \forall t \in I$$

Intuitively f provides the behaviour of the derivative with respect to time of the solution $v(t)$ which means that $f(t, v(t))$ describe the slope of the straight line which is tangent in time t to the function $v(t)$.

It should be noticed that in general the equation (1) admits infinite different solutions since the derivate of a constant is 0 and hence all the functions that have f as derivative are infinite and depends from a scalar c $\in \mathbb{R}$. Therefore, to uniquely identify one function as the solution, another condition is required. More specifically it is required that the solution $v(t)$ in a certain t_0 takes a specific given value v_0 (*Initial Condition*). The union of the two conditions: having a derivative that behaviour like f and interpolate the point $v(t_0) = v_0$ provides the so called **Initial Value Problem** or **Cauchy Problem**.
A lot of interesting theory and theorems have been developed in this field, but unfortunately they are beyond the scope of this thesis.
The last thing that is important to know is the following theorem of existance and uniqueness:

Let Ω be an open set of $\mathbb{R}$x$\mathbb{R}^n$ and $f \in C(\Omega, \mathbb{R}^n)$. Let then $(t_0, v_0) \in \Omega$, if f is globally Lipschitz in Ω with respect to y, then the Cauchy Problem:

$$\begin{cases} v' = f(t, v(t)) \\ v(t_0) = v_0 \end{cases}$$

has a unique solution $v(t)$ in Ω.

By the way, in this thesis we do not face in detail any problem of existence nor uniqueness

of the ODEs' solutions, the only thing that is important to remember is that if f is sufficiently smooth (Lipschitz continuous) and an initial condition is provided, this is sufficient to state that a unique solution exists in a certain domain.

1.6 The Threshold Phenomenon

To better understand the disease and hence planning what to do to stop its spreading, it's crucial to know what are the states of the system in which the the virus doesn't spread anymore and there is no more movements of people from each compartment, in other words which states of the system we want to pursue to declare that the virus spreading comes to an end. Thanks to the equations, it is possible to well define these states which are called the "stationary points" of the system and are founded letting all the three ODEs equal to 0.

$$\begin{aligned} \frac{\mathrm{d}S}{\mathrm{d}t} &= -\beta SI = 0 \\ \frac{\mathrm{d}I}{\mathrm{d}t} &= \beta SI - \gamma I = 0 \\ \frac{\mathrm{d}R}{\mathrm{d}t} &= \gamma I = 0 \end{aligned} \tag{1.4}$$

It's easy to check that this is true if

$$\frac{\mathrm{d}I}{\mathrm{d}t} = I(\beta S - \gamma) = 0 \tag{1.5}$$

- for $I = 0$. This is quite obvious since this means that there aren't any infected people in the population and hence the virus can't spread.

- for $\beta S - \gamma = 0 \rightarrow S = \frac{\gamma}{\beta}$. This is very interesting because it states that if the number of Susceptible at the initial time are less than a certain threshold which is $\frac{\gamma}{\beta}$, the spread dies out. This is why it is called the threshold phenomena. This is because γ, as mentioned above, is the coefficient that quantify "how fast" an individual heal from the virus and hence become recovered; on the other hand β is the coefficient that quantify the power of spreading of the virus. If S is less than that ratio, the virus dies out because people heal faster than those who become infected.

Vaccination is a measure that impacts on the number of Susceptible people in the sense that, theoretically, vaccinated people cannot be infected and hence should be removed from the compartment S of the model.
For instance the threshold founded above provides the sufficient number of people that should be vaccinated in order to eradicate the disease.

1.7 R_0 Index

It is now easy to introduce the so called *basic reproductive ratio* coefficient R_0 which is nothing else than the inverse of the threshold founded in the previous paragraph: $(\frac{\gamma}{\beta})^{-1} = \frac{\beta}{\gamma}$.
In this sense it is possible to re-express the threshold phenomenon using this quantity.
Let's suppose that everyone at the beginning is susceptible $S = 1$.
Using the formula $S = \frac{\gamma}{\beta} \rightarrow S = \frac{1}{R_0} \rightarrow 1 = \frac{1}{R_0}$ it's ease to check that:

- if $R_0 > 1$, since $S = 1$, it is greater than the threshold and hence the virus spreads.
- if $R_0 \leq 1$ instead, since $S = 1$, it is less than the threshold and hence the virus dies out.

To get a full understanding of what R_0 wants to model, since it depends on the transmission rate and on the γ, it is necessary to explicitly define these two quantities.
For instance γ is the average recovery rate, which is the inverse of the average infected period T that is actually the average of days that people remained infected $T = \frac{1}{\gamma}$. This quantity can be observed from the real cases. In the case of SARS-CoV-2 it has been observed no longer than 20 days [1].
β instead is a more complex term and is defined as follows ([2]):
Let's Y be the number of infected people, N the total population size.
Considering a susceptible with an average k contacts per unit of time, of these a fraction $I = \frac{Y}{N}$ are contacts with infected people. Thus, during a small time interval (from t to $t + \delta t$) the number of contacts with infected people is given by the term $k\frac{Y}{N}\delta t$. Let's define the probability of being infected from each contact with an infected person as c and consequently the probability of not being infected as $1-c$. Assuming that the contacts are independent one-another, the probability that a susceptible individual escapes infection following $k\frac{Y}{N}\delta t$ contacts is:

$$1 - \delta q = (1 - c)^{(k\frac{Y}{N})\delta t}$$

It is here that Keeling and Rohani [2] introduced for the first time the term β as $\beta = -k\log(1 - c)$. From the definition, since c is a probability and hence $0 \leq c \leq 1$, it easy to derive that $0 \leq 1 - c \leq 1$ and follows $\log(1 - c) \leq 0$. So the minus that is present in front makes the $\log(1 - c)$ positive and hence:

- β is directly proportional to k.
- β is directly proportional to c, which we can see as a kind of viral load of the disease.

After all these assumptions it immediatly follows that:

$$R_0 = \beta T \propto kcT \tag{1.6}$$

Now it's interesting to see why limiting the number of each individual's contacts per unit of time is crucial for the management of the pandemic. Since T and c are parameters that depend exclusively on the biological structure of the virus, the only coefficient that can be artificially increased or decreased by restrictive measures is k. The usage of masks and other protection measures can reduce maybe the viral load of the disease and hence the c for a single contact, but since c depends on the biological structure of the disease and masks do not mutate the virus, it is more correctly to see that as $c - \epsilon$ where ϵ is the power of protection measures that are take into account foreach contact.

