University of Padova

Department of Mathematics "Tullio Levi-Civita"

Master Thesis in Data Science

A data-driven epidemic model to analyse the course of covid-19 in the veneto region

Supervisor
Nicolò Navarin

Co-supervisor
Alessandro Sperduti

Master Candidate
Claudia Cozzolino

Matriculation N°
1227998

Academic Year

2020-2021

University of Padova

Department of Mathematics "Tullio Levi-Civita"

Master Thesis in Data Science

A data-driven epidemic model to analyse the course of COVID-19 in the Veneto region

Supervisor
Nicolò Navarin

Co-supervisor
Alessandro Sperduti

Master Candidate
Claudia Cozzolino

Matriculation N°
1227998

Academic Year
2020-2021

DEDICATED

TO ALL COVID-19 VICTIMS AND THEIR FAMILIES,
TO ALL THOSE WHO SACRIFICE THEMSELVES TO FIGHT THE PANDEMIC.

Abstract

The current COVID-19 pandemic is an unprecedented global health crisis, with severe economic impacts and social damages. Mathematical models are playing an important role in this ongoing emergency, providing scientific support to inform public policies worldwide. In this thesis work, an epidemic model for the spread of the novel Coronavirus disease in the Veneto region has been proposed. Starting from the available local Health System data to examine past year contagion numbers and other features potentiality, a SEIQRD (Susceptible Exposed Infected Quarantined Removed Deceased) compartmental schema has been designed generalizing the classic SIR model. Then, the infection dynamics have been practically implemented in two versions: as a Deterministic Equation-based formulation and as an Agent-based model. While the former has been maintained simple and computationally inexpensive in order to serve as a baseline and to quickly provide parameter estimates, for the latter a detailed meta-population of agents with personalized attributes and network of contacts has been developed to recreate as realistic as possible simulations. Once these models have been trained and validated, they could became valuable tools for various types of analysis and predictions. In particular, the agent-based version, thanks to its flexibility as well as to its higher resolution, could be exploited for exclusive a posteriori evaluations of the effectiveness of the adopted containment measures in reducing the pandemic in Veneto.

Contents

List of Figures

List of Tables

1
Introduction

As of July, 2021, Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) has infected more than 180 million individuals, reporting more than 3.9 million victims worldwide. Since December 2019, when the virus was firstly identified from a cluster of viral pneumonia cases in Wuhan, China, the situation rapidly evolved becoming a large-scale pandemic.

All Countries with the coordination of the World Health Organization (WHO), are facing important decisions to fight the spread of the epidemic, for which a vaccine has been only developed in recent months. It has been shown that non-pharmaceutical measures like social distancing, face masks, home isolation as well as school and commercial activity closure are essential to reduce contagion numbers effectively [23] [25] [35] [47].

However, after more than a year spent in a pandemic state, profound economic and social damages are evident [18] [39], calling for interventions seeking a trade off between sanitary and financial guarantees. Accurately monitoring case statistics, modeling the epidemiological curves and forecasting the outbreak loads in the health and economic systems are then fundamental to assist and inform policymakers, possibly providing significant thresholds and guideline directions. The entire scientific community has therefore mobilized to make its contribution, not only from a medical or biological point of view, but also from the mathematical aspect.

The purpose of this work is to contribute to the field by producing models for the COVID-19 epidemic in Veneto, a region in northern Italy, an area particularly affected by the pandemic. After a careful and precise analysis of the available local data, provided by the healthcare system of the Veneto region, two models have been created experimenting with different levels of flexibility and numerical tractability. A reference SEIQRD compartmental schema has been designed as a customized generalization of the classic SIR model.

In practice, it has been firstly implemented as a totally Deterministic Equation-based model. Despite its simplistic assumptions, it resulted to be suitable for validating the compartments and transitions choices as well as for producing initial estimates of the parameters, thanks to its low computational burden.

Then, a more complex Agent-based version has been realized, generating a meta-population of individuals with specific personalized attributes such as age and presence of chronic diseases, both variables turned out to be strongly related to the severity of Coronavirus disease. Moreover, in order to overcome the typical differential equations limitation of homogeneous mixing and to obtain more realistic simulations, in this version the dynamics of contagion have been based on networks. Each agent has in fact connections defined by a multi-layered social graph representing plausible household, work, school or other generic community contacts.

Both models have been fitted on the epidemiological curves observed in Veneto, allowing at least part of the parameters to vary over time. In this way, they could better encode the alternating stringency of the containment measures and non-pharmaceutical strategies adopted during the last year. The validation and the training of such models have certainly constituted the most complicated and challenging phase of the entire work, especially as regards the Agent-based one, whose simulations generally require long execution times.

However, once the optimal parameters have been found, these models become useful tools for carrying out unique retrospective analyses on the effectiveness and the timeliness of the applied measures in reducing the effects of the pandemic. As a matter of fact, knowing the parametrized probabilities and rates values in each period, it is possible to simulate various scenarios conducting a What-If analysis. In addition, these models could be used for forecasting purposes, predicting the progress of infections, deaths or hospitalizations.

To conclude, thanks to their flexibility and their detailed structure, they could be possibly exploited by the scientific community to gain new understandings on the transmission dynamics or to provide basis to policymakers.

The thesis is organized as follows. First of all, the basic notions and examples of epidemiological modeling are introduced in Chapter 2, where the state of the art on COVID-19 models is also presented. Chapter 3 describes the overall available data, as well as the performed preprocessing and exploratory analysis. In Chapter 4 the technical details on the adopted modeling methodologies are given, while in Chapter 5 the corresponding fitting and simulation results are shown and discussed. Finally, the concluding remarks are reported in Chapter 6.

2
Epidemiological modeling

An epidemiological model is first of all a mathematical model: a description of a system, a set of rules or of interacting and organized components, using mathematical tools with a certain degree of complexity and approximation. Models are generally developed to help explain a system, to study the effects of its various components as well as to make predictions about their behaviour. From a biological and medical point of view this numerical approach is essential to estimate critical quantities, gain new understanding, organize natural data and seek optimal intervention strategies. Additionally, it is fundamental to simulate experiments not easily reproducible *in vivo* such as the spread of infectious diseases in populations, the primary concern of epidemiological models.
The aim of this chapter is exactly that to present the fundamental basis of most epidemiological models: the Compartmental model, starting from its simplest formulation as SIR (Susceptible–Infectious–Recovered) to its generalized variants. The main popular implementation strategies are then described, while highlighting the major advantages and limitations, with special attention to their recent applications to SARS-CoV-2.

2.1 Compartmental Models

Compartmental models are a class of very flexible modelling technique. As the name suggests, the elements of the system under study are divided into groups, compartments, representing a certain state or condition. The possible variations between compartments are instead modeled with transitions, usually in the form of rates. Although they can be generalized to many contexts, they are especially applied in the epidemiological field, for which they have become famous.

2.1.1 The SIR Model

One of the most adopted epidemic model is the SIR compartmental model proposed by Kermack and McKendrick in 1927 [34], for its description here [40] is closely followed. Considering a disease spreading in a population of size N, it splits its individuals into nonintersecting compartments. In one of the simplest scenarios, there are three such classes:

- **Susceptible**: individuals who have no immunity to the infectious agent, who are healthy but can contract the disease if exposed. The size of this class is usually denoted by S.
- **Infectious**: individuals who are currently infected and can transmit the infection to susceptible individuals who they contact. The size of the class of infectious individuals is denoted by I.
- **Recovered**: individuals who are recovered from the infection and cannot contract the disease again, hence those individuals could neither infect nor been infected, no more affecting in any way the transmission dynamics. The class of recovered individuals is usually denoted by R.

The number of individuals in each of these classes changes with time, that is, $S(t)$, $I(t)$, and $R(t)$ are functions of time t. The total population size N is the sum of the sizes of these three classes and it is assumed costant: $N = N(t) = S(t) + I(t) + R(t)$.

Regarding the epidemiological dynamics, the model represents them with a system of ODEs that describe how each class changes over time. Note that S, I and R must be integers, but assuming that the size of the population N is large enough, the condition could be relaxed to treat them as continuous variables.

Starting to mathematically modelling, when a susceptible individual enters into contact with an infectious individual, that susceptible individual becomes infected with a certain probability and moves from the susceptible class into the infected class. Hence, the susceptible population decreases in a unit of time by all individuals who become infected in that time, while the compartment of infectives increases by the same quantity. The number of individuals who become infected per unit of time in epidemiology is called **incidence**. The rate of change of the susceptible could be then written as the derivative

$$S'(t) = -\text{incidence}.$$

Now the problem moves to rightly represent the incidence: considering one infectious individual, it could be assumed that cN is the number of contacts per unit of time this infectious individual makes. Here it is assumed that the number of contacts made by one infectious individual is proportional to the total population size with per capita contact rate c. The ratio $\frac{S}{N}$ is instead the probability that a contact is with a susceptible individual. Thus, $cN\frac{S}{N}$ is the number of contacts with susceptible individuals that one infectious individual makes per unit of time. However, not every contact with a susceptible individual necessarily leads to transmission of the disease. Suppose p is the probability that a contact with a susceptible individual results in transmission. Then, pcS is the number of susceptible individuals who become infected per unit of time per infectious individual. Consequently, βSI is the number of individuals who become infected per unit of time, the incidence, where $\beta = pc$. As a result, the following differential equation for susceptible individuals is obtained:

$$S'(t) = \beta IS.$$

If we define $\lambda(t) = \beta I$, then the number of individuals who become infected per unit of time is equal to $\lambda(t)S$. This function is called the **force of infection**, where the coefficient β is the constant of proportionality called the **transmission rate** constant. The number of infected individuals in the population $I(t)$ is called instead the **prevalence** of the disease.
At this point, having the susceptible individuals who become infected moved to the compartment I, it is known that those subjects could recover or die. In particular, in the Kermack–McKendrick SIR Epidemic Model it is assumed that they leave the infected class at constant per capita probability per unit of time γ, called the **recovery rate**. That is, γI is the number of infected individuals per unit of time who recover. So,

$$I'(t) = \beta IS - \gamma I.$$

Finally, individuals who recover leave the infectious compartment and move to the recovered one with the same rate

$$R'(t) = \gamma I.$$

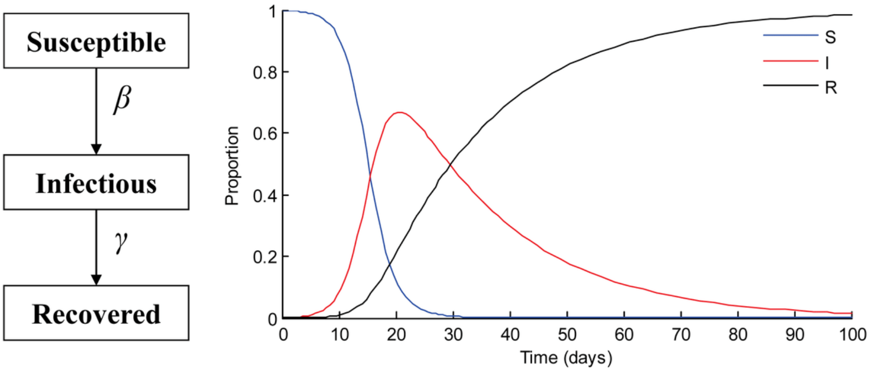

Figure 2.1: Flowchart of the Kermack–McKendrick SIR epidemic model (*left*) and an example of the typical behaviour of the three compartment sizes in function of time, proportionally to the population size N (*right*).

Thus, the whole model is given by the following system of ODEs:

$$\begin{cases} \frac{dS}{dt} = -\beta SI \\ \frac{dI}{dt} = \beta SI - \gamma I \\ \frac{dR}{dt} = \gamma I \end{cases} \tag{2.1}$$

To be mathematically well defined, this system is equipped with initial conditions $S(0)$, $I(0)$, and $R(0)$. Moreover, a differential equation model such as the model in 2.1 is well posed if through initial condition, there exists a unique solution. Because the dependent variables in the model denote physical quantities, it also required that solutions that start from nonnegative initial conditions remain nonnegative for all times in order to ensure mathematical acceptability and biological significance.

2.1.2 Mathematical Properties of the SIR Model

Constant population size

First of all observe that the population size $N(t)$ is constant at every instant of time t, as initially assumed. Denoting by N the total population size at time zero, it is known that

$$N = N(0) = S(0) + I(0) + R(0).$$

By definition, at each time t, the sum of the three not intersecting compartments gives the size of the total population at that time

$$N(t) = S(t) + I(t) + R(t)$$

Obtaining the derivatives and substituting the right hand terms for all the three equations in system, it follows that

$$N'(t) = S'(t) + I'(t) + R'(t) = -\beta SI + \beta SI - \gamma I + \gamma I = 0.$$

Hence, $N(t)$ is constant and equal to its initial value, $N(t) = N$.

Analysis of the three classes dynamics

Now, let's analyse the dynamic for each of the classes. Firstly note that, because $S'(t) < 0$ for all t, the number of susceptible individuals is always declining, independently of the initial condition $S(0)$. Since $S(t)$ is monotone and positive, the limit

$$\lim_{t \to \infty} S(t) = S_\infty$$

exists and it is finite. With an analogous reasoning, the number of recovered individuals has a monotone behaviour, independently of the initial conditions: since $R'(t) > 0$ for all t, the number of recovered individuals is always increasing. Given also that the number of recovered is bounded by N, the limit

$$\lim_{t \to \infty} R(t) = R_\infty$$

exists and it is finite. On the other hand, the number of infected individuals may be monotonically decreasing to zero, or may have nonmonotone behavior by first increasing to some maximum level, and then decreasing to zero. A necessary and sufficient condition for an initial increase in the number of infecteds can be easily determined forcing the derivative of $I(0)$ to be strictly positive

$$
\begin{aligned}
&I'(0) = \beta S(0)I(0) - \gamma I(0) > 0\\
&I'(0) = (\beta S(0) - \gamma)I(0) > 0\\
&\iff \beta S(0) - \gamma > 0\\
&\iff S(0) > \frac{\gamma}{\beta}
\end{aligned}
$$

In other words, if $S(0) > \frac{\gamma}{\beta}$, there is a sudden increase in the prevalence and then a decline to zero, generating a full-fledged **epidemic** or **outbreak**, otherwise, $I(t)$ monotonically decreases to zero resulting in no epidemic. If almost everyone is initially susceptible, i.e. $S(0) \simeq N$, then a newly introduced infected individual can be expected to infect other people at rate βN during his infectious period which lasts $\frac{1}{\gamma}$ (see estimation below). Thus, this first infective individual can be expected to infect:

$$\mathcal{R}_0 = \frac{\beta N}{\gamma}$$

individuals. The number $\mathcal{R}_0$ is called **basic reproduction number**, and it is undoubtedly one of the key parameter when analysing the spread of an infectious disease.
To determine the values of the two limits S_∞ and R_∞, let's instead divide the equation for S by the equation for R

$$\frac{dS}{dR} = -\frac{\beta}{\gamma}S,$$

then solving for S in dR,

$$S(t) = S(0)e^{-\frac{\beta}{\gamma}R(t)}.$$

Now recalling that the size of each of the three compartment is always smaller than the size of the total population, the following bound holds

$$S(t) = S(0)e^{-\frac{\beta}{\gamma}R(t)} \geq S(0)e^{-\frac{\beta}{\gamma}N} > 0 \quad \forall t \in [0, \infty)$$

It is hence possible to conclude that $S_\infty > 0$. This quantity is called the **final size of the epidemic**. It is interesting to highlight that this means that the epidemic does not end because all susceptible individuals have been infected and are now immune, but on the contrary that some individuals are always able to escape the disease.
Another important result to show is that, according to the SIR model, at a certain time the epidemic dies out, i.e.

$$\lim_{t\to\infty} I(t) = I_\infty = 0.$$

To see this, start integrating the first equation

$$\begin{aligned}\int_0^\infty S'(t)dt &= -\beta\int_0^\infty S(t)I(t)dt\\ S_\infty - S(0) &= -\beta\int_0^\infty S(t)I(t)dt\\ S(0) - S_\infty &= \beta\int_0^\infty S(t)I(t)dt\end{aligned}$$

then, recalling that S is nonnegative and monotonically decreasing, the right hand term can be bounded as

$$S(0) - S_\infty = \beta\int_0^\infty S(t)I(t)dt \geq \beta S_\infty\int_0^\infty I(t)dt.$$

This last inequality implies that $I(t)$ is integrable on $[0, \infty)$, hence, being also nonnegative and definitively decreasing, its limit for $t \to \infty$ is 0.
In Figure 2.1 a graphical example is given to summarize the typical dynamics just described.

System solution

To solve the system, first notice that the variable R does not participate in the first two equations. Consider thus only the equations for S and I, which are coupled, and leave out the equation for R, which could be always obtained in this model from the relation $R = N - S - I$:

$$\begin{aligned}S'(t) &= -\beta SI\\ I'(t) &= \beta SI - \gamma I.\end{aligned}$$

Dividing the two equations, obtain

$$\frac{I'}{S'} = \frac{\beta SI - \gamma I}{-\beta SI} = -1 + \frac{\gamma}{\beta S}.$$

Then, separating the variables and integrating at each member

$$\begin{aligned}I' &= \Big(-1 + \frac{\gamma}{\beta S}\Big)S'\\ I &= -S + \frac{\gamma}{\beta}\ln(S) + c\end{aligned}$$

where c is an arbitrary constant. Thus, the orbits of the solution are given implicitly by the equation

$$I + S - \frac{\gamma}{\beta}\ln(S) = c.$$

Since the Kermack-McKendrick model is equipped with initial conditions $S_0 = S(0)$ and $I_0 = I(0)$, from the above equality computed in (S_0, I_0) an explicit expression of the constant c can be deduced:

$$c = I_0 + S_0 - \frac{\gamma}{\beta}\ln(S_0).$$

Hence, the solution could be finally written as

$$I(S) = I_0 + S_0 - S + \frac{\gamma}{\beta}\ln\left(\frac{S}{S_0}\right) \tag{2.2}$$

Even if this is an exact solution, it only gives I as a function of S and not as a function of t: particularly, it does not give any indication of the time taken to reach any particular points on the orbits. Unfortunately, despite the simplicity of the SIR model, it is impossible to obtain an exact solution for $I(t)$. In that case, it is necessary to proceed with numerical strategies, like the Euler's method, which description is not covered because out of the purposes of this work.
It is worthwhile to mention that Equation 2.2 also allows to compute the maximum number of infected individuals that is attained. This number occurs when $I'(t) = 0$, that is, when $S = \frac{\gamma}{\beta}$. Substituting in S, it is trivial to conclude that

$$I_{\text{max}} = I_0 + S_0 - \frac{\gamma}{\beta} - \frac{\gamma}{\beta}\ln(S_0) + \frac{\gamma}{\beta}\ln\left(\frac{\gamma}{\beta}\right)$$

where I_{max} is the maximum number of infected individuals reached in the epidemic, i.e. the maximum severity of the epidemic. Being able to estimate this values for a newly occurring infectious disease could be important to know when the number of infections will begin to decline, but also to forecast the impact of the spreading disease on the sanitary or economic systems.

Parameters Estimation

For many diseases, information about the mean duration of the exposed period or the infectious period is available, for instance in medical literature is known that the duration of the infectious period for influenza is 3-7 days with a mean around 4-5 days. In the SIR model this knowledge could be easily exploited to help better estimate the recovery rate γ.

Start by assuming that there is no inflow in the infectious class and a certain number of individuals I_0 have been put in the infectious class at time zero. Then the differential equation that gives the dynamics of this class is given by

$$I'(t) = -\gamma I, \quad I(0) = I_0.$$

This first order differential equation can be easily solved: the number of people in the infectious class at time t is therefore given by

$$I(t) = I_0 e^{-\gamma t},$$

equivalently, the ratio

$$\frac{I(t)}{I_0} = e^{-\gamma t}$$

gives the proportion of people who are still infectious at time $t \geq 0$, or, mathematically speaking, it gives the probability of being still infectious at that time. At this point, the fraction of individuals who have left the infectious class could be obtained as

$$F(t) = 1 - e^{\gamma t}, \;\; t \geq 0.$$

This is clearly a probability distribution, the corresponding probability density function is given by $\gamma e^{-\gamma t}$. Hence the length of the infective period is distributed exponentially with parameter γ and known expected value $\frac{1}{\gamma}$. To conclude, having in practice the mean duration of the infection in a subject, i.e. the mean time spent in the infectious class, the recovery rate γ can be estimated as the reciprocal of this value.

Regarding the parameter of transmission rate β, the estimation is quite a bit more difficult. However, the equation of the orbit found in the previous paragraph could be additionally exploited to get some interesting approximation. Recalling that $\lim_{t\to\infty} I(t) = 0$, while $\lim_{t\to\infty} S(t) = S_\infty > 0$ gives the final number of susceptible individuals after the epidemic is over, imposing the passage by both (S_0, I_0) and $(S_\infty, 0)$ it could be written that

$$\begin{aligned} & I_0 + S_0 - \frac{\gamma}{\beta}\ln(S_0) = c = 0 + S_\infty - \frac{\gamma}{\beta}\ln(S_\infty) \\ \iff & I_0 + S_0 - S_\infty = \frac{\gamma}{\beta}(\ln(S_0) - \ln(S_\infty)) \end{aligned}$$

Therefore,

$$\frac{\beta}{\gamma} = \frac{\ln\left(\frac{S_0}{S_\infty}\right)}{S_0 + I_0 - S_\infty}.$$

Now, assuming that $I_0 \simeq 0$ and $S_0 \simeq N$, the just obtained equation could be rearranged including the key term $\mathcal{R}_0$

$$\frac{\beta}{\gamma} = \frac{\ln\left(\frac{S_0}{S_\infty}\right)}{N - S_\infty} \iff \mathcal{R}_0\left(1 - \frac{S_\infty}{N}\right) = \ln\left(\frac{S_0}{S_\infty}\right).$$

Practically speaking, the real importance of this relation lies in the fact that, contrary to the contact rate β, the quantities S_0 and S_∞ may be estimated with a quite good accuracy by serological studies. Access these data, however, is possible only after the epidemic has run its course, thus making this estimate of $\mathcal{R}_0$ feasible just in a retrospective way.
An alternative approach to avoid extracting β from data, is to approximate the second equation in System 2.1 with:

$$I' = (\beta N - \gamma)I.$$

From this approximation, which is valid only in the early period of the spreading, it is immediately get the following

$$I(t) = I_0 e^{(\beta N - \gamma)t},$$

meaning that, initially, the number of infectives grows exponentially with **initial exponential growth rate**

$$r = \gamma(\mathcal{R}_0 - 1).$$

Since r may be determined experimentally when an epidemic begins, and N, γ may be measured as well, also β can be indirectly calculated as

$$\beta = \frac{r + \gamma}{N}.$$

A word of warning should be spent to underline that $\mathcal{R}_0$ and r are not the same parameter, even if both are strength measure of the spreading. First of all note that while the former is an unitless quantity, consequently it does not provide any information about time, the latter measures how fast the spreading runs in time. Introducing the concept of **generation time** GT, i.e. the amount of time between an individual is infected by an infector and the time that the infector was infected [16], it is possible to link the two parameters. As a matter of fact, in literature, many authors have provided mechanical ways to combine those three quantities, for example as $\mathcal{R}_0 = 1 + rGT$ [37]. Nevertheless, many problems arise with those relation, either practically and theoretically speaking. Firstly, in order to obtain meaningful relations, often many limiting assumptions should be made, like the one that GT not vary in time, a too strong hypothesis. Additionally, generation times are not trivial to calculate, since a detailed contact-tracing would be needed. For these reasons, a common solution is to introduce the so-called **serial interval**, definable as the time between the instants when an infector and an infectee become symptomatic.

Interpretation of R_0

As already said, $\mathcal{R}_0$ is one of the key parameter when analysing the spread of a disease. The reason of its importance lies in the following property:

> let $(S(t), I(t))$ be the solution of the System 2.1, defined respectively the susceptible, infectious fraction as
>
> $$s(t) = \frac{S(t)}{N}, \quad i(t) = \frac{I(t)}{N},$$
>
> if $\mathcal{R}_0 \leq 1$, then $i(t)$ decreases to 0 as $t \to \infty$. Otherwise, if $\mathcal{R}_0 > 1$, then $i(t)$ first increases up to a maximum value
>
> $$i_{\max} = i_0 + s_0 - \frac{\gamma}{\beta}(1 + \ln(\mathcal{R}_0)),$$
>
> and then decreases to 0 as $t \to \infty$. The susceptible fraction $s(t)$ is instead a decreasing function and the limiting value s_∞ is the unique root in $\left(0, \frac{\gamma}{\beta}\right)$ of the equation
>
> $$i_0 + s_0 - s_\infty + \frac{\gamma}{\beta} \ln\left(\frac{s_\infty}{s_0}\right) = 0.$$

Epidemiologically, these results are reasonable: if enough people are already immune so that a typical infective initially replaces itself with no more than one new infective, the number of infectives decrease, leading to no epidemic. On the contrary, if a typical infective initially replaces itself with more than one new infective, then infectives initially increase so that an epidemic outbreak occurs. The speed at which an epidemic progresses depends on the characteristics of the disease.

2.1.3 Generalizations and Variants of the SIR Model

From the definitions and descriptions just discussed, it is evident that the Kermack–McKendrick SIR model is based on several assumptions:

- there are no vital dynamics, i.e. births and deaths in the population;
- the population is closed, in the sense that either no new individual can enter the population and no one can leave it, therefore eliminating phenomena such as due to tourism or immigration;
- after the exposition to the infection, all recovered individuals have complete immunity and cannot be infected again.

