A DATA DRIVEN EPIDEMIC MODEL TO ANALYZE AND FORECAST THE DYNAMICS OF COVID-19

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF APPLIED MATHEMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

RZA HASANLI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
SCIENTIFIC COMPUTING

AUGUST 2021

Approval of the thesis:

A DATA DRIVEN EPIDEMIC MODEL TO ANALYZE AND FORECAST THE DYNAMICS OF COVID-19

submitted by **RZA HASANLI** in partial fulfillment of the requirements for the degree of **Master of Science in Scientific Computing Department, Middle East Technical University** by,

Prof. Dr. A. Sevtap Selçuk Kestel _____
Director, Graduate School of **Applied Mathematics**

Assoc. Prof. Dr. Hamdullah Yücel _____
Head of Department, **Scientific Computing**

Prof. Dr. Ömür Uğur _____
Supervisor, **Scientific Computing, METU**

Res. Assist. Dr. Cansu Evcin
Co-supervisor, **Department of Mathematics, Tekirdağ Namık Kemal University** _____

Examining Committee Members:

Assist. Prof. Dr. Meltem Gölgeli
Department of Mathematics,
TOBB University of Economics and Technology _____

Prof. Dr. Ömür Uğur
Institute of Applied Mathematics, METU _____

Assist. Prof. Dr. Önder Türk
Institute of Applied Mathematics, METU _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: RZA HASANLI

Signature :

ABSTRACT

A DATA DRIVEN EPIDEMIC MODEL TO ANALYZE AND FORECAST THE DYNAMICS OF COVID-19

Hasanli, Rza
M.S., Department of Scientific Computing
Supervisor : Prof. Dr. Ömür Uğur
Co-Supervisor : Res. Assist. Dr. Cansu Evcin

August 2021, 52 pages

Due to recent evolution of the COVID-19 outbreak, accurate mathematical modelling to capture the dynamics of disease transmission is of vital importance. Since the availability and quality of data differs from region to region, it is very difficult to develop an accurate model from the global perspective. Nevertheless, local predictive models can be developed by collecting data from certain regions. In this thesis, a modified version of Susceptible-Exposed-Infected-Recovered-Dead (SEIRD) differential model is proposed for the analysis and forecast of COVID-19 spread. Parameter estimation of the model is the first step of analysis which is carried out by fitting the model to available data as good as possible in the sense of least squares. In second step of analysis, simulations are performed by using the optimal values of parameters. Through numerical simulations, the effects of public measures, such as isolation, social distancing on the dynamics of COVID-19 outbreak are observed. The model is applied for the COVID-19 outbreak in Turkey, Italy, and Spain. Time to reach the peak, total infected, recovered and dead cases are compared with real data and found to be in good agreement for all the countries.

Keywords: corona virus, parameter estimation, nonlinear regression, mathematical epidemic model, reproduction number

ÖZ

COVID-19 DİNAMİKLERİNİ ANALİZ VE TAHMİN ETMEK İÇİN VERİYE DAYALI EPİDEMİK MODELİ

Hasanli, Rza
Yüksek Lisans, Bilimsel Hesaplama Bölümü
Tez Yöneticisi : Prof. Dr. Ömür Uğur
Ortak Tez Yöneticisi : Arş. Gör. Dr. Cansu Evcin

Ağustos 2021, 52 sayfa

COVID-19 salgınının son zamanlardaki gelişimi nedeniyle, hastalık bulaşma dinamiklerini yakalamak için doğru matematiksel modelleme hayati önem taşımaktadır. Verilerin mevcudiyeti ve kalitesi bölgeden bölgeye farklılık gösterdiğinden, küresel perspektiften doğru bir model geliştirmek çok zordur. Buna rağmen, belirli bölgelerden veri toplanarak yerel tahmin modelleri geliştirilebilir. Bu tezde, COVID-19 yayılımının analizi ve tahmini için Duyarlı-Maruz Kalmış, Bulaşıcı-Kurtarılmış-Ölü (SEIRD) diferansiyel modelinin modifiye edilmiş bir versiyonu önerilmektedir. Parametre tahmini, modelin mevcut verilere en küçük kareler anlamında en iyi şekilde uydurulmasıyla gerçekleştirilen analizin ilk adımıdır. Analizin ikinci adımında, parametrelerin optimal değerleri kullanılarak simülasyonlar gerçekleştirilmektedir. Sayısal simülasyonlar aracılığıyla izolasyon, sosyal mesafe gibi kamusal önlemlerin COVID-19 salgınının dinamikleri üzerindeki etkileri gözlemlenmektedir. Model, Türkiye, İtalya ve İspanya'daki COVID-19 salgını için uygulanmaktadır. Zirveye ulaşma süresi, toplam enfekte, iyileşen ve ölü vakalar gerçek verilerle karşılaştırılmakta ve tüm ülkeler için iyi bir uyum içinde olduğu bulunmaktadır.

Anahtar Kelimeler: korona virüs, parametre tahmini, doğrusal olmayan regresyon, matematiksel epidemik modeli, bulaştırma katsayısı

To My Family

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my supervisor Prof. Dr. Ömür Uğur for his invaluable guidance and continual encouragement through all the stages of writing my thesis.

I would also like to thank my co-supervisor Res. Assist. Dr. Cansu Evcin for her kind help and advice.

Finally, I would like to express my thanks to my family for their continuous support and encouragement which helped me in completion of this thesis.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

SIR	Susceptiple-Infected-Recovered
SEIR	Susceptiple-Exposed-Infected-Recovered
SIRD	Susceptiple-Infected-Recovered-Dead
SEIRD	Susceptiple-Exposed-Infected-Recovered-Dead
DFE	Disease-Free Equilibrium
RMSE	Root-Mean-Square Error

CHAPTER 1

INTRODUCTION

COVID-19 is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1, 20]. The first case of the disease was reported in Wuhan, the capital of Hubei Province in the People's Republic of China in December 2019 [38]. Due to its rapid progress, COVID-19 disease is declared as a Global Pandemic by the World Health Organization on March 11, 2020 [2] resulting in more than 17 million active cases and 4 million deaths in the world, by 20th of August, 2021 [6].

The recent diffusion of COVID-19 disease has increased the interest of scientific community in the epidemic models. Since then researchers are making efforts for developing new refined mathematical models to analyze the current situation and predict the future scenarios. Modelling the dynamics of COVID-19 pandemic is important for creating short-term and long-term strategy for controlling the outbreak. There are several kinds of models that have been proposed to analyze the spread of epidemics, which can be divided into two categories: collective models and networked models. Collective models use a limited number of variables and parameters to describe the evolution of epidemics. They generally include growth models, Richards models, logistic models, and compartmental models such as Susceptible-Infected-Recovered (SIR) models and Susceptible-Exposed-Infectious-Removed (SEIR) models.

In the literature several epidemiological models have been proposed to analyze and forecast the COVID-19 outbreak in different countries. The SIR model and its modified versions, such as SEIR model have been commonly used in the modelling of epidemics. In his paper Khajanchi et al. [22] applied an extended version of SEIR

model to investigate the transmission dynamics of COVID-19 and perform short-term forecasts on the data of India. A discrete-time SIR model including dead compartment was studied by Anastassopoulou et al. [7]. Wu et al. [35] applied a SEIR model to study human-human transmission dynamics of COVID-19 based on the data from Wuhan, China. A stochastic transmission model was also proposed by Kucharski et al. [23] to investigate the dynamics of COVID-19. Fanelli et al. extended the SIR model by adding dead compartments to analyze and forecast COVID-19 outbreak in China, Italy and France [19]. Calafiore et al. proposed a modified version of the SIR model to describe total infected individuals [11]. Piccolomini et al. developed a variant of the SEIRD model by introducing a time-dependent transmitting rate [30].

In this thesis, we propose a data driven epidemic model from a modified version of SEIRD incorporating the public measures through time-dependent parameters. First, the model parameters are estimated by fitting the model to actual data. Then, these optimal values of parameters are used in simulation for prediction. In addition, model parameters, such as dynamic reproduction number is investigated to provide insights in the analysis of COVID-19. With this thesis, we aim at answering the following questions:

- How do the public measures affect the COVID-19 outbreak?
- How do the model parameters, such as the transmission rate and effective daily reproduction number behave after implementation of the public measures?
- How is the performance of proposed model for describing COVID-19 epidemic with the local data of Italy, Turkey, and Spain?

The rest of the thesis is organized as follows: In Chapter 2, we give a theoretical information about the existing compartmental models. In Chapter 3, differential equation solver and optimization algorithms for parameter estimation of epidemic models are briefly discussed. In Chapter 4, we make a numerical experiment with classical SEIR model on COVID-19 data of Turkey to describe the evolution of epidemic. Then, we extend it by introducing time-dependent parameters to formulate the proposed model. We validate the model on COVID-19 data of Italy and make a specific analysis for Turkey and Spain, in Chapter 5. Finally, we give a short summary and concluding

remarks in Chapter 6.

CHAPTER 2

BACKGROUND

2.1 Introduction

Mathematical models have an important role to describe a diverse range of infectious diseases. The models provide meaningful quantitative outcomes such as basic reproduction numbers, thresholds, and contact numbers. Among these mathematical models, deterministic and stochastic approaches are very popular. Additionally, among the deterministic models, compartmental epidemic models are the most used for the study of diseases by the researchers. The models are widely used to describe the epidemics both in time and space and gain information about the transmission dynamics. Epidemic modelling also helps to clarify the trends and make general forecasts. Their importance can be summarized as below:

- Epidemic models can estimate parameters which are important for insights of the transmission mechanisms of epidemic.

- Epidemic models can provide more insights about the thresholds properties.

- Epidemic models can help to model and simulate outbreaks, and find out the effectiveness of the control strategies.

The main objective of this chapter is to provide theoretical information about existing compartmental epidemic models, such as the SIR model and its extended versions.

2.2 Existing Models

Compartmental models simplify the mathematical modelling of diseases. The population is divided into compartments with labels. The order of the labels usually shows the flow pattern between the compartments. A few existing models from which the proposed model is derived are described here. The first simplest model that is used to predict the disease spread is the SIR model.

2.2.1 The SIR Model

In the classic SIR model proposed by Kermack and McKendrick [21], a population is divided into susceptible, infected, and recovered compartments and considered as closed. It initially considers only a small fraction of the population as infected. This small portion of the population infects others as the time passes.

Let $S = S(t)$ be the susceptibles those who can acquire the disease, but presently do not have it, $I = I(t)$ be the infectives those who have the disease and can transmit it to the susceptibles, and $R = R(t)$ be the removed class those who had the disease, are now recovered at time t. The SIR model can then be described as

$$\begin{aligned} \frac{dS}{dt} &= -\beta \frac{SI}{N}, \\ \frac{dI}{dt} &= \beta \frac{SI}{N} - \gamma I, \\ \frac{dR}{dt} &= \gamma I, \end{aligned} \tag{2.1}$$

where β is the transmission rate of the infection, γ is the rate of recovery, and N is the total number of population. The interpretation of the SIR model is as follows:

- The number of susceptible individuals decrease in proportion to their number multiplied by the average proportion of infected individuals I/N.
- The number of infected individuals therefore increases at the same rate adjusted by γI for the fact that some of them are recovering.
- The number of those who recovered increases due to the decrease in the number of infected individuals.

If we add the expressions in (2.1), then

$$\frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt} = 0. \tag{2.2}$$

This implies

$$S(t) + I(t) + R(t) = constant = N. \tag{2.3}$$

Therefore, S, I, and R are bounded by N. A flow chart for the model is generally given as in Figure 2.1.

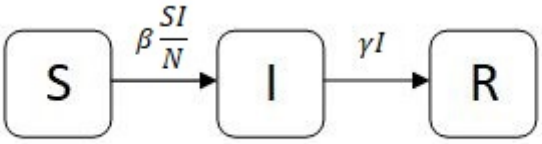


Figure 2.1: The SIR Model.

The classic SIR model is a very simple model. It does not consider any incubation period and even counts dead individuals as a compartment of R. Therefore, the model may not be suitable for describing many diseases.