1.8 Introducing Demography

How the model changes if for instance the phenomena is observed for a more long-term period?
If a more long-term phenomena is observed, considering a closed population can be a too restrictive assumption because demographic processes play their role in the dynamics of the system and hence they should be modelled. The simplest and most common way of introducing demography into the SIR model, according to Matt J Keeling and Pejiman Rohani [2], is to consider in the model two additional rates: the birth (or immigration rate) ν as well as the death (or emigration) rate μ. Hence, we get a new system of equations:

$$\begin{aligned} \frac{\mathrm{d}S}{\mathrm{d}t} &= N\nu - \beta SI - \mu S \\ \frac{\mathrm{d}I}{\mathrm{d}t} &= \beta SI - \gamma I - \mu I \\ \frac{\mathrm{d}R}{\mathrm{d}t} &= \gamma I - \mu R \end{aligned} \tag{1.7}$$

where the quantities μS, μI, μR reflects the people that are removed from the system because of death. It is easy to check that in this case $\frac{\mathrm{d}S}{\mathrm{d}t} + \frac{\mathrm{d}I}{\mathrm{d}t} + \frac{\mathrm{d}R}{\mathrm{d}t} \neq 0$, in fact:

$$\begin{aligned} \frac{\mathrm{d}S}{\mathrm{d}t} + \frac{\mathrm{d}I}{\mathrm{d}t} + \frac{\mathrm{d}I}{\mathrm{d}t} &= N\nu - \beta SI - \mu S + \beta SI - \gamma I - \mu I + \gamma I - \mu R \\ &= N\nu - \mu S - \mu I - \mu R \\ &= N\nu - (S + I + R)\mu \end{aligned} \tag{1.8}$$

this becomes zero iif $\mu = \nu$ in each time step.

1.9 Other Models

Now that is clear how the compartment based models work and which are their main features, it is possible to custom each model adding or removing compartments in order to better describe the phenomena and consider all the assumptions that we want to make. If, for instance, a disease causes death it is crucial to know the rate and the speed in which infected people die: in this case it is reasonable to add a new compartment of dead people and study the flux of individuals from the infected compartment to the dead compartment. This model is so called SIRD.

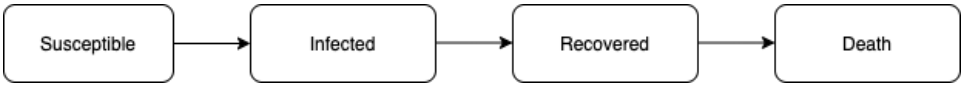

Figure 1.3: SIRD model

Another example that could be made is considering a virus where individuals do not recover, such as meningitis, plague and veneral diseases. In these particular diseases the only two stages that are taken into account is when an individual can infect and when cannot.

It should be noticed that after the infectious period the individual become again

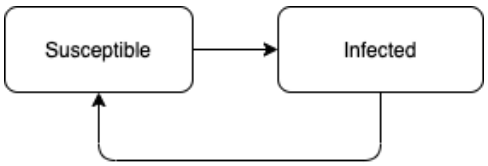

Figure 1.4: SIS model

susceptible and this means that individual could be infected again. This model is called SIS.

As last example it is reported one of the possible compartment models that can be used to model the HIV/AIDS. Since the HIV/AIDS has been studied for so long and hence many clinical data are available, the model can be more complex and can take into account several different factors and parameters and therefore better describe the phenomena. The model reported here is so called SICA [3] , in fact it has 4 different compartments where:

- **S** are susceptible individuals;
- **I** are HIV-infected individuals with no clinical symptoms of AIDS (the virus is living or developing in the individuals but without producing symptoms or only mild ones) but able to transmit HIV to other individuals

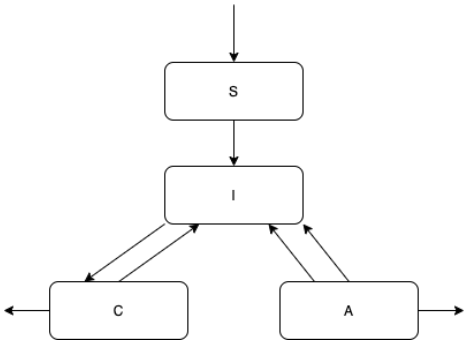

Figure 1.5: SICA model

- C are HIV-infected individuals under ART treatment (the so called chronic stage) with a viral load remaining low;
- **A** are HIV-infected individuals with AIDSclinical symptoms.

Chapter 2

Time forced and SEIRD model

The previous chapter introduced basic ideas and concepts founding the art of modelling infection diseases. However many simplification assumptions were made in order to make the model easier to understand and more comprehensible. For instance one important feature that should be modelled is the potential mutation of the virus from one season to another, hence the model parameters which describe its behaviour can change over time. This is the case of influenza: the viral load is not constant over the weeks as you can see in 2.1, in some months the positive cases grow exponentially and in others the disease totally disappear. This cyclic behaviour is repeated, with some small variations, for each year - as plotted in 2.2.

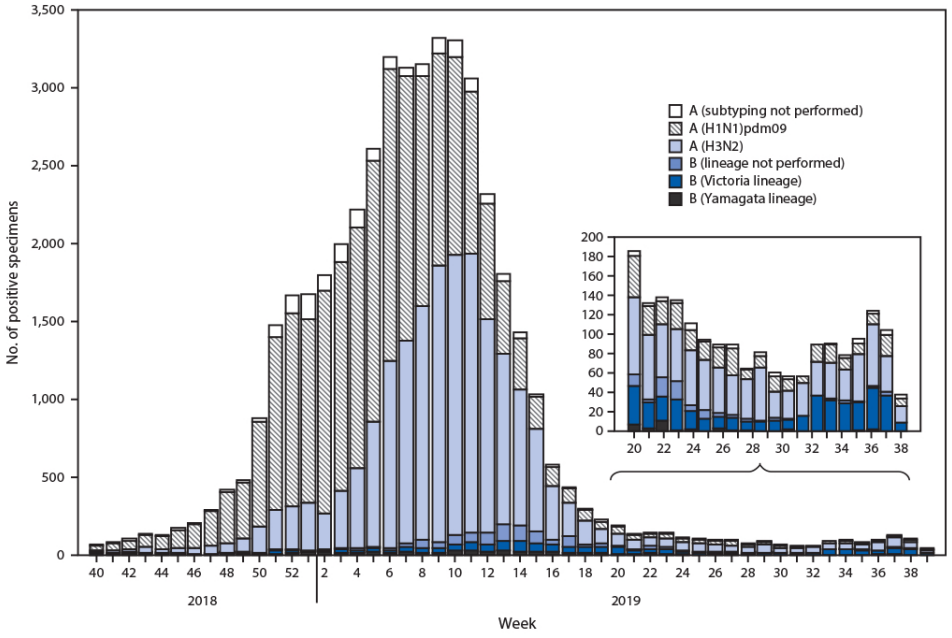

Figure 2.1: Influenza cases in US per week

Source: Centers for Disease Control and Prevention [4]