These assumptions seem very restrictive, but within certain limits and scenarios, they can be satisfied. For example, most of the diseases typical of childhood years, commonly called childhood diseases, lead to permanent immunity and can be thus suitably modelled by the SIR epidemic model (e.g. chickenpox, smallpox, rubella, etc.).

However, when the duration of the epidemic outbreak is quite long, when the pandemic is global in scope or when the hypotheses do not hold from a medical - clinical point of view, the SIR simplifications are too limiting. For these reasons, many variants of this model have arisen in the literature over time [6][15]. A short summary for the main generalizations are here listed.

- **The SIR model with vital dynamics:** introducing B, the **birth rate**, and μ, the **natural mortality rate**, respectively the number of births and deaths per unit time, the classical system is modified as follows

$$\begin{cases} \frac{dS}{dt} = B - \beta SI - \mu S \\ \frac{dI}{dt} = \beta SI - \gamma I - \mu I \\ \frac{dR}{dt} = \gamma I - \mu R \end{cases} \tag{2.3}$$

  Note that usually, it is assumed that the birth rate depends on the total population size, i.e. $B = \Lambda(N)$. Moreover, it is important to underline that here the fatality parameter refers to deaths due to natural causes and not due to the disease in study.

- **The SIS model:** after the exposition to infection, in this variant, individuals could only become susceptible again. This is the case, for example, for the common cold and influenza, which do not confer any long-lasting immunity. The model system reduces to

$$\begin{cases} \frac{dS}{dt} = -\beta SI + \gamma I \\ \frac{dI}{dt} = \beta SI - \gamma I \end{cases} \tag{2.4}$$

Recalling that $N = S + I$ is constant, the system can be rewritten in the form of a logistic differential equation

$$\frac{dI}{dt} = (\beta N - \gamma)I\left(1 - \frac{1}{N - \frac{\gamma}{\beta}}\right),$$

for which it is possible to easily find an analytical solution by separation of variables given an initial condition I_0

$$I(t) = \frac{I_0 K}{I_0 + (K - I_0)e^{-rt}},$$

where $r = \beta N - \gamma$ and $K = N - \frac{\gamma}{\beta}$. Similarly to what done for the SIR, also for the SIS model it is possible to give explicit conditions in the basic reproduction number for the prediction of the epidemic outbreak: if $\mathcal{R}_0 < 1$ then all solutions with non-negative initial value approach the limit zero as $t \to \infty$, while if $\mathcal{R}_0 > 1$ then all solutions with non-negative initial values except the constant solution $I \equiv 0$ approach the limit K.

- **The SIRD model:** a new compartment D is defined to model potential fatal infection. An infected subject could hence move to class R or D, respectively if completely recovered or deceased due to the disease. These design settings could be of particular interest for infections with very high mortality percentages, like Ebola virus. The new system of equations is

$$\begin{cases} \frac{dS}{dt} = -\beta SI \\ \frac{dI}{dt} = \beta SI - \gamma I - \mu I \\ \frac{dR}{dt} = \gamma I \\ \frac{dD}{dt} = \mu I \end{cases} \tag{2.5}$$

  Again, μ parameterizes the number of deaths per unit time, but with the difference that, in this variant, only deceases due to the infection are considered.

- **The SEIR model:** considering that, for many common infections there is a significant incubation period during which individuals who have been infected are not yet infectious, a popular modification includes a compartment E for exposed but not yet contagious subjects. The time interval between the instant when an individual is infected and the one when he or she becomes infectious is called **latent period**. Thus, introducing a new parameter ϵ, where ϵ^{-1} is the mean value of the latent period, assumed exponentially distributed, the SEIR model is described by the following ODEs system

$$\begin{cases} \frac{dS}{dt} = -\beta SI \\ \frac{dE}{dt} = \beta SI - \epsilon E \\ \frac{dI}{dt} = \epsilon E - \gamma I \\ \frac{dR}{dt} = \gamma I \end{cases} \tag{2.6}$$

for which similar mathematical properties can be obtained as those previously shown for SIR and SIS.

- **The MSIR model:** while implementing a model with vital dynamics, it could be important to include the **passive immunity** phenomenon. As a matter of fact, for many infections, like measles, babies do not directly born into the susceptible compartment. On the contrary, since for the first few months of life they are immune to the disease thanks to the protection from maternal antibodies, passed across the placenta or additionally through colostrum, they initially belong to a new class, say M, for maternally derived immunity. Only when these passive antibodies are gone, the infant becomes susceptible to the disease, moving from the passively immune state M to the susceptible state S with a per capita rate δ. In the case of infants without any passive immunity, because their mothers were not exposed to the infection, the class S is directly entered, so that they could immediately be infected. Mathematically this is translated as

$$\begin{cases} \frac{dM}{dt} = B - \delta M - \mu M \\ \frac{dS}{dt} = \delta M - \beta SI - \mu S \\ \frac{dI}{dt} = \beta SI - \gamma I - \mu I \\ \frac{dR}{dt} = \gamma I - \mu R \end{cases} \tag{2.7}$$

where B and μ are respectively the birth rate and the natural mortality rate with the same interpretation of System 2.3.

As anticipated, this is only a list of few representatives of the most popular variants, it is easy to imagine that an huge variety of generalizations can be designed. It is possible to derive many other combinations as a mixture of these models, e.g. SEIS, MSIRS, etc. Other ways contemplate changes in the effects of the vital dynamics or vertical transmission of the infection from parents to their offspring, true for diseases like AIDS and Hepatitis B [11]. Finally, a common approach is to add new *ad hoc* compartments, e.g. V for vaccinated, Q for quarantined, etc. or classes representing also non-human species acting as transmission vector, such as mosquitoes in malaria [3].

2.2 Practical Implementation

Despite the simplicity and compactness of its definition, the practical development of a compartmental model cannot be considered equally trivial. In fact, this places mathematicians and more in general researchers in front of a challenge: implement a model that is as much more representative of the real phenomenon while computationally and analytically tractable. According to the adopted solution, the methods can be classified as deterministic or stochastic, networks-based or agent-ased epidemic models. On the following pages a short survey is presented.

2.2.1 Deterministic versus Stochastic Models

One of the first aspects to consider during the designing phase is undoubtedly that of choosing whether to proceed in a deterministic or stochastic way. Recall that in a **deterministic model** every set of variable states is uniquely determined by the parameters in the model and by the initial value of the variables themselves. On the other hand, **stochastic models** are characterized by randomness, hence variable states are described by probability distributions instead of being simply constant.

In the case of compartmental models, their original deterministic nature, which formulation corresponds to the one just presented above, makes them quite simple to treat both analytically and computationally. Nevertheless, as expected, their major downside is that they could not be very realistic, since what they aim to represent, e.g. inter-human contacts or exposition to viruses, is intrinsically ruled by randomness. Made these premises, their powerful remains undeniable: the classical deterministic SIR model without births and deaths is still among the most studied and it is generally quite effective at describing the dynamics of a range of infections in many populations [4]. Even today this strategy could be appealing when interested in more modest extensions of the SIR model that focus only on specific aspects of the complex problem, maintaining a relatively small parameter set and deterministic equations. A successful example is given in [29] where the authors adopted this method to study the transmission dynamics within and between households, assuming them as the key mechanism for the spread and persistence of infection.

However, there are cases in which deterministic models are insufficient, making necessary the choice of a stochastic strategy. Firstly, if considering an epidemic outbreak in a small community, like a day-care center, an office or a school, it seems reasonable to assume some uncertainty

in the final number of infected. Therefore, it is important to stress that the deterministic approaches are valid and reasonable only in case of sufficiently large populations and relatively short epidemic, for this reason they should be used and interpreted with caution [7].
Moreover, even when the community is large and $\mathcal{R}_0 > 1$, it should be possible that, by chance, the epidemic never takes off if the outbreak is initiated by only one or a few initial infectives. These arguments motivate the definition of stochastic version epidemic model. Statistically, a further evidence that is often brought in favor of their use is that they enables parameter estimates from disease outbreak data to be equipped with standard errors [12].
In practice, the main differences are, as said above, that the state function $S(t)$, $I(t)$, etc. as well as the parameters like the contact rate or the recovery time are now random variables. Regarding the mathematical tools, discrete time Markov chain, continuous time Markov chain, stochastic differential equations, branching process and Poisson process constitutes the fundamental basis, which allow to confer unique properties to the stochastic models. As a matter of fact the probability of disease extinction, the probability of disease outbreak, the quasistationary probability distribution, the final size distribution or the expected duration of an epidemic could be derived, showing in some case that, even under the same hypothesis, stochastic models can exhibit different asymptotical behaviours from their deterministic counterpart. Of course there are many possible types of stochastic epidemic model, the decision of which type of model to choose, or a new one to invent, depends on the specific questions to be explored and on the available data. Please refer to [1][2][14][28] for more technical details and for many nice examples starting from the original SIR definition.

2.2.2 Network Models

The typical example of epidemic modeling for a society subgroup, like school students, provided above, also reveals another weakness of deterministic models. The assumption of a homogeneous uniformly mixing cannot be easily accepted for small population size. In **homogeneous mixing** the contacts of a person are assumed equally randomly distributed among all others in the population. One immediate implication of this hypothesis is that the force of infection λ is the same for all individual classes, contrary to what happens in real populations, where the mixing is more likely to be heterogeneous and contacts are not random. As a matter of fact, recent research works have shown that the social contact networks have community structure in which nodes usually tends to have more links within a cluster than that of between communities [51]. They also observed that, even when there is a significant number of infected individuals in a community, the contacts that actually transmit diseases between susceptible and infected do not grow quickly. This phenomenon, called crowding or protection effect, makes clear that the linear force of infection used in the standard SIR model has serious limitation under the typical scenario. An **heterogeneous mixing** assumption is therefore preferable, since it allows λ to reflect the social structure, such as age-related changes in the degree of mixing and contact or differences in behavior, e.g. due to non pharmaceutical interventions, like isolation and curfew, which can alter the standard contact patterns, all important factors for understanding disease spread [20].

Practically speaking, this can be done switching to network epidemic models. A network, or a graph, is a structure made of a set of objects eventually paired according to some relation. These objects are mathematically abstracted in vertices, also called nodes, while the related pairs of vertices are called edges. In network epidemic models then, the whole population is represented by a graph whose vertices are the subjects and whose edges describes physical contacts or social relationships. In other words, the population is made of single individuals instead of compartments allowing to consider different interactions types from the classical homogeneous mixing. These individual-level models usually make analysis difficult and simulations computationally intensive, but offer a totally different way of describing biological populations which seems to better fit epidemiological data coming from the real world.

Another attractive reason for studying epidemic models on networks is to better understand what network features affect spreading the most. In this way it is possible to gain knowledge and make hypothesis on how to reduce the spreading adopting suitable public health measures such as vaccination, isolation, travel restrictions, etc., testing their effectiveness by incorporating them into the model.

Reformulating the problem, new questions thence arise: how to build a realistic network, in particular, how assign links, how are node degrees, i.e. numbers of edges connected to each vertex, distributed and what do they depend on?

The first factor to consider is whether the underlying social structure is known or not. Although very rare, there are cases in which contacts between individuals belonging to the population under study were directly observed or estimated from data taken on a sample. In such works, researchers proceeded by means of prospective survey [42], contact diaries [45], available socio-demographic data [24][37], all tools that constitute milestones in this field before the advent of modern contact tracing, possible only with more advanced technology or in more systematic and meticulous settings.

However, in most cases this is not feasible, which explains why a **random network model** is often more advocated.

Several artificial generated networks have been proposed in the field of disease transmission. Each of these idealized networks can be defined in terms of how individuals are distributed in space, both in geographical or social terms, and how connections are established. The most popular types among them are now presented specifying assumptions as well as the major implications for epidemic spread. For the main notions and results, [33] has been closely followed.

- **Random network:** in this network, the spatial position of individuals is irrelevant, and connections are formed at random. As a matter of facts, this type of models has the least possible structure. For example, in the famous **Erdős-Rényi random graph** it assumes that every pair of individuals is connected to each other, independently, with probability δ/n, where n is the number of nodes and δ is the unique parameters representing the mean degree. Given these assumptions, it is easy to deduce that the number of neighbours any individual has is distributed as a Binomial $Bin(n-1, \delta n)$ for which it is mathematically known that asymptotically, as $n \to \infty$, it tends to the Poisson distribution $Pois(\delta)$. The random network is therefore characterized by a lack of clustering and by homogeneity of individual-node network properties. From disease spreading point of view, despite an effective rescaling in the growth rate can be observed, the epidemic dynamics for this particular settings remains analogous to a an SIR model with a homogeneously mixed population.

- **Lattice:** these models rely on very different assumptions. Individuals are positioned on a regular grid of points, usually two dimensional, in which only adjacent individuals are linked to mimic spatial localized contacts. Lattices are therefore homogeneous at the individual level and highly clustered thanks to the localized nature of its connections. Similarly to all networks, lattice models show a reduced initial growth of infection compared with random-mixing models, usually presenting a stronger effect because the spatial clustering of contacts causes a more rapid saturation of the local environment. Another common feature is the power-law, at which the frequency distributions of both epidemic sizes and epidemic durations have been proved to obey to.

- **Small-world network:** this model tries to overcome the limitation of the previous ones. While lattices display high clustering but long path lengths, i.e. a large number of steps required to move between two randomly selected individuals, random networks have short path lengths, but low clustering. Small-world networks tries instead to offer a trade-off solution by adding few random connections to a lattice like network. These rare long-range connections have a very important effect when simulating the spread of infection, since that allow the infection to reach all individuals in the population quite quickly. With the support of Percolation theory it has been proved that, even with a few distant links, there are significant changes in epidemic behaviour, which could potentially dramatically increase the likelihood of an epidemic. Nevertheless, since these long-range connections are less frequent, the transmission of infection still remains predominantly localized, keeping true the strong saturation property.

- **Scale-free network:** in this category of graphs another important standard network measures is considered: the degree distribution of node-individuals. In the topologies described untill now the number of contacts per subject is almost uniform. However, observing real social networks, it is often the case that many individuals have a small number of neighbours, while a few of them have significantly more connections. Since these highly connected individuals, called **super-spreaders**, could be fundamental protagonists in the spread and maintenance of infection, the inclusion of such disproportionately connected nodes in epidemic models is necessary. Scale-free networks provide an easy way to achieve this behaviour of heterogeneity. The most famous model among them is the **Preferential attachment model**, also called **Barabási-Albert model** by the names of its inventors. It is built dynamically, adding one by one new individuals to an already formed network with a connection mechanism that simulates the natural formation of social contacts. Each new node connects preferentially to individuals that already have a large number of contacts. This produces a graph where the number of contacts per individual takes a power-law distribution.

- **Exponential random graph:** this class of models provides a way to explicitly construct very flexible networks with a given set of properties. Its is inspired by statistical physics and allows for penalizing or favoring more or less any network feature such as individual edges, or summary statistics like the mean degree, number of triangles, low degree correlation or higher moments of the degree distribution. Exponential random graphs have the simple property that the probability of connection between two nodes is independent of the edges presence between any other pair of distinct nodes. This allows the likelihood of any nodes being connected to be calculated conditionally on the graph equipped with certain network properties. In practice, there is no direct method for generating such networks. They are instead obtained by starting with an initial network and then adding or deleting edges based on techniques like Markov Chain Monte Carlo until the chain is close to stationarity. A a range of plausible networks are thus produced, guaranteeing that they agree with the provided information.

Other well known structures notable to mention are the **Spatial network**, the **Configuration model**, the **Poissonian random graphs**, as well as the **Random block models**.

The type of network topology to be used should be carefully chosen depending on the context and disease. At that point, given the graph of contacts, the transmission model can be designed properly setting the parameters or their probability distributions in case of stochastic approach. Many re-adaptations of the SIR on network are possible. Two famous examples are the Reed-Frost discrete time epidemic model and then a continuous time Markovian model.

Another important aspect to take into consideration is the variability of edges in time: while for short time spans outbreaks a **static network** may be sufficient, when interested in longer periods, a **dynamic** model, where not only nodes can appear or disappear mimicking the vital dynamics, but also connections could be dropped or created, might be preferred.
In order to draw realistic conclusions, it is finally necessary to validate the models, preferably fitting them on specially collected contacts network or disease data to infer parameters with proper statistical methods. Of course, if the entire underlying graph is observed, it is often straightforward. However, as mentioned above, this only happens in rare situations. In fact, more complex inference strategies are usually performed on the few available egocentric or outbreak data [13].

2.2.3 Agent Based Models

The last practical methodology to implement an epidemic model here presented is the **Agent-Based Model** (ABM), or **Individual-Based Model** (IBM), approach.
Agent-based modeling and simulation is a quite new strategy that has gained increasing attention over the last two decades, catching on in various fields of application, ranging from economy and social sciences to biology and epidemiology. The reasons for its success are of different nature:

1. the systems to analyze and model are becoming more and more complex in terms of their interdependencies and details, requiring new solutions.

2. Some systems have always been too complex for classical equation based methods to adequately model. As a matter of fact, to guarantee analytical and computational tractability, these traditional approaches rely on tight assumptions. ABMs are instead able to relax some of them, providing a more realistic view.

3. Nowadays data are being collected and organized into databases at finer levels of granularity. Micro-data can hence support individual-based simulations including many variables and details.

4. Finally, but most importantly, computational power is advancing rapidly, large-scale microsimulation models, not plausible just a few years ago, can be now ran within minutes or hours.

Let's now describe the main ingredients and characteristics of ABMs. The main reference for this part is [38].

The primary protagonists are, of course, the **agents**. Despite, there is no universal agreement on the precise definition of the term "agent", general guidelines could be given still allowing an extremely high degree of flexibility. Agents consist of any type of independent components, having diverse, heterogeneous, and dynamic **attributes** and **behavior**. Regarding the latter, it is intended as the representation of a process that links the agent's sensing of its environment to its decisions and actions. Its description can range from simple if-then rules to complex behavioral models from the fields of cognitive science or artificial intelligence. According to some authors, the component's behavior must also be **adaptive** in order for it to be considered an agent. In this sense, the agent label is reserved only to components that can learn from their environment and dynamically change their behaviors in response to experiences, hence providing adaptation. Another characteristic usually associated to agents is **autonomy**, they should then be active responders and planners rather than purely passive components.
From a more practical point of view, agents are usually described by the following properties:

- **autonomous** and **self-directed:** an agent can function independently in its environment and in its interactions with other agents, generally from a limited range of situations that are of interest;

- **modular** or **self-contained:** an agent is an identifiable, discrete individual with a set of characteristics or attributes, behaviors, and decision-making capability;

- **social:** an agent interacts with other agents according to a given protocol or mechanism (e.g. contention for space and collision avoidance, agent recognition, communication and information exchange, influence and other domain-or application-specific rules);

- **live in an environment:** an agent is situated, in the sense that its behavior is situationally dependent, which means it is based on the current state of its interactions, not only with other agents, but also with the surrounding environment;

- **explicit goals:** an agent may have a role or tasks that drive its behavior, this allows it to continuously compare the outcomes of its moves to its goals, giving it a benchmark for possibly modifying them as in a reinforcement learning setting;

- **learn** and **adapt:** an agent may have the ability to change its behaviors based on its experiences. Individual learning and adaptation requires an agent to have **memory**, usually in the form of a dynamic agent attribute;

- **resource attributes:** an agent may finally have specific finite attributes that indicate its current stock of one or more resources (e.g. energy, wealth, information, etc.).

Once the agents were defined properly setting their type, attributes and behavioural rules according to the application interests, the **agent relationships** should be added. As a matter of fact, the second fundamental ingredient of an ABM is the method controlling which agents interact, when they interact, and how they interact. Many topologies are available to this purpose, in particular, a list of the most commonly used schemes for representing social agent interaction is presented below.

- **"Soup" Model:** in the "soup" or aspatial model, agents have no location and the model has no specific spatial representation. In general, pairs of agents are randomly selected for interaction and then returned to the soup from which they came.
- **Cellular Automata:** this configuration represents agent interaction patterns and available local information by using a lattice. Specifically the cells immediately surrounding an agent are its neighborhood, agents can move from cell to cell on this grid, usually only horizontally or vertically, with the rule of no more than one agent per cell at the same time.
- **Euclidean Space:** in this model, agents can freely roam in 2D, 3D or higher dimensional spaces.
- **Geographic Information System (GIS):** according to this settings, agents move over a realistic spatial landscape, which considers geographical boundaries, natural elements as well as artificial building or services (schools, shops, etc.)
- **Network topology:** it allows an agent's neighborhood to be defined more generally and sometimes more accurately. As already seen, network topology may be either static or dynamic, where links could be determined step by step according to the mechanisms included in the model.

No matter what topology is used to connect the agents, in an agent-based model the essential idea is that interaction as well as information transfer remains **local**.
Given the fundamentals of agent-based modelling, it is clear that ABMs present multiple appealing features when dealing with epidemiological models: randomness, heterogeneous mixing and heterogeneous population. They are therefore capable of overcoming all the major limitations and fragility of compartmental models and, more in general, of traditional models described by differential equations, like SIR [43]. Indeed, ABMs provide a quite simple way to model a system in which individuals, as single entities, could present important differences both in attributes and behaviour, allowing to include a range of possible factors, ranging from age or gender to cultural, economical or health status, all potential determinants in the spread

of an infectious disease in a population. Moreover, ABMs are also more suitable in situations in which it has been observed that individuals tend to adapt their behaviour to the epidemic, due to social pressure, fear of contagion or after the introduction of control measures as well as vaccination programs [31].
Lastly, it is worthwhile to mention that the compartmental, equation-based and agent-based approaches are not necessarily mutually exclusive, in fact there are successful examples in literature where an hybrid version is implemented seeking to maintain the qualities from both the methods [10].

To conclude, a great variety of epidemiological models have been developed over the years, starting from the simple deterministic formulation of Kermark and McKendric, increasing the realism by adding a more and more detailed social structure. The cost of this gained precision is that these models are often analytically intractable, hence researchers could only rely on complex microsimulations, which can be difficult to parametrize, require a computationally intensive analysis and might hinder even the basic understanding of the causal factors of the studied behaviour. Therefore, despite the increasing computational power available to researchers today, analytically tractability remains an invaluable property to guarantee interpretable tools to understand the role that different determinants play in the spread of the infection. Finally, the fact that very little computational power is required for their analysis and practical implementation makes simple compartmental model still very attractive for policy information [44], in particular when quick results are needed, as in the case of new infection outbreaks, like in the current Coronavirus emergency.

2.3 Models for COVID-19

Since the very first days from the discovery of the novel coronavirus circulation in early 2020, mathematical methods, statistical analysis and forecasting were fundamental tools to inform policy makers and researchers. The common will to fight a devastating pandemic has led to an extraordinary production in scientific literature, in particular for the epidemic modeling field. For a complete understanding it is important to briefly recall the main characteristics about this recent Coronavirus disease (principally referring to the last World Health Organization reports).
Firstly, COVID-19 is a contagious disease with a very heterogeneous symptomatology. It has

been observed that at least a third of infected people are asymptomatic, i.e. do not develop any noticeable symptoms. On the other hand, among the remaining cases, the majority are paucisymptomatic, i.e. develop mild to moderate symptoms (including fever, cough, headache, fatigue, breathing difficulties, loss of smell and taste up to mild pneumonia), while 14% develop severe symptoms (dyspnea, hypoxia), and 5% suffer critical symptoms (respiratory failure, shock, or multiorgan dysfunction). In addition, older people, and those with pre-existing medical problems like cardiovascular disease, diabetes, chronic respiratory disease, and cancer seem to be more likely to develop serious illness.

Regarding the infection spread, transmission occurs when people are exposed to virus-containing respiratory droplets or airborne particles exhaled by an infected person. It has been observed that the risk of infection is highest when people are in close proximity for a long time, but particles can even be inhaled over longer distances, particularly in poorly ventilated and crowded indoor spaces, where they can remain suspended in the air for minutes to hours. Touching a contaminated surface or object may also lead to infection although this seems not to substantially contribute to the overall transmission.

After an incubation period, infected subjects can transmit the virus to another person, starting from up to two days before symptoms onset and they could remain infectious till to ten days after, in moderate cases, and up to twenty days for more severe cases.

Since its first detection in China, SARS-CoV-2 has rapidly spread in few months all over the World, in particular in Italy, which was one of the first Western countries to face the health emergency and is still one of the most severely affected nations. The lack of preparation for the outbreak of COVID-19 beside the high rate of transmissibility, unknown complications, and inappropriate medications led indeed to a world-wide disaster paralyzing the health care and economic systems even in developed countries. The main motivations for epidemic model studies should be hence searched in the urgent need to respond as precisely as possible to questions on the dynamics of diffusion as well as on the effectiveness of proposed containment measures, while providing guideline thresholds for the general public safety.

The nature and purpose of the models have physiologically evolved with the progress of the epidemic, depending on the acquired medical knowledge and most importantly from the available data.

Although almost a century has passed since its definition, the SIR model has been one of the mostly adopted, at least as a starting point [41]. The simplicity of such ordinary differential Equation-based model made it very appealing, in particular in the first phase of the epidemic, when advanced Artificial Intelligence-based models could not be properly validated due of the absence of sufficient training datasets and Agent-based modeling could be not exploited because the detailed population-level parameters such as rates of contacts, distancing and virus infectivity parameters they rely on were not yet known.