2.2.2 The SEIR Model

As mentioned above, the SIR model is not capable of capturing the incubation period, i.e., the time elapsed before developing symptoms. This leads to SEIR model [13] which considers a transition from susceptible to exposed individuals, denoted by $E = E(t)$. The SEIR model can be described as

$$\begin{aligned} \frac{dS}{dt} &= -\beta \frac{SI}{N}, \\ \frac{dE}{dt} &= \beta \frac{SI}{N} - \alpha E, \\ \frac{dI}{dt} &= \alpha E - \gamma I, \\ \frac{dR}{dt} &= \gamma I, \end{aligned} \tag{2.4}$$

where, additionally, α is the incubation rate (transition rate from E to I). The flow chart of the SEIR model is given in Figure 2.2.

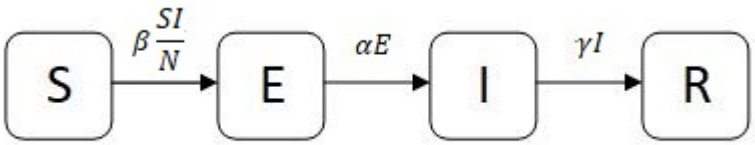


Figure 2.2: The SEIR Model.

2.2.3 The SIRD Model

For deadly diseases, adding another compartment, namely Deaths (D), to the SIR model, we get the simplest SIRD model which can be described as follows:

$$\begin{aligned} \frac{dS}{dt} &= -\beta\frac{SI}{N}, \\ \frac{dI}{dt} &= \beta\frac{SI}{N} - \gamma I - \delta I, \\ \frac{dR}{dt} &= \gamma I, \\ \frac{dD}{dt} &= \delta I, \end{aligned} \tag{2.5}$$

where, additionally, δ is the death rate. The flow chart of the SIRD model can be seen in Figure 2.3.

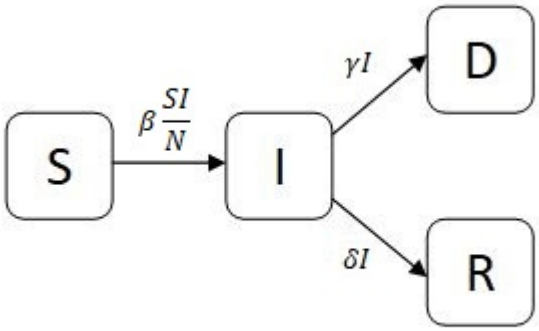


Figure 2.3: The SIRD Model.

Many extensions of these models which include extra compartments, such as vaccinated individuals (V), cross-immune individuals (C), quarantined individuals (Q), hospitalized individuals (H) [14, 12], etc. have been proposed. Most of the studies consider the SEIR model with a deterministic approach by fixing the parameters to

delineate the spread of diseases [36, 29].

2.3 The Basic Reproduction Number

The basic reproduction number $\mathcal{R}_0$ is the number of secondary infections that one infected person would produce in a fully susceptible population. It provides a threshold condition for the stability of the disease-free equilibrium (DFE). When $\mathcal{R}_0 < 1$, the DFE is stable meaning that the disease dies out. On the other hand, the DFE is unstable when $\mathcal{R}_0 > 1$ causing an epidemic. Thus, $\mathcal{R}_0 = 1$ acts as a threshold between the disease dying out or causing an epidemic. It is a dimensionless number that can be expressed as,

$$R_0 = \begin{pmatrix} number\,of\,contacts \\ per\,unit\,time \end{pmatrix} \times \begin{pmatrix} probability\,of\,transmission \\ per\,contact \end{pmatrix} \times \begin{pmatrix} duration\,of \\ infection \end{pmatrix}. \quad (2.6)$$

For SIR model, an epidemic occurs if the number of infected individual increases, that is, when $\frac{dI}{dt} > 0$. Equivalently, when

$$\frac{\beta SI}{N} - \gamma I > 0, \quad (2.7)$$

or

$$\frac{\beta S}{\gamma N} > 1 \quad (2.8)$$

holds. Substituting $S \approx N$ considering nearly everyone is susceptible at the outset of the epidemic, we get

$$\mathcal{R}_0 = \frac{\beta}{\gamma} > 1. \quad (2.9)$$

Reproduction number for disease models can be calculated in a more formal approach. In their paper van den Driessche and Watmough [34], give a way of determining $\mathcal{R}_0$ for a compartmental model by using the next generation matrix. Let

$\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$, with each $x_i \geq 0$ be the number of individuals in each compartment, where the first $m < n$ compartments contain infected individuals. Assume also that DFE x_0 exists and is stable in the absence of disease. In this setting, $\mathcal{F}_i$ is the rate of new infections in compartment i, and $\mathcal{V}_i$ is the rate of the transfers of infections from one compartment to another. Under the assumptions in [34], the next generation matrix G is defined as $G = FV^{-1}$, where

$$F = \left(\frac{\partial \mathcal{F}_i(x_0)}{\partial x_j}\right), \quad \text{and} \quad V = \left(\frac{\partial \mathcal{V}_i(x_0)}{\partial x_j}\right), \quad 1 \leq i, j \leq m. \tag{2.10}$$

Then, $\mathcal{R}_0$ is determined as the spectral radius of the matrix $G = FV^{-1}$.

As an example, let us consider the classic SEIRD model. To calculate the next generation matrix for the SEIRD model, we need to specify the number of ways that new infections can arise and move between compartments. Firstly, we extract the infected parts of this model. These are:

$$\begin{aligned} \frac{dE}{dt} &= \beta\frac{SI}{N} - \alpha E, \\ \frac{dI}{dt} &= \alpha E - \gamma I - \delta I. \end{aligned} \tag{2.11}$$

The DFE of the SEIRD model is given by: $(S^*, E^*, I^*, R^*, D^*) = (N, 0, 0, 0, 0)$. Therefore, the corresponding F and V matrices are formulated as

$$F = \begin{pmatrix} 0 & \beta \\ 0 & 0 \end{pmatrix}, \quad V = \begin{pmatrix} \alpha & 0 \\ -\alpha & \gamma + \delta \end{pmatrix}. \tag{2.12}$$

Then, the next generation matrix is

$$G = FV^{-1} = \begin{pmatrix} \frac{\beta}{\gamma+\delta} & \frac{\beta}{\gamma+\delta} \\ 0 & 0 \end{pmatrix}, \tag{2.13}$$

and the basic reproduction number is the dominant eigenvalue of G, which is

$$\mathcal{R}_0 = \frac{\beta}{\gamma + \delta}. \tag{2.14}$$

It should be noted that, the basic reproduction number of the epidemic models without vital dynamics is constant over time.

In order to detect changes in disease transmission, estimation of the effective reproduction number, $\mathcal{R}_e(t)$, is important. The effective reproduction number, on the other hand, is the number of secondary infections that one infected person would produce through the entire duration of disease. Typically, it is product of the basic reproduction number and the proportion of susceptible population. It can be expressed in terms of $\mathcal{R}_0$ as below:

$$\mathcal{R}_e(t) = \mathcal{R}_0 \times \begin{pmatrix} proportion\, of \\ susceptiple\, population \end{pmatrix}. \tag{2.15}$$

It should be noted that $\mathcal{R}_e(t)$ changes over time in the epidemic models described so far and also indicates whether the disease is dying out or increasing over time. For the SIR model, $\mathcal{R}_e(t)$ can be written as

$$\mathcal{R}_e(t) = \frac{\beta S(t)}{\gamma N(t)}. \tag{2.16}$$

CHAPTER 3

PARAMETER ESTIMATION FOR EPIDEMIC MODELS

3.1 Introduction

In epidemic models, the transition from one compartment to another compartment is expressed as a derivative with respect to time. Therefore, the compartmental type of epidemic models, like SIR, SEIR, and SIRD are governed by a system of nonlinear differential equations. These models are initiated from the physical feature of infectious diseases and then used to determine the unknown parameters of the model using the available data. Every model is described by the epidemic parameters. The values of epidemic parameters sometimes can be determined from the characteristic of the disease or the demographic information. However, there are numerous parameters that can not be estimated directly from this information. Hence, such model parameters are needed to be estimated by fitting the model predictions to the available data.

In the spread of diseases, generally, the parameters appearing in their respective models are not constants, but time-dependent. In the literature, Saikia et al. considers the time-dependent parameters to describe the first phase of the COVID-19 pandemic in India studying the SEIR model [31]. The model proposed by de León et al. [15] also includes transmission, recovery, and death rates as functions of time to describe the spread of COVID-19 in Mexico by extending the SEIR model. They calibrate the model parameters to the available data by minimizing the sum of squared errors.

We observe that tuning the unknown parameters in compartmental models is one of the main challenges. The emphasis of this thesis is to determine these parameters from the data. For this reason, an efficient nonlinear differential equation solver with

an optimization approach is essential. There are a number of algorithms for nonlinear optimization in the literature and they are mostly based on deterministic methods and stochastic methods. Deterministic models are usually gradient based and fast in computing. The main drawback of gradient based is that the optimal values of unknown parameters are highly dependent upon initial conditions. On the other hand, stochastic methods requires long computational running time for searching the global optima [25].

The main object of this chapter is to a give brief information about the differential equation solver and optimization algorithms for parameter estimation of epidemic models.

3.2 Differential Equation Solver

For the parameter estimation problem in epidemic models, the differential equation solver is very important. There are many methods in the literature to solve the differential equations with given initial conditions. Among them, Runge-Kutta methods are very accurate and have better convergence compared to other methods at the same step size.

3.2.1 Runge-Kutta Methods

Runge-Kutta methods are the most popular methods to solve differential equations. Among them, the fourth-order Runge-Kutta method is the most widely used method, but higher order Runge-Kutta methods also exist. For the purpose of this thesis, we use the fifth-order Runge-Kutta (RK45) method which is provided by SciPy. This method uses the Dormand-Prince pair of formulas [18] with 6 stages. For the steps, the fifth-order accurate formula is used, but the error is controlled by the fourth-order accuracy method.

If we consider the given initial value problem as

$$\dot{y} = f(t, y, x), \quad y(t_0) = y_0 \tag{3.1}$$

where y is the state vector depending on time t and the vector of parameters x. Let $a_{21}, a_{31}, a_{32}, \ldots, a_{51}, a_{52}, \ldots, a_{54}, b_1, b_2, \ldots, b_5, c_1, c_2, \ldots, c_5$ be the real coefficients. Then the fifth-order Runge-Kutta method with 6 stages for this problem can be expressed as

$$y_{n+1} = y_n + h \sum_{i=1}^{6} b_i k_i, \tag{3.2}$$

where h is the step size and

$$\begin{aligned} k_1 &= f(t_n, y_n), \\ k_2 &= f(t_n + c_2 h, y_n + a_{21} h k_1), \\ &\vdots \\ k_6 &= f(t_n + c_6 h, y_n + a_{61} h k_1 + a_{62} h k_2 + \cdots + a_{65} h k_5). \end{aligned} \tag{3.3}$$

Dormand-Prince method [18] is the most popular member of the Runge-Kutta family for solving ordinary differential equations. This method uses the coefficients for the fifth-order solution. The values of coefficients for the Dormand-Prince method with 6 stages are given in Table 3.1.

Table 3.1: Dormand-Prince method tableau.

c_i	a_{ij}					b_i	b_i^*
0						$\frac{5179}{57600}$	$\frac{35}{384}$
$\frac{1}{5}$	$\frac{1}{5}$					0	0
$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$				$\frac{7571}{16695}$	$\frac{500}{1113}$
$\frac{4}{5}$	$\frac{44}{45}$	$-\frac{56}{15}$	$\frac{32}{9}$			$\frac{393}{640}$	$\frac{125}{192}$
$\frac{8}{9}$	$\frac{19372}{6561}$	$-\frac{25360}{2187}$	$\frac{64448}{6561}$	$-\frac{212}{729}$		$-\frac{92097}{339200}$	$-\frac{2187}{6784}$
1	$\frac{9017}{3168}$	$-\frac{355}{33}$	$\frac{46732}{5247}$	$\frac{49}{176}$	$-\frac{5103}{18656}$	$\frac{187}{2100}$	$\frac{11}{84}$

where b_i^* are the coefficients that are used in error correction which can be defined for 6 stages as

$$\epsilon_{n+1} = y_{n+1} - y_{n+1}^* = h \sum_{i=1}^{6} (b_i - b_i^*) k_i \tag{3.4}$$

3.3 Optimization Algortihms

In this section, different constrained optimization algorithms are briefly discussed to estimate the parameter values of epidemic models.