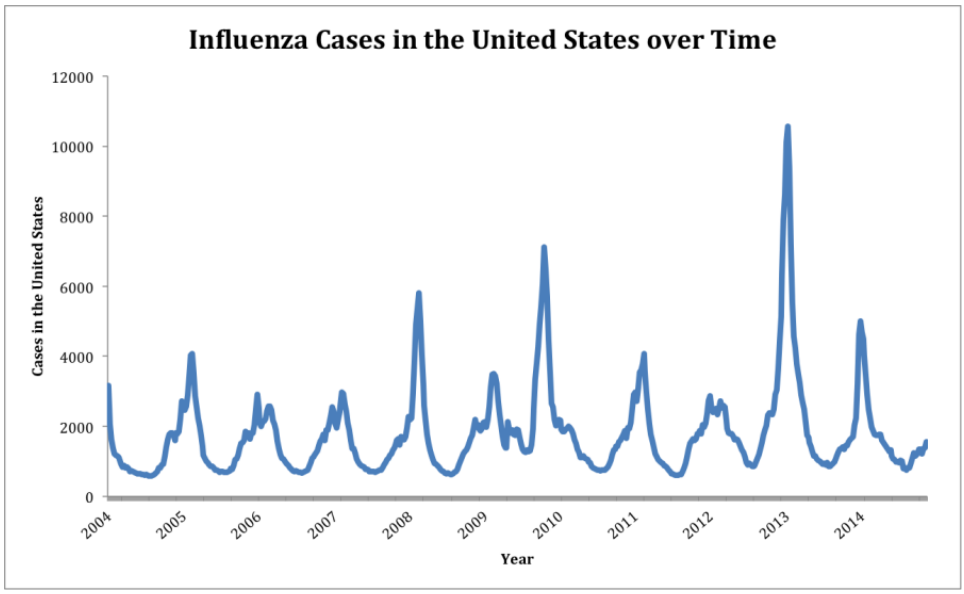

Figure 2.2: Influenza cases in US per year [5]

Source: Google Flu trends

Another important example of this kind are the diseases transmitted by childhood people: statistical studies have revealed a huge correlation with the scholastic period, the peak of infections coincides exactly with the beginning of schools and in the same way the end of contagion coincides with the ending of the school period and this is repeated every year.

2.1 Time forced models

For this reason it becomes necessary to introduce new techniques that allow modellers to include and describe seasonal variations of transmission rates and hence time forced models were introduced. The basic idea consists on considering the β transmission rate as time dependent and hence modelling it as a function that changes through time. A consequence of that is obviously the rise of another problem: how to choose the β function. This choice basically depends on the features that are needed: the frequency of peaks, the interval of periodicity, derivatives properties and so on. A key study was made in this field by Bailey (1975) who incorporated seasonality in the SIR model explained above (actually he used a simplified version) with the primary aim of establishing the amplitude of variation in contact rates necessary to produce the observed 80% fluctuation in epidemics. He has chosen the transmission rate function as:

$$\beta(t) = \beta_0(1 + \beta_1 \cos(\omega t)) \tag{2.1}$$

where

- β_0 denotes the average transmission rate,
- ω is the period of the forcing
- β_1 is the amplitude of seasonality which is restricted to the unit interval

Many different functions have been used in later studies but the detailed explanation of these methodologies and their implications is beyond the scope of this thesis. However, it's interesting to notice that time forced models become very popular in the recent years because the rising of computational power makes possible all the complex added calculation that were brought by using a time variant β transmission rate.

2.2 SEIRD time forced model

The model that has been used for the numerical analysis described in this thesis is deeply based on the work done by Prof. Piccolomini and Prof. Zama in their publication "Monitoring Italian COVID-19 spread by a forced SEIRD model" [8]. They used a compartmental SEIRD model, which means that includes five different compartments:

- Susceptible: people that can be infected
- Exposed: people that exhibit no obvious signs of infection and the abundance of pathogen may be too low to allow further transmission.
- Infected: people that are positive and can infect others.
- Recovered
- Dead: people dead due to SARS-CoV-2.

Clear idea of what is the difference between Infected and Exposed individuals is given by the following plot 2.3
The main reasons that lead to the choice of this model can be summarized as the following ones:

- Covid cause death and hence the Dead compartment was required
- The choice of the Exposed compartment is motivated by the fact that this group provides a better control on the infection transmission rate: adding this additional class as middleware between Susceptible and Infected provides a better idea on the flux of people between classes, hence a much more clear understanding on the behaviour of the disease.

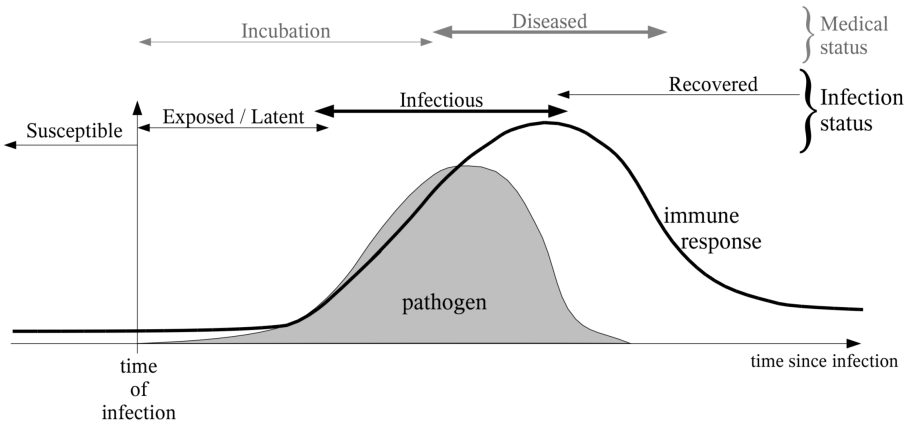

Figure 2.3: Time Line infection

Source: Matt J. Keeling and Pejman Rohani, "Modeling Infectious Diseases"

- the choice of the time-forced model is motivated by the fact that COVID-19 and influenza viruses have a similar disease presentation. That is, they both cause respiratory disease, which presents as a wide range of illness from asymptomatic or mild through to severe disease and death. Secondly, both viruses are transmitted by contact, droplets and fomites [10]. Since it is well known that influenza is a seasonality disease, it was reasonable to assume that COVID-19 has a pretty similar behaviour.

As written in the paper mentioned above, the system of equations in the SEIRD model is the following:

$$\begin{aligned} \frac{\mathrm{d}S}{\mathrm{d}t} &= -\frac{\beta}{N}SI \\ \frac{\mathrm{d}E}{\mathrm{d}t} &= \frac{\beta}{N}SI - \alpha E \\ \frac{\mathrm{d}I}{\mathrm{d}t} &= \alpha E - \gamma I \\ \frac{\mathrm{d}R}{\mathrm{d}t} &= (1-\epsilon)\gamma I \\ \frac{\mathrm{d}D}{\mathrm{d}t} &= \epsilon\gamma I \end{aligned} \tag{2.2}$$

where:

- N is the total number of population
- β is the infection rate
- α is the incubation rate
- γ is the recovery rate, as explained for the SIR model, which is also equal to $\frac{1}{T}$ where T is the average infection period

- ϵ is the fraction of all removed individual who die

It can be noticed that, for simplicity, it has been assumed that the population is closed since it holds:

$$\frac{\mathrm{d}S}{\mathrm{d}t} + \frac{\mathrm{d}E}{\mathrm{d}t} + \frac{\mathrm{d}I}{\mathrm{d}t} + \frac{\mathrm{d}R}{\mathrm{d}t} + \frac{\mathrm{d}D}{\mathrm{d}t} = 0 \tag{2.3}$$

