# ACIT5900

**MASTER THESIS**

**in**

## Applied Computer and Information Technology (ACIT)

**August 2021**

**Mathematical Modelling and Scientific Computing**

# History of epidemiological models and their application to Covid-19 data

**Aafreen Aafreen**

**Department of Computer Science**

**Faculty of Technology, Art and Design**

OSLOMET

# Abstract

The history of infectious diseases has always been important to us due to their impact on humanity. Infectious diseases lead to the emergence of epidemiological models that were essential to understand and find answers to end an epidemic. Since as early as 1760 human diseases were analyzed using so-called epidemiological models. The main objectives of this thesis are (i) to examine and review the history of these epidemiological models and (ii) to discuss the parameters which can be derived from them such as the reproduction number. In particular, I focus on the SIR model to understand qualitative features of the spread of the disease in specific cases, namely in Norway and the UK. I implement the SIR model as proposed in the paper "Estimation of Time-Dependent Reproduction Number for Global COVID-19 Outbreak" authors Petrova, T.; Soshnikov D.; Grunin A [18], additionally exploring the effect of the recovery rate. Finally, I compare the results.

# Table of Contents

# List of Figures



# List of Tables

# 1. Introduction and Background

An epidemic model is known to analyse the nature of the spread of infectious diseases and has been capturing everyone's attention including mine since the rise of the Covid- 19 pandemic and is the motivation behind this thesis. The main objective of this thesis is to examine and review these models and discuss their evolution and derive some useful information, if possible. In the review of models for disease spreading, we describe the compartmental models while highlighting the SIR model and further discuss the extensions of SIR models as well as the state-of-the-art models.

To understand the epidemic model we implement the simplest compartmental model known as the SIR model derived from the paper "Estimation of Time-Dependent Reproduction Number for Global COVID-19 Outbreak" by authors Petrova, T.; Soshnikov D.; Grunin A[18] and make comparisons with some real estimates available for public use and draw conclusions about effectiveness, limitations and relevance of the basic SIR model in today's pandemic crisis the world has been dealing with.

For centuries, infectious diseases have had a tremendous impact on us and since recorded human history it has been witnessed that epidemics caused many deaths, wiping out large numbers of human populations before vanishing and recurring again in the later years perhaps with less severity due to developed resistance over a period of time. This paved the way for the emergence of epidemiological models over time.

The "Spanish" flu epidemic for instance claimed more than 50 million lives all over the world between 1918-19. Another well-known epidemic in the history of infectious diseases known as black deaths spread from Asia to Europe in the year 1346 recurring several times in the 14th century, took over one-third of the entire population of Europe between the period 1346 - 1350. The plague reappeared again for another 300 years in several parts of Europe, notably known as the Great Plague of London between 1665-1666 before disappearing permanently. Many such infectious diseases appeared in the next hundreds of years to become epidemic or endemic, causing many deaths before they disappeared [1].

Endemic diseases have been incredibly common in developing countries, taking many lives every year due to diseases like measles, malaria, cholera and typhoid etc. In the 1980s, measles caused over 2,600,000 deaths annually but was reduced to 160,000, due to the relavant vaccine development. In 2011, according to the World Health Organization (WHO) annual reports there were about "1,400,000 deaths due to tuberculosis, 1,200,000 deaths due to HIV/AIDS, and 627,000 deaths due to malaria (but other sources estimate the number of malaria deaths to have been more than 1,000,000)" [1]

One of the most recent infectious diseases in the history of epidemics that killed millions of humans on a global scale is Covid-19. It was first witnessed to have emerged in Wuhan, China in December 2019 and then, later on, spread rapidly to the rest of the world. Today, it has been more than a year since the emergence of Coronavirus disease also known as Covid-19. Despite the mass vaccination rollout, we are still witnessing the virus reappearing again in the form of new variants. Although there is abundantly collected data available now from all around the world over the internet, with many experts analyzing the data, nations have been adopting different strategies such as lockdowns, social distancing, use of masks and development of vaccines. The question however remains about the effectiveness of the epidemic models used. Are they enough to cover the entire dynamics of the disease transmission?