Let $(t_1, y_1), (t_2, y_2), \ldots, (t_n, y_n)$ be the n set of observations and $f(t, x)$ be the objective function, where $x = (\theta_1, \theta_2, \ldots, \theta_n)$ is the vector of unknown parameters. Then, the parameter estimation can be stated as

$$x^* = \underset{x \geq 0}{\operatorname{argmin}} f(x), \tag{3.5}$$

where x^* is the vector of parameters that minimizes the objective function f.

SciPy Optimization Toolbox provides many methods to minimize objective functions subject to constraints [4]. For the purpose of this thesis, we use L-BFGS-B, TNC, Trust-Region Constrained, and Trust Region Reflective algorithms.

3.3.1 The L-BFGS-B Algorithm

The L-BFGS-B algorithm is an extension of L-BFGS (Limited-Memory Broyden-Fletcher-Goldfarb-Shanno) algorithm to handle bound constrained problems [10, 37]. It uses the trust region techniques while updating Hessian and line search algorithms by BFGS method. The working principle of the L-BFGS-B algorithm for one iteration is as follows:

1. Estimate the Cauchy point for

   $$\Phi(\lambda) = m(\mathbf{x_k} - \lambda \mathbf{g_k}), \tag{3.6}$$

   where λ is the steplength and $m(\mathbf{x})$ being quadratic form as

   $$m(\mathbf{x}) = f(\mathbf{x_k}) + \mathbf{g_k^T}(\mathbf{x} - \mathbf{x_k}) + (\mathbf{x} - \mathbf{x_k})^{\mathbf{T}} \mathbf{B_k} (\mathbf{x} - \mathbf{x_k})/2, \tag{3.7}$$

   where $\mathbf{g_k}$ is the gradient at point $\mathbf{x}_k$ and $\mathbf{B_k}$ is the approximate Hessian of $f(\mathbf{x})$ at iteration k. Identify active set $A(\mathbf{x})$ and inactive set $I(\mathbf{x})$.

2. Minimize the function in (3.7) for the unconstrained variables to find a search direction.

3. Perform a line search along the new search direction to minimize $f(\mathbf{x})$.

4. Use the L-BFGS method to update the Hessian [27] and check the convergence.

The L-BFGS-B method is affordable for solving large scale problems. It requires roughly $(12 + 2m)N$ memory where m is the number of BFGS updates and N is the size of the model. Generally, $m = 5$ is a typical choice.

There are three different terminating criteria for L-BFGS-B algorithm. The method stops when the maximum number of iterations is reached or the objective function becomes smaller than the tolerance or the norm of the projected gradient is smaller than the threshold value.

3.3.2 TNC Algorithm

TNC method uses a truncated Newton algorithm [26, 28] to minimize a function subject to bound constraints. This algorithm also called Newton Conjugate-Gradient and uses the gradient information. Newton-CG method implements conjugate gradient algorithm to invert the local Hessian. It is based on fitting the function in a quadratic form:

$$f(\mathbf{x}) \approx f(\mathbf{x_0}) + \nabla f(\mathbf{x_0})(\mathbf{x} - \mathbf{x_0}) + \frac{1}{2}(\mathbf{x} - \mathbf{x_0})^T \mathbf{H}(\mathbf{x_0})(\mathbf{x} - \mathbf{x_0}), \tag{3.8}$$

where $\mathbf{H}(\mathbf{x_0})$ is the Hessian. If the Hessian is positive definite, then the optimal solution can be found by equating the gradient of the quadratic form to zero, resulting in

$$\mathbf{x_{opt}} = \mathbf{x_0} - \mathbf{H}^{-1} \nabla f. \tag{3.9}$$

TNC method differs from the Newton-CG method by allowing each variable to be given upper and lower bounds.

3.3.3 Trust-Region Constrained Algorithm

Trust-Constrained method uses a trust-region algorithm for constrained optimization. If we consider the trust-region subproblem, then the objective function can be approx-

imated by using the quadratic model given as

$$\begin{aligned} \min m_k(p) &= f_k + g_k^T p + \frac{1}{2} p^T B_k p, \\ \text{s.t. } ||p|| &\leq \Delta_k, \end{aligned} \tag{3.10}$$

where Δ_k is the trust region radius, g_k is the gradient, and B_k is the Hessian at the point x_k. The critical issue is to update the size of the trust-region radius at every iteration. In this case, the ratio ρ_k between the actual reduction and predicted reduction guides us in determining the radius of trust-region.

$$\rho_k = \frac{f(x_k) - f(x_k + p_k)}{m_k(0) - m_k(p_k)}. \tag{3.11}$$

Depending on the constraints, it switches between two implementations. It is the most versatile algorithm for constrained minimization and the most suitable for large scale problems. It implements Byrd-Omojokun Trust-Region SQP method [24] for equality constraints, and switches to trust-region interior point method [9] for inequality constraints. This interior point algorithm solves inequality constraints by adding slack variables and solving a series of equality-constrained barrier problems with progressively smaller values of the barrier parameter.

3.3.4 Trust Region Reflective Algorithm

Trust Region Reflective Algorithm uses the first order necessary conditions for bound constrained optimization problem [8]. If we consider the minimization problem stated as

$$\min f(x),\ x \in \mathcal{F} = \{x \mid l \leq x \leq u\} \tag{3.12}$$

where l and u are the lower and upper bounds, respectively. Then the first order necessary conditions for x_* to be a local minimum are:

$$\begin{aligned} g(x_*)_i &= 0, \quad \text{if } l_i < x_i < u_i, \\ g(x_*)_i &\leq 0, \quad \text{if } x_i = u_i, \\ g(x_*)_i &\geq 0, \quad \text{if } x_i = l_i, \end{aligned} \tag{3.13}$$

where $g(x)$ is the gradient of f. Now, define a vector $v(x)$ as follows:

$$v(x)_i = \begin{cases} u_i - x_i, & \text{if } g_i < 0 \text{ and } u_i < \infty, \\ x_i - l_i, & \text{if } g_i > 0 \text{ and } l_i > -\infty, \\ 1, & \text{otherwise.} \end{cases} \tag{3.14}$$

Next, define a matrix $D = diag(v(x)^{1/2})$, then the first order optimality can be expressed as $D(x_*)^2 g(x_*) = 0$. Introducing the change of variables in the original trust-region subproblem as $x = D\hat{x}$, the equivalent trust-region problem becomes as

$$\min m(p) = \frac{1}{2} p^T B p + g^T p, \quad \text{s.t. } ||D^{-1} p|| \leq \Delta. \tag{3.15}$$

Therefore, Trust Region Reflective algorithm solves the trust-region subproblem with a special diagonal matrix. The trust-region shape is determined by the distance from the bounds and the direction of the gradient. The algorithm is quite robust for unbounded and bounded problems.

CHAPTER 4

MODEL FORMULATION AND METHODS

The main object of this chapter is to simulate the classical SEIR model and evaluate its performance on COVID-19 data of Turkey, and then extend it to propose a new model for describing the current COVID-19 transmission dynamics.

4.1 Data Requirements and Format

The dataset is obtained from the repository of DataHub [3]. It lists the confirmed cases, reported recoveries and deaths of COVID-19 time series disaggregated by country. In order to include the infected cases, we subtract recovered and dead cases from the confirmed cases for each day. A snapshot of the data that is being used is shown in Figure 4.1:

	Date	Country	Confirmed	Recovered	Deaths	Infected
0	2020-03-25	Turkey	2433	26	59	2348
1	2020-03-26	Turkey	3629	26	75	3528
2	2020-03-27	Turkey	5698	42	92	5564
3	2020-03-28	Turkey	7402	70	108	7224
4	2020-03-29	Turkey	9217	105	131	8981

Figure 4.1: COVID-19 data of Turkey.

4.2 Numerical Experiment with the SEIR Model

We set the parameters manually focusing on available data [3]. Assuming the illness lasts an average of 14 days (at least as long as the mild form lasts, which accounts for up to 80% of the cases), we find the value of $\gamma = 1/14 = 0.0714$. Taking into account the average duration of the incubation period as 3 days, we find the value of $\beta = 3/14 = 0.2143$ and $\alpha = 1/3 = 0.33$. Let us also take the population of Turkey to be equal to 82×10^6. We use the data on Turkey as of March 20, 2020 for the initial conditions:

$$S_0 = 82 \times 10^6 - 355,$$
$$I_0 = 355,$$
$$E_0 = 20I_0.$$

We take the value of E_0 arbitrarily, without loss of generality. As a result of modelling by fifth-order Runge-Kutta method (RK45) provided by SciPy with given initial conditions from March 20 to April 18, we get the following graph in Figure 4.2.

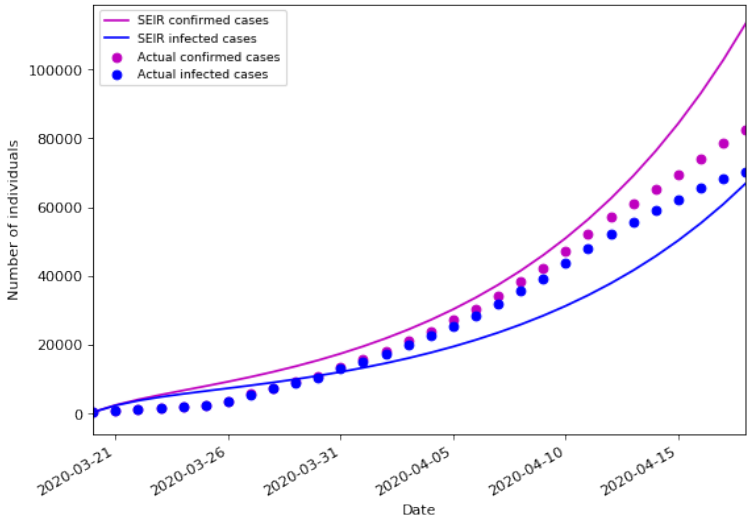


Figure 4.2: The SEIR model simulation with "historical" parameters on COVID-19 data of Turkey.

It can be observed that the results do not fit the data with these "historical" parameters and here the optimization methods come to our aid.

4.3 Optimizing the SEIR Model

Optimization methods are the algorithms that allow you to find the minimum of an objective function. In this case, we face with a nonlinear regression problem: how to choose the vector of parameters $\mathbf{x} = (\beta, \alpha, \gamma, q)^T$ of the differential equations so that the set of points for solving the differential equations is as close as possible to the set of observations, where we introduce parameter q to make assumption for the initial condition of exposed individuals as $E_0 = qI_0$. We use the weighted sum of squares as a measure of the model error. Then, the objective function takes the form as

$$f(\mathbf{x}) = w_1 \sum_{j=1}^{M} (C_j - \mathcal{C}_j)^2 + w_2 \sum_{j=1}^{M} (I_j - \mathcal{I}_j)^2 + w_3 \sum_{j=1}^{M} (R_j - \mathcal{R}_j)^2, \tag{4.1}$$

subject to

$$\mathbf{x} = (\beta, \alpha, \gamma, q)^T \geq 0.$$

where w_1, w_2, and w_3 are the weights, C_j and $\mathcal{C}_i$ are estimated and exact cumulative confirmed cases, I_j and $\mathcal{I}_j$ are estimated and exact infected cases, and R_j and $\mathcal{R}_j$ are estimated and exact recovered cases, respectively. Recalling the SEIR model in (2.4), C_j, I_j and R_j are the model estimates for $I + R$, I and R, respectively, at time j. By fitting the model to the actual data using equal weights, we get the solution shown in Figure 4.3.