As expected, many variants have been proposed as an alternative to the classic SIR and SEIR definitions, introducing newer compartments and topologies better describing SARS-CoV-2 infection. Reflecting the observed course, a common approach has been to separately model different sub-groups distinguishing on the stage and severity of the disease, for example including asymptomatic, symptomatic, hospitalized and deceased compartments [36]. In addition, quarantined, isolated or positive detected via test have been also commonly represented especially when assessing the quality of preventive measures and non-pharmaceutical interventions is the core purpose [27]. Only lately, vaccinated term has been also added to the model to predict how vaccination, with specially crafted drugs like AstraZeneca, Pfizer, Moderna, etc., could control the epidemic [19].

However, as already discussed in the previous sections, the SIR/SEIR models have numerous limitations and their validity is based on assumptions, such as closed population, homogeneity or constant parameters, all too restricting for the study of COVID-19 epidemic, especially for its long duration and global scope. For these reasons, it has been often used in its stochastic or Agent-based versions, which are more complex and flexible. In particular, the very characteristics of the virus, of the associated disease and of the pandemic evolution, led the researchers to focus on techniques that would allow to include individual as well as local variability: age-stratified compartments and parameters [47], meta-populations with distinct socio-economical roles and contact types like schools, households, work spaces [23], or multiple communities representing different clusters or geographical areas [8].

The correct representation of population connectivity has been another fundamental point given that Coronavirus transmission is by close contact and that the implemented containment methods and interventions like mask and social distancing try to directly modify it. Network based models are then ideal to generate more realistic mixing, for example estimating the underlying graph from social network data or contact tracing [9]. In literature, several proposals have further added to the level of individual contacts, e.g. for mobility and community interactions in order to inspect possible spreading paths between regions or to test the effectiveness of travel restriction [21] [25] [35].

The SIR model is also not very accurate in predicting for wider temporal horizons. For these purposes it has been therefore preferred to exploit more sophisticated strategies requiring more detailed knowledge of the biomedical and epidemiological aspects. Technically speaking, many

authors have proceeded with standard time-series techniques, like ARIMA models or smoothing splines [22] [26] more suited for prediction. Recently, even Deep Learning approaches have been tested for forecasting, benefiting from the natural memory properties for sequences of the Recurrent Neural Networks, for example by embedding long short term memory (LSTM) into compartmental formulation in order to design an hybrid model [48] [50] or designing a spatio-temporal Graph Neural Network with mobility data [32].
In conclusion, the range of models produced in this last year is very rich and varied. As often happens, there does not exist a single technique or proposal better than others, but the modeling choices must be guided based on the scenario of interest, the specific purpose or questions and on the available data.

3
Data presentation

Prior to the model design and implementation, an extensive work has been done to properly clean, prepare and analyze in depth the data under investigation. The purpose of this chapter is to present the dataset, with its main potentialities and problems, and to summarize the main studies and results obtained during the exploratory data analysis phase.
First of all, the data here examined were made available by Azienda Zero, the Veneto regional health authority, to the Cardio-Thoraco-Vascular Sciences and Public Health Department of the University of Padova. The records were provided in three different databases:

1. the first set contains personal and clinical information of patients tested positive to SARS-CoV-2 by molecular swab performed in Veneto between the 21st of February 2020 and 23rd February 2021. In almost all cases, for each patient only the date of first positivity is reported, together with any other possible dates of hospitalization or re-positivization. For this reason, among the total $328,832$ records, only 516 are related to the same patients appearing twice. In particular, it has been noticed that, for these cases, the presence of duplicated subjects is due to significant modification between the different rows, such as hospitalization or clinical status updates. Probably they were automatically collected in the dataset by the system.

2. The second database contains instead the records of all molecular swabs, both positive and negative, performed in Veneto region between the 1st of January 2020 and 1st April 2021. In total $4,511,302$ tests were collected. The available features are much less and sparse with respect to the previous set, appearing more linked to swabs information rather than to the subject on which that are performed.

3. The last database is essentially similar to the second as concern its fields, but it contains $3,800,129$ records about antigenic swabs made, always in Veneto, between the 1st of January 2020 and 30th March 2021.

The three sets are thus structurally inhomogeneous, however a subject code is always reported as primary key in order to ensure unique distinction of each patient within each database. Moreover, considering that the majority of the collected information comes from an application manually compiled by the health workers, many physiological errors, inconsistencies, missing values and redundancies have been found. The most worth to mention common problems are: outcome results transcribed with different expressions, identical record repetition, patients neither domiciled nor resident in Veneto, patients with infeasible or not given date of birth, swab report dates switched with collection dates. A thorough cleaning has been then carried out by crossing all the available features in order to minimize all the errors losing as little information as possible. After that, the three databases have been manipulated, standardized and merged together in order to create an unique repository from which extrapolate the history of each subject in terms of swabs, in particular, the number of molecular and antigenic swabs, both positive or negative to which he/she subjected to, the first date and the duration of his/her positivity period, the date of the definitive recovery etc.
The overall database resulting from these phases contains $8,205,939$ swab records, of which $4,416,677$ molecular and $3,789,262$ antigenic, relating to $2,350,259$ distinct subjects.
At this point, an in-depth exploratory and statistical analysis has been carried out. The main results are presented in the following sections.

3.1 Swab tests analysis

With the intention to proceed from the general to the particular, the analyses carried out on the overall dataset are firstly reported in this section. In this initial part, the investigations focused more on the swabs tests information rather than on the tested patients.
To start, the number of swabs performed over the weeks has been determined, both distinguishing by type (molecular or antigenic) and by outcome (positive or negative). The graphs in Figure 3.1, 3.2 show the the absolute frequencies. Figure 3.3 shows instead the percentage numbers of positive swabs per week, both on the total and on the different types of tests (note that in this, as well as in all the next lineplots, unless otherwise specified, the observations have been interpolated with cubic splines in order to make the graph smoother).

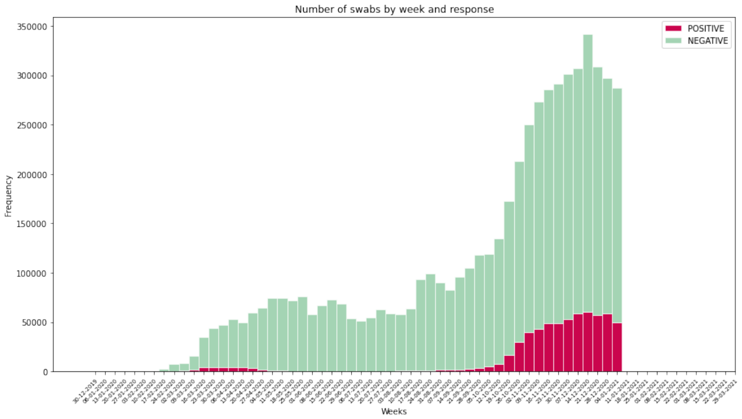

Figure 3.1: Weekly swab tests occurrences by response.

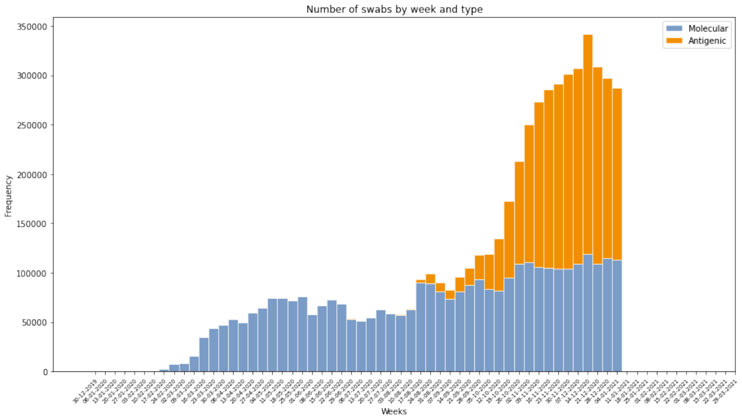

Figure 3.2: Weekly swab tests occurrences by type.

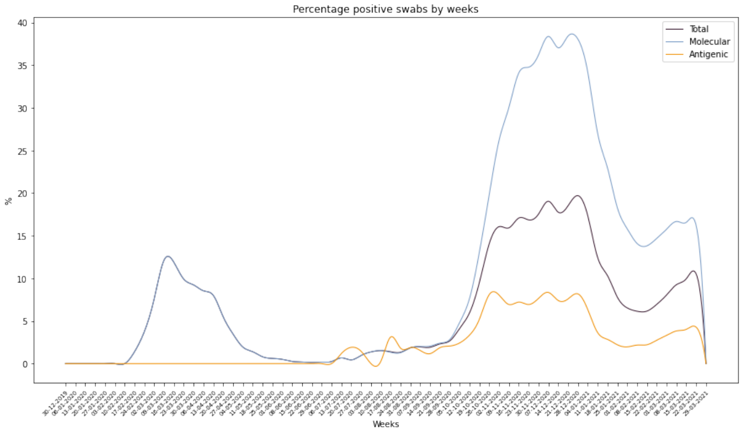

Figure 3.3: Weekly percentage of positive swab tests by type.

From these preliminary images it can be already deduced that, as expected, the number of total swabs performed increases consistently over time, reaching a maximum of nearly $350,000$ units per week thanks to the systematic introduction of antigenic tests. In addition, it can be noted that the percentage of positive outcomes is higher in the second peak rather than in the first one, at the beginning of COVID-19 outbreak. Comparing the prevalence by exam typology, it is also interesting to observe that in this late period the percentage of positives is larger for molecular swabs, for which it approaches a rate of 40%. A possible explanation could be that antigenic swabs, given their speed and convenience, have been frontier tests, used not only to assess suspicious subjects, but widely adopted during monitoring and screening activities in specific subgroups or samples of the population.

3.2 Tested population analysis

The exploration then moved on understanding what were the characteristics of the overall population tested: age, presence of pre-existing diseases, clinical course of COVID-19 and possible associations between the various factors.

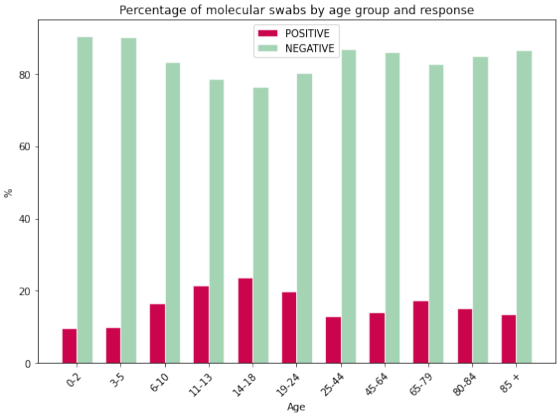

Figure 3.4: Percentage of molecular swabs by age group and response.

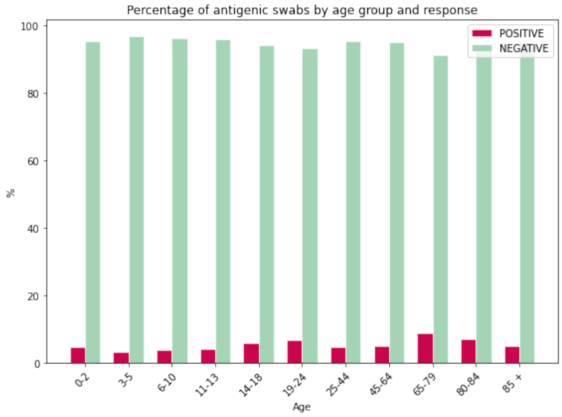

Figure 3.5: Percentage of antigenic swabs by age group and response.

3.2.1 Age

The analysis firstly focused on the occurrences by age group. In order to have an unbiased size reference, the age distribution among the patients subjected to at least one swab has been firstly compared with that of the total population of Veneto (source: ISTAT data), both depicted in

Figure 3.6. The histogram in Figure 3.7 shows instead the absolute occurrences by age and by swab type. Finally, Figures 3.4 and 3.5 represent respectively the percentage number of molecular and antigenic tests divided by age group and by outcome.

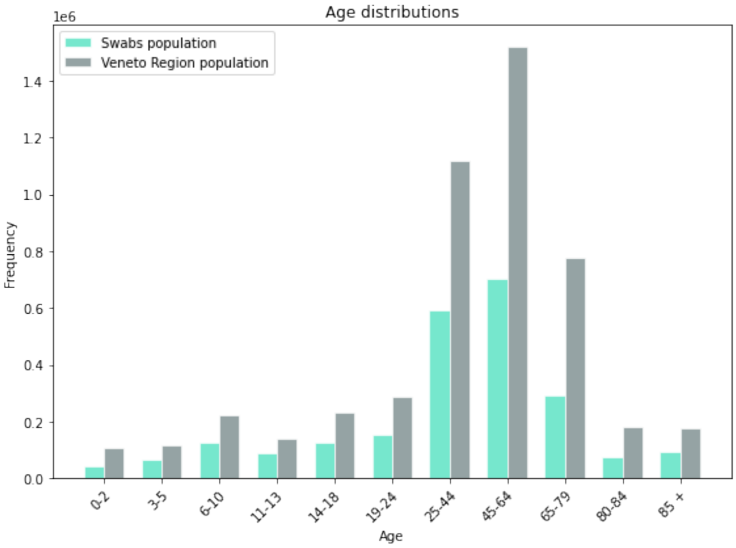

Figure 3.6: Age distribution of tested versus total population in Veneto.

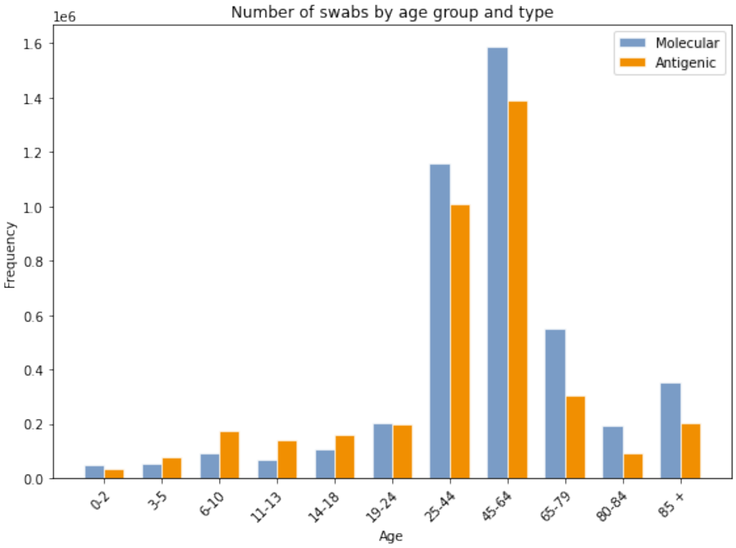

Figure 3.7: Absolute frequencies of swab tests by age class and type.

From these images it seems that the percentage of positives within the antigen tests is lower and less variable with respect the various age categories. On the contrary, in molecular tests, different age groups correspond to different positivity percentages: in particular, the positivity rate shows a bimodal trend with a first peak around $14 - 18$ years and a moderate second spike at $65 - 79$.

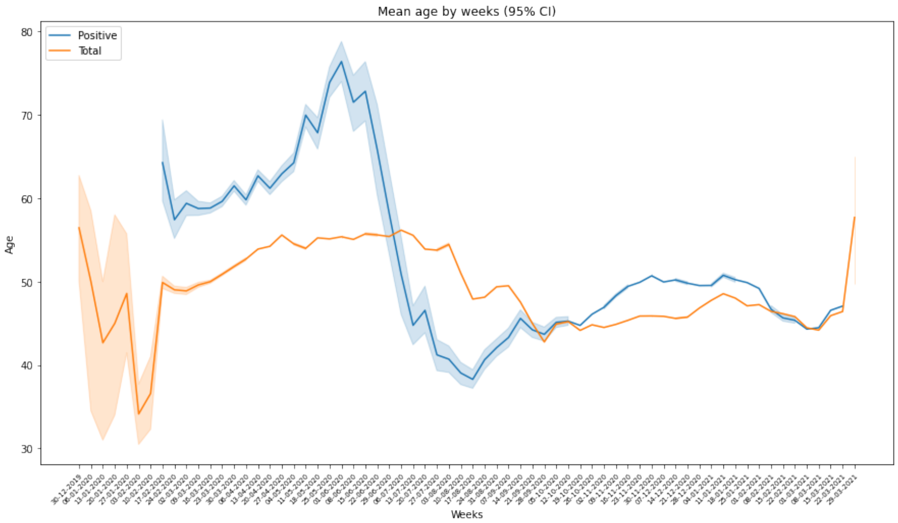

Figure 3.8: Average age of tested and positive tested individuals by week. Bounds represents punctual 95% confidence interval for the statistic.

Suspecting that age has been among the most crucial factors during the pandemic, the weekly mean ages of examined patients has been determined. From Figure 3.8 it can be observed that, in the first weeks of 2020, the average age of the population subjected to swabs shows a very high variability due to the small sample size of performed swabs in that period. Starting from March, at the outbreak beginning, when tests started to be routinely performed, there is instead a certain stability, together with a low variability, around $45 - 55$ years old.

A very different behavior is exhibited by the weekly average ages of patients tested positive to COVID-19. While in the first months of the epidemic the statistic is close to 60 year, with the introduction of containment measures and non pharmaceutical interventions (masks, distancing, etc.) it grows up to 75. Around June 2020, just after the end of the total lock-down, the average age drops abruptly till August, where the minimum value of 37 is recorded. Finally it starts increasing again getting closer and closer to the average age of the overall population subjected to swabs (please note that the observations at the very time extremes, both left and right, are of no significance as very noisy and incomplete).

Motivated by this triphasic trend, the analyses have been refined splitting the data according to three macro-intervals (for convenience, the extremes have been chosen in correspondence to beginning or end of the months):

1. from the 30th of December 2019 to the 31st May 2020;
2. from the 1st of June 2020 to the 31st August 2020;
3. from the 1st of September to the 1st April 2021.

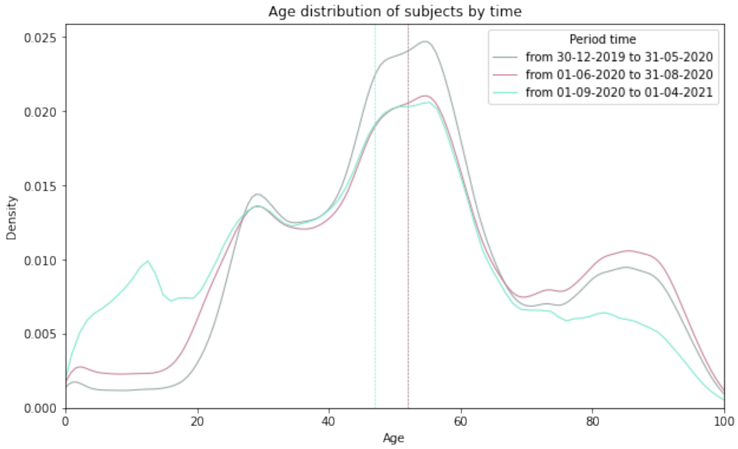

Figure 3.9: Age distribution of tested population during the three different macro-intervals.

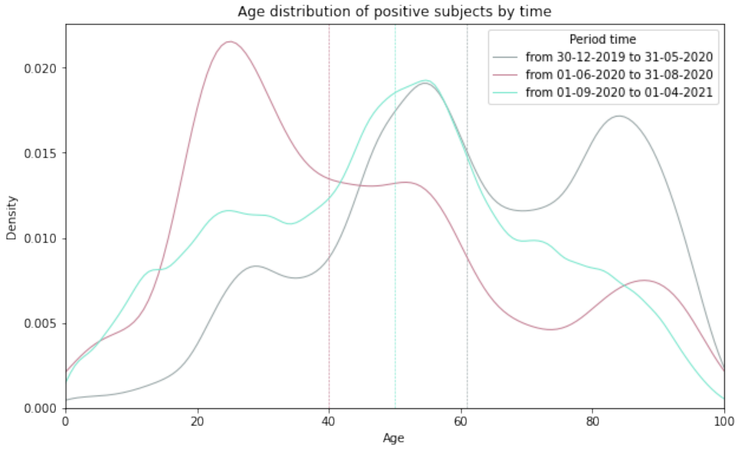

Figure 3.10: Age distribution of positive tested population during the three different macro-intervals.

The age distribution, both for all tested subjects and for only those resulting positive, has been then estimated from data via a KDE method for each of these periods. The results are shown in Figures 3.9 and 3.10 in which the median value is also indicated with a dashed vertical line. Considering the overall swabs, the centrality measures seems stable, only the tails slightly vary: the number of younger individuals subjected to swab increases over time, vice versa for those aged over 80, whose relative values decrease. On the contrary, for tests with positive outcome, the distribution seems to change more significantly over the three time intervals. A reduction in age is, particularly evident during the summer months (median 40, mode 25 years, against respectively 62 and 57 years recorded in the first peak of infected).

3.2.2 Presence of chronic pathologies

From this section onward, the described results refer to the analysis carried out only on the first database, i.e. the set of records related to positive molecular swabs. As anticipated, in this database numerous features are available with both personal and clinical information. For the sake of completeness, it should be precised that the investigations considered only the subjects, thus eliminating the possible duplicates of a person with more swabs records, with domicile or residence in a Venetian province. As a matter of fact, there are 5, 862 individuals with both residence and domicile outside the Veneto region (e.g non-EU citizen or coming from neighboring municipalities). Discarding these observation, the actual number of analyzed subjects is 322, 015.

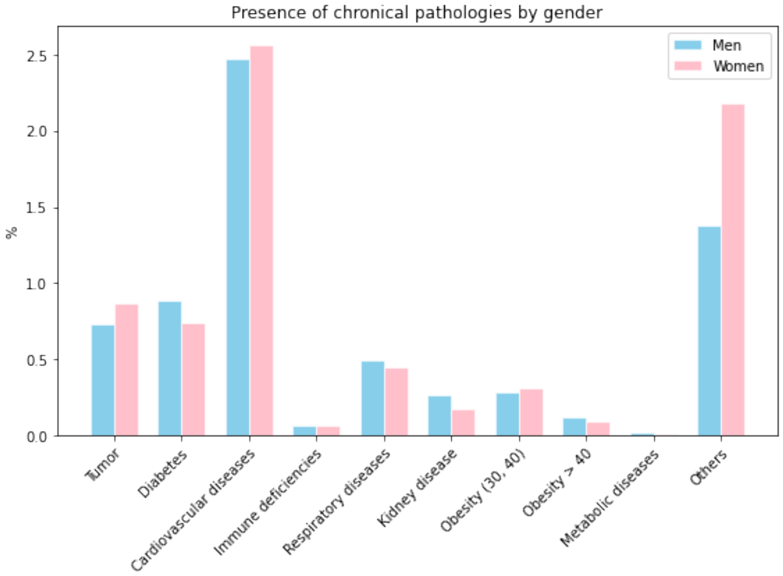

Figure 3.11: Prevalence percentages of principal chronic conditions and diseases in positives.

To start, the data relating to clinical history have been firstly taken into account, which could be interesting for a potential correlation study with hospitalizations. In detail, for each recorded patient it is known whether he/she is pathological and if suffering from certain disorders or diseases such as diabetes, immunodeficiency, obesity etc. Calculating the percentages of pathological subjects in the sample of positives it emerges that for both genders less than 10% has at least one chronic pathology.

Going deeper into the available categories of conditions, Figure 3.11 represents the percentages of subjects, again positive and diversified by gender, affected by specific pathologies. Although it would be necessary to evaluate these percentages with those of the entire population, from this graph it is possible to deduce that the most common pathologies among COVID-19 infected are cardiovascular diseases, diabetes and cancer.

The age distribution of pathological and non-pathological positives have been also compared (see Figure 3.12). As expected, in the first case the distribution is left skewed, while the second is more symmetrical. The values of the medians, represented with the vertical dashed lines, further confirm this observation.

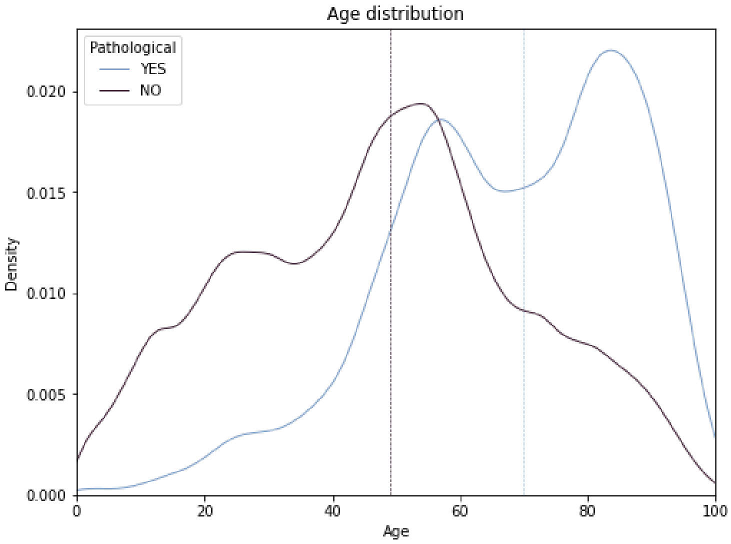

Figure 3.12: Age distribution pathological versus non-pathological positive tested population.

3.2.3 Clinical status

Subsequently, the subjects have been divided into seven macro categories according to the severity of their clinical condition during the period of COVID-19 positivity: asymptomatic, mildly

symptomatic, severe symptomatic, ordinary hospitalization, hospitalized in sub-intensive care area, hospitalized in intensive care and finally deceased. To simplify, an unique status have been hence assigned to each subject, referring only to the most critical occurred.
In Figures 3.13, 3.14 the absolute frequencies and the percentages of occurrences of such conditions are presented distinguishing by different age groups. From these graphs it can be deduced that, as already known, at younger ages (less than 25) the symptomatic rate is very low and only in rare cases the symptoms are severe or require hospitalization. Between 25 and 65 years, the number of symptomatic cases start to increase, with a consequent raise in hospitalizations, which however only rarely is of intensive type. Finally, among the over 65s, the number of hospitalizations in critical areas as well as deaths become significant, monotonically increasing as age grows.