The important role of modelling the data is to understand the way the disease spreads and to predict the course of the disease. Epidemic modelling has been an important tool that has been evolving throughout history exactly to do this. It all began with the first-ever understanding and discovery of living microscopic organisms that were responsible for the spread of any disease, dated back as early as (384 BCE–322 BCE) through Aristotle's writings in ancient Greece. The earliest record of disease analysis started through quantifying the number of deaths and documenting the causes behind it. The first mention of such a thing was found in a book called “Natural and Political Observations made upon the bills of Mortality” written by a scientist called John Graunt in 1662. While Gruant's work may have been the beginning of the quantification and analysis, the very first mathematician to demonstrate the nature of epidemiology through mathematical models was Daniel Bernouli through his work on “inoculating against smallpox” in 1760. The purpose of the creation of such a mathematical concept was to defend and prove that vaccination could improve life expectancy by almost 3 years. Bernoulli's work laid a foundation for what we know today about germ theory and mathematical models of epidemiology [2][1].

In the year 1906, it was W.H Hamer who first suggested ‘mass action law’ for the rate of new infections, and proposed that the number of individuals who are either susceptible and infected affects the rate at which the infection spreads. This is the concept that is relevant even to this day and served as the basis for all the mathematical models that evolved through time [1].

Another noteworthy work done in the field of malaria was by Dr Ross. In 1902 through his work on “The dynamics of the transmission of malaria between mosquitoes and humans”. He was even awarded the second Nobel prize in medicine. In the year 1911, Dr Ross came up with a simple compartmental model that involved mosquitoes and humans. The model proved that it was generally not necessary to entirely eliminate the mosquitoes to put an end to malaria, which was a notion believed by everyone at the time. Instead, the model successfully demonstrated that mere reduction of the mosquito population below a certain level would suffice to keep malaria disease under control. It was around the same time that the notion of basic reproduction was introduced and has been used in most epidemiological models even today to understand the spread of disease. Of course, the model proposed by Dr Ross turned out to be a huge success in controlling the malaria disease [1].

The next coming years compartmental model formulated by Kermack–McKendrick epidemic model (1927) and the Reed–Frost epidemic model (1928) emerged around the same time, where both the

models describe the relationship between individuals that are susceptible, infected and immune. The mathematical model formulated by Kermack and McKendrick in 1927 made predictions that were very close to the actual nature of epidemics and was quite successful in predicting the behaviour of the outbreak[1].

The Kermack and McKendrick model is based on assumptions made when it comes to the rate of transmission between different compartmental classes within the population. Kermack–McKendrick theory in its earliest form was formulated as a partial differential equation that models the infected population with respect to age-of-infection while using simple compartments for people who are susceptible (S), infected (I), and recovered/removed (R) [3].

Both the endemic and epidemic models by Kermack and Mckendrick from the years 1927 and 1933 were built upon the research done by Ronald Ross and Hilda Hudson. Although their theories were meant to be crude and based on their observations back in the day compared to the compartmental models. Their model became one of the milestones in epidemiology and is still relevant today for pandemics like Covid-19. It is an interesting fact to note that in the history of epidemics it was public health physicians such as Sir R.A. Ross, W.H. Hamer, A.G. McKendrick, and W.O. Kermack who laid the foundations of the entire approach to epidemiology based on compartmental models not by mathematicians[3].

More recently, epidemiological agent-based models (ABMs) have emerged as an alternative to compartmental models. The agent-based model for epidemiology is quite sought after by the government departments for policymaking as they capture the real world's complexity quite well and are known to have great efficiency and performance[4].

# 1. Compartmental Epidemic Models

Developing '*compartmental models'* has been a widely used technique in the history of epidemic modelling when it comes to studying disease transmission in a population. A compartmental model in general, analyses a disease pattern in a population by dividing it into several compartments based on assumptions closest to the disease's nature and the pace at which the infectious diseases spread. The very first theory to describe and predict the transmission pattern of infectious diseases in a given population over a period of time was hypothesized by W.O. Kermack and A.G. McKendrick in 1927, 1932, and 1933 in a sequence of three papers [3][5][6].

## 2.1 The SIR Model

SIR is one of the simplest compartmental epidemic models that was first derived from models by Kermack and McKendrick. It is a mathematical representation that models the spread of the disease. The model is based on assumptions that disease is spread through human contacts and classifies the population into three different classes that may change with time t [7].