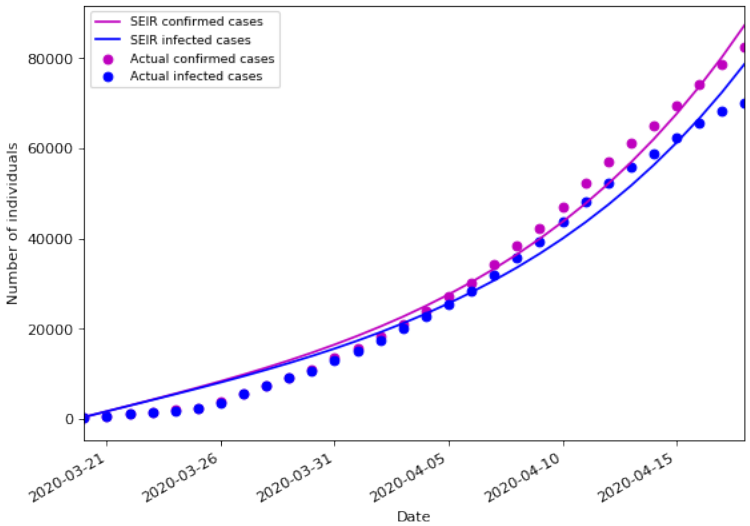


Figure 4.3: The SEIR model fitting on COVID-19 data of Turkey.

At first glance, it can be observed that everything seems to be fine. The optimal values

of the parameters obtained by L-BFGS-B algorithm are:

$$\beta = 9.99 \times 10^{-1},$$
$$\alpha = 8.63 \times 10^{-3},$$
$$\gamma = 1.06 \times 10^{-2},$$
$$q = 4.19 \times 10^{2},$$

and the value of exposed individuals is $E_0 = 419 I_0 = 148745$ which is dubious with 355 registered cases on March 20. We also find that the average time for transition to the active phase of the disease is equal to $1/\alpha \approx 115$ days which does not reflect the real scenario.

Why is it so? Everything becomes clear if we analyze the shape of the solution curve on a longer period of time. To do this, we take the SEIR model with these optimal values of parameters and simulate over 6 months in order to predict the future. Solution curves for the total confirmed and infected cases are shown in Figure 4.4:

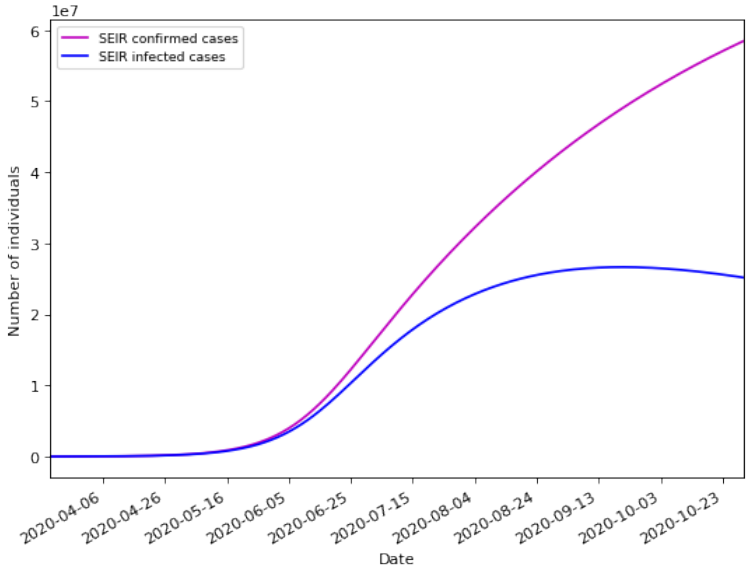


Figure 4.4: The SEIR model simulation with optimized parameters on COVID-19 data of Turkey.

It can be seen that the curves for confirmed and infected cases increase exponentially at the beginning of lockdown, then start to decrease exponentially with a low rate on July. In addition to that, the forecast with optimal parameters is terrifying - almost 97% of the country's population will be infected by October at the same year.

In Figure 4.5, three European countries are compared with Turkey in terms of the number of cases. It can be seen that the development rate of the epidemic for Turkey is lagging about a month, and the growth in the total number of cases is almost linear in all countries, in contrast to the results obtained in the SEIR model. This raises three questions:

1. If the growth of the epidemic is linear, why can't we still predict anything with confidence for a month or a year ahead?

2. Why does the growth rate of the epidemic slows down to linear?

3. How should we change the classical SEIR model to be more relevant?

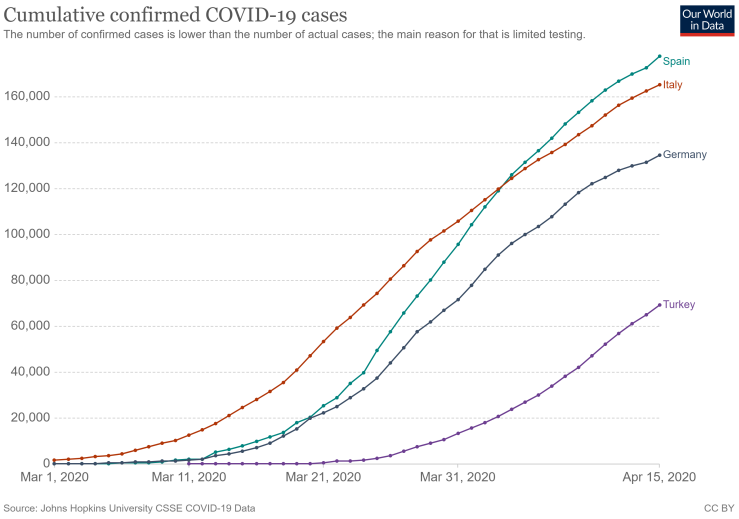


Figure 4.5: Cumulative Confirmed COVID-19 cases [5].

We start by answering the first question. When we predict something, we face with an unpleasant task: the data on which we build the model is not ideal, and the model built on its basis does not reflect the future. Moreover, the model that reflects the real situation is also very limited. The sudden development of an epidemic in a new big city, the use of a more effective method of treatment, a change in the way of collecting information - all these can introduce so many errors that a long-term forecast will be absolutely far from reality in the data.

4.4 Modifying the SEIR model

Let us try to answer the question why the growth of the epidemic slows down to linear. With the same number of infected people, the scale of the population starts to play a significant role associated with the limited social interaction. More precisely, remember that the number of cases in the SEIR model is directly proportional to the average number of cases I/N in the population. The model works well in small populations, where everyone can communicate with each other, and the infected individuals are evenly distributed. In reality, especially on a large scale population, if you take two accidental infected people, it turns out that they have never talked or seen each other, and in general, they live in different cities. All that unites them is the social connections leading to the transmission of the virus.

We make a geometric interpretation to prove the linear epidemic growth theorem in a large population. Let the graph of social connections be represented in the form of d dimensional lattice. In reality, with an average number of 20 social contacts daily [16], the dimension can be roughly estimated as $d \approx 4$. Each carrier n of the infection generates a d dimensional hypercube growing around it from secondary infected. The edge of the cube has a length of $n^{\frac{1}{d}}$, and if the contact with an infected person leads to illness with probability of P, then every day each edge is lengthened on average by $2P$. Thus, we obtain a growth model expressed by the recurrence relation as

$$n_{k+1} = (n_k^{\frac{1}{d}} + 2P)^d. \tag{4.2}$$

Imagine n as a function of time and denote $n_k = n(t_k), n_{k+1} = n(t_{k+1}), \ldots$. Differentiating the previous expression with respect to time, we get:

$$n'_{k+1} = n'_k \left(1 + \frac{2P}{n_k^{1/d}}\right)^{d-1}. \tag{4.3}$$

It can be seen that for large values of n_k derivatives $n'_{k+1} \approx n'_k$ are approximately equal, hence the growth is linear.

Turning to the world statistics on the epidemic shown in Figure 4.6, the same thing can be seen: after exponential growth over a short period of time, the growth of the epidemic is linear starting from April, despite the fact that the classical models promise us a continuation of the exponential rise in the number of cases.

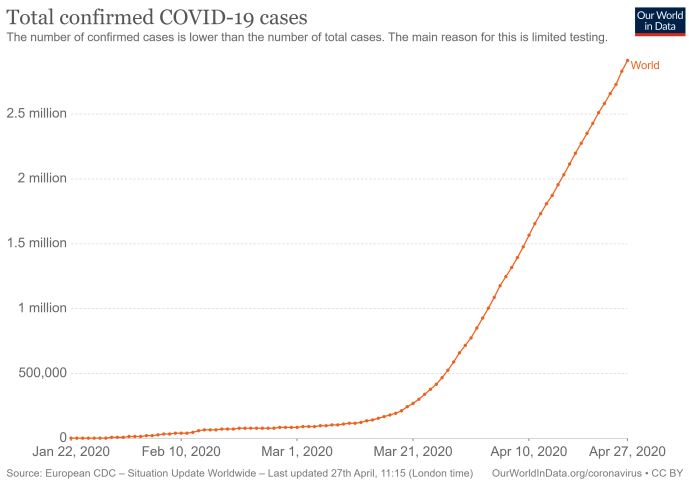


Figure 4.6: Total Confirmed COVID-19 cases [5].

Now let us modify the classical SEIR model. It can be immediately seen that the SEIR model is not very suitable for describing COVID-19, because the hidden carriers of the infection E are not contagious in this model. This shortage can be corrected by introducing an additional parameter θ characterizing the ratio of the transmission ability of the exposed to the infected compartments [32]. The corresponding parts of the model can then be expressed as:

$$\begin{aligned} \frac{dS}{dt} &= -\beta \frac{S(I+\theta E)}{N}, \\ \frac{dE}{dt} &= \beta \frac{S(I+\theta E)}{N} - \alpha E. \end{aligned} \tag{4.4}$$

When $\theta = 0$, the infection ability of the patients in the latent period is ignored. When $\theta = 1$, it implies that the infection ability of the exposed individuals is the same as the infected individuals. Additionally, in order to incorporate public measures and restrictions, we assume that the model parameters have exponentially decreasing behaviour. Therefore, we express the model parameters with these natural conditions:

$$\begin{aligned} \beta(t) &= \beta_0 e^{-\mu t}, && 0 < \beta_0 < 1 \text{ and } 0 \leq \mu, \\ \theta(t) &= \theta_0 e^{-\xi t}, && 0 < \theta_0 < 1 \text{ and } 0 \leq \xi, \\ \alpha(t) &= \alpha_0 e^{-\sigma t}, && 0 < \alpha_0 < 1 \text{ and } 0 \leq \sigma, \\ \gamma(t) &= \gamma_0 e^{-\eta t}, && 0 < \gamma_0 < 1 \text{ and } 0 \leq \eta, \\ \delta(t) &= \delta_0 e^{-\rho t}, && 0 < \delta_0 < 1 \text{ and } 0 \leq \rho, \end{aligned} \tag{4.5}$$

where β_0, μ, θ_0, ξ, α_0, σ, γ_0, η, δ_0, and ρ are the parameters that can be obtained by

fitting the model to the observations. Including the dead (D) compartment in the modified SEIR model and replacing the constant model parameters by time-dependent parameters in (4.5), we propose a modified version of the SEIRD model for the current COVID-19 epidemic which can be expressed as:

$$\begin{aligned}
\frac{dS}{dt} &= -\beta(t)\frac{S(I+\theta(t)E)}{N}, \\
\frac{dE}{dt} &= \beta(t)\frac{S(I+\theta(t)E)}{N} - \alpha(t)E, \\
\frac{dI}{dt} &= \alpha(t)E - \gamma(t)I - \delta(t)I, \\
\frac{dR}{dt} &= \gamma(t)I, \\
\frac{dD}{dt} &= \delta(t)I,
\end{aligned} \tag{4.6}$$

with initial conditions $(S(t_0), E(t_0), I(t_0), R(t_0), D(t_0)) = (S_0, E_0, I_0, R_0, D_0)$ for some initial time t_0. Furthermore, as the initial condition for the E compartment is unavailable, we assume that

$$E_0 = qI_0, \quad 0 \leq q, \tag{4.7}$$

holds true for some parameter q to be optimized later. The flowchart of the modified SEIRD model is given in Figure 4.7. While developing the proposed model, we also make the following assumptions:

1. Total population (N) is constant throughout the simulation of the model.
2. Zoonitic transmission is not considered.
3. The proposed model assumes the continuation of existing control measures such as lockdown and quarantines and does not consider a scenario of the second surge of the epidemic.