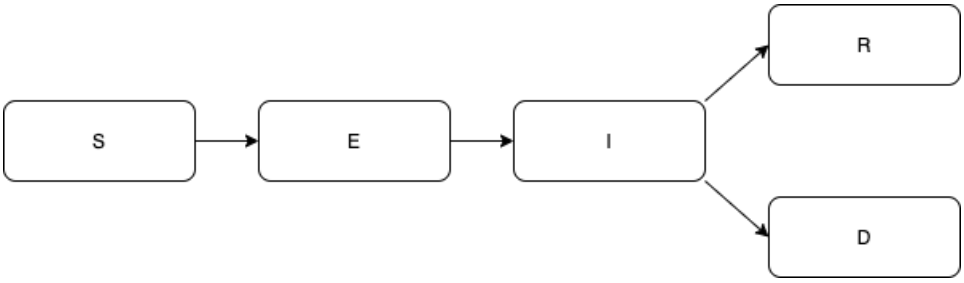

Figure 2.4: SEIRD model

Chapter 3

Numerical Results

In this chapter will be discussed all the attempts that were made in order to estimate as accurate as possible the parameters of the SEIRD model mentioned above and hence try to find the best values of parameters that describe the COVID-19 spreading. The technique used for finding the best values of parameters is through an iterative optimization process that changes the parameters at each iteration in order to minimize the loss function, which is essentially the MSE obtained from the real values and the estimated ones.

3.1 Epidemic Data

The data used for this analysis were taken from the dataset of the Italian Civil Protection Department, which is daily updated. An automated script is developed in python that every time the algorithm runs, it takes automatically the latest data from the Civil Protection online repository. All the implemented code is open source and can be found in the github repo [6]. Below you can find the code of the Datamanagement.py which retrieves the data from the dataset of the Italian Civil Protection and manipulates it in order to create the dataframe *data* which is ready to be processed.

```
EMILIA_ROMAGNA = 8
LOMBARDIA = 3
csv_url = "https://raw.githubusercontent.com/pcm-dpc/COVID-19/master/dati-regioni/dpc-covid19-ita-regioni.csv"

req = requests.get(csv_url)
url_content = req.content
csv_file = open('./region.csv', 'wb')
csv_file.write(url_content)

"read_saved_file"
df = pandas.read_csv('./region.csv', index_col=None);

"Choose the columns"
columns = ['data', 'totale_positivi', 'dimessi_guariti', 'deceduti', 'tamponi', 'totale_casi']
df_regione = pandas.DataFrame(columns=columns);
for index, row in df.iterrows():

    "Select the region to analyze"
    if (row['codice_regione'] == EMILIA_ROMAGNA):
        df_regione = df_regione.append(row[columns]);
```

```
data = df_regione.loc[:, ['totale_positivi', 'dimessi_guariti', 'deceduti']].astype('float').values
```

3.2 Cauchy Problem initialization

To solve the system 2.2, as explained in the Chapter One, are necessaries five different initial conditions in order to create one different Cauchy Problem foreach compartment and therefore having one solution foreach. Starting from an initial time $t = t_0$, which can be set manually from the constants.py file, the values of the populations $S(t_0), E(t_0), I(t_0), R(t_0), D(t_0)$ are assigned accordingly on the basis of the available data. In particular:

- $I(t_0)$ is get from the infected column of the csv at time t_0.
- $R(t_0)$ same as Infected.
- $D(t_0)$ same as Infected.
- $E(t_0)$: the calculation of E_0 is very trivial since no data is available to retrive the initial condition of the number of Exposed people at time t_0. Several strategies were attempted and are discussed in detailed in further paragraphs. The first attempt was setting $E(t_0)$ be equal to $I(t_0)$ multiplied by a constant which initially was set equal to 10 (an empirical approach that works pretty well).
- $S(t_0)$ is calculated from the total number of population of that region which is known. $S(t_0) = N - S(t_0) - E(t_0) - I(t_0) - R(t_0) - D(t_0)$.

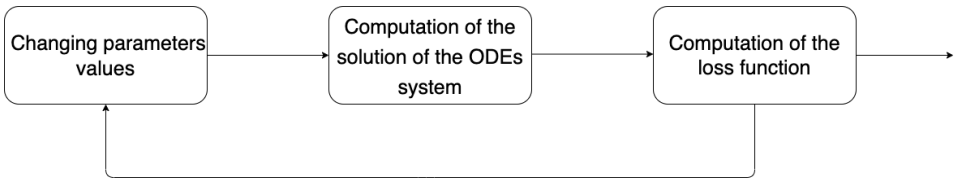

Figure 3.1: Iteration scheme of the optimization process

3.3 Interval Variations

To improve the model flexibility of the parameter estimation the integral period is splitted in many sub-intervals. In each subinterval a new estimation of parameters in performed and the results are stored in arrays of values.

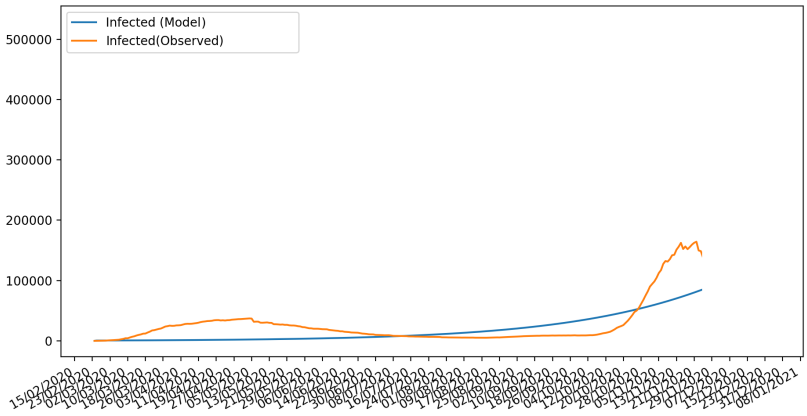

Figure 3.2: Large Error estimation without splitting the integration period

Splitting the integral period is convenient and necessary since the 2.2 model with the same constant coefficient applied to the entire integral period does not fit the available data as you can see from 3.2

After several attempts in which it was taken a fixed time interval, a variational approach has been later preferred.

In order to create a fair policy to choose the interval size we proceeded as follows:

Five different coefficient were set: three different standard interval lenghts MIN INTERVAL, MEDIUM INTERVAL and MAX INTERVAL and two critical thresholds ERROR RANGE MIN and ERROR RANGE MAX. The choice of which interval lenght use for the next interval estimation is based on the average infected error (AVI) which is computed as follows:

Let O = Sum of Infected Observed during the n^{th}interval

Let E = Sum of Infected Estimated during the n^{th} interval

Let S = n^{th+1} interval size

$$\text{AVI} = \frac{\text{O} - \text{E}}{n^{th} \text{ interval size}}$$

- if IVP $\leq$ ERROR RANGE MIN then S = MAX INTERVAL
- if IVP $\geq$ ERROR RANGE MAX then S = MIN INTERVAL
- otherwise S = MEDIUM INTERVAL

AVI performs the role of a precision coefficient for the current estimation. If AVI is high it means the estimation is bad, maybe because the infected people observed suddenly increased, and hence a shorter interval is needed to catch these variations.

The choice of this policy is totally arbitrary but this seemed to work pretty well.