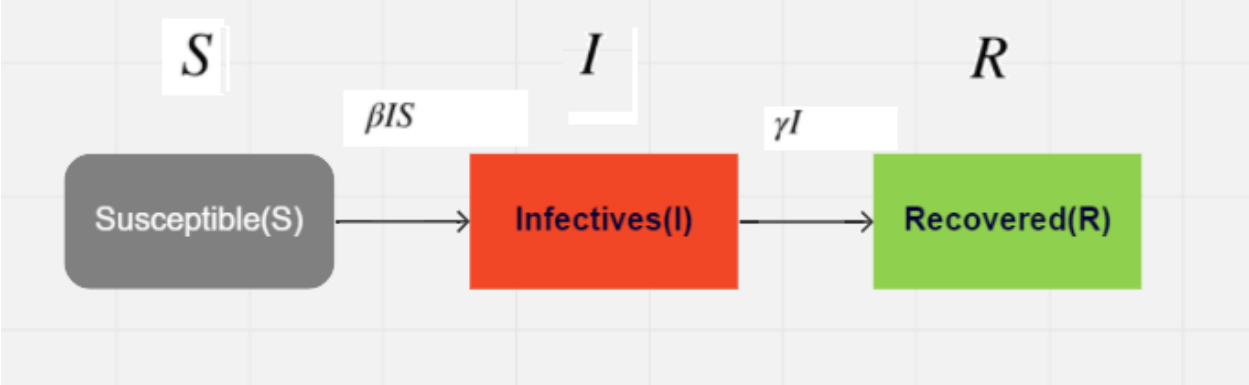


Figure 1: The picture depicts the working model of how the SIR model works where S(t) represents the number of susceptible to disease which means those who have not been infected I(t): stands for the number of individuals that have been infected and R(t): is for the number of individuals that have recovered from the disease [7].

Here the entire population is divided into three classes that go by the abbreviation *S*, *I*, and *R* where $S(t)$ stands for 'susceptibles', representing the part of the population who are not yet infected but susceptible to the disease at a given time $t$. $I(t)$ represents the number of infected cases, people in this category are likely to spread the infection to the susceptibles through social interactions or physical contact. Finally, R(t) stands for recovered individuals who will no longer be infected again as shown in figure 1.

Since the total population is a constant as we are not considering newborns and the population N given by N = S + I + R.

Assuming the virus infection starts with at least one infected individual, it is inevitable that there will be a change in each of these variables given the fact that the disease is contagious and also that most of the healthy population has a certain amount of immunity to fight the disease as well. These changes in numbers in each category is determined by the contact rate of the disease and recovery rate[7].

The change in each class S, I, R is given by following three differential equations:

$$\frac{dS}{dt} = \quad - \beta SI \tag{1}$$

$$\frac{dI}{dt} = \beta SI - \gamma I \tag{2}$$

$$\frac{dR}{dt} = \gamma I \tag{3}$$

*Parameter* $\beta$*:* describes the effective contact rate of the disease: an infected individual comes into contact with other individuals per unit time (of which the fraction that is susceptible to contracting the disease is S/N). [7]

*Parameter* $\gamma$**:** is the mean recovery rate: that is, $1/\gamma$is the mean period during which an infected individual can pass it on. Note that the parameters βand γare often based on assumptions because depends on factors in the real world that are often complex and non-stationary[2]

Under no restrictions by the government such as lockdowns, social distancing and masks etc, the disease spreads exponentially:

Assuming $S \approx N$ (everyone is still assumed to be susceptible);

$$I(t) \approx I(0) \quad \exp[(\beta N - \gamma)t] = I(0) \exp[\gamma(\beta N/\gamma - 1)t] \tag{4}$$

*Basic Reproduction number*

$R_0$stands for the basic reproductive number. The basic reproduction number$R_0$in short, is the expected number of secondary cases caused by an infection in a population that is susceptible to disease. It is the ratio of rates at which the virus reproduces through human contact[21]. Therefore the general formula to calculate $R_0$is

$$R_0 = \frac{Rate\ of\ Infection}{Rate\ of\ recovery} = \frac{\beta SI}{\gamma I} = \frac{\beta S}{\gamma}$$

At the beginning of the pandemic, the number of susceptible S = total population N, so that

$$R_0 = \beta N/\gamma$$

A reproductive number is greatly influenced by human to human contact because by nature pathogens need a new host to survive and multiply. It is important to note that $R_0$ should not be confused with "basic reproductive rate", rather it is a dimensionless number that helps with calculating and predicting disease dynamics[21].