The parameters of the modified SEIRD model are determined by solving the non linear least square problem with positive constraint. In this model, we define the state variables as $\mathbf{u}(t) = (S(t), E(t), I(t), R(t), D(t))$ depending on the vector of parameters $\mathbf{x} = (\beta_0, \mu, \theta_0, \xi, \alpha_0, \sigma, \gamma_0, \eta, \delta_0, \rho, q)^T$, and the observed data given at times $t_i, i = 1, 2, \ldots, n$ as the vector $\mathbf{y}_i = (\mathcal{I}_i, \mathcal{R}_i, \mathcal{D}_i)$ composed of infected, recovered,

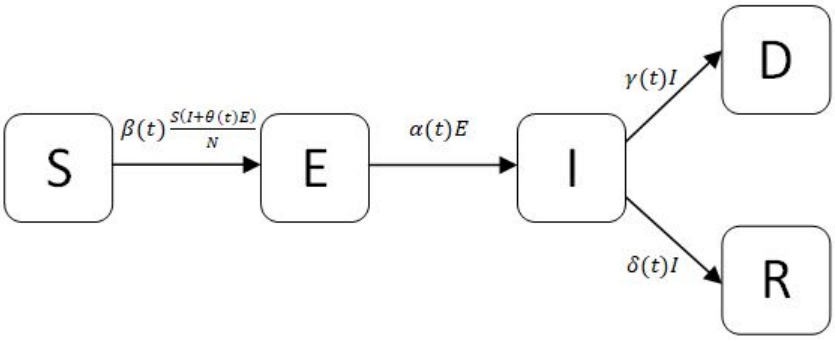


Figure 4.7: Flowchart of the modified SEIRD model.

and dead cases, respectively. Let $F_i(\mathbf{u}, \mathbf{x}) = (I_i, R_i, D_i)$ be the function computing the numerical solution $\mathbf{u}$ of the system of differential equations (4.6) for infected, recovered, and dead cases at time i, then the objective function of our model with the weights w_1, w_2, and w_3 takes the form as:

$$\min f(\mathbf{x}) = w_1 \sum_{i=1}^{n} (I_i - \mathcal{I}_i)^2 + w_2 \sum_{i=1}^{n} (R_i - \mathcal{R}_i)^2 + w_3 \sum_{i=1}^{n} (D_i - \mathcal{D}_i)^2, \qquad (4.8)$$

where n is the number of observations.

4.5 Computation of the Daily Reproduction Number of the Proposed Model

The basic reproduction number $\mathcal{R}_0$ is the most important parameter to determine dynamics of COVID-19. In real world scenarios, with development of the epidemic, more and more public measures are taken to control the spread which gradually reduces the reproduction number. Thus, we try to give a formulation of $\mathcal{R}_0$ for the modified SEIRD model. Firstly, we assume that all parameters in the system are constant and we apply the results of Diekmann, Heesterbeek and Metz [17] and Van den Driessche and Watmough [34]. The disease-free equilibrium of the modified SEIRD model is $(S^*, E^*, I^*, R^*, D^*) = (N, 0, 0, 0, 0)$ and the infectious part of the system is determined by the following equations:

$$\begin{aligned} \frac{dE}{dt} &= \beta \frac{S(I + \theta E)}{N} - \alpha E, \\ \frac{dI}{dt} &= \alpha E - \gamma I - \delta I. \end{aligned} \qquad (4.9)$$

Then, the next generation matrix G is determined as

$$G = FV^{-1} = \begin{pmatrix} \frac{\beta}{\gamma+\theta} + \frac{\beta\theta}{\alpha} & \frac{\beta}{\delta+\gamma} \\ 0 & 0 \end{pmatrix}, \tag{4.10}$$

where

$$F = \begin{pmatrix} \beta\theta & \beta \\ 0 & 0 \end{pmatrix}, \quad \text{and} \quad V = \begin{pmatrix} \alpha & 0 \\ -\alpha & \gamma+\delta \end{pmatrix}. \tag{4.11}$$

As a result, $\mathcal{R}_0$, the basic reproduction number, is given by the dominant eigenvalue of G as

$$\mathcal{R}_0 = \frac{\beta}{\gamma+\delta} + \frac{\beta\theta}{\alpha}. \tag{4.12}$$

Here, the first term represents the contributions from the infectious compartment I as in the classical SEIRD model, wheras the second term represents the contributions from the mildly infectious E compartment in the modified SEIRD model, which results in an increase in the value of $\mathcal{R}_0$.

After deriving the $\mathcal{R}_0$ of the proposed model, we follow the works of de León et al. [15] and Tang et al. [33], and we replace the constant parameters of the model with the time-dependent versions of them given in Equation (4.5). Then, we define

$$\mathcal{R}_d(t) = \frac{\beta(t)}{\gamma(t)+\delta(t)} + \frac{\beta(t)\theta(t)}{\alpha(t)} \tag{4.13}$$

as the effective daily reproduction ratio, to measure the 'daily reproduction number', the number of new infections produced by a single infected individual per day.

We plot $\mathcal{R}_d(t)$ vs t for some sets of parameters given in Figure 4.8. As expected, it decreases exponentially behaving similar to time-dependent transmission rate given in the same figure.

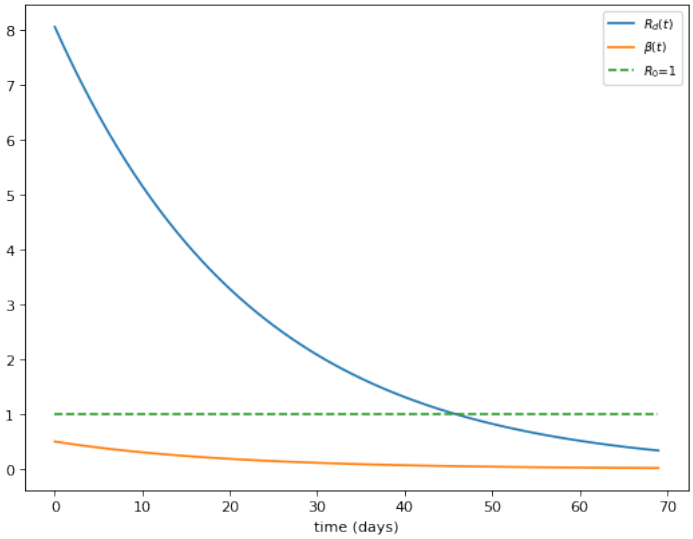


Figure 4.8: $\mathcal{R}_d(t)$ vs t and $\beta(t)$ vs t for $\beta_0 = 0.5$, $\mu = 0.05$, $\theta_0 = 0.6$, $\xi = 0.$, $\alpha_0 = 0.33$, $\sigma = 0.$, $\gamma_0 = 0.05$, $\eta = 0.001$, $\delta_0 = 0.02$, and $\rho = 0.02$.

CHAPTER 5

RESULTS AND DISCUSSION

In this chapter, we present the results obtained by a simulation of the proposed model. All the model parameters are estimated from fitting of the model to the available data. We use equal weights to minimize the objective function throughout the simulation. The governing differential equations are solved by RK45 method provided by SciPy with given initial conditions. It is observed that the prediction results are highly dependent on number of priori dataset. For this reason, different subsets of measurements on COVID-19 data of Italy are fitted to the modified SEIRD model for identification of the model parameters. Then, we simulate the model with optimal values of parameters up to days when daily confirmed cases decline significantly. We also test the model for other countries to collect some useful information about the parameters, the evolutions of active and confirmed cases, the peak values of infected cases, and the estimation of the effective daily reproduction ratio. The flowchart of the procedure is given in Figure 5.1.

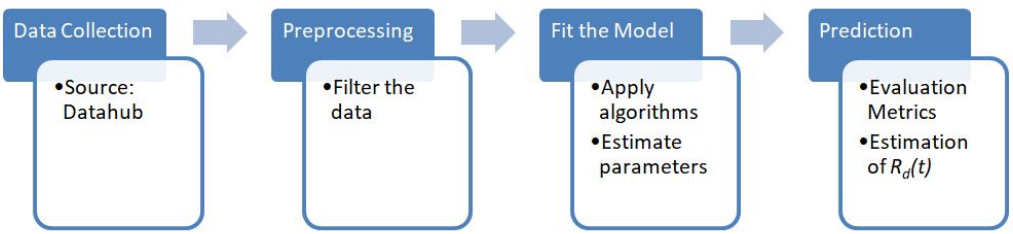


Figure 5.1: Flowchart of the research process.

Evaluation metrics are very important to understand the performance of a model. There are many different evaluation metrics, however, for the purpose of this thesis, we use only Root-Mean-Square Error (RMSE) and R-Square.

RMSE is the standard deviation of the prediction errors and a measure of the goodness of fit. It is commonly used in regression problems to verify the experimental results. Mathematically, it is expressed as

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2},$$

where y_i is the i^{th} observed value in the dataset, $\hat{y}_i$ is the i^{th} predicted value from the model, and N is the number of observations in the dataset.

R-Square is another metric to calculate the relative error. It is a statistical measure that represents the proportion of the variance for a dependent variable that can be explained by the model. Mathematically,

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2},$$

where $\bar{y}$ is the average value of the observed data.

5.1 Validation Against COVID-19 Data of Italy

It is important to validate the model with real data before applying for prediction of an epidemic evolution. For this reason, validation is carried out by using the modified SEIRD model and taking different subsets of measurements from March 1 to May 9 of the year 2020:

- **S1** 30 days measurements: 01/03/2020-30/03/2020
- **S2** 40 days measurements: 01/03/2020-09/04/2020
- **S3** 50 days measurements: 01/03/2020-19/04/2020

As of May 9, the daily confirmed cases declined significantly, so we set it as ending point of our model.

L-BFGS-B, TNC, Trust-Region Constrained, and Trust Region Reflective (TRF) optimization algorithms are used to fit the model to the observation data. RMSE values for different cases using three subsets of measurements are given in Table 5.1.

Table 5.1: RMSE values for different cases in Italy.

	RMSE		
Algorithms	**Infected**	**Recovered**	**Dead**
L-BFGS-B	806.37	280.73	110.35
TNC	673.56	323.77	2854.93
Trust-Constr	1153.68	221.64	84.56
TRF	659.16	227.85	105.47

(a) 30 days fitting.

	RMSE		
Algorithms	**Infected**	**Recovered**	**Dead**
L-BFGS-B	734.50	292.62	279.30
TNC	893.56	555.86	3023.35
Trust-Constr	797.47	289.73	295.78
TRF	860.40	295.26	289.60

(b) 40 days fitting.

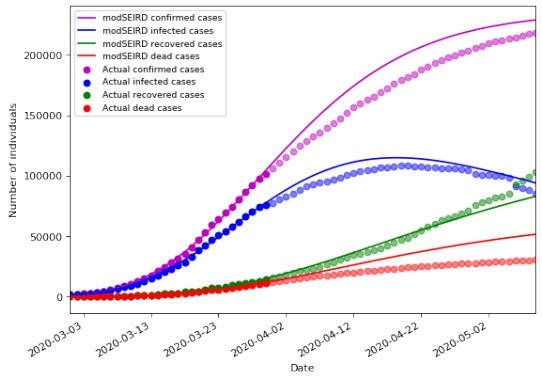
	RMSE		
Algorithms	**Infected**	**Recovered**	**Dead**
L-BFGS-B	1264.49	512.65	385.68
TNC	1722.65	1193.72	4427.96
Trust-Constr	1130.55	512.27	363.65
TRF	1223.02	511.56	372.94

(c) 50 days fitting.