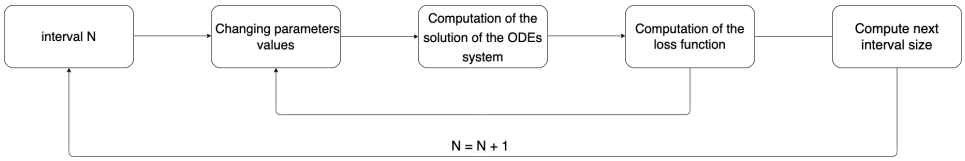

Figure 3.3: Updated Iteration Scheme with Variational Interval approach

It should be noticed that using the variational approach all the parameters are time-forced with a piecewise function that is defined with the values of the parameters estimated in each interval:
Let k be the k^{th} interval, and t $\in (t_k, t_{k+1}]$it holds:

$$\begin{aligned}\alpha(t) &= \alpha_k \\ \epsilon(t) &= \epsilon_k \\ \beta(t) &= \beta_k\end{aligned}$$

For the simplicity, γ is assumed to be fixed since it has been seen from medical observation that the average infection period is 20 days and hence γ is taken $\frac{1}{20}$.
As already mentioned in the previous chapter, for the SEIRD model used in this thesis a time-forced approach is chosen for the transmission rate. Two different functions were proposed:

$$\beta(t) = \beta(t_k)(1 - \rho_k \frac{t - t_k}{t}) \quad t \in (t_k, t_{k+1}] \text{ and } \rho_k \in (0, 1) \tag{3.1}$$

$$\beta(t) = \beta(t_k)(e^{-\rho_k(t-t_k)}) \quad t \in (t_k, t_{k+1}] \text{ and } \rho_k \geq 0 \tag{3.2}$$

And hence β depends to a value ρ which is estimated as a piece-wise constant (in each interval) function.

Hence the system2.2 becomes:

$$\begin{aligned}\frac{\mathrm{d}S}{\mathrm{d}t} &= -\frac{\beta(t)}{N}SI \\ \frac{\mathrm{d}E}{\mathrm{d}t} &= \frac{\beta(t)}{N}SI - \alpha(t)E \\ \frac{\mathrm{d}I}{\mathrm{d}t} &= \alpha(t)E - \gamma I \\ \frac{\mathrm{d}R}{\mathrm{d}t} &= (1 - \epsilon(t))\gamma I \\ \frac{\mathrm{d}D}{\mathrm{d}t} &= \epsilon(t)\gamma I\end{aligned} \tag{3.3}$$

3.4 Parameter Estimation and Prevision

In order to estimate the parameters α_k, ϵ_k and ρ_k we fit the solution of 3.3 to the measured data of the infected, recovered and dead population (I_o, R_o, D_o). Then we estimate the parameters of the models with respect to the both β functions by solving a non linear least squared problem.
Mathematically the problem can be formulated as follows: Let $\mathbf{u}$(t) be the multi-value function solution of the system 3.3:

$$u(t) = (S(t), E(t), I(t), R(t), D(t)) \quad t \in [t_0, \text{maxT}] \tag{3.4}$$

and let:

$$\alpha = (\alpha_1, \alpha_2, ...\alpha_p)$$
$$\epsilon = (\epsilon_1, \epsilon_2, ...\epsilon_p)$$
$$\rho = (\rho_1, \rho_2, ...\rho_p)$$

the vectors containing the values of parameters for all the intervals p. We define then a restriction $\mathbf{v}$ of $\mathbf{u}(t,\alpha,\epsilon,\rho,\gamma)$ to the three measured populations $(I(t),R(t), D(t))$. And hence:

$$\text{v(t, } \alpha, \epsilon, \rho, \gamma) = \text{(I(t), R(t), D(t)) with t} \in [t_0, \text{ maxT}]$$

For each day t(i) $\in \mathbf{t}= (t_0, t_1, t_2, ..., t_{maxT})$ (time discretization) we compute the restricted function $v(t(i),\alpha,\epsilon, \rho, \gamma)$. In the same way we collected all the observed data on daily bases from the repository with the same density and number of records of the solution calculated from the system above. Let $\mathbf{Y} = (I_o(t(i)), R_o(t(i)), D_o(t(i)))$ for each day t(i) $\in \mathbf{t} = (t_0, t_1,t_2, ...t_{maxT})$. The parameters $\boldsymbol{\alpha}, \boldsymbol{\epsilon}, \boldsymbol{\rho}$ are estimated by solving the following non linear least squared problem:

$$\min_{\boldsymbol{\alpha},\boldsymbol{\epsilon},\boldsymbol{\rho}} = \|v(\mathbf{t}, \boldsymbol{\alpha}, \boldsymbol{\epsilon}, \boldsymbol{\rho}, \gamma) - \mathbf{Y}\|_2^2 \tag{3.5}$$

3.5 Computational Strategies

Several critical decisions were taken during the coding phase. I am reporting here some of the most important.
To compute the solution of the system, we used a on ODE solver (*odeint*) that provides the solution of the Cauchy Problem using a variable step Runge Kutta method.
The calculation of the mean squared instead is done using a *minimize* function from the *lmfit* [7]library which allowed us to add constraints to the parameters. The minimize

function performs a gradient-based optimization for the Least Squared Error function. After few initial attempts we noticed that the time-forced model was very sensible to the initial values given to the parameters and for this reason we decided to split the estimation in two parts:

- the first phase which performs an estimation with a longer integration period considering constant parameters.
- the second phase performs the iterative interval estimation explained above and takes as initial parameters the outcomes of the first phase.

A more comprehensive idea is provided by the schema 3.4. During the second phase estimation, since the model required initial values for each compartment, we decided to apply a "teaching force" technique, forcing the value of Exposed at each iteration to take the value of Infected (at the first time of the interval) multiplied by a constant value equal to 10.
The forecast compute the solution of the system with the parameters equal to the output of the second phase estimation.

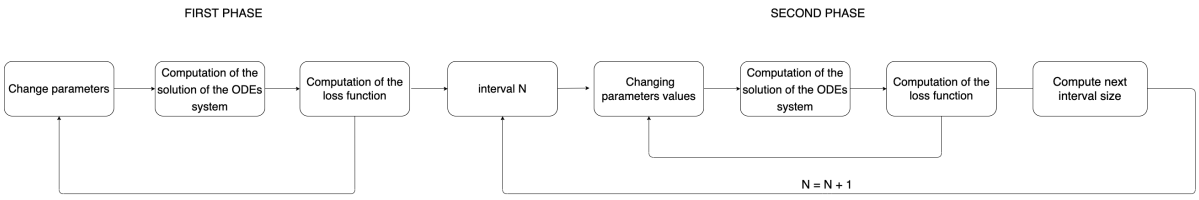

Figure 3.4: Definitive Schema of the optimization process used

3.6 Results

Below you can find the main results that we have obtain during the analysis and prevision of COVID-19 spread for two italian regions: Emilia Romagna and Lombardia.
Settings used:
MIN INTERVAL = 7 days
MEDIUM INTERVAL = 10 days
MAX INTERVAL = 15 days
Days of prevision: 30 days
Days dedicated to the first phase: 30 days