During an endemic, the reproductive number $R_0$ could be very crucial to determine if the epidemic will die out or spread. Using the equation (4) if $R_0 = \beta N/\gamma < 1$ then *I(t)* decreases approaching zero, which means the epidemic will eventually die out. However, when $R_0 = \beta N/\gamma > 1$, then *I(t)* will increase exponentially leading to endemicity. Figures 2(a), 2(b), 2(c) and 2(d) demonstrate how the reproduction numbers affect the outcome of the epidemic or pandemic.

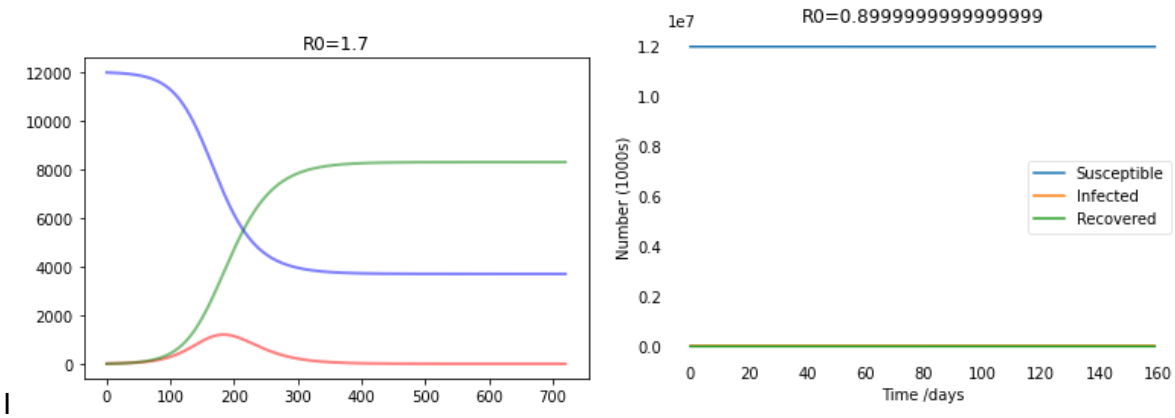


Figure 2(a) and 2(b) : SIR model plots for R0 = 1.7 and 0.89 values.

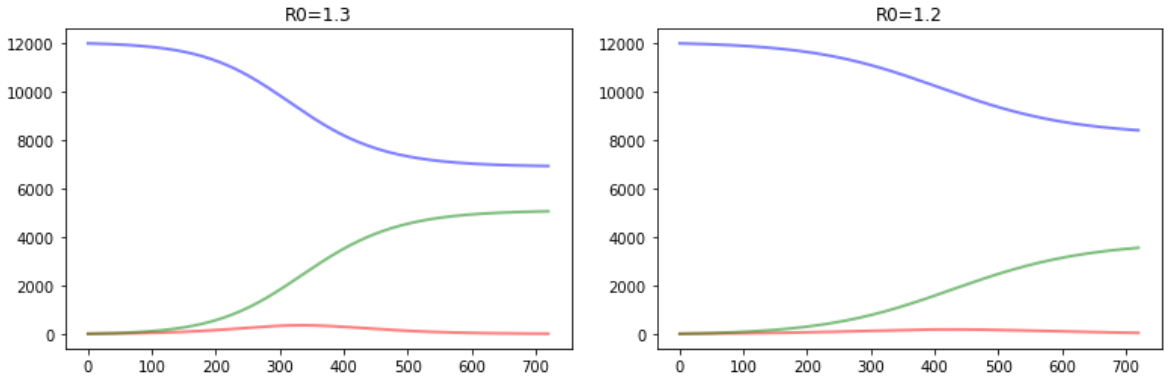


Figure 2(c) and 2(d) : SIR model plots for R0 = 1.3 and 1.2 values.