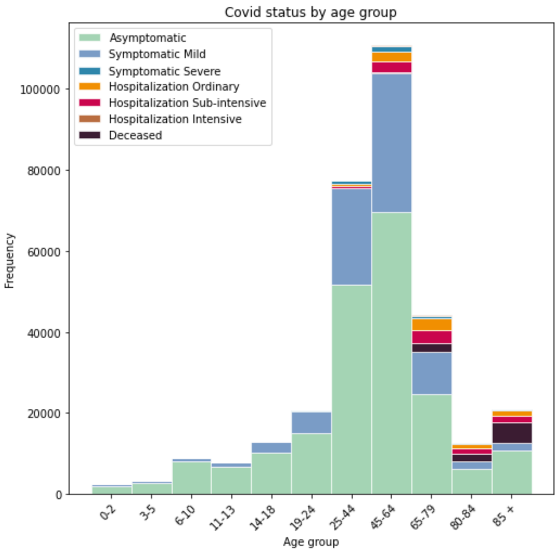

Figure 3.13: Absolute frequencies of clinical status during positivity by age class.

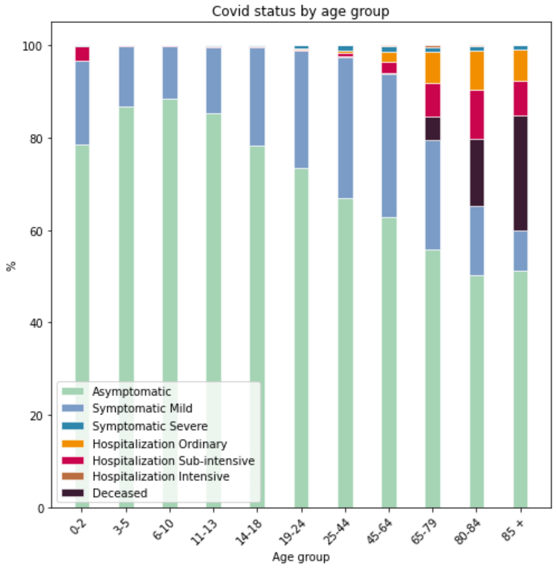

Figure 3.14: Percentages of clinical status during positivity by age class.

The histograms in Figures 3.15 and 3.16 show instead how the frequencies and percentages of the various clinical status vary over the course of the weeks, starting from the date of first confirmed Italian case (21st February 2020). In absolute terms, while hospitalizations and deaths are more stable, a significant increase in the number of asymptomatic and pauci-symptomatic patients is evident. This phenomenon could be easily explained by to the increase in the number of performed tests over time. In percentage terms, this behavior is reflected in a decrease over time both in hospitalizations, even in critical areas, and in deaths. By analyzing in more detail the trend over time of the most severe clinical status (Figure 3.17) it could be in fact observed that the number of hospitalizations remains almost constant between the first and second epidemic peak, while the number of deaths shows an increase of 100%. To have a fair comparison, recall that an increase of 500% occurred in the total number of swabs, as already shown in Figure 3.2. Finally, note once again, that in both the mentioned graphs the summer period curve (from June to August) is substantially flat, both in terms of positive swabs and hospitalizations or deaths.

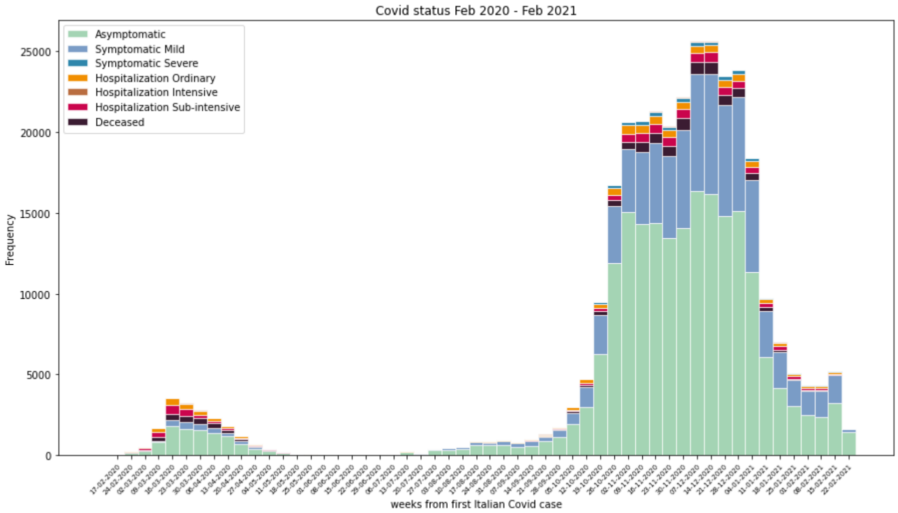

Figure 3.15: Absolute frequencies of clinical status during positivity by week.

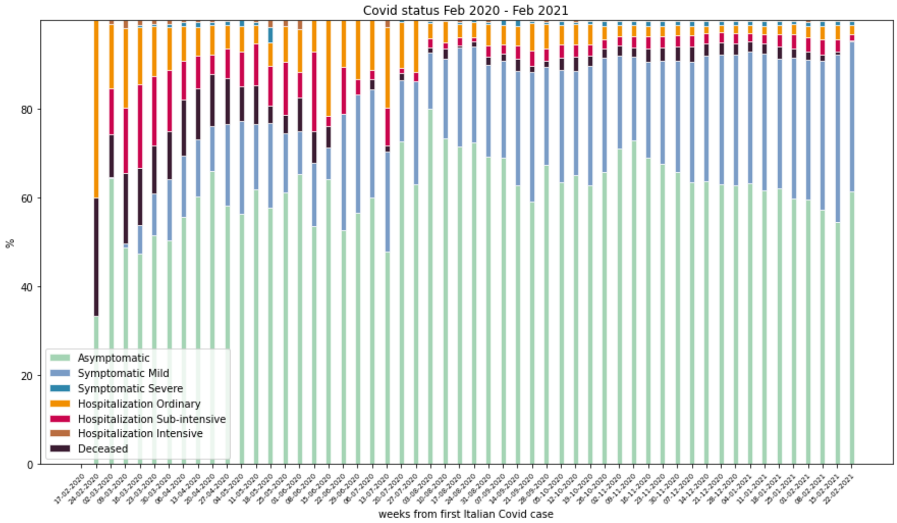

Figure 3.16: Percentages of clinical status during positivity by week.

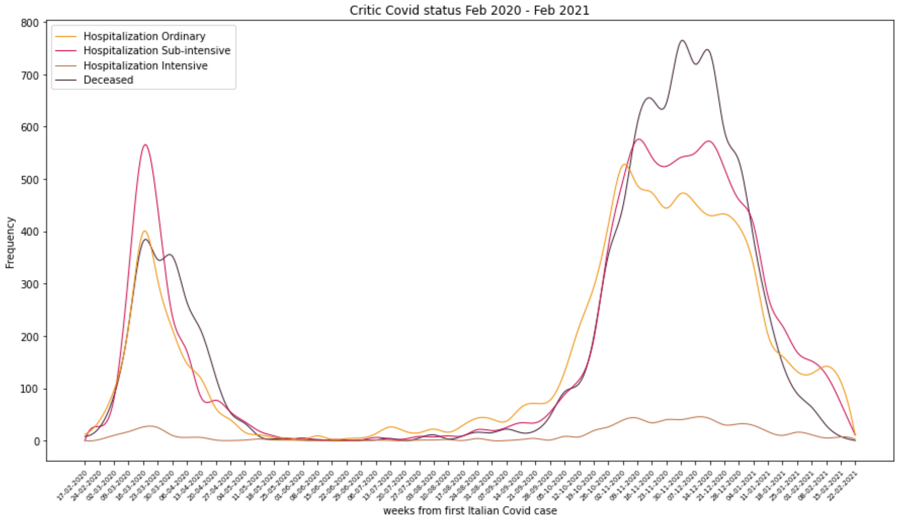

Figure 3.17: Absolute frequencies of hospitalizations and deceases by week.

3.2.4 Positivity duration

An important aspect to highlight, both from a epidemiologic and a healthcare point of view, regards the duration of the positivity interval of COVID-19 patients.

Figures 3.18 and 3.19 show, respectively, an histogram related to observed time span values and the corresponding estimated probability. Please take in mind that for this study only positive subject with at least a molecular test have been considered. Moreover the duration of positivity has been obtained as the number of days between the first swab with positive outcome and the one of first negativization.

Looking at the images it is easy to deduce that two weeks are usually sufficient to obtain a definitive negativization (median positivity period around 15 days). In addition, the associated density probability is right-skewed (skeweness $= 7.63$), with a right tail significantly heavier than the left, indicating an unneglectable number of healings much longer than the common observed.

In Figures 3.20 and 3.21 it is instead represented the positivity duration distributions as different categories vary: age and clinical state. In the first image it is evident how the positivity period length of a patient is proportionally related with his/her age, showing up to 10 days of gap between the mode values in young people (about 12 days in age class $6-10$ years) and elderlies (about 21 days for the over 85s).

Almost the same considerations could be done for clinical states: patients with important symp-

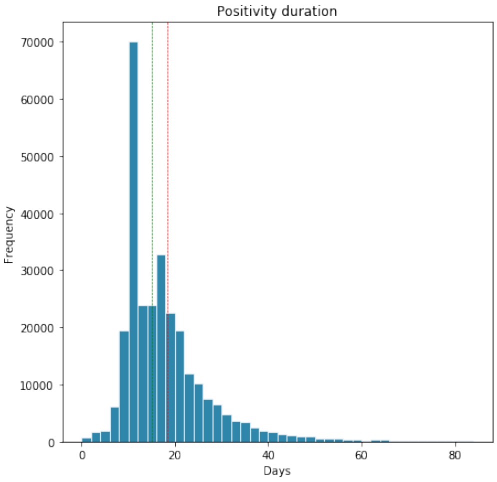

Figure 3.18: Absolute frequency histogram of COVID-19 positivity duration (in days).

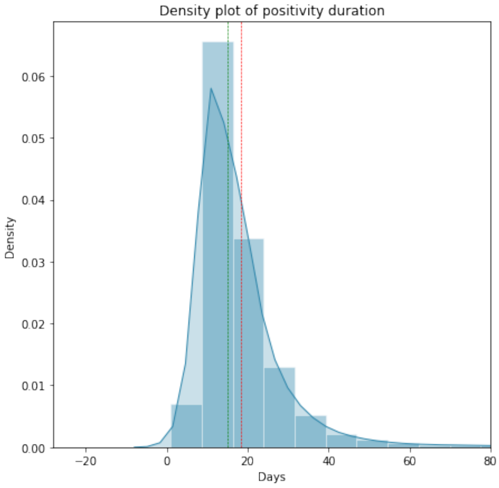

Figure 3.19: Estimated density probability of COVID-19 positivity duration (in days).

toms present a longer positivity time. As a matter of fact, the curves exhibit a decrease in skeweness as the severity of symptomatology increases, with a flattening of the median on the average. Interestingly, by considering the positivity duration trend, it seems to distinguish two clusters having almost overlapping curves: non-severe (asymptomatic, pauci-symptomatic and symptomatic patience) and severe (sub-intensive care unit, ICU and deceased). A singularity in this quite monotone behavior regards the positivity duration of dead patients. It seems in fact just slightly shorter than that for ICU subjects. A reasonable explanation could be that death factor reduces the right tail containing the longest positivity intervals.

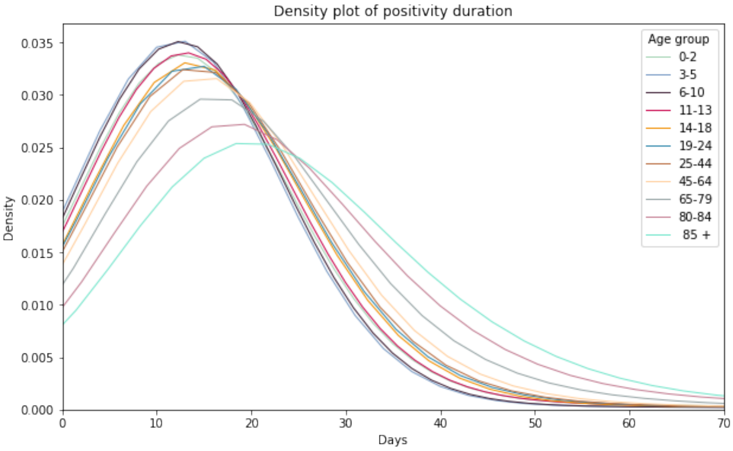

Figure 3.20: Estimated density probabilities of positivity duration (in days) for the different age classes.

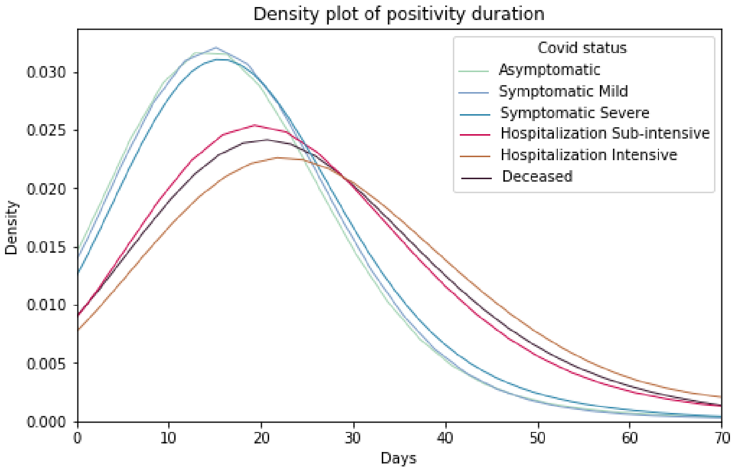

Figure 3.21: Estimated density probabilities of positivity duration (in days) for the different clinical states.

3.2.5 Association of chronic pathologies with clinical status

A further step has been the assessment of any possible relation between the severity of clinical condition and the presence of pathologies. As a matter of fact, calculating the percentages for each type of clinical status both among pathological and non-pathological subjects, a positive correlation seems to subsist (see Figure 3.22). In the barplots depicted in Figure 3.23 these percentages are further refined distinguishing by type of disease or disorder. As expected, hospitalization and death rates are higher, up to three times, for unhealthy individuals. Looking specifically at the disorder types, it can be noticed that subjects suffering from severe obesity, diabetes, metabolic or cardiovascular diseases are - in proportion - the most hospitalized in intensive or sub-intensive care areas. As concerns deaths, it is interesting to observe that mortality is below 5% in non-pathological subjects, while it is around 20% in patients with diabetes, respiratory or metabolic diseases and obesity, reaching up to 43% for people with kidney disease. To rigorously test this suspected association between pre-existing pathological conditions and greater risk of hospitalization or death, a Logistic Regression has been then performed for each of the binary variables representing the following states:

- whether the subject was generally hospitalized or not;
- whether the subject was hospitalized in sub-intensive or intensive care units or not;
- whether the subject deceased of COVID-19 infection or not.

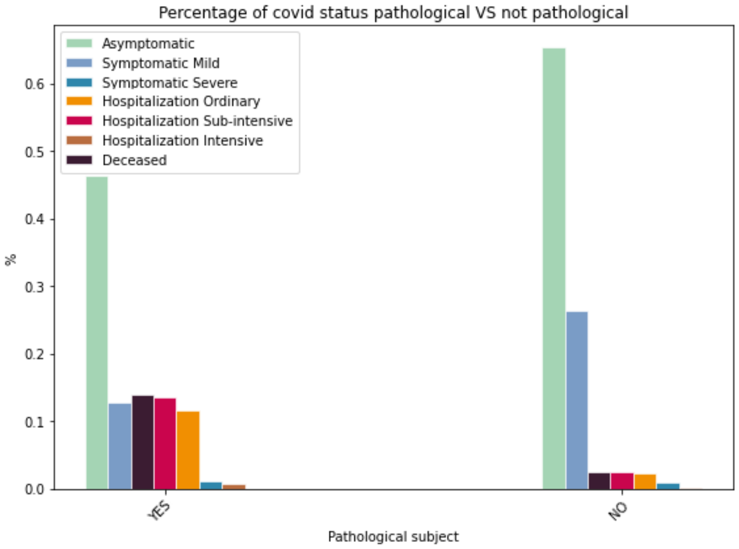

Figure 3.22: Comparison of clinical status percentages in pathological versus non-pathological subjects.

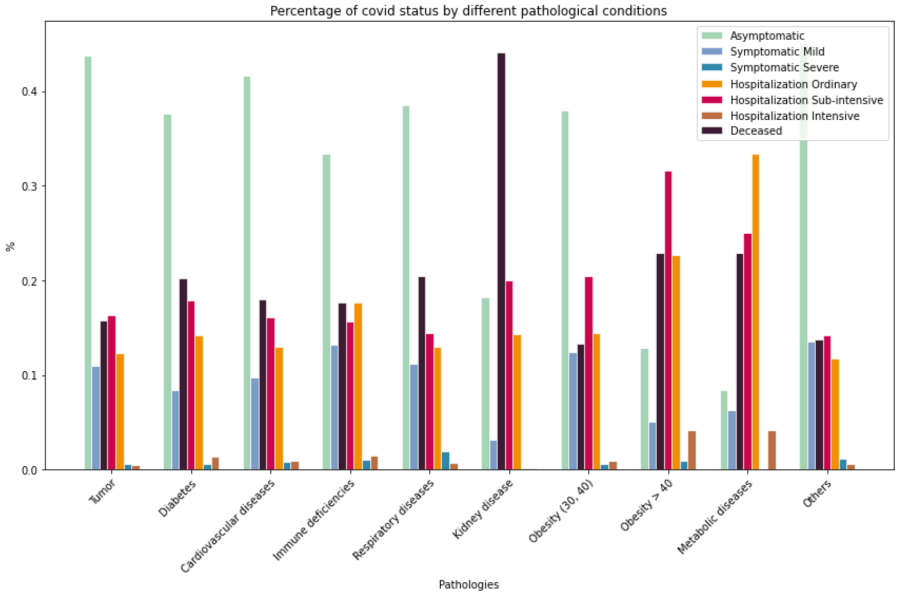

Figure 3.23: Comparison of clinical status percentages among different pre-existing pathological conditions.

The independent variables considered are: gender, age and the possible presence of (at least) one of the different pathologies annotated in the database. Tables 3.1, 3.2 and 3.3 summarize the results obtained by the three different models. In the case of hospitalization, both ordinary and in critical areas, it can be noted that the association with various pathologies is always significant (p value < 0.05) except for kidney diseases. Regarding critical hospitalization in sub-intensive or intensive areas, the correlations look very similar, however the significance in the correlation with cardiovascular or respiratory diseases and immunodeficiency is lost. Lastly, for the death variable, again almost all covariates are significant in the model, except in this case the presence of metabolic diseases and tumors.

Finally, it can be noted that, for all the fitted regressions, age and gender have very low p value, indicating a strong correlation, respectively, between having an elderly age and being male with being hospitalized or deceased due to COVID-19 infection. To conclude, a word of warning should be spent: it is important to remember that, as shown in Figure3.12, pathological subjects are more frequently older than 60 years. Moreover, the same subject can have more than one pathology. Regression, being a quite simple model, could then be partially affected by potential collinearity between the different variables, thus making the produced estimates not always reliable.

Parameter	Estimate	Std. Error	z value	$P(>\|z\|)$
(Intercept)	-6.0359813	0.0346133	-174.383	$< 2e-16$
Gender (M)	0.6675209	0.0172746	38.642	$< 2e-16$
Age	0.0472151	0.0004588	102.920	$< 2e-16$
Chronic pat. (Yes)	1.0858865	0.0529421	20.511	$< 2e-16$
Tumor (Yes)	0.1137400	0.0552832	2.057	0.039647
Diabetes (Yes)	0.2049454	0.0533589	3.841	0.000123
Cardiovascular dis. (Yes)	-0.1534317	0.0482027	-3.183	0.001457
Immune def. (Yes)	0.8049120	0.1714105	4.696	$2.66e-06$
Respiratory dis. (Yes)	0.2258442	0.0709723	3.182	0.001462
Kidney dis. (Yes)	-0.0794344	0.0935577	-0.849	0.395858
Metabolic dis. (Yes)	0.4334227	0.0799977	5.418	$6.03e-08$
Obesity BMI $30-40$ (Yes)	1.6877930	0.1267510	13.316	$< 2e-16$
Obesity BMI > 40 (Yes)	1.7627082	0.3543682	4.974	$6.55e-07$
Others (Yes)	0.2998589	0.0462308	6.486	$8.81e-11$

Table 3.1: Logistic Regression coefficients estimates for generic hospitalization.

Parameter	Estimate	Std. Error	z value	$P(>\|z\|)$
(Intercept)	-6.4458454	0.0450921	-142.948	$< 2e - 16$
Gender (M)	0.6033411	0.0226399	26.649	$< 2e - 16$
Age	0.0436977	0.0005985	73.010	$< 2e - 16$
Chronic pat. (Yes)	0.9202806	0.0657675	13.993	$< 2e - 16$
Tumor (Yes)	0.1600888	0.0663979	2.411	0.01591
Diabetes (Yes)	0.2053499	0.0634610	3.236	0.00121
Cardiovascular dis. (Yes)	0.0189364	0.0600309	0.315	0.75243
Immune def. (Yes)	0.3473677	0.2143526	1.621	0.10512
Respiratory dis. (Yes)	0.0883396	0.0859035	1.028	0.30378
Kidney dis. (Yes)	-0.0713022	0.1098011	-0.649	0.51610
Metabolic dis. (Yes)	0.4425574	0.0917432	4.824	$1.41e - 06$
Obesity BMI $30 - 40$ (Yes)	1.3633186	0.1319264	10.334	$< 2e - 16$
Obesity BMI > 40 (Yes)	0.8960875	0.3663225	2.446	0.01444
Others (Yes)	0.2349282	0.0556509	4.221	$2.43e - 05$

Table 3.2: Logistic Regression coefficients estimates for sub-intensive or intensive care units hospitalization.

Parameter	Estimate	Std. Error	z value	$P(>\|z\|)$
(Intercept)	-12.644054	0.097843	-129.228	$< 2e - 16$
Gender (M)	0.876102	0.025678	34.119	$< 2e - 16$
Age	0.123706	0.001126	109.840	$< 2e - 16$
Chronic pat. (Yes)	0.484169	0.079416	6.097	$1.08e - 09$
Tumor (Yes)	0.094009	0.074613	1.260	0.207688
Diabetes (Yes)	0.263789	0.070051	3.766	0.000166
Cardiovascular dis. (Yes)	-0.159370	0.072477	-2.199	0.027884
Immune def. (Yes)	0.939991	0.242704	3.873	0.000108
Respiratory dis. (Yes)	0.748789	0.090565	8.268	$< 2e - 16$
Kidney dis. (Yes)	1.284249	0.102791	12.494	$< 2e - 16$
Metabolic dis. (Yes)	-0.185186	0.121107	-1.529	0.126238
Obesity BMI $30 - 40$ (Yes)	1.166812	0.171859	6.789	$1.13e - 11$
Obesity BMI > 40 (Yes)	1.103265	0.441432	2.499	0.012444
Others (Yes)	0.179513	0.062880	2.855	0.004306

Table 3.3: Logistic Regression coefficients estimates for decease.

3.2.6 Contacts and chains of infection

Within the first database, the one relating to individuals tested positive to a molecular swab, there are two potentially interesting features: the subject identifier code of the hypothetical in-

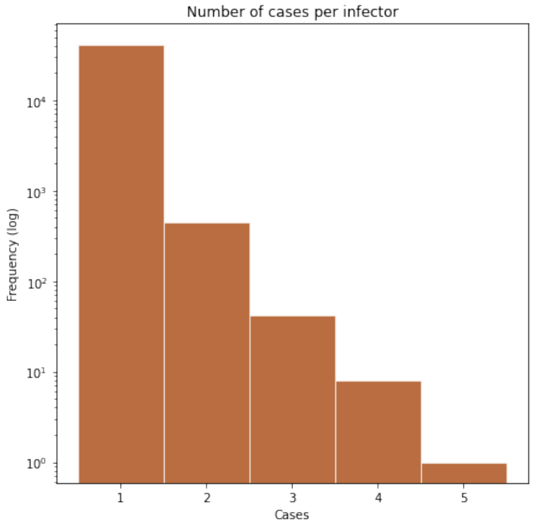

Figure 3.24: Log-scale occurrences of number of generated cases by same infector.