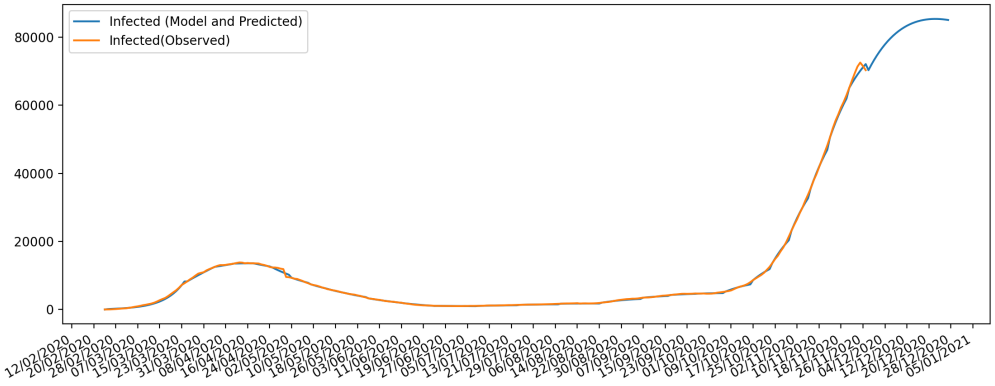

Figure 3.5: Estimation and Prevision of Infected people in Emilia Romagna using time-forced beta 3.2

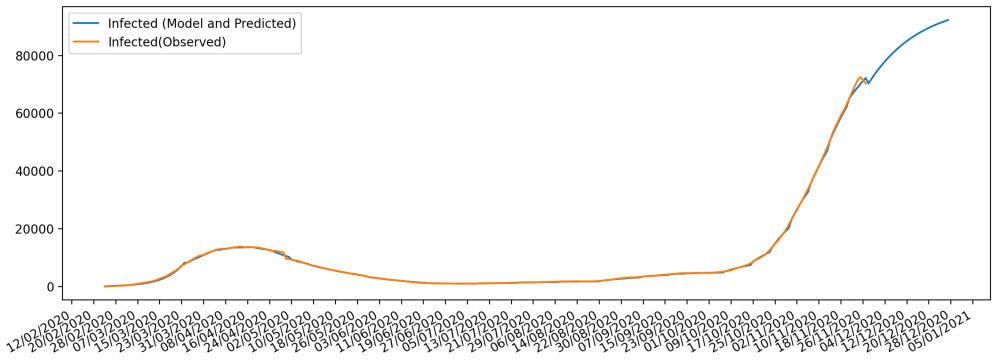

Figure 3.6: Estimation and Prevision of Infected people in Emilia Romagna using time-forced beta 3.1

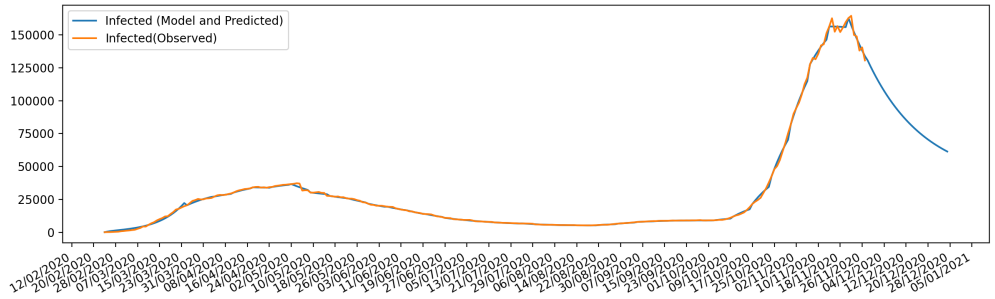

Figure 3.7: Estimation and Prevision of Infected people in Lombardia using time-forced beta 3.2

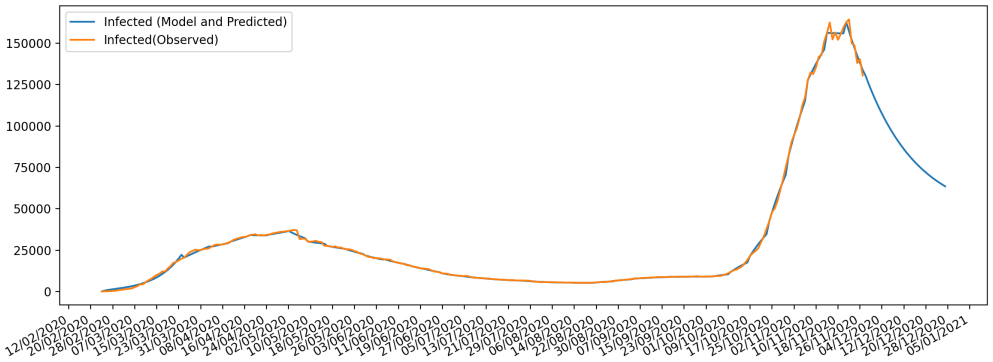

Figure 3.8: Estimation and Prevision of Infected people in Lombardia using time-forced beta 3.1

We can notice that in general the model fits the observed data both using the 3.1 or 3.2 time-force beta functions. However it is possible to see that, in the case of Emilia Romagna 3.5, the exponential beta time-forced 3.2 leads to a peak more rapidly than the beta rational one. On the other hand instead, for the Lombardia region, the curve tends to decrease more rapidly for the 3.1 beta function. So in general, time-forced beta 3.1 tends to describe the disease's dynamic more linearly than the exponential one.

Since in the previous results we have considered a quite long period (almost 10 months) and for definition our SEIRD model assumes a closed population, we then tried to see how the forecast change if a smaller total period is considered. More specifically, in the following charts we have set as day one of estimation the 22nd of June and hence limiting the total integral period to approximately 5 months.
For exponential beta (3.11 and 3.9) the results of the forecast are very closed to the results reported above 3.7 and 3.5. Concerning the rational beta 3.1 instead the results depends on the regions: in the case of Lombardia the outcomes seem similar to 3.8, on the other hand for Emilia Romagna 3.10, results are very different. The model did't manage to catch the trend inversion of the last days.

To measure the accuracy of the model, below you can find a table of the forecast made from the 24th October to the 6th of November for Emilia Romagna.

- Forecast using beta exponential (3.2): infected that are forecasted by the model using the beta exponential (3.2)

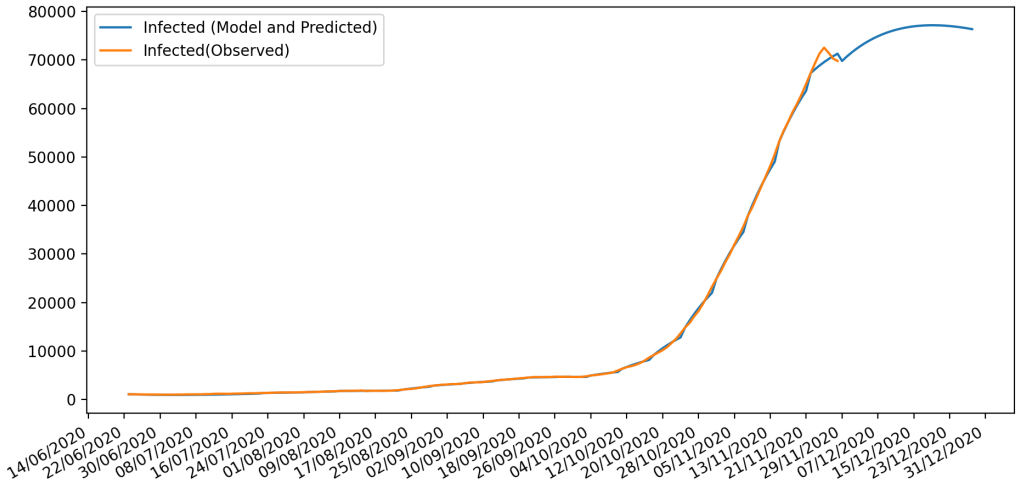

Figure 3.9: Estimation and Prevision of Infected people in Emilia Romagna using time-forced beta 3.2 starting from 22 June