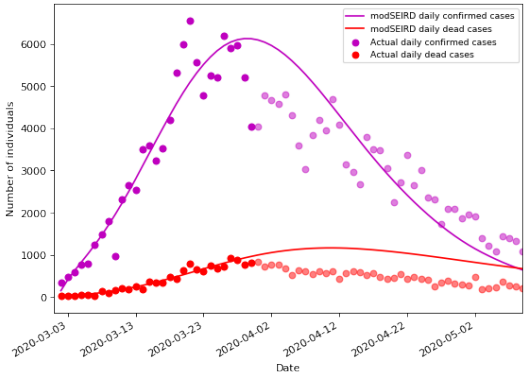
As expected, RMSE values increase as we add more data points for training. TNC method shows poor performance and does not converge to solution. RMSE values obtained for infected, recovered and dead cases using L-BFGS-B, Trust-Region Constrained and Trust Region Reflective methods are comparable. But, we prefer to use L-BFGS-B method for the rest of numerical simulations, since it converges to solution faster and uses less memory than Trust-Region Constrained method. Additionally, the optimal values of parameters obtained by L-BFGS-B method are more meaningful than the values obtained by Trust-Region Constrained and Trust Region Reflective methods.

By updating our model dynamically using different subsets of measurements, we

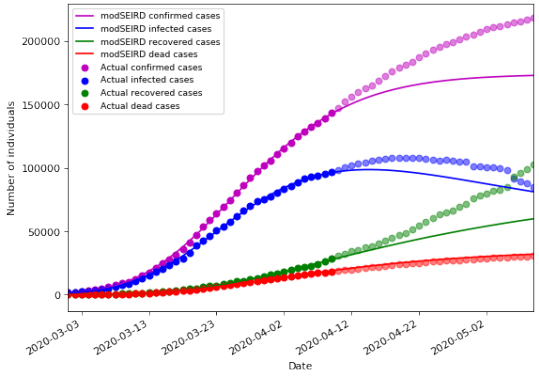
make predictions until the ending point of our model. The simulation and prediction results of the modified SEIRD model using L-BFGS-B algorithm are illustrated in Figure 5.2.



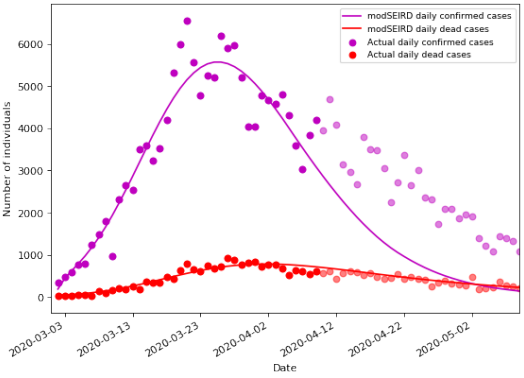
(a) Simulations for confirmed, infected, recovered and dead cases using 30 days training data.



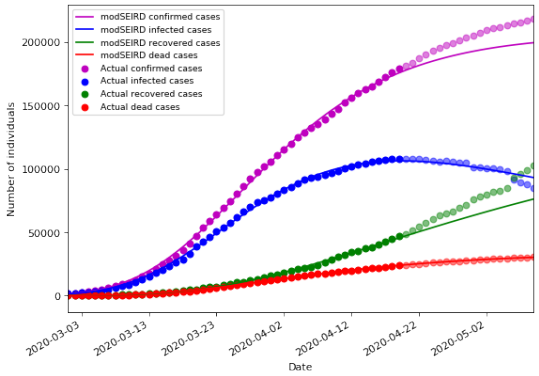
(b) Simulations for daily confirmed and dead cases using 30 days training data.



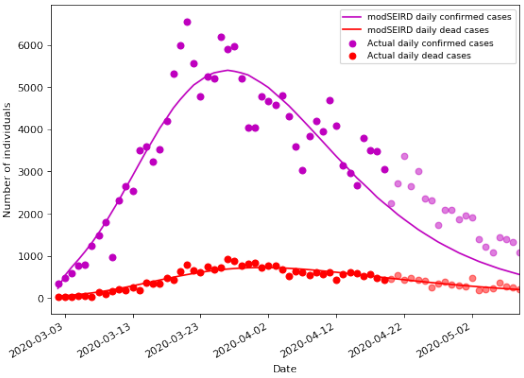
(c) Simulations for confirmed, infected, recovered and dead cases using 40 days training data.



(d) Simulations for daily confirmed and dead cases using 40 days training data.



(e) Simulations for confirmed, infected, recovered and dead cases using 50 days training data.



(f) Simulations for daily confirmed and dead cases using 50 days training data.

Figure 5.2: Simulation and prediction results for different cases in Italy.

As can be seen, the model perfectly fits three different subsets of measurements for

trained data, but the predictions slightly differ. The model predicted on 30 March slightly overestimates the confirmed cases. In the second subsets of measurements, the model underestimates the cases, since the infected cases show a sudden increase at the prediction date. However, the model trained at the late stage of the epidemic is capable of fitting on previous data and also predict the development of epidemic accurately. The evolutions of cumulative confirmed and infected cases for the model trained at this stage are compared in Table 5.2.

Table 5.2: Comparison of cumulative confirmed and infected cases for Italy using 50 days trained data.

	Confirmed cases		**Infected cases**	
Days	**Actual data**	**modSEIRD**	**Actual data**	**modSEIRD**
1	1694	1694	1577	1577
10	10149	11867	8794	10041
20	47021	47644	38549	38134
30	101739	100153	75528	74511
40	143626	146111	96877	98960
50	178972	175664	108257	106321
60	203591	191674	104657	102267
70	218268	199591	84842	93116

The turning point of the model trained at the late stage is March 27. Beginning on March 1, Italian government implemented extensive shutdown measures, which were imposed to all of Italy on March 10. These measures started to reduce the number of reported daily cases approximately 3 weeks later. The model exactly estimates time to reach the peak of infected cases. According to the model, the infected cases show a peak on April 19 with about 106321 infected person while the true value is 108257. RMSE and R-Square values of confirmed, infected, recovered, and dead cases for 20 days predictions are given in Table 5.3.

Table 5.3: RMSE and R-Square values of confirmed, infected, recovered, and dead cases for 20 days predictions using the model trained at the late stage of the epidemic in Italy.

Cases	**RMSE**	**R-Square**	**% Error**
Confirmed	12759.21	-0.30	<8.55
Infected	3144.71	0.79	<9.75
Recovered	13895.38	0.21	<26.
Dead	105.69	0.99	<0.52

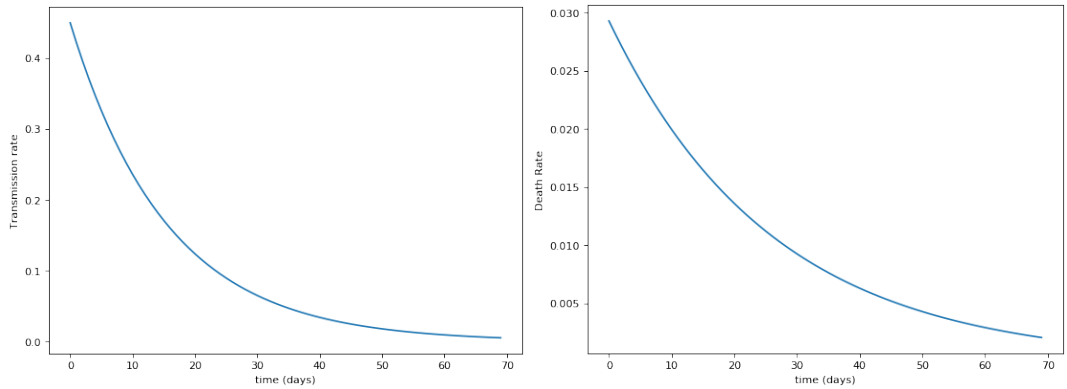
From the table, it can be seen that the model can not predict the confirmed cases accurately, since there is a sudden increase in recovered cases. Overall, the model learned at this stage is capable of predicting the infected and dead cases fairly close to the real scenario.

The values of optimized parameters of the model are given in Table 5.4. The parameters learned at this stage can accurately reflect dynamics of COVID-19. It can be seen that the incubation rate is found to be constant by the optimizer.

Table 5.4: Optimized parameters for Italy.

Parameter	**Value**
β_0	4.49×10^{-1}
μ	6.45×10^{-2}
θ_0	1.91×10^{-1}
ξ	8.12×10^{-1}
α_0	8.21×10^{-1}
σ	0.
γ_0	1.84×10^{-2}
η	3.37×10^{-3}
δ_0	2.92×10^{-2}
ρ	3.83×10^{-2}
q	6.65×10^{-8}

Next, we simulate the time-dependent transmission, death and effective daily reproduction ratio for 70 days in order to collect some meaningful quantitative information about the dynamics of COVID-19. The illustration of time-dependent transmission and death rate for Italy is given in Figure 5.3



(a) (b)

Figure 5.3: Estimation of the (a) transmission rate and (b) death rate for Italy.

The infection rate and death rate at the beginning of restrictions are found to be $\beta(0) = 0.449$ and $\delta(0) = 0.029$, respectively. At the end of the time period, the infection and death rate reduce significantly because of the strong public measures taken by Italian government. If we look at the effective daily reproduction ratio estimation given in Figure 5.4, it can be observed that $\mathcal{R}_d(t)$ is 9.51 at the beginning of epidemic. It exponentially decreases to below 1 after 47 days.

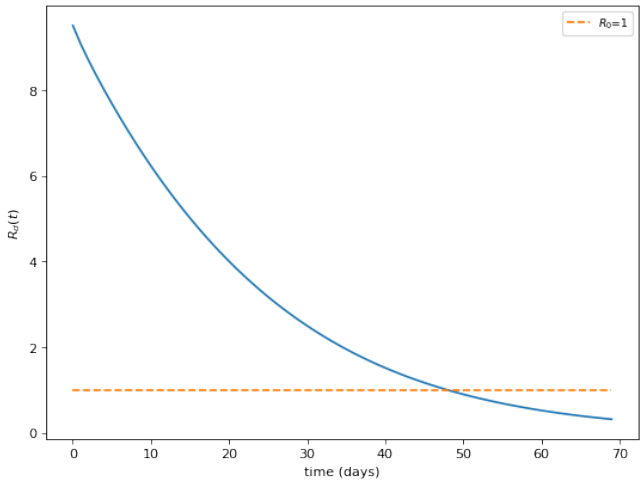


Figure 5.4: Estimation of $\mathcal{R}_d(t)$ for Italy.

We plot $\mathcal{R}_d(t)$ vs t for different intensities of the transmission rate by taking other parameters to be optimal values. If we consider the parameter μ as the intensity rate of the public measures, we can see the importance of this value in the evolution of epidemic for Italy in Figure 5.5.

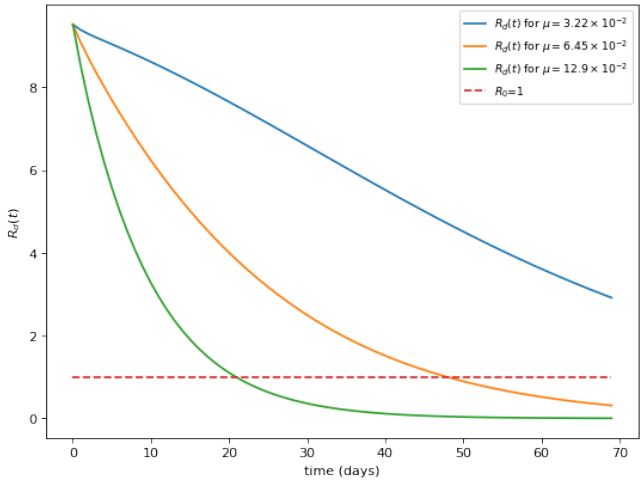


Figure 5.5: $\mathcal{R}_d(t)$ estimation for different values of μ in Italy by taking other parameter values of the model to be optimal.

From the figure, it can be seen that by increasing the rate of restrictions twice, the effective daily reproduction ratio can be reduced to below 1 about 4 weeks earlier for Italy. Otherwise, softening these restrictions would cause an increment in daily cases, and the epidemic would die too late.

5.2 The Modified SEIRD Model for Turkey

Since the government of Turkey started to take public measures starting from mid-March, we set 20 March as an initial point of our model. As of 3 May, the daily confirmed cases reduced significantly, we set this as the ending point of our model. We train our model at the late stage of epidemic using 35 days observation data. The simulation and prediction results using L-BFGS-B optimization algorithm are illustrated in Figure 5.6 and Figure 5.7.