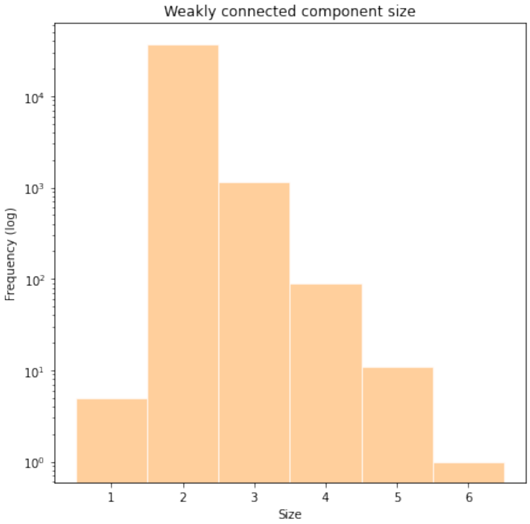

Figure 3.25: Log-scale occurrences of weakly connected component size in the reconstructed directed graph of infection.

fector and the nature of this infection contact, i.e. whether infected and infector are co-housing or not. The usability of this information is however severely limited by two factors. Firstly, a very low percentage of the entire records has at least one of these two fields filled in (only $42,036$ out of $322,015$ rows), secondly, within these small set, there are $18,773$ records reporting an infector code which cannot be traced back to one of subjects at disposal in the database.
Despite the weakness of this information, these contact tracking data have been anyway explored in order to save as most as possible useful notions on infection transmission.
To begin, the number of cases generated by each of the annotated infector has been calculated (the histogram in Figure 3.24 shows the log-frequencies). Most of the individuals appear to have caused at most two infections, coherently with the historically estimated values of the effective reproduction number $\mathcal{R}_t$, although there are also rarer cases of more prolific spreaders with up to five positivized contacts.
With the available information, the chains of infection have been then reconstructed. Practically speaking, this has been done building a directed graph, where nodes represent the positive individuals, and directed edges are placed to symbolize the causal relation between infector and infected nodes. To detect possible correlations, paths or clusters, the weekly connected components have been then determined. Figure 3.25 shows the histogram, again in log-scale, relating to the size of the infection chains that can be reconstructed from the graph as connected components. As it can be seen, the results are unfortunately not very significant. There are few

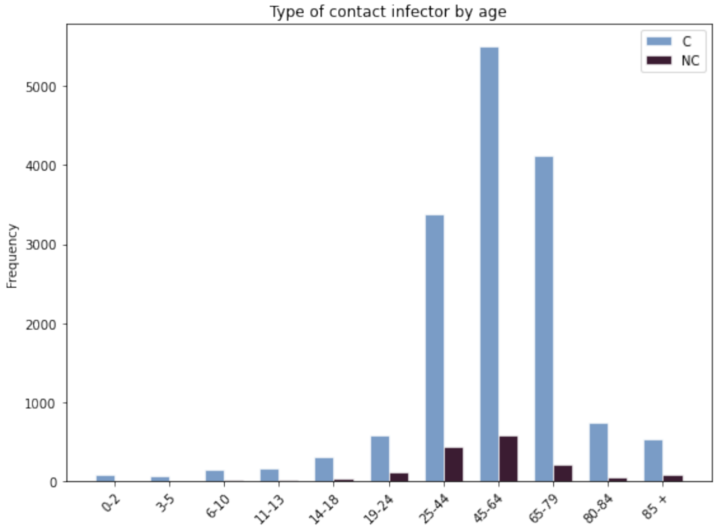

Figure 3.26: Absolute frequencies of infector by age group and contact type.

chains of unitary size, relating to patients who are annotated as infectors of themselves (clearly errors due to manual filling). The majority of chains are of size 2 or 3, although in rare cases it is possible to reconstruct a path of length 6.

Considering the type of contact, from Figure 3.26, it can be seen that most of the infections were caused by cohabiting infectors aged between 25 and 79 years.

To better examine how the ages of the infector-infected contacts are distributed, the matrices of occurrences have been calculated. In detail, each element $M(i, j)$ of each table corresponds to the number of contacts between an infector belonging to age group i and an infected subject in age group j. The results are represented to the left of the following page: in Figure 3.27 the total occurrences in all contacts, in Figure 3.29 those between cohabitants only and finally in Figure 3.31 the non cohabitants.

Such matrices are useful to determine the quality and quantity of annotations on the tracking, however, they are difficult to interpret as they depend on the size of the subpopulation which is not uniform among the different age categories. An attempt of normalization is given in Figures 3.28, 3.30 and 3.32. Mathematically, each element $M(i, j)$ has been divided by the area of the corresponding rectangle (i, j), i.e. the product between the number of infected subjects in range i and that of infected subjects in range j. Then each row has been divided by the sum of all its elements, so as to make the matrix stochastic and therefore interpretable in probabilistic terms. It is important to specify that while the couples are counted allowing repetitions, the subjects are instead considered unique, trying to model the infected-infected population in a more precise way.

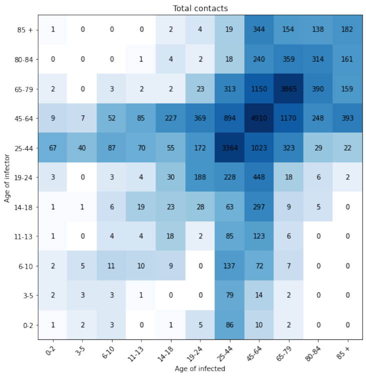

Figure 3.27: Absolute frequencies in total contacts.

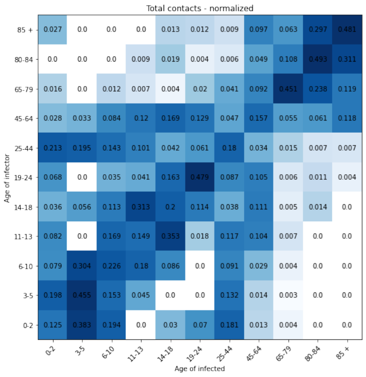

Figure 3.28: Normalized frequencies in total contacts.

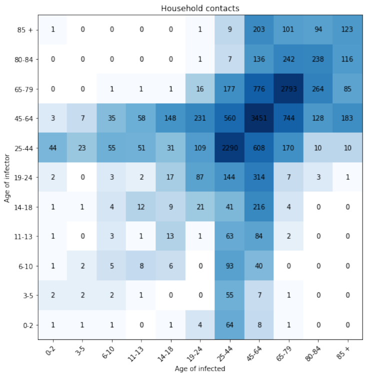

Figure 3.29: Absolute frequencies in household contacts.

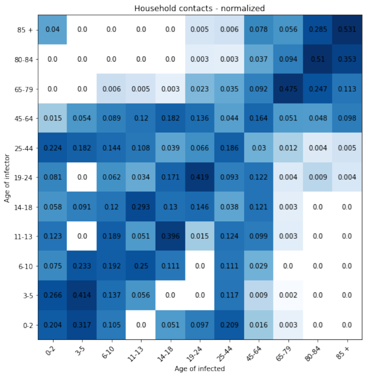

Figure 3.30: Normalized frequencies in household contacts.

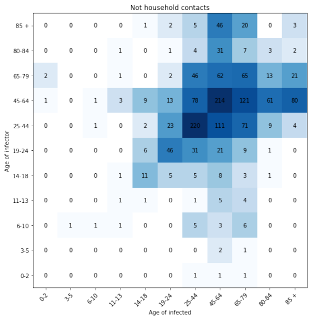

Figure 3.31: Absolute frequencies in non-household contacts.

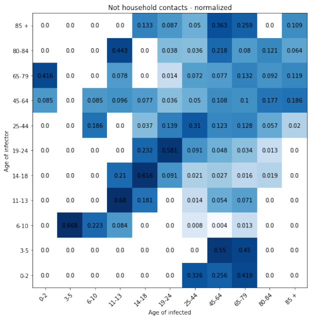

Figure 3.32: Normalized frequencies in non-household contacts.

It is interesting to note that in all cases the matrices exhibit a greater number of contacts in the diagonal, which represents relationships between individuals belonging to the same age group. This behaviour is particularly marked in the household contacts, where plausibly it reflects interactions between siblings of close ages or couples. In this matrix it can be also seen that larger values are even present in the off-diagonal, in correspondence to parent-child couples (e.g. $25-44$ with $0-2$). As concern non-cohabitants contacts, the matrix presents the most important values especially in correspondence of the first entries of the diagonal, probably representing infection transmission between schoolmates or work colleagues, while as the age increases peer contacts seem to decrease.

In general, the obtained contact maps patterns quite resemble those already existing in literature, such as the average daily contacts matrices estimated for Italy in the POLYMOD data [42], nevertheless, it will be necessary to operate carefully when using those produced by the Venetian data as for some ages couples the sample size may be not sufficient to give meaningful results.

3.2.7 Incidence

The last quantity which has been extrapolated and analyzed by positive patients data is the weekly incidence of COVID-19. In order to compare the epidemic trend between different geographical areas and various population subgroups, such as students, workers, etc. the calculation of incidence has been repeated for all the Venetian provinces or municipalities as well as for each age class. In compact notation, for each Venetian territory i, for each age group j and for each week k, the incidence has been calculated as the quantity

$$I(i,j,k) := \frac{NC(i,j,k)\cdot 10^5}{P(i,j)} \tag{3.1}$$

where $NC(i,j,k)$ is the number of new positive cases counted in week k with ages among class j and domiciled in i, while $P(i,j)$ is is the total number of inhabitants belonging to the age group j and domiciled in province/municipality i (demographic data from ISTAT census 2020). The scaling to 100 thousands has been instead chosen since it is the commonly adopted reference number.

The graphs depicted in Figure 3.33 represents the incidence trends in each age class comparing by provinces. For the 563 municipalities instead, the total incidence has been instead summarized in an animated cloropleth map. Four instantaneous examples are given in Figure 3.34.

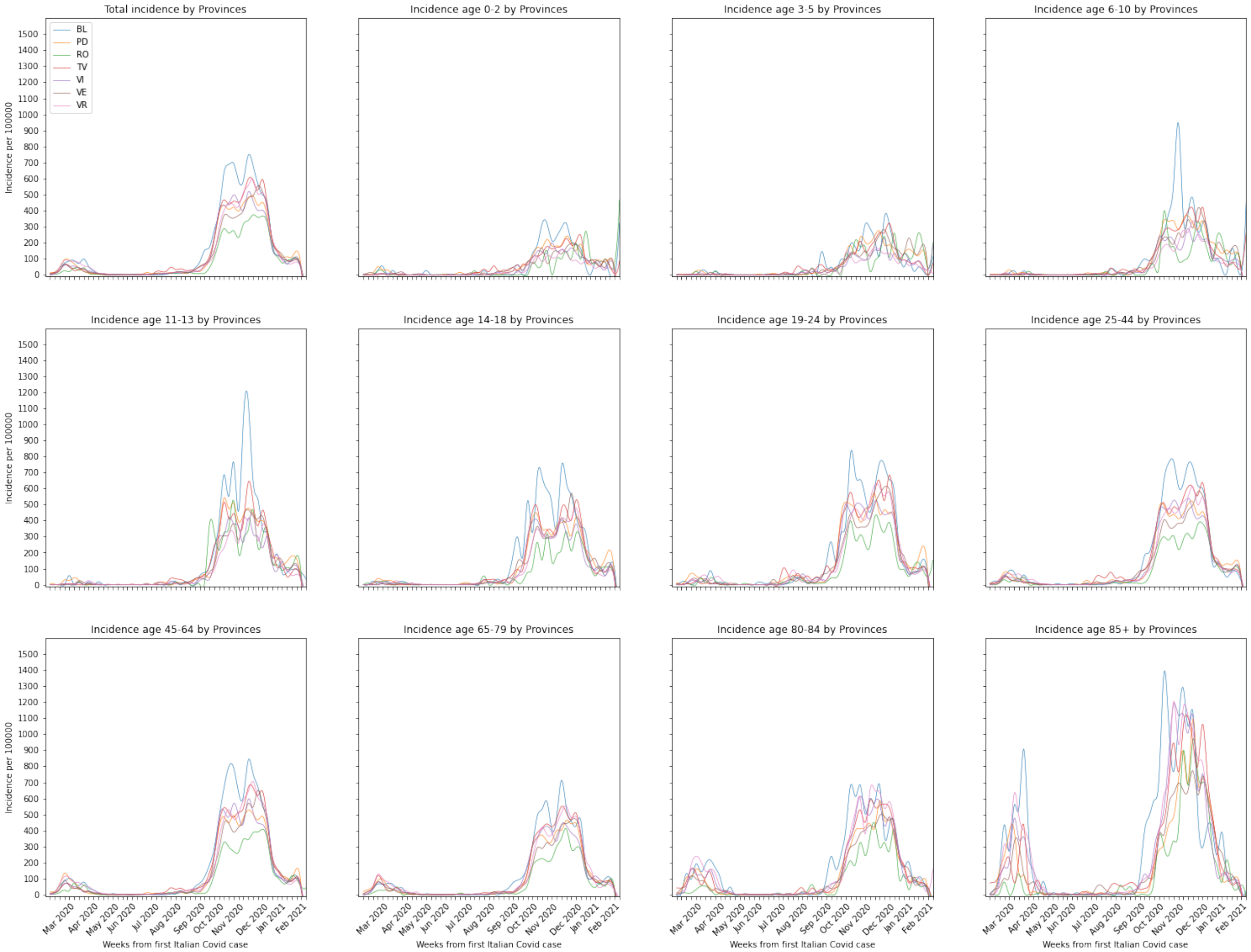

Figure 3.33: Weekly incidence of COVID-19 in the Venetian provinces for each age group.

The incidence measure confirms many deductions already presented above, such as the presence of two separate peaks or that elderlies were initially most affected by the infection. However, it also points out the importance, not only of the age factor, but also of the possible local territorial differences in the spread of the epidemic due to geophysical condition, like population density, or socio-cultural aspects, in particular in the initial or intermediate phases of the epidemic.

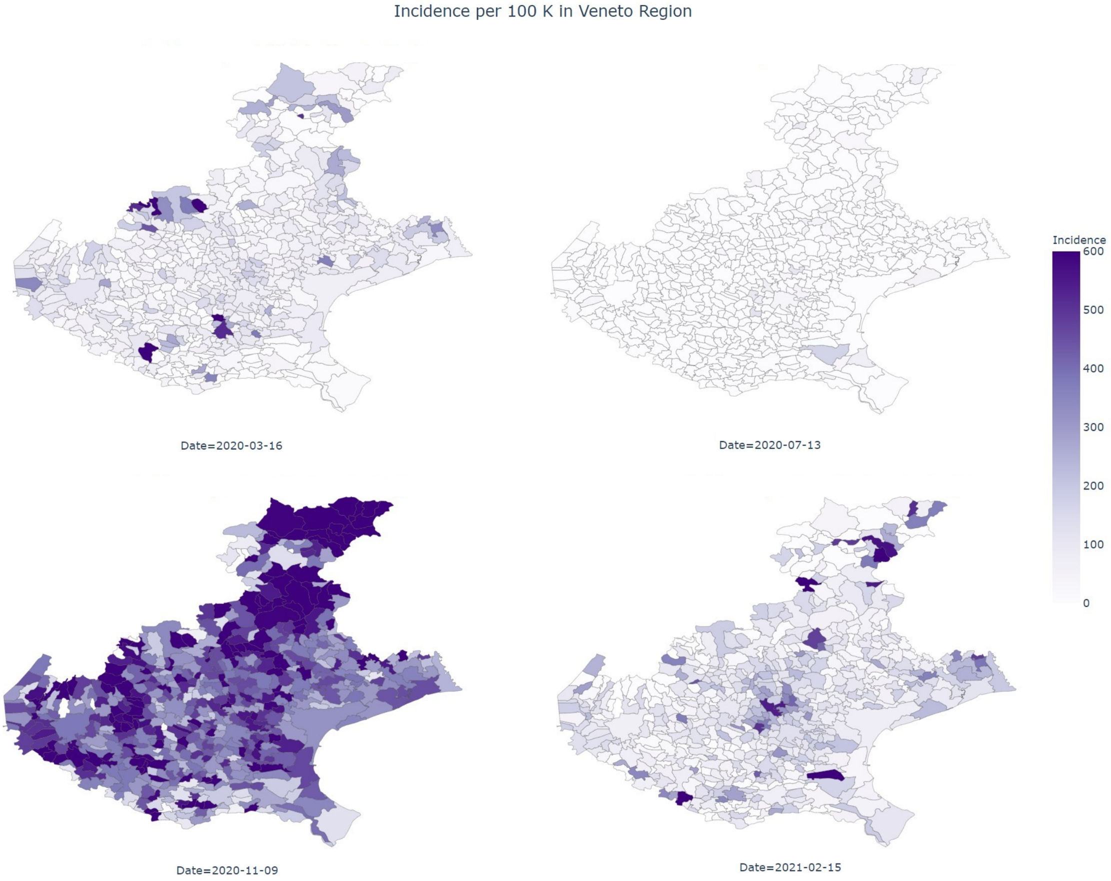

Figure 3.34: Incidence in the Venetian municipalities in four different representative time instants of COVID-19 epidemic.

4
Methods

In this chapter the implemented methodologies will be accurately described. The main focus are the designed epidemiological models, as the very purpose of this work. Details will be also given regarding the adopted strategies for the parameters fitting as well for the validation on the Veneto region case.

4.1 Deterministic Model

To begin the experimentation, a deterministic Equation-based model (DM) have been firstly developed as a generalization of the standard SIR. In this setting no stochasticity is allowed in order to have a single direct result, avoiding the necessity to repeat the simulations multiple times to have a summarizing behaviour. Indeed, the purpose and motivations of this type of model in this study are several. First of all, have a baseline in order to both validate the compartmental scheme and have a comparison reference in performances as well as starting estimates for parameters when further more complex models will be developed. In addition, its simplicity of implementation and its short execution time, thanks to neglectable computational burden, make it very suitable for an initial experimental phase.

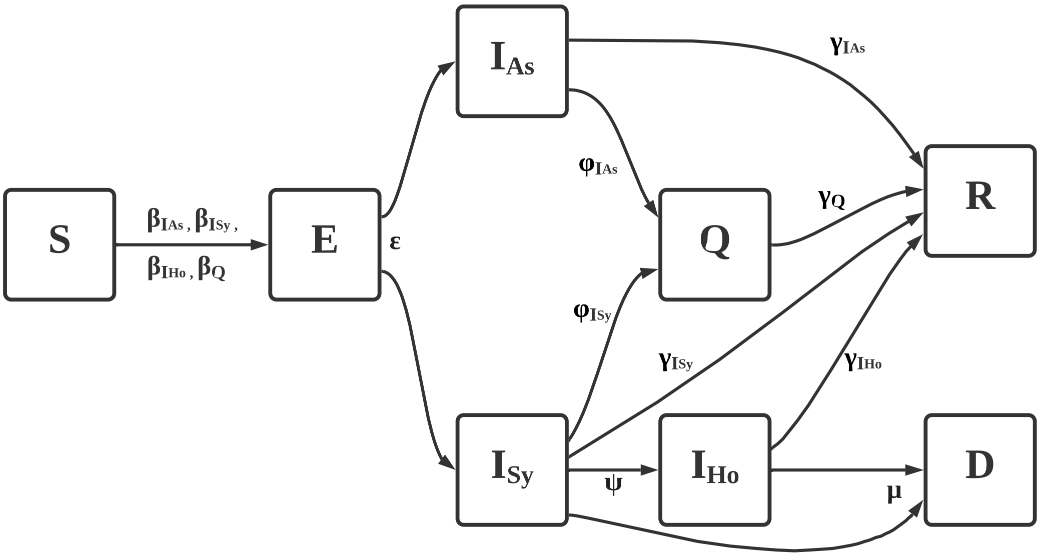

Figure 4.1: SEIQRD compartmental model: Susceptible (S), Exposed (E), Infected Asymptomatic (I_{As}), Infected Symptomatic (I_{Sy}), Infected Hospitalized (I_{Ho}), Quarantined (Q), Removed (R), Deceased (D).

4.1.1 Compartmental formulation

Let's hence describe the included compartments and the represented dynamics. As usual, S denotes the class of susceptible, healthy individuals with no immunity and that can contract the virus if exposed to infected person and E represents exposed but not yet contagious subjects. Once the incubation phase ends, exposed individuals become infectious with rate ϵ and could either develop symptoms (I_{Sy}) or be asymptomatic (I_{As}) according to probabilities P_{As} and P_{Sy}. More severe symptomatic patients could eventually be hospitalized with probability P_{Ho}, moving to I_{Ho} with rate ψ. Critical symptomatics could die by the viral disease both when hospitalized or not respectively with probabilities $P_{D|Ho}$ and $P_{D|Sy}$, moving to D with fatality rate μ. Infected asymptomatic and infected symptomatic not hospitalized could be uncovered and home quarantined with probabilities $P_{Q|As}$ and $P_{Q|Sy}$, moving to Q with rates ϕ_{As}, ϕ_{Sy}. Individuals in I_{As}, I_{Sy}, I_{Ho}, Q all contribute to the infection dynamics with differently scaled transmission rates β_{As}, β_{Sy}, β_{Ho}, β_Q. Finally, infected subjects move to R when they recover, when they are discharged from the hospital or when their quarantine period ends with recovery rates γ_{As}, γ_{Sy}, γ_{Ho}, γ_Q.

Note that, here it has been assumed that infected hospitalized once discharged are certainly also healed and no longer contagious, therefore they do not need any further home quarantine

period. The reason for this choice has been to make the roles of Q and I_{Ho} almost equivalent from an epidemiological point of view: in both cases the subjects are certainly tested, traced by the health system and can potentially infect only a very limited number of contacts. This decision has allowed to simplify some necessary transitions.
The formal mathematics of the DM is shown in System 4.1.1, while a graphical schema of the compartments is represented in Figure 4.1. A summary of all the parameters is also given in Table 4.1.

System 4.1.1

$$
\begin{cases}
\frac{dS}{dt} = -\frac{S}{N}(\beta_{As}I_{As} + \beta_{Sy}I_{Sy} + \beta_{Ho}I_{Ho} + \beta_{Q}Q) \\
\frac{dE}{dt} = \frac{S}{N}(\beta_{As}I_{As} + \beta_{Sy}I_{Sy} + \beta_{Ho}I_{Ho} + \beta_{Q}Q) - (P_{As} + P_{Sy})\epsilon E \\
\frac{dI_{As}}{dt} = P_{As}\epsilon E - (1 - P_{Q|As})\gamma_{As}I_{As} - P_{Q|As}\phi_{As}I_{As} \\
\frac{dI_{Sy}}{dt} = P_{Sy}\epsilon E - (1 - P_{Q|Sy} - P_{D|Sy} - P_{Ho})\gamma_{Sy}I_{Sy} - P_{D|Sy}\mu I_{Sy} - P_{Q|Sy}\phi_{Sy}I_{Sy} - P_{Ho}\psi I_{Sy} \\
\frac{dI_{Ho}}{dt} = P_{Ho}\psi I_{Sy} - (1 - P_{D|Ho})\gamma_{Ho}I_{Ho} - P_{D|Ho}\mu I_{Ho} \\
\frac{dQ}{dt} = P_{Q|As}\phi_{As}I_{As} + P_{Q|Sy}\phi_{Sy}I_{Sy} - \gamma_{Q}Q \\
\frac{dR}{dt} = (1 - P_{Q|As})\gamma_{As}I_{As} + (1 - P_{D|Sy} - P_{Q|Sy} - P_{Ho})\gamma_{Sy}I_{Sy} + (1 - P_{D|Ho})\gamma_{Ho}I_{Ho} + \gamma_{Q}Q \\
\frac{dD}{dt} = \mu(P_{D|Sy}I_{Sy} + P_{D|Ho}I_{Ho})
\end{cases}
$$

It is easy to understand that, by definition, this SEIQRD model suffers from all the main limitations of a simple Equation-based model, such as closed population assumption, no vital dynamics and above all homogeneous mixing. However, its number of parameters and transition edges should ensure some flexibility. Pleas note that all the splitting edges are parametrized with probabilities summing up to 1 in order to mathematically represent disjoint possibilities as XOR split.
Once defined the system of ODEs representing the transition flows from each compartment, two steps have been necessary to have a practically usable model: find valid parameters values and obtain the solution.
For the latter task, the `odeint` function from the Python `scipy.integrate` library has been exploited. This module provides a pre-implemented way to numerically solve stiff or non-stiff systems of first-order ordinary differential equations using `lsoda` from the FORTRAN library `odepack`. In particular, it returns the solution for the initial value problem given the initial conditions. In the case of interest, aiming to model the epidemic in Veneto starting from the very

beginning, while it was reasonable to set $I_{Ho\,0}$, Q_0, R_0, D_0, respectively the initial number of hospitalized, quarantined, removed and deceased to zero, the real starting sizes for E_0, $I_{As\,0}$ and $I_{Sy\,0}$ were unknown. For that reason, these quantities have been treated ad hyperparameters during the training phase. Regarding N, the population size, it has been approximated to $4,900,000$ individuals (source ISTAT census).

4.1.2 Parameters estimation from available data

From the investigations carried out so far on the data provided by the Veneto region, it has been clear that part of the parameters necessary to the model could be potentially directly observable from the data itself.

In particular, as seen in the previous chapter, features from more than three hundred thousand patients tested positive to molecular tests for SARS-CoV-2 are available for this study. Therefore, taking advantage of the information transmitted by the regional Health System regarding the clinical status during positivity, the dates of the last positive and the first negative swab, any dates of hospitalization, discharge or death, as well as dates of start and end of quarantine, it has been possible to explicitly estimate the value of the following parameters: P_{As}, P_{Sy}, P_{Ho}, $P_{D|Sy}$, $P_{D|Ho}$, γ_{As}, γ_{Sy}, γ_{Ho}.

For γ_Q, the recovery rate from the home isolated compartment, on the other hand, it has been assumed that it is constantly equal to the reciprocal of the minimum quarantine period, which by law was almost always 14 days during the first period of the pandemic.

Table 4.2 shows the obtained statistics and, where meaningful, the computed 95% confidence intervals.