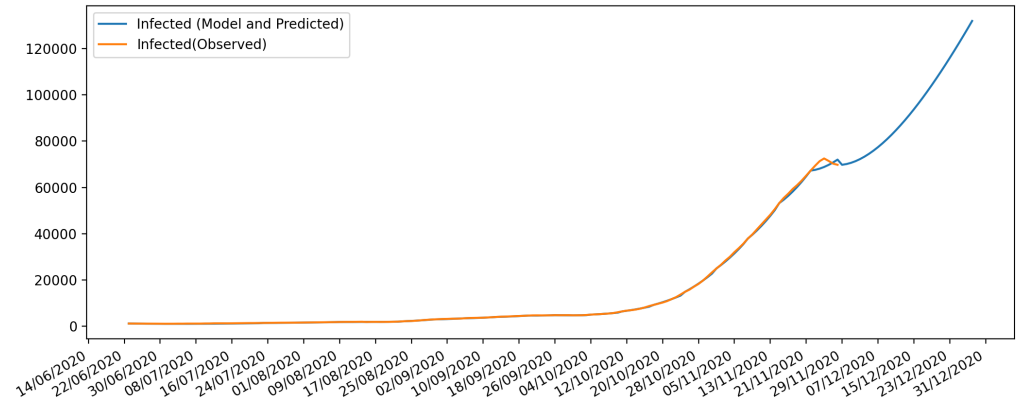

Figure 3.10: Estimation and Prevision of Infected people in Emilia Romagna using time-forced beta 3.1 starting from 22 June

- Forecast using beta rational (3.1): infected that are forecasted by the model using the beta rational(3.1)

- Observed: infected provided by the Italian Government

We can clearly see from the table 3.1 that the model generally tends to understimate the infected population. This is very probably due to the particular period taken into account

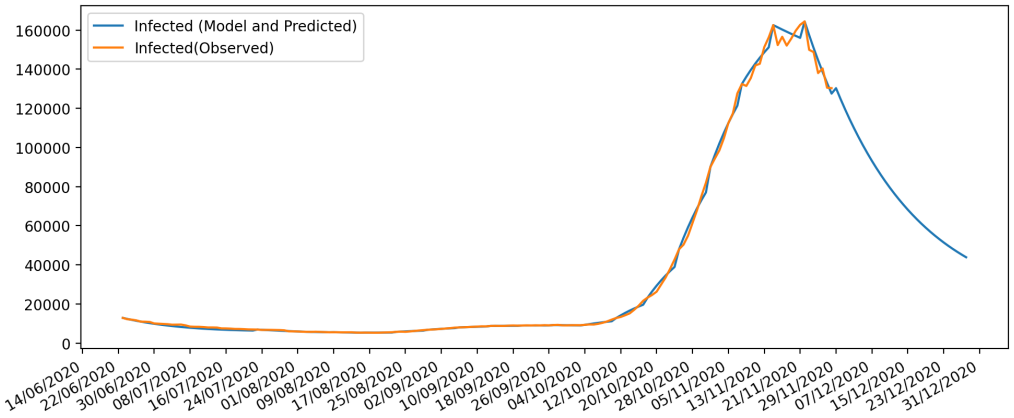

Figure 3.11: Estimation and Prevision of Infected people in Lombardia using time-forced beta 3.2 starting from 22 June

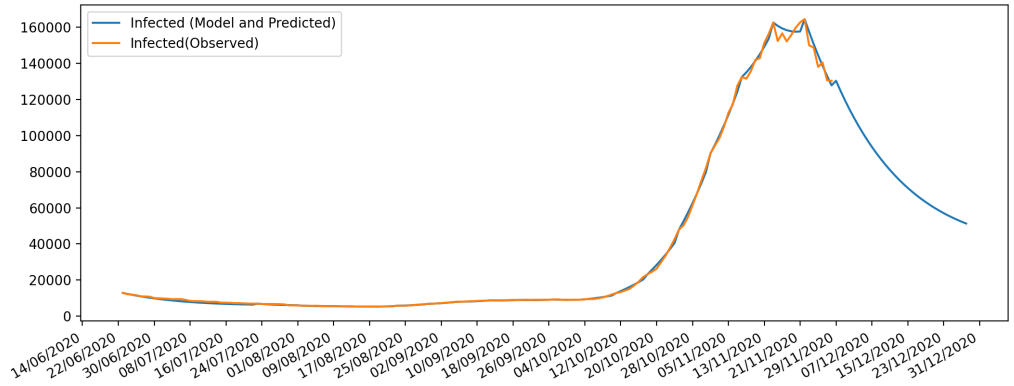

Figure 3.12: Estimation and Prevision of Infected people in Lombardia using time-forced beta 3.1 starting from 22 June

for the forecast: the last two weeks of October the infected population rapidly increased and the model didn't manage to catch these changes. Below you can find another forecast we made in the (following) period from the 13th November to the 25th.

It is possible to see that this time (Table 3.2) the forecast, even at the 10th day, is very closed to the real value observed. One possible explaination could be given considering the rate in which the infected are increasing that is very closed to the one of the previous time interval (see Table 3.1): that means that the model already estimated its parameters in order to catch that particular rate of change and hence the forecast improved accordingly.

Table 3.1: Forecast from 23rd to 6th of November

date	observed	forecast using beta exponential (3.2)	forecast using beta rational (3.1)
2020-10-24	13642	13439	13414
2020-10-25	14828	14301	14263
2020-10-26	15769	15101	15064
2020-10-27	17080	15843	15821
2020-10-28	18230	16531	16536
2020-10-29	19713	17167	17212
2020-10-30	21421	17753	17852
2020-10-31	23213	18293	18457
2020-11-01	24917	18789	19031
2020-11-02	26492	19243	19575
2020-11-03	28348	19658	20091
2020-11-04	29974	20036	20581
2020-11-05	31976	20378	21046
2020-11-06	33730	20687	21489

Table 3.2: Forecast from 14th to 25th of November

date	observed	forecast using beta rational (3.1)	forecast using beta exponential (3.2)
2020-11-14	50562	51087	51171
2020-11-15	53201	53852	53977
2020-11-16	55429	56461	56587
2020-11-17	57268	58924	59009
2020-11-18	59319	61250	61254
2020-11-19	61009	63447	63331
2020-11-20	62934	65524	65249
2020-11-21	65080	67488	67016
2020-11-22	67274	69345	68640
2020-11-23	69380	71102	70129
2020-11-24	71344	72766	71491
2020-11-25	72526	74342	72731