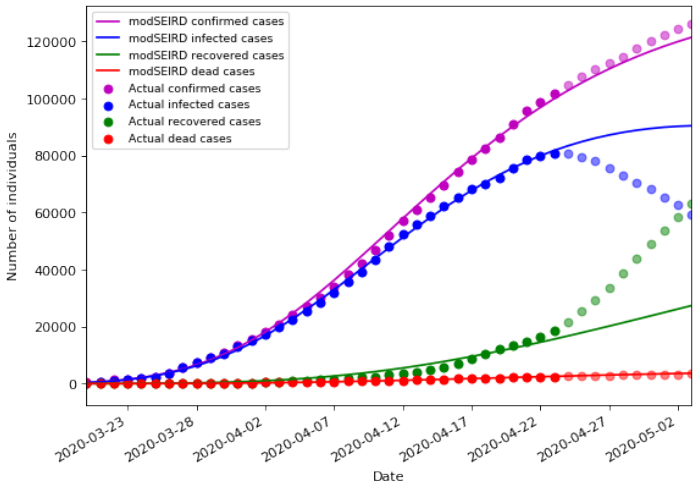


Figure 5.6: Simulation and prediction results for confirmed, infected, recovered and dead cases using 35 days training data in Turkey.

The model interpolates well to the training data. If we look at RMSE and R-Square values for different cases given in Table 5.5, it can be observed that RMSE values for different cases are reduced and R-Square is improved significantly compared to the classical SEIR model with constant coefficients.

From the Figure 5.7, it can be seen that the turning point of the model is 13 April. Beginning on March 16, the government of Turkey started to make partial shutdowns, which were extended to all of country on March 23. These measures started to reduce

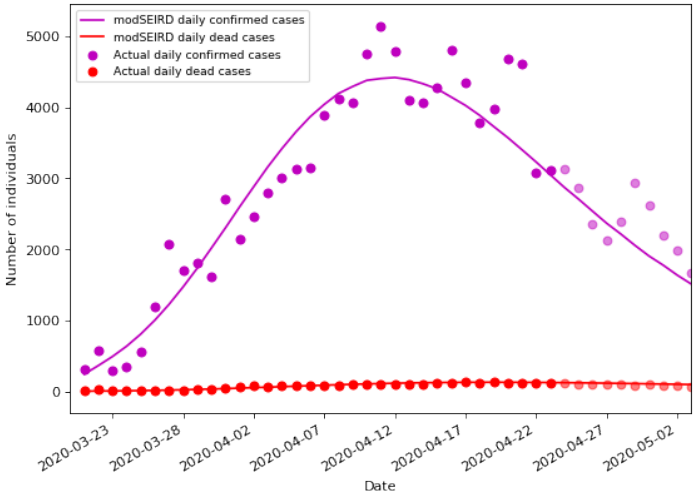


Figure 5.7: Simulation and prediction results for daily confirmed and dead cases using 35 days training data in Turkey.

Table 5.5: RMSE and R-Square values of confirmed, infected, recovered, and dead cases for 35 days fitting in Turkey.

Cases	**RMSE**	**R-Square**
Confirmed	1203.08	0.99
Infected	711.00	0.99
Recovered	1222.97	0.94
Dead	46.39	0.99

daily reported cases about 4 weeks later. The model overestimates time to the outbreak peak. Since there is a sudden change in the number of infected and recovered individuals, the model is not capable of predicting the development of these cases accurately. RMSE and R-Square values of different cases for 10 days predictions can be seen in Table 5.6.

Table 5.6: RMSE and R-Square values of different cases for 10 days predictions in Turkey.

Cases	**RMSE**	**R-Square**	**% Error**
Confirmed	3167.90	0.78	<3.61
Infected	18921.71	-6.46	<52.03
Recovered	22169.57	-1.69	<56.64
Dead	162.62	0.60	<7.39

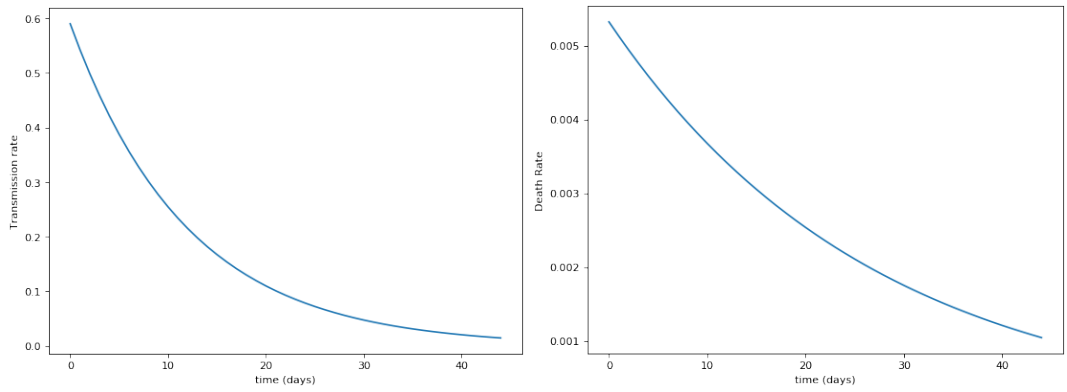
The model predicts the confirmed cases with less than 3.61% error compared to the actual numbers. However, the model overestimates the infected cases and underestimates the recovered cases which results in high percent error with the actual numbers.

The values of optimized parameters of the model trained at this stage of epidemic are given in Table 5.7:

Table 5.7: Optimized parameters for Turkey.

Parameter	**Value**
β_0	5.89×10^{-1}
μ	8.40×10^{-2}
θ_0	9.99×10^{-1}
ξ	4.39×10^{-1}
α_0	8.96×10^{-1}
σ	0.
γ_0	1.34×10^{-2}
η	0.
δ_0	5.32×10^{-3}
ρ	3.70×10^{-2}
q	5.56×10^{-1}

From the table, it can be seen that the incubation and recovery rate are found to be constant by the optimizer. At the beginning of lockdown, the transmission and death rate for Turkey are found to be $\beta(0) = 0.589$ and $\delta(0) = 0.0053$, respectively. The illustrations of the transmission and death rate for 45 days are shown in Figure 5.8.



(a) (b)

Figure 5.8: Estimation of the (a) transmission rate and (b) death rate for Turkey.

In Figure 5.9, the effective daily reproduction ratio estimation for Turkey is shown.

At the beginning of the containment phase of an outbreak, the effective daily reproduction ratio is found to be 32.04. This is much higher than the daily reproduction numbers obtained for Italy and Spain. The model predicts that this value is reduced

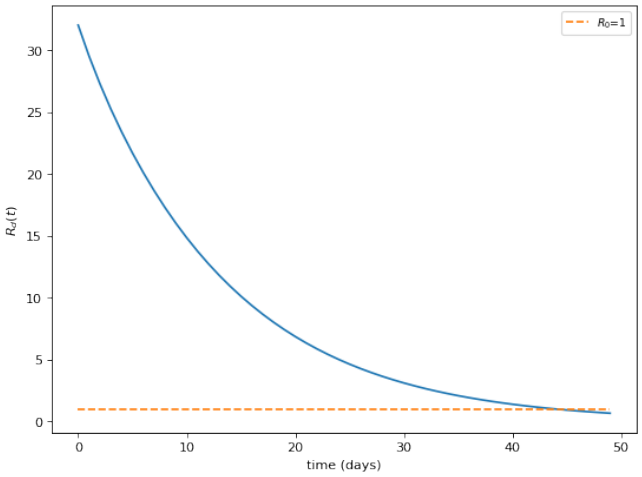


Figure 5.9: Estimation of $\mathcal{R}_d(t)$ for Turkey.

to below 1 after 44 days by implementing public measures at the same intensity.

5.3 The Modified SEIRD Model for Spain

We set March 10 as the initial point of our model. As of May 18, the daily reported cases declined to single digits, so we set it as the ending point of our model. We train the model using 50 days training data by taking into account the late stage of epidemic. The simulation and prediction results for different cases can be seen in Figure 5.10 and Figure 5.11.

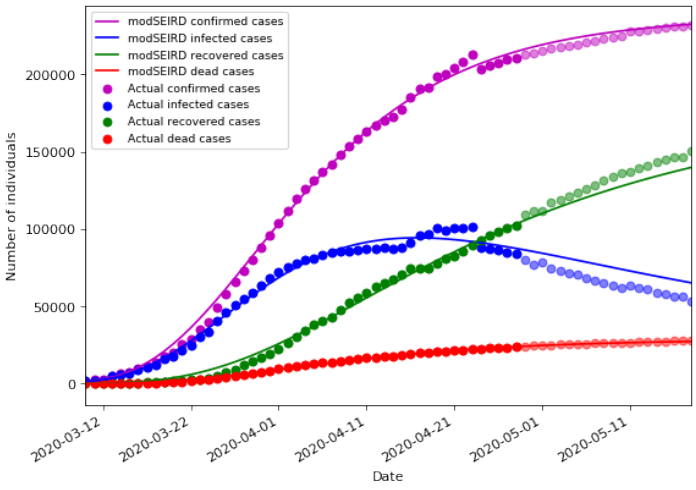


Figure 5.10: Simulation and prediction results for confirmed, infected, recovered and dead cases using 50 days training data in Spain.

From the Figure 5.10, it can be seen that there is a sudden development and decline

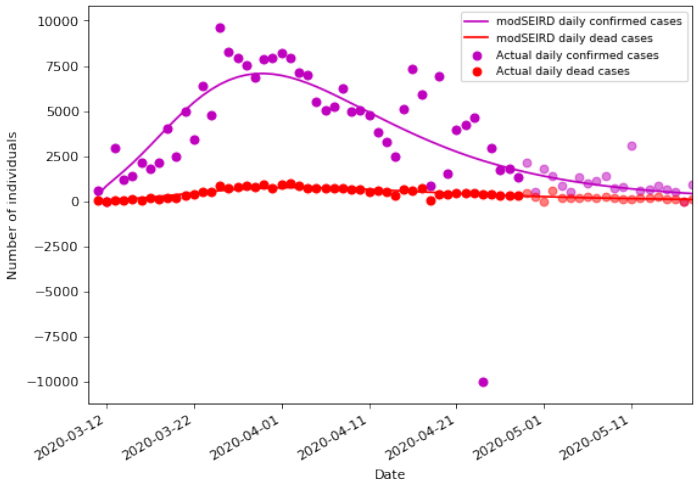


Figure 5.11: Simulation and prediction results for daily confirmed and dead cases using 50 days training data in Spain.

in the numbers of the infected and recovered cases which makes data noisy. However, the model tries to fit the training data accurately by minimizing this error. This can also be observed from RMSE and R-square values for different cases given in Table 5.8.

Table 5.8: RMSE and R-Square values of confirmed, infected, recovered, and dead cases for 50 days fitting in Spain.

Cases	**RMSE**	**R-Square**
Confirmed	3159.53	0.99
Infected	3622.32	0.98
Recovered	2527.68	0.99
Dead	383.78	0.99

From the Figure 5.11, it can be observed that the turning point of the model is March 29. Beginning on March 10, the Spanish government implemented partial lockdown measures which were extended to all country on March 14. These measures took effect in a daily reports about 2 weeks later. The model estimates peak infected cases on April 16 with an error of 7% compared with the real number. Since the data is too noisy, the model is not capable of predicting the epidemic development accurately. The performance of the model for 20 days predictions is described in Table 5.9.

The model predicts the infected cases poorly, since there is sudden change that is very difficult to describe its variability. But, the predicted values of confirmed cases are less than an error of 0.62% compared to the actual numbers.

Table 5.9: RMSE and R-Square values of different cases for 20 days predictions in Spain.