4.1.3 Parameters fitting

For the remaining unknown parameters, which are the most interesting and challenging, a training strategy has been adopted.

Time dependence

The first important factor to take into consideration has been the time dependence and variability of the parameters values. In Figure 4.2 an example of solution with fixed parameters is depicted. It can be seen that the produced solution returns functions of two types: monotonic “s”-shaped curves or bell shaped curves. Even testing different combinations of initial conditions, it has been clear that keeping all parameters constant over time, it would have been

Parameter	Description
β_{As}	Transmission rate in infected asymptomatic
β_{Sy}	Transmission rate in infected symptomatic
β_{Ho}	Transmission rate in infected hospitalized
β_Q	Transmission rate in home quarantined
ϵ	Inverse of latent, incubation period duration
ϕ_{As}	Quarantining rate for infected asymptomatic
ϕ_{Sy}	Quarantining rate for infected symptomatic not hospitalized
ψ	Hospitalization rate for infected symptomatic
γ_{As}	Recovery rate in infected asymptomatic
γ_{Sy}	Recovery rate in infected symptomatic not hospitalized
γ_{Ho}	Recovery rate in infected hospitalized
γ_Q	Inverse of quarantine period duration
μ	Fatality rate
P_{As}	Probability to be infected asymptomatic ($P(I_{As}\|I)$)
P_{Sy}	Probability to be infected symptomatic ($1 - P_{As}$)
P_{Ho}	Probability to be hospitalized when infected symptomatic ($P(I_{Ho}\|I_{Sy})$)
$P_{Q\|As}$	Probability to be home quarantined when infected asymptomatic ($P(Q\|I_{As})$)
$P_{Q\|Sy}$	Probability to be home quarantined when infected symptomatic not hospitalized ($P(Q\|I_{Sy} - I_{Ho})$)
$P_{D\|Sy}$	Probability to decease when infected symptomatic not hospitalized ($P(D\|I_{Sy} - I_{Ho})$)
$P_{D\|Ho}$	Probability to decease when infected hospitalized ($P(D\|I_{Ho})$)

Table 4.1: Deterministic model parameters description.

Parameter	Estimate	95% CI
P_{As}	0.645	$(0.6435, 0.6468)$
P_{Sy}	0.355	$(0.3531, 0.3565)$
P_{Ho}	0.223	$(0.2203, 0.2252)$
$P_{D\|Sy}$	0.024	$(0.0237, 0.0255)$
$P_{D\|Ho}$	0.271	$(0.2655, 0.2764)$
$1/\gamma_{As}$	17.859 days	$(17.788, 17.931)$
$1/\gamma_{Sy}$	17.762 days	$(17.674, 17.850)$
$1/\gamma_{Ho}$	26.094 days	$(25.758, 26.429)$
$1/\gamma_{Q}$	14 days	-

Table 4.2: Known or observable parameters in Veneto population data.

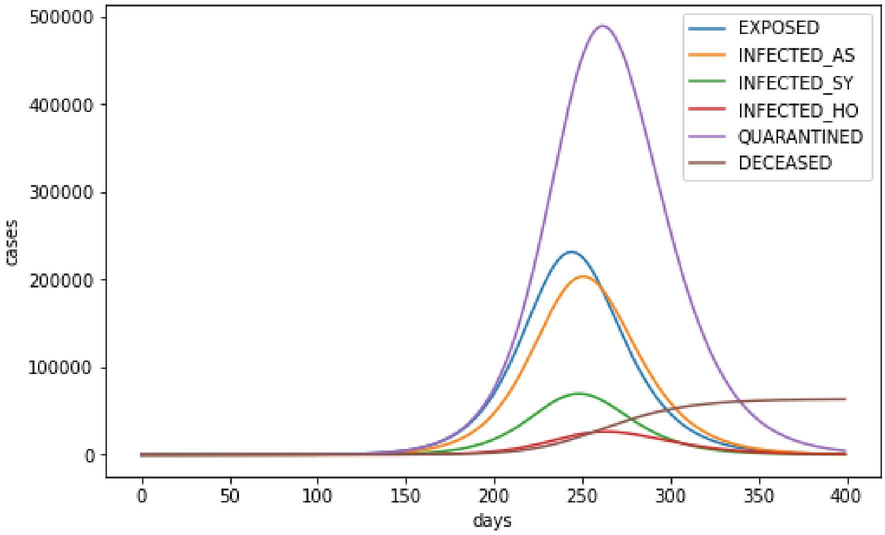

Figure 4.2: Example of system solution with constant parameters.

impossible to obtain behaviors with two peaks like those observed in Veneto (see Chapter 3). As expected then, to have reliable estimates, the fitting should have been done separately on different time intervals. At this point the choice of periods has been crucial for the success and meaningfulness of the results. Even if a trial and error approach has been followed, these decisions have been mostly guided by knowledge on the available time-series and historical facts, such as the presence of peaks or the activation of specific containment measures.

Training data and performance metric

The other fundamental ingredients for the fitting have been, as usual, a performance measure and a training set. It has been therefore necessary to understand which of the various information from the provided databases were the most decisive and, above all, the most reliable. As already discussed in the previous Chapter, the records of positive tested cases seem incomplete, especially at the beginning of the pandemic, when the made swab exams were fewer compared to later stages. It is therefore not certain that the reported values are the actual ones, in particular for the asymptomatic infected who are the most difficult to detect. On the other hand, the observations reported for hospitalizations, quarantined and deaths due to, or with, the novel Coronavirus disease can be considered more correct. It is obvious instead that the numbers of exposed persons are impossible to observe. As a matter of fact, by definition an exposed subject is not yet infected, hence it does not show any symptom and would result negative to test, being impossible to be traced by the health mechanism.

It has been hence decided to exploit as training data the daily observation of hospitalized, deceased, quarantined and all the positive subjects, respectively represented by symbols I_{Ho}, D, Q and P. In the case of model predictions, this quantity of positive tested has been approximated as the sum of all infected and quarantined individuals.

Concerning the performance metric, given a time interval $[t_i, t_f]$, it has been defined by an *ad hoc* Mean Weighted Squared Relative Error as follows:

$$\text{MWSRE}(t_i, t_f) = \sum_{c\,\in\{I_{Ho},D,Q,P\}} \left(\omega(c) \cdot \frac{\sum_{t=t_i}^{t_f} \left(\frac{\bar{y}_c(t,x)-y_c(t)}{y_c(t)}\right)^2 \delta(\bar{y}_c(t,x) - y_c(t))}{(t_i - t_f) \sum_{t=t_i}^{t_f} \delta(\bar{y}_c(t,x) - y_c(t))} \right)$$

where $y_c(t)$ and $\bar{y}_c(t,x)$ respectively denote real and predicted values, depending on trainable parameters x, for the size of class c at time $t \in [t_i, t_f]$. Note that, each compartment c taken into account has been weighted by $\omega(c)$ in order to give more importance to errors made on the number of deceases and infected hospitalized, which are more truthful. On the other hand, the choice of the relative error has been made to eliminate the differences in the orders of magnitude between the different types of compartment sizes. In addition, a second weighting component δ is included to encourage the model to overestimate rather than underestimate. In practice δ has been defined as a step function similar to $\text{sign}(\cdot)$. Again, all of these weights has been decided as hyperparameters, prior to the actual fitting.

Optimization algorithm

The last choice has regarded the optimization algorithm to perform the training in practice. Also for this task a pre-implemented Python module has been used: `scipy.optimize`. Note that, since the parameters are rates or probabilities, i.e. with support in $[0, 1]$, it has been necessary to use a method to minimize a multivariate objective subject to constraints. Technically speaking, the function `minimize` with argument `method` set to `'trust-constr'` has been exploited. In short, it defines a Trust Region method for constrained optimization by including a barrier penalty to encourage solution points in the desired support space, please refer to [17] for more details.

4.2 Network Agent Based Model

COVID-19 strength of transmission, similarly to other airborne diseases, radically changes depending on the context environment, the duration and type of contact. Consequently it could be incorrect to assume that contagion is equally likely between family contacts, work, school or other community spaces.

It is also worthwhile to remember the role of super-spreaders and super-spreader events resulting in clusters of cases, as happened in the South-Korean church in the early 2020. They constitute a concrete major risk factor for epidemic spread.

Furthermore, from the medical knowledge acquired so far, as well as from the analyses on the Veneto region here produced, it is clear that the severity of the disease, the risk of critical hospitalization or the risk of fatality are strongly correlated with age.

All these considerations have been the motivation for designing of a more complex Agent-based model (ABM), possibly able to overcome the transmission and population homogeneity limitations of an Equation-based version.

4.2.1 Meta-population setting

To simulate the mechanisms of infection as reliable as possible, a high resolution synthetic population has been firstly created.

Each agent of the meta-population is defined as a subject of age class a, which could be eventually a pathological patient depending on $p \in \{\text{YES, NO}\}$. Here pathological means that the individual has at least one disease commonly classifiable as chronic such as diabetes, hypertension, cancer, etc. The adopted age groups are the same eleven as those of the studies in the previous chapter: $0-2$, $3-5$, $6-10$, $11-13$, $14-18$, $19-24$, $25-44$, $45-64$, $65-79$, $80-84$, $85+$. In this way students of various grades, workers and the elderly are precisely represented by the model.

Regarding contacts, agents are connected each other by edges defined in a multi-layers network, please refer to Figure 4.3 for a summarizing schema. The first layer gives the household connections (orange). The second connection type represents workplaces: agents in adult ages are supposed to have both generic contacts at work (light blue), both close contacts with colleagues (blue). On the other hand, the third layer models schools: children and young agents attend a school (shaded area) and belong to one of its classes (light green). They are also supposed to have close contact only with some friends among the classmates (dark green). For simplicity here an agent is defined student when aged between 3 and 24 years, while it is defined a working adult when aged from 25 to 64 years. Finally, the last edge type tries to include all the possible remaining common and habitual contacts in the community (magenta): friends, sport, shops, cafes, etc.

In practice, the values of a and p have been randomly assigned to each agent. The reference discrete distributions were been priorly obtained respectively from the last ISTAT census on Veneto and from a 2014 survey on the CSD Longitudinal Patient Database, an Italian general practice registry [5] (Tables 6.2 and 6.1 in Appendix). Note that, the pathologies presence probability here varies depending on the different age groups.

For the network instead, to implement household, work and school contacts, random samplings respectively of size h, w and s agents have been iteratively performed until all the population has been covered. Each set of nodes has been then connected following the simple complete graph as reference topology rule. Moreover, each completely connected group of co-working agents have been further randomly partitioned into non-intersecting subgraphs of size w_c to simulate different stronger contacts between close colleagues. Analogously for school, firstly students have been partitioned into classes of size s_c and then these have been randomly divided into small groups of s_f friends. In this way each agent has a sort of local context of belonging, but its number of really significant and more likely infectious contacts remains limited.

To produce a more realistic population, the sample sizes have been also generated stochastically. The household composition h varies in $[1, 5]$ according to the observed distribution in the Veneto region (as always obtained from ISTAT census data, here reported in Table 6.3 in Appendix). A correction mechanism has been also implemented in order to allow only adult aged individuals to be assigned to one member family (single) and to ensure that children have at least a parent.

The remaining sizes w, s, w_c, s_c, s_f have been instead defined as random variables uniformly or normally distributed, whose statistics have been left as hyperparameters.

Finally, as regards the community layer, the Barabási-Albert model has been preferred as the most suitable random graph for social network. As a matter of fact, thanks to its preferential attachment property it encourages non-uniform degrees and the presence of hub nodes which could potentially be super-spreaders. Again, m, the number of new edges to be added at a time by the building algorithm, has been treated as hyperparameter.

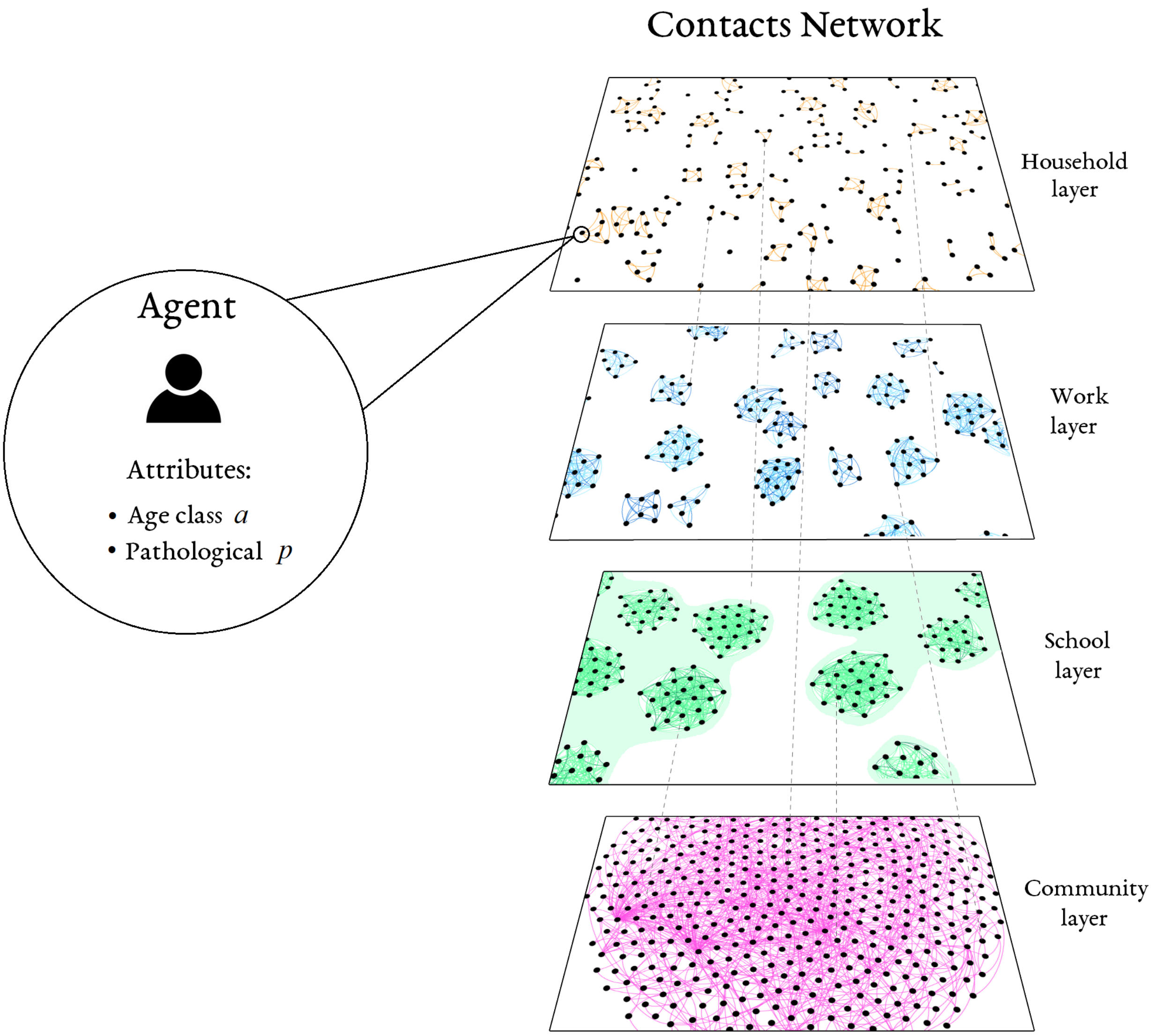

Figure 4.3: Meta-population setting in the Agent Based model: agents attributes and multi-layered network of contacts.

4.2.2 Infection dynamics

As regards the contagion transmission dynamics, the Agent-based model follows exactly the same compartmental scheme as the deterministic one (see Figure 4.1). The real important difference lies in how those parameters are formulated and how the mechanisms of infection is implemented in practice.

Although all agents follow the same infection dynamics rules, each of them now has part of the parameters strictly personalized, depending on his age a, on the possible presence of chronic pathologies p and on a stochastic component. The choice of which parameters to customize has been mostly guided by the available data, reflecting whether it would be possible to estimate in such detail from the Veneto observations or if there would already exist known values in the literature.

Indeed, as already explained above, the available health system databases have allowed to explicitly estimate some of the probabilities and rates. By further splitting the records in order to divide by age groups and pathological or non-pathological subjects, it has been possible to obtain a and p specific approximation for the proportions P_{As}, P_{Sy}, P_{Ho}, $P_{D|Sy}$, $P_{D|Ho}$ (see Table 6.5 in Appendix).

Similarly, for the recovery rates $1/\gamma_{As}$, $1/\gamma_{Sy}$, $1/\gamma_{Ho}$ the mean and the standard deviation could be calculated (Table 6.6 in Appendix). Starting from those values, a random recovery time is here assigned to each agent sampling from a $Gamma$ probability distribution having the corresponding statistics.

Note that for $1/\gamma_Q$ instead, again the value of 14 days has been adopted, as commonly defined by law.

Finally, differently from the DM, here also ϵ has been allowed to vary dependently on the age. Exploiting the estimation provided in [30] (reported in Table 6.4 in Appendix), the incubation time is assigned to each agent sampling from a $Gamma$ distribution.

For all the remaining, ϕ_{As}, ϕ_{Sy}, ψ, μ, $P_{Q|As}$ and $P_{Q|Sy}$, it has been proceeded as usual by simply adopting constants to be properly trained.

Concerning the mechanisms that simulate contagion, the Agent-based model does not include β transmission rates, but uses probabilities. At each time step, each infected agent in I_{As}, I_{Sy}, I_{Ho} or Q, can only infect its susceptible neighbouring agents in the underlying network graph according to specific probabilities.

In detail, individuals in Q can only transmit the virus to co-housing contacts, this eventuality is regulated by probability $P_{h|Q}$. Hospitalized patients are instead supposed to potentially infect

solely through the community layer edges with probability $P_{o|Ho}$. Finally, for subjects in I_{As} and I_{Sy} it is allowed to spread COVID-19 in all the designed environments, respectively with probabilities $P_{h|As}$, $P_{h|Sy}$ for household members, $P_{w|As}$, $P_{w|Sy}$ in workplaces, $P_{w_c|As}$, $P_{w_c|Sy}$ for close work colleagues, $P_{s|As}$, $P_{s|Sy}$ in schools, $P_{s_c|As}$, $P_{s_c|Sy}$ for classmates, $P_{s_f|As}$, $P_{s_f|Sy}$ for school friends and $P_{o|As}$, $P_{o|Sy}$ for other community contacts.
Note that since the multi-layered social network is constant, in order to include in the model also contacts of non-repetitive or habitual nature, the possibility of completely random connections sampled among the overall population has been added in the case of asymptomatic and symptomatic infected. As usual, infection force is parametrized by probabilities: $P_{c|As}$, $P_{c|Sy}$.

4.2.3 Model implementation and fitting

In practice, the `mesa` module, the Python 3-based counterpart to NetLogo or MASON, has been exploited to develop the Agent-Based model.
Technically speaking, each agent is spatially placed in its correspondent node of the underlying social network, given as a `networkx` graph object, while time is controlled by a `StagedActivation` scheduler.
Precisely, each step of the simulation corresponds to a real day and includes two stages. Firstly, the variables of each agent are checked and eventually updated if it is time to move to a different compartment. Then, once a day, individuals in I_{As}, I_{Sy}, I_{Ho} or Q can infect their contacts: for every susceptible and allowed neighbour, a Bernoulli experiment is conducted with the respective probability, where the success outcome corresponds to virus transmission. Please note that at each stage, agents are activated randomly to avoid possible artefacts due to order bias.
As regards the choice of parameters, it has been proceeded in a similar way to the Deterministic model. As already discussed, part of the values have been directly extrapolated from the data of the Veneto region and from recent publications, while for the remaining a training procedure has been necessary. However, since a single simulation of the model with a fairly high number of agents, at least around 10^5, takes quite long time, to speed up the fitting it has been decided to use the results of the Equation-based model for the parameters ϕ_{As}, ϕ_{Sy}, ψ, μ, $P_{Q|As}$ and $P_{Q|Sy}$.
Another important clarification to be made is that also in this case the parameters strictly linked to the infection dynamics can vary over time. Hence, as expected, in order to have an unbiased comparison and to be able to exploit previous estimates, the time periods to be used must be the same as in the Deterministic model.

Summing up, the only trainable parameters left behind have been the transmission probabilities $P_{h|As}$, $P_{h|Sy}$, $P_{w|As}$, $P_{w|Sy}$, $P_{w_c|As}$, $P_{w_c|Sy}$, $P_{s|As}$, $P_{s|Sy}$, $P_{s_c|As}$, $P_{s_c|Sy}$, $P_{s_f|As}$, $P_{s_f|Sy}$, $P_{o|As}$, $P_{o|Sy}$, $P_{c|As}$, $P_{c|Sy}$, $P_{h|Q}$ and $P_{o|Ho}$.

Although the fitting setting is the same as the first model, the agent-based version has required some fundamental changes. First of all, since in this case the results are run dependent, before calculating the MWSRE measure it is necessary to repeat several times the simulation in order to obtain an average behaviour. Furthermore, since it is computationally impossible to run a simulation with a number of agents equal to the Veneto population, the predicted numbers should be further re-proportioned.

It is therefore evident that a single function evaluation is numerically expensive and slow. Moreover, the ABM does not produce continuous values, but discrete curves. For these reasons, it has been decided to choose as numerical optimization algorithm a different method than Trust Region. In details, the derivative-free optimization solver for constrained problems COBYLA (Constrained optimization by linear approximation) [46] has been adopted. It works by iteratively approximating the actual constrained optimization problem with linear programming problems, therefore it is able to speed up the training, but it is less precise. Concluding the practical implementation overview, note that also for this operation, the built-in library function `minimize` has been used.

5
Experiments and results

In this chapter, the results of the different designed modelling strategies on Veneto will be presented, while highlighting strengths and weaknesses of each method. The conducted simulation and forecasting experiments will be also described, discussing the real capabilities and applicability of the models to the COVID-19 case.

5.1 Deterministic model results

To start, the results obtained with the deterministic-differential equation model are introduced. However, before moving on the actual discussion on fitting, it is worthwhile to remember which hyperparameters have been left to a trial and error approach.
As a matter of fact, one of the most critical decision has been the choice of reasonable time intervals in which varying the parameters. Please remember that these trainable parameters have been designed to be piecewise constant functions. First of all, a criterion based on the reconstruction of the different dates of activation or variation of the implemented protection or restriction measures has been adopted. Then the hypothesized periods have been validated by comparing the trend of the curves in the observed data. After several attempts, it has been decided to split over the following eight macro-intervals:

- **Period 1**, from 17th February 2020 to 10th March 2020, includes the beginning of the epidemic. Starting from the week of the first COVID-19 case in Veneto in the municipality of Vo', several ordinances were introduced at the local level with restriction measures mainly addressed to school, public events and travelling.

- **Period 2**, from 11th March 2020 to 10th May 2020, represents the total lockdown phase. As a matter od fact, in those months the containment measures were extended to the whole national territory with the Prime Ministerial Decree (DPCM) of 8th March. Schools, bars, restaurants, work activities and non-essential shops were completely closed. In addition, transports were reduced to minimum and travelling was allowed only for basic needs. Face mask, social distancing and sanitization became mandatory.

- **Period 3**, from 11th May 2020 to 14th June 2020, contains a few post lockdown weeks, when a gradual reopening of activities began and regional, national travel restrictions decayed.

- **Period 4**, from 15th June 2020 to 5th September 2020, spans instead on summer weeks, during which the low number of infections allowed generalized reopenings and reduced the obligation to wear face mask only to indoor spaces.

- **Period 5** goes from 6th September 2020 to 2rd November 2020. It includes schools reopening of all levels and the new increment in the number of positive cases, with consequent reintroduction of limitations, such as gyms closure and limited hours for bars and restaurants.

- **Period 6**, from 3rd November 2020 to 23th December 2020, contains the days when the second wave of the infection reported its maximum peak. Closure of secondary schools and travel limitations between municipalities were added to the previous interventions.

- **Period 7** approximatively represents the Christmas break, from 24th December 2020 to 18th January 2021. In those days, severe measures similar to those of the first lockdown were implemented throughout Italy, with the exception of some rules to allow visits to family members on holidays.

- **Period 8**, from 19th January 2021 to 13th February 2021, finally includes the remaining available days, when, with the beginning of the new year, there were again a relaxation of the rules and a partial reopening of high schools. Note that the very last ten observations, until February 23rd, were not included in the training in order to allow prediction experiments on them.

A second challenging decision regarded the starting values to be passed as initial condition to the ODEs System. While at $t = 0$, day 17th February 2020, it was reasonable to assume zero cases in Veneto for $I_{Ho\,0}$, Q_0, R_0, D_0, the real starting sizes for E_0, $I_{As\,0}$ and $I_{Sy\,0}$ were unknown. Therefore different combinations of values have been manually tested. It was interesting to discover that with null values, or very low numbers for all three compartments, it is

impossible to reproduce the first phase of exponential growth and the outbreak observed in Veneto at the end of February. Indeed, it has been empirically determined that a fair fitting could result only by setting $E_0 \in [50, 100]$ and $I_{As\,0}, I_{Sy\,0} \in [3, 10]$. On the other hand, for the fitting of periods from 2 to 8, it has been chosen to assume as initial conditions the solutions found at the previous step in correspondence of the relative time t_i .
For the remaining hyperparameters required to train, such as the values in the weighting functions δ and ω or the optimization constraints, refer to the Appendix.
Given all the ingredients, the optimal parameters values returned by the fitting procedure for each macro-interval are finally shown in Table 5.1. The estimated curves on hospitalized, deceased, quarantined and positive cases are also represented in Figure 5.1 with the corresponding real observations.