3.7 An alternative approach for Exposed Population

After these first results, we tried an alternative approach for calculating the Exposed Population in each interval of estimation. Which is the following:

$$\text{Exposed}_{t(k)} = \text{Infected}_o(t(k))e^{\rho_k} e^{\frac{\text{total positive swabs}}{\text{total swabs done}}}$$

Let:

- A = e^{ρ_k}
- B = $e^{\frac{\text{total positive swabs}}{\text{total swabs done}}}$

Concerning the second coefficient (B) the idea was to find a coefficient that basically provides us the accuracy of the observed infected people in the total population (how accurate are the measurements made by the Italian Government) and use that coefficient to find the exposed population in the interval k. Therefore we decided to use the rate between total positive swabs and total swabs done.
In particular we follow these intuitions:
Let P = total positive swabs
Let D = total swabs done

- if $\frac{P}{D} \to 0$ it means that the number of total swabs is much more greater than the positive swabs and hence the virus is under control in the sense that the number of infected people observed very probably reflects the reality.
- $1 > \frac{P}{D} >> 0$ means that the number of infected people observed does not reflect the reality. E.g if the rate of positive swabs and total swabs is near 20% means that the infected people observed are heavly understimated.

For these reasons we introduced a new penalty coefficient (B) for the calculation of the Exposed People. This particular coefficient has the properties that if the exponent tends to 0, the whole exponential approaches to 1. Otherwise there is a penalty multiplication factor that is bounded by e.
For the first factor (A) instead, the intent was trying to find a way to model the restrictions made by the government.
The intuition in this case is the strong correlation between restrictions and the contacts that each infected person can have during the day. Since β, as mentioned in the Chapter One is directly proportional to the number of contacts that each person has in δt and since the time-forced β directly depends on the coefficient ρ for definition, the idea is that in a certain way ρ could model the degree of freedom of circulation (and therefore

the contacts per day foreach infected person). Considering the $\beta(t)$ time-forced function 3.1, since $0 \leq \rho \leq 1$:

- $\rho = 0$ means total lockdown
- $\rho = 1$ means total absense of restrictions

The charts 3.13 and 3.14 simulate these situations.
Therefore if there is a total lockdown and many swabs are done, the number of Exposed at time $t(k)$ coincides exactly with the infected at time $t(k)$. This will cause a drastically decrease of new infected cases. Otherwise a penalty multiplication factor is introduced which is upper bounded by e^2.

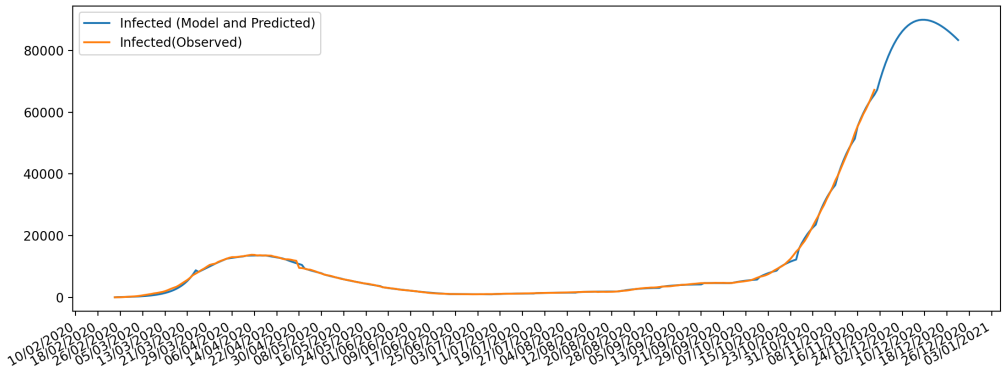

Figure 3.13: Emilia Romagna prediction with ρ of the penalty factor set to 1

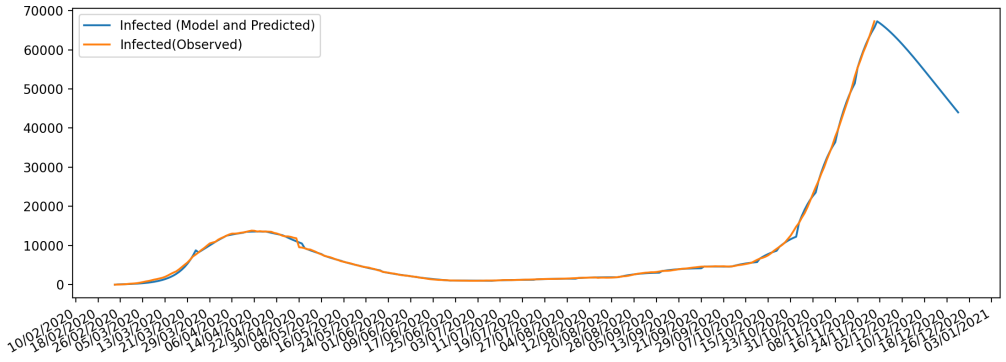

Figure 3.14: Emilia Romagna prediction with ρ of the penalty factor set to 0

Chapter 4

Conclusions

In conclusion it could be said that this SEIRD model, as we can see from the results reported above, describes pretty well the COVID-19 spread and hence it is a good model since it usefull - in the sense that it could be taken into account for some predictions and other analysis. However it reveals some weaknesses in transparency. In fact it is not very clear which time-forced beta function should be used and why. Moreover the calculation of exposed people used for the forecasts mentioned above is too empiric and again it is not very clear why it performs so well.
For future works it could be very interesting to explore how to combine in a better way the ideas mentioned in the last paragraph and try to model how restrictions made by government impact the curve and therefore try to get the ideal timing to perform restrictions and in which modalities.
A possible idea, that I would pursue, would be to add an extra Asymptomatic Compartment and change the β function as dependent to k (contacts per person in a unit of time) and c (viral load of the disease), as written in the R_0 Section of Chapter One, and try to estimate these two additional parameters. If k is known then it is possible to model the government restrictions and hence try to make previsions not just on infected people but on the actions that are necessary to make for flattening the curve below the limit of the intensive care units available.

Bibliography

[1] https://www.cdc.gov/coronavirus/2019-ncov/hcp/duration-isolation.html

[2] *Modeling Infectious Diseases In human and animals*, Matt J Keeling and Pejiman Rohani, *Princeton Press*

[3] *A SICA compartmental model in epidemiology with application to HIV/AIDS in Cape Verde*, Cristiana J. Silva, Delfim F. M. Torres

[4] https://www.cdc.gov/mmwr/volumes/68/wr/mm6840a3.htm

[5] https://harkeraquila.com/22246/science-and-technology/google-flu-trends/

[6] https://github.com/pietromiotti/thesis/tree/lmfit

[7] https://lmfit.github.io/lmfit-py/

[8] *Monitoring Italian COVID-19 spread by a forced SEIRD model*, Elena Loli Piccolomini, Fabiana Zama

[9] *Appunti sulle equazioni differenziali ordinarie*, Antonio Ambrosetti, *Springer*

[10] *https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19-similarities-and-differences-with-influenza*