Cases	**RMSE**	**R-Square**	**% Error**
Confirmed	2129.95	0.87	<0.62
Infected	9217.76	-0.46	<20.52
Recovered	7694.12	0.62	<7.57
Dead	701.22	0.57	<3.73

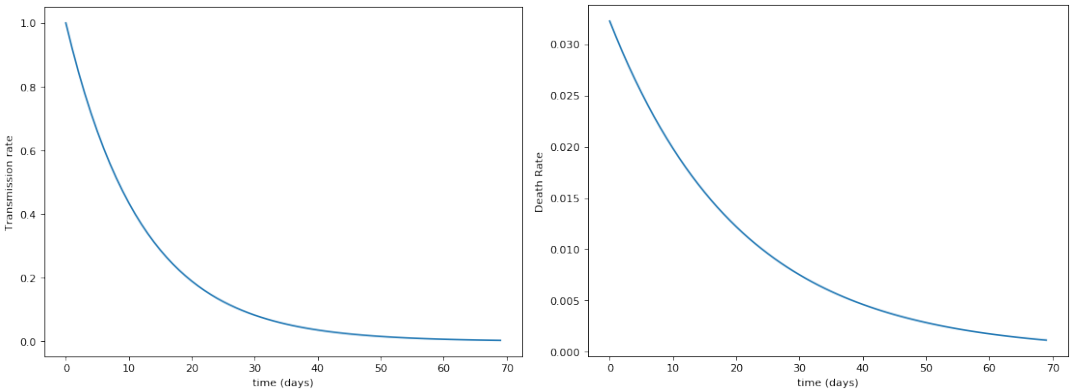
Optimized parameter values of the model using 50 days training data are given in Table 5.10.

Table 5.10: Optimized parameters for Spain.

Parameter	**Value**
β_0	9.99×10^{-1}
μ	8.32×10^{-2}
θ_0	2.91×10^{-1}
ξ	9.95×10^{-1}
α_0	4.54×10^{-1}
σ	7.07×10^{-2}
γ_0	5.23×10^{-2}
η	1.35×10^{-2}
δ_0	3.22×10^{-2}
ρ	4.86×10^{-2}
q	8.29×10^{-5}

Transmission rate of the infection for Spain is much higher than the rates of previous two countries at the beginning of quarantine measures. The mortality rate is found to be $\delta(0) = 0.032$ which is slightly higher than the value obtained for Italy. In Figure 5.12, we plot the time-dependent transmission and death rate estimations for Spain.

Figure 5.13 shows the estimated effective daily reproduction ratio for Spain. At the beginning of lockdown, the daily reproduction number is 12.46 which is higher than the value obtained for Italy. According to the model, this ratio is efficiently reduced to below 1 after 40 days. This implies that Spanish government implemented public measures, like isolation, quarantine, and public closings too early and extensively for the first wave of COVID-19 epidemic.



(a) (b)

Figure 5.12: Estimation of the (a) transmission rate and (b) death rate for Spain.

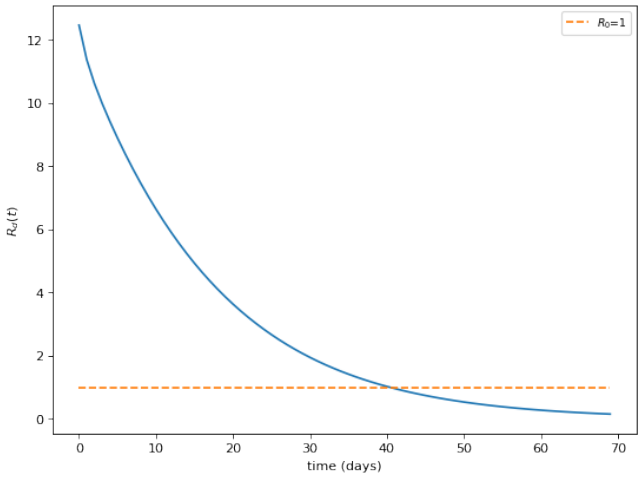


Figure 5.13: Estimation of $\mathcal{R}_d(t)$ for Spain.

CHAPTER 6

CONCLUSION AND FUTURE WORK

We have proposed a fully new modified version of the SEIRD model to predict COVID-19 epidemic evolution in particular countries. Our model takes into account the exposed individuals as contagious and incorporates government and social distancing measures through the time-dependent parameters. We consider the closed population in our model, and the birth, mortality and others are not taken into account. Overall, our proposed model tries to capture the COVID-19 transmission dynamics accurately. The model can be predictive when the implementation of social measures begins to ameliorate the epidemic. During this phase, we use time-dependent parameters to model these measures. If these measures are taken too early and extensively, then the epidemics can pass the turning point and result in a subsequent reduction in daily reported cases. On the other hand, if the countries reduce these measures too early, the cumulative reported cases may not flatten but instead continue growing linearly at a low rate. In this case, our model should be updated with new optimized parameters in order to predict the next surge of the epidemic.

Another limitation of this thesis is the untimely analysis, which is an analysis based on data retrieved up to the late stage of epidemic. If we look at the current data, the trend of COVID-19 cases is slightly different from the predictions made in this paper. However, the results of this thesis show that the country-specific characteristics of a transmission rate of the infection and effective daily reproduction number which depends on time.

As a future scope, population density instead of homogenous population can be considered to develop the model. Hospitalization rates, use of intensive care units (ICUs),

unreported cases and others can be included to get more detailed analysis of COVID-19. Additionally, all the parameters can be thought as time and compartment dependent functions to get more accurate results.

REFERENCES

[1] Coronavirus disease 2019 (COVID-19)—symptoms and causes, Mayo Clinic, accessed 1 August 2021, https://www.mayoclinic.org/diseases-conditions/coronavirus/symptoms-causes/syc-20479963.

[2] Coronavirus disease (COVID-19) pandemic, accessed 1 August 2021, Online Available at : https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

[3] Novel Coronavirus 2019, accessed 1 August 2021, Online Available at: https://datahub.io/core/covid-19.

[4] Optimization, https://docs.scipy.org/doc/scipy/reference/tutorial/optimize.html.

[5] Our world in data, accessed 1 August 2021, Online Available at: https://ourworldindata.org/.

[6] Worldometer, accessed 1 August 2021, Online Available at: https://www.worldometers.info/coronavirus/.

[7] C. Anastassopoulou, L. Russo, A. Tsakris, and C. Siettos, Data-based analysis, modelling and forecasting of the COVID-19 outbreak, PLOS One, 15(3), p. e0230405, 2020.

[8] M. A. Branch, T. F. Coleman, and Y. Li, A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems, SIAM Journal on Scientific Computing, 21(1), pp. 1–23, 1999.

[9] R. H. Byrd, M. E. Hribar, and J. Nocedal, An interior point algorithm for large-scale nonlinear programming, SIAM Journal on Optimization, 9(4), pp. 877–900, 1999.

[10] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, A limited memory algorithm for bound constrained optimization, SIAM Journal on Scientific Computing, 16(5), pp. 1190–1208, 1995.

[11] G. C. Calafiore, C. Novara, and C. Possieri, A modified SIR model for the COVID-19 contagion in Italy, In: 2020 59th IEEE Conference on Decision and Control (CDC), IEEE, pp. 3889–3894, 2020.

[12] R. Casagrandi, L. Bolzoni, S. A. Levin, and V. Andreasen, The SIRC model and influenza A, Mathematical Biosciences, 200(2), pp. 152–169, 2006.

[13] G. Chowell, Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts, Infectious Disease Modelling, 2(3), pp. 379–398, 2017.

[14] B. J. Coburn, B. G. Wagner, and S. Blower, Modeling influenza epidemics and pandemics: insights into the future of swine flu (H1N1), BMC Medicine, 7(1), pp. 1–8, 2009.

[15] U. A.-P. de León, Á. G. Pérez, and E. Avila-Vales, A data driven analysis and forecast of an SEIARD epidemic model for COVID-19 in Mexico, arXiv preprint arXiv:2004.08288, 2020.

[16] S. Y. Del Valle, J. M. Hyman, H. W. Hethcote, and S. G. Eubank, Mixing patterns between age groups in social networks, Social Networks, 29(4), pp. 539–554, 2007.

[17] O. Diekmann, J. A. P. Heesterbeek, and J. A. J. Metz, On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations, Journal of Mathematical Biology, 28(4), pp. 365–382, 1990.

[18] J. R. Dormand and P. J. Prince, A family of embedded Runge-Kutta formulae, Journal of Computational and Applied Mathematics, 6(1), pp. 19–26, 1980.

[19] D. Fanelli and F. Piazza, Analysis and forecast of COVID-19 spreading in China, Italy and France, Chaos, Solitons & Fractals, 134, p. 109761, 2020.

[20] N. E. Huang, F. Qiao, and K.-K. Tung, A data-driven model for predicting the course of COVID-19 epidemic with applications for China, Korea, Italy, Germany, Spain, UK and USA, medRxiv, 2020.

[21] W. O. Kermack and A. G. McKendrick, A contribution to the mathematical theory of epidemics, Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, 115(772), pp. 700–721, 1927.

[22] S. Khajanchi, K. Sarkar, and J. Mondal, Dynamics of the COVID-19 pandemic in India, arXiv preprint arXiv:2005.06286, 2020.

[23] A. J. Kucharski, T. W. Russell, C. Diamond, Y. Liu, J. Edmunds, S. Funk, and R. M. Eggo, Early dynamics of transmission and control of COVID-19: a mathematical modelling study, The Lancet Infectious Diseases, 20(5), pp. 553–558, 2020.

[24] M. Lalee, J. Nocedal, and T. Plantenga, On the implementation of an algorithm for large-scale equality constrained optimization, SIAM Journal on Optimization, 8(3), pp. 682–706, 1998.

[25] M. L. Michelsen, P. G. Thomsen, and M. R. Kristensen, Parameter estimation in nonlinear dynamical systems, Master's Thesis, Department of Chemical Engineering Technical University of Denmark, 2004.

[26] S. G. Nash, Newton-type minimization via the Lanczos method, SIAM Journal on Numerical Analysis, 21(4), pp. 770–788, 1984.

[27] J. Nocedal, Updating quasi-Newton matrices with limited storage, Mathematics of Computation, 35(151), pp. 773–782, 1980.

[28] J. Nocedal and S. Wright, *Numerical Optimization*, Springer Science & Business Media, 2006.

[29] G. Pandey, P. Chaudhary, R. Gupta, and S. Pal, SEIR and regression model based COVID-19 outbreak predictions in India, arXiv preprint arXiv:2004.00958, 2020.

[30] E. L. Piccolomini and F. Zama, Preliminary analysis of COVID-19 spread in Italy with an adaptive SEIRD model, arXiv preprint arXiv:2003.09909, 2020.

[31] D. Saikia, K. Bora, and M. P. Bora, COVID-19 outbreak in India: an SEIR model-based analysis, Nonlinear Dynamics, pp. 1–25, 2021.

[32] P. Shi, S. Cao, and P. Feng, SEIR transmission dynamics model of 2019 nCoV coronavirus with considering the weak infectious ability and changes in latency duration, MedRxiv, 2020.

[33] B. Tang, N. L. Bragazzi, Q. Li, S. Tang, Y. Xiao, and J. Wu, An updated estimation of the risk of transmission of the novel coronavirus (2019-nCov), Infectious Disease Modelling, 5, pp. 248–255, 2020.

[34] P. Van den Driessche and J. Watmough, Reproduction numbers and subthreshold endemic equilibria for compartmental models of disease transmission, Mathematical Biosciences, 180(1-2), pp. 29–48, 2002.

[35] J. T. Wu, K. Leung, and G. M. Leung, Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study, The Lancet, 395(10225), pp. 689–697, 2020.

[36] Z. Yang, Z. Zeng, K. Wang, S.-S. Wong, W. Liang, M. Zanin, P. Liu, X. Cao, Z. Gao, Z. Mai, J. Liang, X. Liu, S. Li, Y. Li, F. Ye, W. Guan, Y. Yang, F. Li, S. Luo, Y. Xie, B. Liu, Z. Wang, S. Zhang, Y. Wang, N. Zhong, and J. He, Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions, Journal of Thoracic Disease, 12(3), p. 165, 2020.

[37] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization, ACM Transactions on Mathematical Software (TOMS), 23(4), pp. 550–560, 1997.

[38] H. Zhu, L. Wei, and P. Niu, The novel coronavirus outbreak in Wuhan, China, Global Health Research and Policy, 5(1), pp. 1–3, 2020.