Parameter	Period 1	Period 2	Period 3	Period 4	Period 5	Period 6	Period 7	Period 8
t_i	0	23	84	119	202	260	311	337
t_f	22	83	118	201	259	310	336	361
β_{As}	0.6184	0.2291	0.2334	0.5719	0.7026	0.3144	0.2165	0.1406
β_{Sy}	0.6022	0.4060	0.4110	0.7359	0.7488	0.5640	0.3957	0.1217
β_{Ho}	0.0785	0.0231	0.0014	0.0025	0.0359	0.0407	0.0286	0.0014
β_Q	0.0726	0.0054	0.0011	0.0016	0.0201	0.0191	0.0112	0.0010
ϵ	0.2650	0.4597	0.1028	0.1030	0.1266	0.4078	0.5350	0.5964
ϕ_{As}	0.3968	0.6331	0.9767	0.7772	0.8250	0.7998	0.7230	0.8444
ϕ_{Sy}	0.4051	0.3861	0.9727	0.8183	0.6906	0.6742	0.6379	0.6447
ψ	0.9510	0.5568	0.0747	0.9025	0.7703	0.6825	0.6879	0.5579
μ	0.0349	0.0603	0.0900	0.0340	0.0468	0.0793	0.1143	0.1642
$P_{Q\|As}$	0.1163	0.6920	0.7425	0.5201	0.6386	0.6962	0.5376	0.6585
$P_{Q\|Sy}$	0.0946	0.6410	0.7349	0.5693	0.5562	0.5453	0.4034	0.3339

Table 5.1: Results of Deterministic model parameters fitting on Veneto population data.

To begin with the results on transmission rates βs, it seems that the model is able to distinguish the different roles of asymptomatic, symptomatic, quarantined and hospitalized infected, giving each of them different weights in the contagion capacity. As a matter of fact, it can be observed that β_{Sy} is almost always greater than or equal to β_{As}, while the values of β_Q and β_{Ho} are definitely lower by at least an order of magnitude.
On the other hand, analysing how transmission rates vary over time, it seems that the obtained values are consistent with the observed phases of growth or decrease and at the same time with the severity of the active restrictions measures on each individual period. Indeed, after the start of the epidemic, all βs importantly decrease thanks to the lockdown. However, during sum-

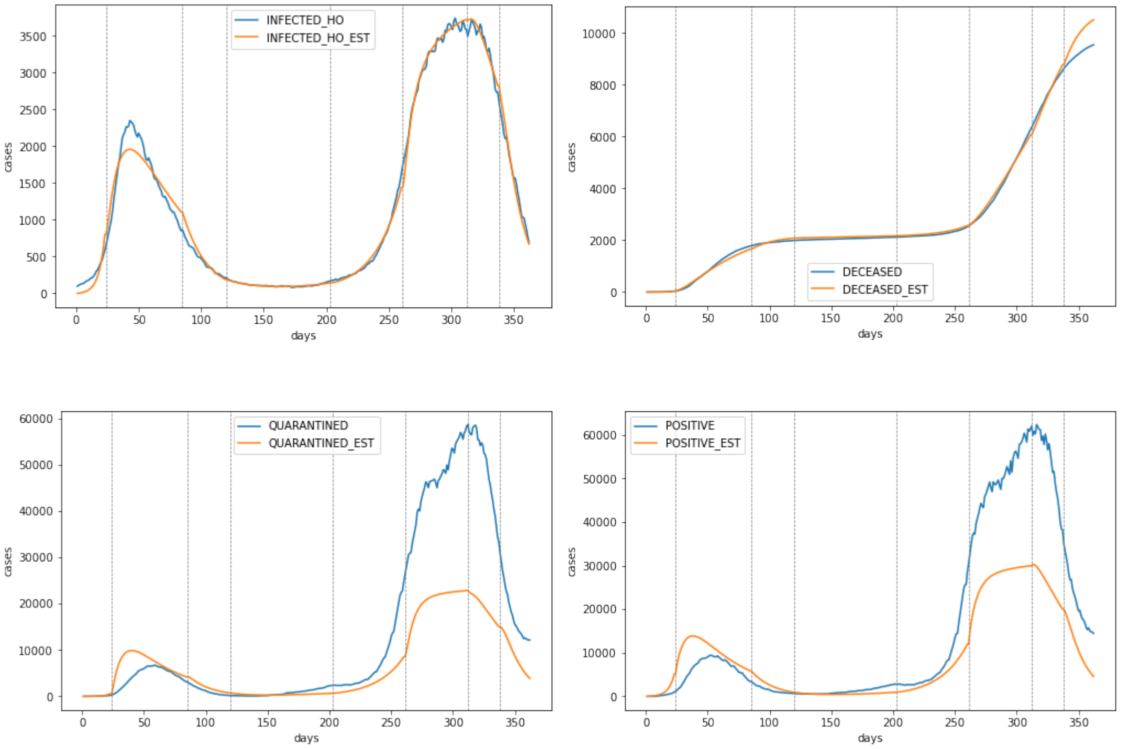

Figure 5.1: Deterministic model estimated numbers versus real numbers of infected hospitalized (*left top*), deceases (*right top*), quarantined (*left bottom*) and positive cases (*right bottom*). Vertical dashed lines separate the different fitted time periods.

mer and especially with school reopening it seems that the virus has begun to strongly circulate again, causing the second wave in late autumn, when the transmission rates diminish with the reintroduction of more stringent restrictions.

Regarding the probabilities of passing into compartment Q, first of all it should be remembered that they represents the possibility of being home isolated. Therefore, in the case of positive symptomatic subjects, the total probability of being uncovered is obtained by adding to $P_{Q|Sy}$ also the probability of being hospitalized $P_{Ho} \approx 0.223$, since in that case a patient would be surely tested to COVID-19.

With this premise in mind, it can be noticed that, with the exception of the first period in which very low values are estimated, the proportion of infected quarantined is around $40 - 70\%$. Moreover, it is interesting to observe that the lowest probabilities occur in summer and in the Christmas season, both holiday-related times of the year during which people may have lowered their attention.

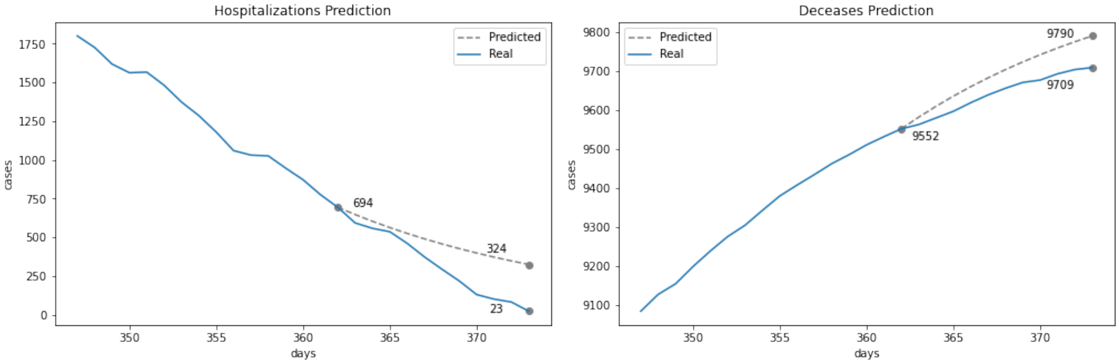

Figure 5.2: Experiment of prediction on hospitalized (*left*) and deceases (*right*) with fitted Deterministic model.

For all the remaining parameters ϵ, ϕ_{As}, ϕ_{Sy}, ψ, μ, it is difficult to give an interpretation of the results. It is also suspected that the model has slightly overfitted on some of them, compensating possible errors or limitations, despite showing curves quite close to the real ones.

In fact, looking at the graphs in Figure5.1, it can be seen that the Deterministic model almost correctly reproduces the numbers of deaths and hospitalized patients, but fails in the estimates of positives and quarantined, especially in the maximum peak phases. While the error in the first wave may still be reasonable assuming that the observations at the start of the pandemic could be incomplete and remembering that the model is encouraged to overestimate by definition of the chosen performance measure weights, it is not clear why the model predictions largely deviates in the second peak.

Possible explanations could be searched in the too limiting deterministic formulation or in the unreliability of parameters estimates directly obtained from the regional health system data, for which it has been assumed constant validity over time. With regard to the latter hypothesis, it should be remembered that in Chapter 3 it was shown that the proportions of asymptomatic, symptomatic and hospitalized patients have changed over time due to the increase in the number of molecular and antigenic tests, in the awareness of the population and in the screening campaigns. In particular, it was observed that the proportion of deceased and hospitalized compared to asymptomatic or mild symptomatic has decreased over the months. Therefore, by forcing instead the model to keep this ratio constant and to optimize the trainable parameters by fitting more importantly on the I_{Ho} and D curves, it is reasonable that it has underestimated the other categories of positives I_{As} and Q.

Although the model could be further perfected, since the training results were generally fair, it has been anyway proceeded to test its forecasting performances.

In detail, it has been decided to try to predict the cases of deaths and hospitalizations for COVID-19 for ten consecutive days. The last solutions found during the fitting procedure has been exploited as initial conditions of the ODEs system, correcting where possible with the real recorded numbers. The obtained prediction results are shown in Figure 5.2.
As expected, the Equation-based model is not able to accurately predict for wide forecasting horizons, showing a worsening in the error as time increases. In the conducted experiments, it seems that the model is able to predict a correct trend. However, the obtained forecasts show quite high relative errors around 13.08 and 0.008 respectively for infected hospitalized and deceases.
Hence, it is not possible to conclude that the results are acceptable, especially when more precise estimates are needed, as in the case of hospitalizations, whose numbers are usually exploited as thresholds by policy makers.

5.2 Agent Based model results

As regards the Agent-based model, the fitting has been decidedly more challenging due to the complex structure of the underlying population as well as of the personalized dynamics of contagion. To ensure a trade-off between reliable results as well as shorter execution times, it has been opted for a number of simulation repetitions equal to 10 and a number of $20,000$ agents. Moreover, parallelization has also been exploited to speed up both the multiple runs of the same experiment and the training procedure.
Since the model fitting has not yet completed on all the available time records, only the achieved partial results are here reported. Similarly to the DM, the optimal estimated values for the parameters and the fitted curves are shown, respectively in Table 5.2 and Figure 5.3. Remember that, since the ABM has a stochastic component, the plots represent the mean behaviour obtained for each instant of time t together with the corresponding 95% confidence interval.
From the plots it can be seen that this method is able to reproduce the trend of COVID-19 cases better than the deterministic counterpart, especially for the numbers of positive and quarantined subjects during the second wave of the epidemic. Nevertheless it is necessary to point out that the ABM is not accurate on low numbers. Indeed, being the meta-population smaller than the real one, each agent here represents about 250 individuals, thus ABM can not encode more subtle phenomena. A possible solution would be to adopt a hybrid approach, as in [10], exploiting the Equation-based model for the first simulation days, or any other period with lower case numbers, and then continue with the Agent-based one.

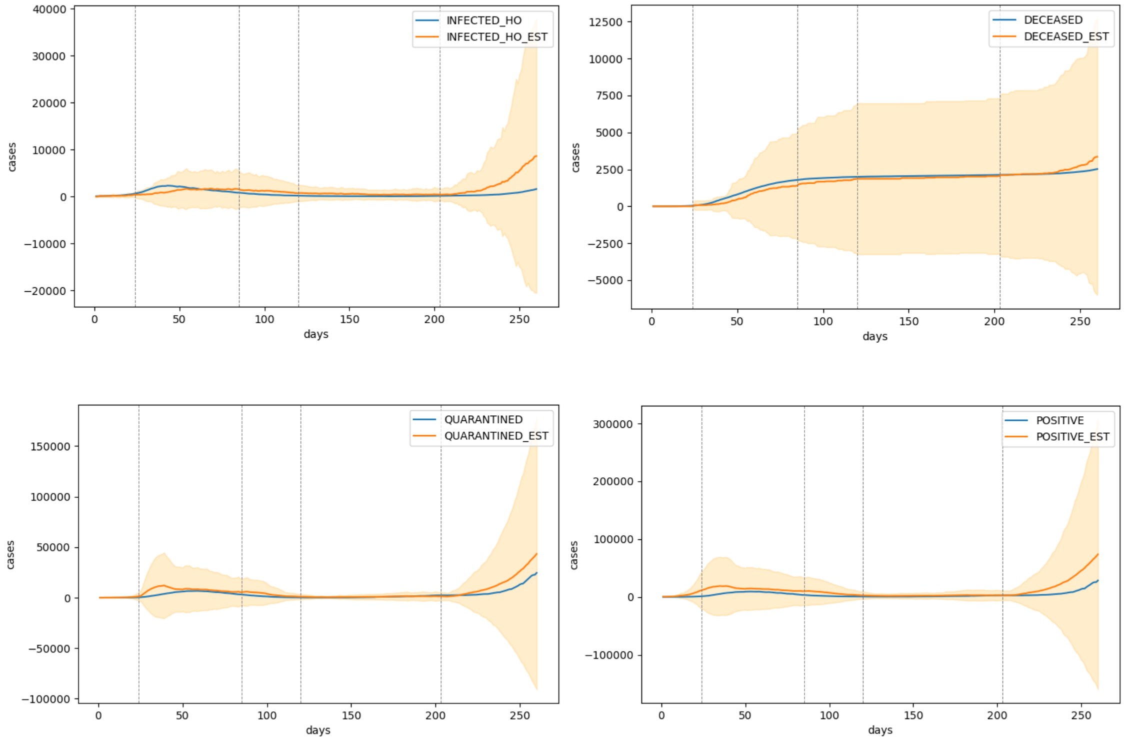

Figure 5.3: Agent-based model estimated numbers versus real numbers of infected hospitalized (*left top*), deceases (*right top*), quarantined (*left bottom*) and positive cases (*right bottom*). Vertical dashed lines separate the different fitted time periods, while the shaded areas give pointwise 95% confidence bounds.

As concern the trainable variables, i.e. the various probabilities of transmission, it is hard to interpret the obtained numerical results and evaluate their significance. As a matter of fact, for some of them, unexpected or even apparently incorrect behaviours can be noted among the different periods. It can not be concluded that the ABM has correctly learned the distinct roles of the different class of infectors as well as of the various types of contacts, as it does not return particular distinguishable patterns. It therefore seems that the model has partially overfitted due to the high degree of freedom allowed by the many trainable probabilities, all equal competitors in the same dynamics of contagion. Limiting the search space by providing more tight constraints for the variables in the optimization problem, exploiting where possible priori domain knowledges, may help the training algorithm to converge to better and more reliable solutions.

Despite it is not possible to determine with certainty whether the resulting parameters are valid or not, given the more than sufficient performance in reproducing the observed curves in the

Parameter	Period 1	Period 2	Period 3	Period 4	Period 5
t_i	0	23	84	119	202
t_f	22	83	118	201	259
$P_{c\|As}$	$5.043e-05$	0.00013806	$< 1e-16$	0.00341038	$8.569e-08$
$P_{o\|As}$	$3.018e-05$	$< 1e-16$	$4.366e-08$	$3.297e-06$	$2.013e-07$
$P_{h\|As}$	$< 1e-16$	$1.345e-05$	$< 1e-16$	0.00946477	0.41788340
$P_{w\|As}$	0.00142420	0.00021126	$1.132e-07$	0.00983549	0.88507141
$P_{w_c\|As}$	0.00010259	$2.821e-06$	$9.447e-07$	0.00443611	$< 1e-16$
$P_{s\|As}$	$7.204e-05$	$< 1e-16$	$9.999e-09$	$1.000e-08$	$< 1e-16$
$P_{s_c\|As}$	0.06904497	$< 1e-16$	$9.999e-09$	$< 1e-16$	$< 1e-16$
$P_{s_f\|As}$	0.15547666	$1.000e-08$	$9.999e-09$	$< 1e-16$	0.00249346
$P_{c\|Sy}$	0.01283323	0.00013608	0.02343780	0.00865469	0.01390100
$P_{o\|Sy}$	$2.179e-07$	$< 1e-16$	$< 1e-16$	$3.638e-05$	$< 1e-16$
$P_{h\|Sy}$	0.00673244	0.07981005	$< 1e-16$	0.00402936	$1.509e-07$
$P_{w\|Sy}$	0.78679119	0.06119351	$< 1e-16$	0.00620262	$< 1e-16$
$P_{w_c\|Sy}$	$8.035e-05$	0.91536280	0.99999999	0.99457201	0.29880006
$P_{s\|Sy}$	0.00462901	$< 1e-16$	$1.000e-08$	$1.000e-08$	$< 1e-16$
$P_{s_c\|Sy}$	0.18031151	$< 1e-16$	$< 1e-16$	$< 1e-16$	0.00013314
$P_{s_f\|Sy}$	0.20030259	$1.000e-08$	$9.999e-09$	$1.000e-08$	$< 1e-16$
$P_{h\|Q}$	0.01413586	0.02758206	$< 1e-16$	0.02025461	0.00058013
$P_{o\|Ho}$	0.00016182	$1.428e-05$	$< 1e-16$	$7.285e-05$	$2.650e-07$

Table 5.2: Results of Agent-based model parameters fitting on Veneto population data.

Veneto region, it has been decided to continue the study experimenting with this fitted version of ABM.

The prediction of deaths and hospitalizations cases for COVID-19 has been hence tested also for this model. The ten-days forecasting results are shown in Figure 5.4. It should be specified that, given the presence of the stochastic component and then the impossibility of reproducing exactly the same values, in this case the forecasting is not done on the absolute numbers, but on the daily increases or variations Δs. In this way, it is possible to obtain more precise numbers using the last known daily data as starting value and simply adding the predictions on the Δs. The predictive abilities of ABM seem decidedly superior to those of DM, showing in the reported test relative errors of 0.047 and 0.023 respectively for the number of hospitalizations and deceases. In addition, apart from some minor fluctuations, the trend generally looks closer to real observations. However, a fair comparison will only be possible once both models have the training completed, so as to be able to repeat the forecasting trial on the same time interval.

The latest carried out experimentation has been a What-if analysis. Given the greater detail of ABM in reproducing both the population and the spread of infections in the various major social environments and dynamics, this model has made possible to simulate various potential pandemic scenarios in Veneto. In particular, the possible effects on the number of positive,

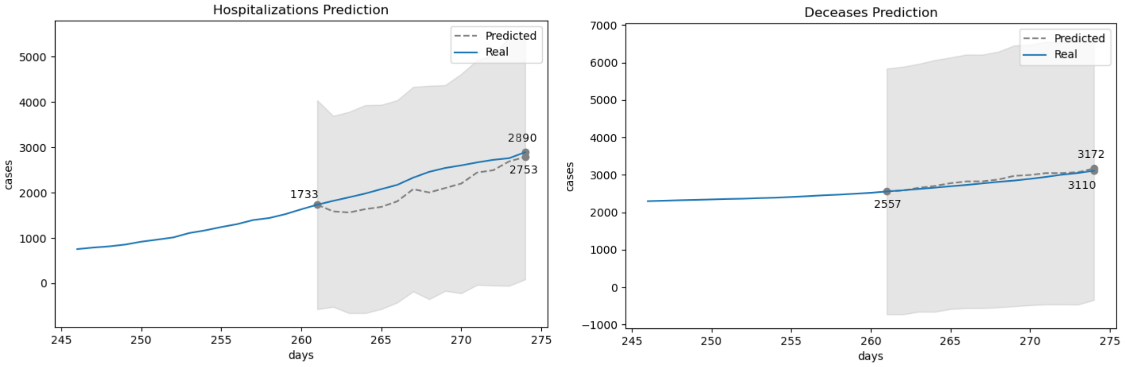

Figure 5.4: Experiment of prediction on hospitalized (*left*) and deceases (*right*) with fitted Agent-based model. As usual, shaded areas give pointwise 95% confidence bounds.

quarantined, hospitalized and death cases have been simulated considering whether different restriction strategies were implemented or not.

Firstly, it has been studied on the synthetic population what would have happened if no measures had been implemented to contrast the spread of the novel Coronavirus. This test allowed not only to demonstrate once again the importance of the public implemented interventions, but also to provide a control example for the time-dependent parameters obtained from the training. In other words, it permits to check whether the separate fitting on various intervals was useful or not. In practice, it has been proceeded by adopting for each time t, the probabilities and rates fitted for the first period, i.e. that corresponding to the beginning of the epidemic. A second attempt was generated instead to prove the possible benefit of a second lockdown after the summer. In fact, it would be particularly interesting to determine whether it could have avoided the second, and most serious, peak of infections registered in the following months. In this case the simulation has been run assuming valid the parameters obtained for Period 2 also for Period 5.

The other experiments concerned instead school protection measures. Indeed, the correlation between schools opening and the increase in infections number is still a hot topic in epidemiology research as well as between politicians. Here the trend of the aforementioned quantities have been then obtain both assuming that schools had not been closed during first the first lockdown phase, both simulating no school reopening after summer 2020. For the former, the probabilities $P_{s|As}$, $P_{s|Sy}$, $P_{s_c|As}$, $P_{s_c|Sy}$, $P_{s_f|As}$, $P_{s_f|Sy}$ trained on Period 1 have been exploited also on Period 2, while for the latter they have been all simply assumed to be zero also in Period 5.

The curves for the various reproduced scenarios are compared to those actually observed in

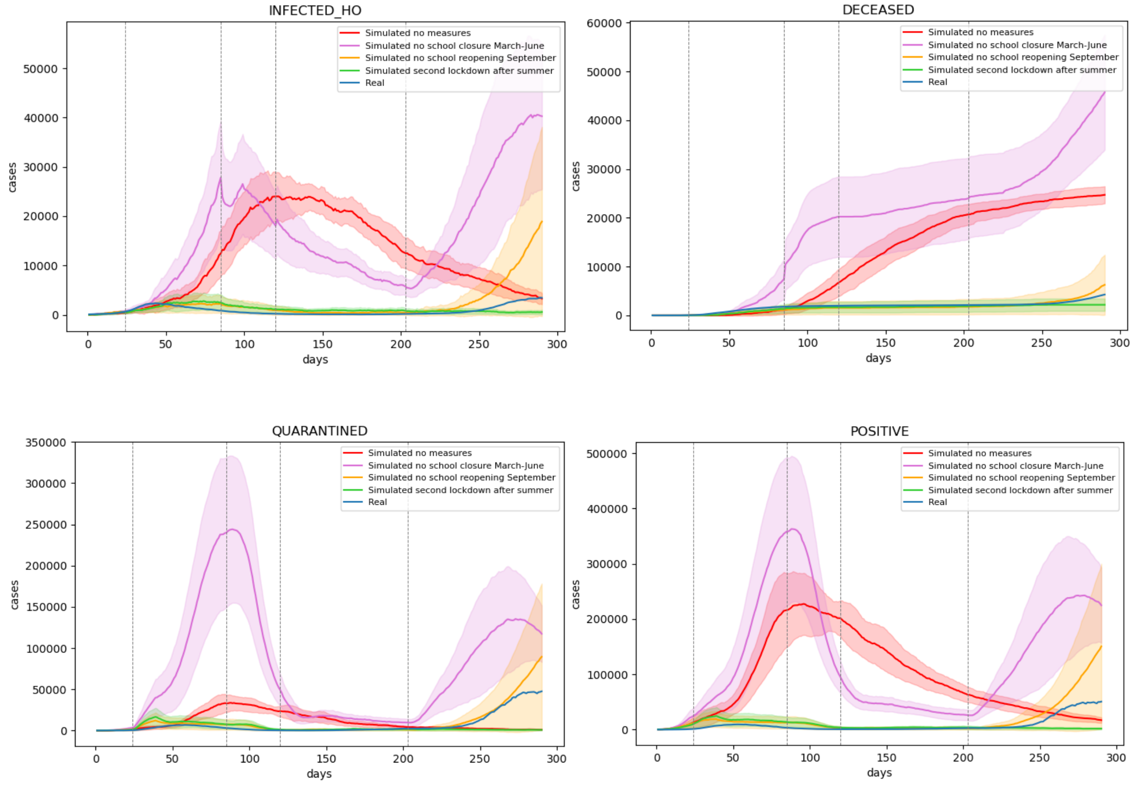

Figure 5.5: Simulation experiments of different pandemic scenarios on the Veneto population by varying the active restrictions and containment measures.

Figures 5.5.
As expected, in the event that no restrictions against the epidemic were implemented (red line), the numbers of infected subjects, as well as those with critical symptoms and deaths, would be higher of several orders of magnitude. An interesting behaviour to highlight is that in this test the curves are unimodal, just like those generally reproduced by an equation method with constant parameters. However the bells here seem flatter, probably due to the protection effect guaranteed by the implementation on network.
Continuing on the second experiment, it is clear that the repetition of a lockdown even after the summer would have benefited the health system, definitively reducing cases to zero (green line). On the other hand, implementing these extreme measures always leads to serious social and economic consequences. From this point of view, a second period of total closure would indeed have been unsustainable.
As regards the restriction on schools, in the case face-to-face teaching modality would not

restart from September (orange line) it can be noticed that the real curves intersect the confidence zone of the reproduced numbers and that the model generally returns coincident or even larger values. Therefore, it cannot be proved any significant difference. The test should perhaps be repeated using more records and continuing the simulation at least until December 2020.
Finally, from the simulating scenario testing the event that schools were never closed (purple line), it turns out that the epidemic curves almost maintain the same patterns as those originally recorded. As a matter of fact, both waves of the epidemic can be still distinguished in the two peaks separated by the summer period, which as usual report instead contained contagions. The important difference lies in the scale of magnitude, now much higher and at the level of the results obtained from the simulation with no active restrictive measures.
To conclude, although most of the reproduced scenarios seem reasonable and realistic, it should be always taken into account that the accuracy of these results is closely linked to that of the parameter values. Therefore the predicted numbers, especially as regards specific dynamics or environments such as schools, could be overestimated or underestimated due to previous overfitting errors of the model during training. A manual validation of the parameter estimates would help to understand if ABM could actually constitute a precise tool to perform What-if analysis for the COVID-19 epidemic in Veneto.

5.3 Code availability

All the code produced for the aforementioned experiments is available at
https://github.com/coclab/COVID-19_Models_Veneto.

6
Conclusion

In this thesis work, two different strategies for epidemiological modeling implementations have been proposed for the same reference SEIQRD compartmental schema. In fact, once designed custom compartments and transitions, an equation-based Deterministic model has been developed and tested against a more complex Agent-based model.
The two methods, *ad hoc* created and trained on the COVID-19 infection cases in the Veneto region, exhibited opposite advantages and limitations, as expected from their complementary nature. While the Equations-based one is immediate to run, but more imprecise in fitting and in predicting, the Agent-based model is very computationally demanding, but much more flexible and accurate in forecasting. To speed up the training of the latter, it has been therefore decided to exploit part of the parameters estimated for the former, which becomes the baseline model.
Once the time-dependent parameters have been validated on the numbers of positive, quarantined, hospitalized and deceased in Veneto, the Agent-based model has been used to experiment with different pandemic scenarios changing the implemented anti-contagion rules. These simulations have once again demonstrated how much containment measures, in particular lockdown, are essential to reduce the pandemic. Moreover, according to these experiments, there seems to be a correlation between schools opening and contagions increase. However, it is not possible to ascertain whether these results are reliable or not, since the estimates obtained for the transmission probabilities in classroom or in other general school-related environments could be inaccurate due to model overfitting.

Although the performances can be considered almost acceptable and satisfactory, the models have not yet reached an optimal level. The developed prototypes could hence constitute a good starting point for further works and improvements. In particular, future researches should consider the implementation of more sophisticated mechanisms to represent additional types of containment measures.

For instance, a multi resolution model with several agents populations running in parallel could be designed to better encode local variabilities among the regional provinces or municipalities encompassing information about population density, proportion of elderly people or prevalences of certain diseases. This type of approach could also include the exchange of transition flows between the different communities, which would result particularly useful to understand how the contagion has moved between the geographical areas and whether the travelling restrictions have been effective.

Other possible enhancements may comprise the addition of the viral load between the agents attributes in order to allow variable personal contagiousness, the insertion of a vaccinated compartment or a modification of the simulation step with night and day shifts to study the effect of curfew.

To conclude, the real challenges would be not only practical, indeed researchers, with the complicity of the regional institutions, should now especially face how to find adequate datasets containing such high resolution demographical and medical records required by this kind of methods.

Appendix

Age distribution in Veneto

Source: ISTAT census

Age group (a)	Frequency	%
$0-2$	105598	2.16
$3-5$	118043	2.41
$6-10$	222031	4.55
$11-13$	141385	2.89
$14-18$	232380	4.76
$19-24$	285686	5.85
$25-44$	1117212	22.89
$45-64$	1521131	31.17
$65-79$	777127	15.92
$80-84$	181729	3.72
85+	176811	3.62
Tot.	4879133	100

Table 6.1: Observed absolute and percentage numbers of subjects by age class in the Veneto population.

Chronic pathologies presence in Italy

Source: 2014 Italian general practice registry [5]

Age group (a)	% Pathological (p =YES)
$0-44$	28
$45-64$	59
$65-79$	84
80+	91

Table 6.2: Observed percentages of subjects with at least one chronic clinical condition by age in the Italian population.

Family composition in Veneto

Source: ISTAT census

Household size (h)	%
1	30.2
2	29.6
3	19.2
4	15.0
5+	6.0

Table 6.3: Observed percentages of households by number of components in the Veneto region.

Incubation period by age $\left(\frac{1}{\epsilon}\right)$

Source: [30]

Age group (a)	Mean (days)	Std.
$0-18$	7.0	4.21
$19-64$	7.7	4.21
65+	9.0	4.21

Table 6.4: Estimates of latent period duration in days depending on age.

Age group (a)	Pathol. (p)	P_{As}	95% CI	P_{Sy}	95% CI	P_{Ho}	95% CI	$P_{D\|Sy}$	95% CI	$P_{D\|Ho}$	95% CI
$0-2$	NO	0.78	(0.77, 0.8)	0.22	(0.2, 0.23)	0.15	(0.12, 0.18)	0.0	(0.0, 0.0)	0.0	(0.0, 0.0)
$0-2$	YES	0.85	(0.69, 1.0)	0.15	(0.0, 0.31)	0.67	(0.13, 1.0)	0.0	(0.0, 0.0)	0.0	(0.0, 0.0)
$3-5$	NO	0.87	(0.85, 0.88)	0.13	(0.12, 0.15)	0.02	(0.01, 0.03)	0.0	(0.0, 0.0)	0.0	(0.0, 0.0)
$3-5$	YES	0.9	(0.71, 1.0)	0.1	(0.0, 0.29)	0.0	(0.0, 0.0)	0.0	(0.0, 0.0)	0.0	(0.0, 0.0)
$6-10$	NO	0.88	(0.88, 0.89)	0.12	(0.11, 0.12)	0.02	(0.01, 0.02)	0.003	(0.0, 0.006)	0.0	(0.0, 0.0)
$6-10$	YES	0.95	(0.85, 1.0)	0.05	(0.0, 0.15)	1.0	(1.0, 1.0)	0.0	(0.0, 0.0)	0.0	(0.0, 0.0)
$11-13$	NO	0.85	(0.85, 0.86)	0.15	(0.14, 0.15)	0.01	(0.01, 0.02)	0.0	(0.0, 0.0)	0.0	(0.0, 0.0)
$11-13$	YES	0.78	(0.64, 0.92)	0.22	(0.08, 0.36)	0.14	(0.0, 0.4)	0.0	(0.0, 0.0)	0.0	(0.0, 0.0)
$14-18$	NO	0.78	(0.78, 0.79)	0.22	(0.21, 0.22)	0.01	(0.01, 0.01)	0.0	(0.0, 0.0)	0.0	(0.0, 0.0)
$14-18$	YES	0.79	(0.69, 0.88)	0.21	(0.12, 0.31)	0.31	(0.09, 0.54)	0.0	(0.0, 0.0)	0.0	(0.0, 0.0)
$19-24$	NO	0.73	(0.73, 0.74)	0.27	(0.26, 0.27)	0.02	(0.02, 0.02)	0.0	(0.0, 0.0)	0.019	(0.0, 0.046)
$19-24$	YES	0.68	(0.61, 0.74)	0.32	(0.26, 0.39)	0.21	(0.1, 0.31)	0.0	(0.0, 0.0)	0.0	(0.0, 0.0)
$25-44$	NO	0.67	(0.67, 0.67)	0.33	(0.33, 0.33)	0.04	(0.04, 0.04)	0.0	(0.0, 0.0)	0.021	(0.012, 0.029)
$25-44$	YES	0.64	(0.61, 0.67)	0.36	(0.33, 0.39)	0.3	(0.26, 0.35)	0.0	(0.0, 0.0)	0.048	(0.01, 0.085)
$45-64$	NO	0.63	(0.63, 0.64)	0.37	(0.36, 0.37)	0.12	(0.11, 0.12)	0.001	(0.001, 0.001)	0.071	(0.064, 0.078)
$45-64$	YES	0.52	(0.5, 0.53)	0.48	(0.47, 0.5)	0.55	(0.52, 0.57)	0.004	(0.002, 0.007)	0.077	(0.062, 0.093)
$65-79$	NO	0.57	(0.57, 0.58)	0.43	(0.42, 0.43)	0.38	(0.37, 0.39)	0.017	(0.015, 0.019)	0.22	(0.21, 0.23)
$65-79$	YES	0.37	(0.35, 0.38)	0.63	(0.62, 0.65)	0.8	(0.79, 0.82)	0.016	(0.011, 0.021)	0.251	(0.23, 0.272)
$80-84$	NO	0.52	(0.51, 0.53)	0.48	(0.47, 0.49)	0.58	(0.56, 0.59)	0.066	(0.059, 0.073)	0.362	(0.345, 0.379)
$80-84$	YES	0.37	(0.34, 0.39)	0.63	(0.61, 0.66)	0.85	(0.83, 0.87)	0.044	(0.032, 0.057)	0.387	(0.354, 0.42)
85+	NO	0.53	(0.52, 0.53)	0.47	(0.47, 0.48)	0.57	(0.56, 0.58)	0.212	(0.203, 0.221)	0.502	(0.488, 0.516)
85+	YES	0.44	(0.42, 0.46)	0.56	(0.54, 0.58)	0.79	(0.77, 0.81)	0.136	(0.12, 0.153)	0.549	(0.522, 0.575)

Table 6.5: Observable proportions in Veneto population data partitioned by age and pathology presence.

Age	Pathol.	$1/\gamma_{As}$			$1/\gamma_{Sy}$			$1/\gamma_{Ho}$		
group (a)	(p)	Mean (days)	95% CI	Std.	Mean (days)	95% CI	Std.	Mean (days)	95% CI	Std.
$0-13$	NO	14.61	$(14.43, 14.78)$	12.01	15.78	$(15.33, 16.24)$	12.15	19.28	$(15.61, 22.94)$	19.3
$0-13$	YES	16.96	$(14.89, 19.02)$	8.67	12.38	$(7.67, 17.08)$	5.63	13.5	$(-0.03, 27.03)$	8.5
$14-18$	NO	16.3	$(16.02, 16.57)$	13.53	17.64	$(17.08, 18.21)$	14.25	20.0	$(10.68, 29.32)$	22.58
$14-18$	YES	19.95	$(15.81, 24.09)$	15.74	16.55	$(13.57, 19.53)$	4.44	24.0	$(-2.88, 50.88)$	21.64
$19-24$	NO	16.6	$(16.35, 16.85)$	14.77	16.72	$(16.38, 17.06)$	12.08	21.16	$(17.67, 24.65)$	16.96
$19-24$	YES	20.39	$(17.68, 23.1)$	14.84	22.75	$(11.91, 33.59)$	35.64	22.27	$(13.53, 31.02)$	13.02
$25-44$	NO	16.78	$(16.64, 16.91)$	14.62	17.01	$(16.85, 17.17)$	12.32	20.39	$(19.2, 21.58)$	18.57
$25-44$	YES	26.98	$(23.92, 30.04)$	42.24	22.89	$(19.75, 26.04)$	26.81	33.62	$(23.81, 43.44)$	54.06
$45-64$	NO	17.4	$(17.29, 17.51)$	14.16	17.72	$(17.59, 17.85)$	11.97	23.45	$(22.88, 24.02)$	18.51
$45-64$	YES	28.1	$(26.36, 29.84)$	41.81	22.07	$(20.89, 23.25)$	18.18	27.37	$(25.43, 29.3)$	31.78
$65-79$	NO	18.77	$(18.58, 18.96)$	13.61	18.75	$(18.5, 19.0)$	12.15	25.16	$(24.68, 25.63)$	16.64
$65-79$	YES	25.61	$(24.04, 27.18)$	27.2	23.83	$(21.95, 25.7)$	17.79	29.85	$(28.19, 31.51)$	29.59
$80-84$	NO	22.27	$(21.79, 22.75)$	17.43	20.47	$(19.82, 21.12)$	13.21	26.59	$(25.68, 27.51)$	19.44
$80-84$	YES	29.49	$(27.2, 31.78)$	27.59	23.35	$(20.86, 25.83)$	12.38	31.45	$(28.76, 34.15)$	30.78
85+	NO	25.03	$(24.64, 25.43)$	18.44	21.57	$(20.89, 22.24)$	13.6	28.42	$(27.58, 29.26)$	19.8
85+	YES	33.96	$(32.05, 35.87)$	35.64	26.6	$(24.26, 28.94)$	13.22	37.21	$(33.81, 40.62)$	42.41

Table 6.6: Observable recovery rates in Veneto population data partitioned by age and pathology presence.

Weighting functions

The weighting functions for Mean Weighted Squared Relative Error to be used for the parameter fitting have been precisely defined as follows:

$$\omega(c) = \begin{cases} 1 & \text{if } c = I_{Ho} \text{ or } c = D \\ 0.05 & \text{if } c = Q \\ 0.01 & \text{if } c = P \end{cases}$$

$$\delta(e) = \begin{cases} 1 & \text{if } e < 0 \\ 0.5 & \text{if } e \geq 0 \end{cases}$$

where c indicates the compartment or the class of observation to consider, while e is the difference between the predicted and the real values.

Optimization problem

The optimization problem involved in the training procedure for the Deterministic model at each time interval $[t_i, t_f]$ has been formulated as follows

$$\begin{aligned} \min_{x} \quad & \text{MWSRE}(t_i, t_f) \\ \text{where} \quad & x = (\beta_{As}, \beta_{Sy}, \beta_{Ho}, \beta_Q, \epsilon, \phi_{As}, \phi_{Sy}, \psi, \mu, P_{Q|As}, P_{Q|Sy}) \\ \text{s.t.} \quad & 0.01 \leq \beta_{As} \leq 1 \\ & 0.01 \leq \beta_{Sy} \leq 1 \\ & 0.001 \leq \beta_{Ho} \leq 0.2 \\ & 0.001 \leq \beta_Q \leq 0.2 \\ & 1/10 \leq \epsilon \leq 1 \\ & 1/10 \leq \phi_{As} \leq 1 \\ & 1/10 \leq \phi_{Sy} \leq 1 \\ & 1/14 \leq \psi \leq 1 \\ & 1/30 \leq \mu \leq 1 \\ & 0.00001 \leq P_{Q|As} \leq 0.8 \\ & 0.00001 \leq P_{Q|Sy} \leq 1 - P_{Ho} - P_{D|Sy} \,. \end{aligned}$$

Analogously, the optimization problem involved in the Agent-Based model fitting at each time interval $[t_i, t_f]$ is

$$\begin{aligned} \min_{\bar{P}} \quad & \text{MWSRE}(t_i, t_f) \\ \text{where} \quad & \bar{P} = (P_{h|As}, P_{h|Sy}, P_{w|As}, P_{w|Sy}, P_{wc|As}, P_{wc|Sy}, P_{s|As}, P_{s|Sy}, \\ & \qquad P_{sc|As}, P_{sc|Sy}, P_{s_f|As}, P_{s_f|Sy}, P_{o|As}, P_{o|Sy}, P_{h|Q}, P_{o|Ho}) \\ \text{s.t.} \quad & 0 \leq \bar{P}_j \leq 1 \, \forall \, j. \end{aligned}$$

Note that sometimes the constraints have been tightened to enforce a priori knowledges, e.g. zero probability on school contacts when learning spaces were closed, to help the algorithm to converge to more reasonable solutions.

Bibliography

[1] Allen, L. (2008). An Introduction to Stochastic Epidemic Models. In: Brauer, F., van den Driessche, P. & Wu, J. (eds) Mathematical Epidemiology. Lecture Notes in Mathematics. *Springer*, 1945.

[2] Allen, L. (2017). A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis. *Infect Dis Model*, 2(12), 128–142.

[3] Anderson, R. M. (1982). *Population Dynamics of Infectious Diseases: Theory and Applications.* Chapman and Hall.

[4] Anderson, R. M. & May, R. (1992). *Infectious Diseases of Humans.* Oxford University Press.

[5] Atella, V., Piano Mortari, A., Kopinska, J., Belotti, F., Lapi, F., Cricelli, C., & Fontana, L. (2019). Trends in age-related disease burden and healthcare utilization. *Aging Cell*, 18(1).

[6] Bailey, N. T. J. (1975). *The mathematical theory of infectious diseases and its applications (2nd ed.).* Griffin.

[7] Bartlett, M. S. (1957). Measles periodicity and community size. *Journal of the Royal Statistical Society*, Series A. 120 (1), 48–70.

[8] Bertuzzo, E., Mari, L., Pasetto, D., Miccoli, S., Casagrandi, R., Gatto, M., & Rinaldo, A. (2020). The geography of COVID-19 spread in Italy and implications for the relaxation of confinement measures. *Nature Communications*, 11(1).

[9] Block, P., Hoffman, M., Raabe, I. J., Dowd Beam, J., Rahal, C., Kashyap, R., & Mills, M. C. (2020). Social network-based distancing strategies to flatten the COVID-19 curve in a post-lockdown world. *Nature Human Behaviour*, 4, 588–596.

[10] Bobashev, G. V., Goedecke, D. M., Yu, F., & Epstein, J. M. (2007). A hybrid epidemic model: Combining the advantages of agent-based and equation-based approaches. *Proceedings - Winter Simulation Conference*, (pp. 1532–1537).

[11] Brauer, F. & Castillo-Chávez, C. (2001). *Mathematical Models in Population Biology and Epidemiology*. Springer.

[12] Britton, T. (2010). Stochastic epidemic models: a survey. *Mathematical Biosciences*, 225(1), 24–35.

[13] Britton, T. (2020). Epidemic models on social networks—with inference. *Statistica Neerlandica*, 74, 222–241.

[14] Britton, T. & Pardoux, E. (2019). *Stochastic Epidemic Models with Inference*. Springer.

[15] Capasso, V. (1993). *Mathematical Structure of Epidemic Systems*. Springer.

[16] Champredon, D., Dushoff, J., Park, S. W., & Weitz, J. S. (2019). A practical generationinterval-based approach to inferring the strength of epidemics from their speed. *Elsevier Epidemics*, 27, 12–18.

[17] Conn, A. R., Gould, N. I., & Toint, P. L. (2000). *Trust region methods*. SIAM.

[18] Cutler, D. M. & Summers, L. H. (2020). The COVID-19 Pandemic and the $16 Trillion Virus. *JAMA*, 324(15), 1495–1496.

[19] Dashtbali, M. & Mirzaie, M. (2021). A compartmental model that predicts the effect of social distancing and vaccination on controlling COVID-19. *Sci Rep*, 11, 8191.

[20] Del Valle, S. Y., Hyman, J. M., & N., C. (2013). Mathematical models of contact patterns between age groups for predicting the spread of infectious diseases. *Math Biosci Eng.*, 10(5–6), 1475–1497.

[21] Della Rossa, F., Salzano, D., Di Meglio, A., et al. (2020). A network model of Italy shows that intermittent regional strategies can alleviate the COVID-19 epidemic. *Nat Commun*, 11(5106).

[22] Ding, Y., Huang, R., & Shao, N. (2021). Time Series Forecasting of US COVID-19 Transmission. *Altern Ther Health Med.*, 27(S1), 4–11.

[23] Ferguson, N. M., Laydon, D., Nedjati-Gilani, G., Imai, N., Ainslie, K., Baguelin, M., Bhatia, S., Boonyasiri, A., Cucunubá, Z., Cuomo-Dannenburg, G., Dighe, A., Dorigatti, I., Fu, H., Gaythorpe, K., Green, W., Hamlet, A., Hinsley, W., Okell, L. C., Elsland, S. V., Thompson, H., Verity, R., Volz, E., Wang, H., Wang, Y., Walker, P. G., Walters, C., Winskill, P., Whittaker, C., Donnelly, C. A., Riley, S., & Ghani, A. C. (2020). Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. *Imperial.Ac.Uk*, (pp. 3–20).

[24] Fumanelli, L., Ajelli, M., Manfredi, P., Vespignani, A., & Merler, S. (2012). Inferring the Structure of Social Contacts from Demographic Data in the Analysis of Infectious Diseases Spread. *PLoS Computational Biology*, 8(9), 35–39.

[25] Gatto, M., Bertuzzo, E., Mari, L., Miccoli, S., Carraro, L., Casagrandi, R., & Rinaldo, A. (2020). Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures. *Proceedings of the National Academy of Sciences May 2020*, 117(19).

[26] Gecili, E., Ziady, A., & Szczesniak, R. D. (2021). Forecasting COVID-19 confirmed cases, deaths and recoveries: Revisiting established time series modeling through novel applications for the USA and Italy. *PLoS ONE*, 16(1), e0244173.

[27] Giordano, G., Blanchini, F., Bruno, R., et al. (2020). Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nat Med*, 26, 855–860.

[28] Greenwood, P. & Gordillo, L. (2009). Stochastic Epidemic Modeling. In: Chowell, G., Hyman, J.M., Bettencourt, L.M.A. & Castillo-Chavez, C. (eds) Mathematical and Statistical Estimation Approaches in Epidemiology. *Springer*.

[29] House, T. & Keeling, M. J. (2008). Deterministic epidemic models with explicit household structure. *Mathematical Biosciences*, 213, 29–39.

[30] Huang, S., Li, J., Dai, C., Tie, Z., Xu, J., Xiong, X., Hao, X., Wang, Z., & Lu, C. (2021). Incubation period of coronavirus disease 2019: new implications for intervention and control. *International Journal of Environmental Health Research*, 0(0), 1–9.

[31] Hunter, E., Mac Namee, B., & Kelleher, J. (2017). A taxonomy for agent-based models in human infectious disease epidemiology. *Jasss*, 20(3).

[32] Kapoor, A., Ben, X., Liu, L., Perozzi, B., Barnes, M., Blais, M., & O'Banion, S. (2020). Examining COVID-19 Forecasting using Spatio-Temporal Graph Neural Networks. *arXiv*.

[33] Keeling, M. J. & Eames, K. T. D. (2005). Networks and epidemic models. *J. R. Soc. Interface*, 2, 295–307.

[34] Kermack, W. & McKendrick, A. (1927). A contribution to mathematical theory of epidemics. *Proc. Roy. Soc. Lond. A*, 115, 700–721.

[35] Lai, S., Ruktanonchai, N. W., Zhou, L., Prosper, O., Luo, W., Floyd, J. R., Wesolowski, A., Santillana, M., Zhang, C., Du, X., Yu, H., & Tatem, A. J. (2020). Effect of non-pharmaceutical interventions to contain COVID-19 in China. *Nature*.

[36] Leontitsis, A., Senok, A., Alsheikh-Ali, A., Al Nasser, Y., Loney, T., & Alshamsi, A. (2021). SEAHIR: A Specialized Compartmental Model for COVID-19. *International journal of environmental research and public health*, 18(5), 2667.

[37] Liu, Q. H., Ajelli, M., Aleta, A., Merler, S., Moreno, Y., & Vespignani, A. (2018). Measurability of the epidemic reproduction number in data-driven contact networks. *Proceedings of the National Academy of Sciences of the United States of America*, 115(50), 12680–12685.

[38] Macal, C. M. & North, M. J. (2009). Agent-based modeling and simulation. *Proceedings - Winter Simulation Conference*, (pp. 86–98).

[39] Mandel, A. & Veetil, V. (2020). The Economic Cost of COVID Lockdowns: An Out-of-Equilibrium Analysis. *EconDisCliCha*, 4, 431–451.

[40] Martcheva, M. (2010). *An Introduction to Mathematical Edipemiology*. Springer.

[41] Moein, S., Nickaeen, N., Roointan, A., et al. (2021). Inefficiency of SIR models in forecasting COVID-19 epidemic: a case study of Isfahan. *Sci Rep*, 11, 4725.

[42] Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G. S., Wallinga, J., Heijne, J., Sadkowska-Todys, M., Rosinska, M., & Edmunds, W. J. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine*, 5(3), 0381–0391.

[43] Nepomuceno, E. G., Resende, D. F., & Lacerda, M. J. (2019). A Survey of the Individual-Based Model applied in Biomedical and Epidemiology. *arXiv: 1902.02784*, 1.

[44] Pellis, L., Ferguson, N. M., & Fraser, C. (2009). Threshold parameters for a model of epidemic spread among households and workplaces. *Journal of the Royal Society Interface*, 6(40), 979–987.

[45] Potter, G. E., Smieszek, T., & Sailer, K. (2015). Modeling workplace contact networks: The effects of organizational structure, architecture, and reporting errors on epidemic predictions. *Network Science*, 3(3), 298–325.

[46] Powell, M. J. D. (2007). A view of algorithms for optimization without derivatives. *Cambridge University Technical Report DAMTP.*

[47] Prem, K., Liu, Y., Russell, T. W., Kucharski, A. J., Eggo, R. M., Davies, N., Jit, M., Klepac, P., Flasche, S., Clifford, S., Pearson, C. A. B., Munday, J. D., Abbott, S., Gibbs, H., Rosello, A., Quilty, B. J., Jombart, T., Sun, F., Diamond, C., Gimma, A., van Zandvoort, K., Funk, S., Jarvis, C. I., Edmunds, W. J., Bosse, N. I., & Hellewell, J. (2020). The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. *The Lancet Public Health*, 2667, 1–10.

[48] Rahmadani, F. & Lee, H. (2020). Hybrid Deep Learning-Based Epidemic Prediction Framework of COVID-19: South Korea Case. *Applied Sciences*, 10(23), 8539.

[49] Zabeo, F. (2020). Stochastic epidemic models on dynamic networks. *Master thesis in Mathematics, University of Padova*.

[50] Zhang, G. & Liu, X. (2021). Prediction and control of COVID-19 spreading based on a hybrid intelligent model. *PLoS ONE*, 16(2), e0246360.

[51] Zhang, Z., Wang, H., Wang, C., & Fang, H. (2015). Modeling Epidemics Spreading on Social Contact Networks. *IEEE Transactions on Emerging Topics in Computing*, 3, 410–419.