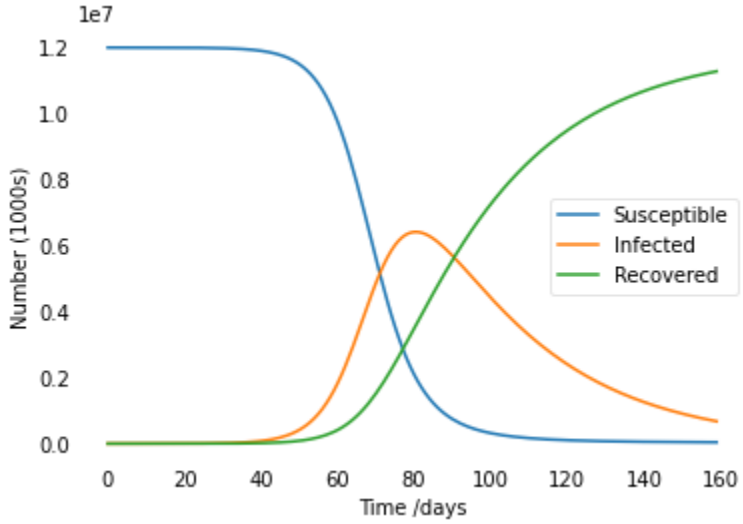


Figure 3 : The plot figure below simulates the disease under discretized parameters and population: population N =12000000, I(0)= 100, β=0.2, 1/γ=30days (infected period) for 160 days.

Figure 3 simulates a scenario where a few cases of infection are introduced to a given population N in two possible ways for instance (i) and when either through local members returning from a trip outside the local population or (ii) a visitor that belongs to a non-local population.

When $R_0 = \beta S/\gamma > 1$,it is observed that *S(t)* decreases and *I(t)* increases over a period of time t and then after a certain point *S(t)* approaches zero that means everyone is infected at this point. As soon as either the number of susceptibles $S(t)$ or Infectives $I(t)$ reaches zero it is considered to exit the model's system

Summing the two equations (1) and (2) we obtain

$$\left(\frac{dS}{dt} + \frac{dI}{dt}\right)' = \quad - \gamma I \qquad (5)$$

From equation (5) one can say that function $(S+I)$ will be a non-negative decreasing function and therefore tends to a limit $S_\infty$ as the time $t$ tends to $\infty$[8].

## 2.2 The SIS Model


Figure 4: The working of the SIS model of how the individuals keep moving between two compartments back and forth

The simplest *SIS* model is essentially, a disease transmission model that involves only two types of individuals, an infectious and a susceptible where individuals can either be infected or be susceptible to the disease through contact. While some types of infectious disease may confer some immunity against the disease after the infected individual has recovered from the illness, however in some cases of infections such as common cold and influenza, individuals don't necessarily grow immunity against it. Even after the recovery from such illnesses hence they are still susceptible to illnesses when it comes to seasonal flu or influenza. The *SIS* model is, therefore, more useful to describe in such cases of diseases that do not offer immunity against illness, therefore the individual's passages between the infective and susceptible category back and forth under this model. Similar to the *SIR* model, during the beginning stages of an epidemic it may seem that there is an exponential growth of infected cases.

The model equations for *SIS* model is :

$$\frac{dS}{dt} = -\beta SI + \gamma I$$
$$\frac{dI}{dt} = \beta SI - \gamma I \quad (6)$$

Similar to the *SIR* model, since the total population is a constant as we are not considering newborns or deaths, therefore, Population N is given by N = S + I. Further, we reduce SIS model equation (6) into a single differential equation, by replacing $S$ by $N - I$ [8]

Therefore we get,

$$I' = \beta I(N - I) - \gamma I = (\beta N - \gamma)I - \beta I^2 = (\beta N - \gamma)I(1 - \frac{I}{N - \gamma/\beta}). \quad (7)$$

Using equation (7), which is a logistic differential equation, if $\beta N - \gamma < 0$ or $R_0 = \beta N/\gamma < 1$, then all solutions with non-negative initial values approaches zero which is called a disease-free equilibrium that means when the number of infected cases reaches 0 i.e I= 0 then the number of susceptibles becomes equal to total population N therefore; S=N. Meanwhile, if $R_0 = \beta N/\gamma > 1$, then all solutions with non-negative initial values approach the limit $N - \gamma/\beta > 0$, with time t then it is called an *endemic* equilibrium[8].

## 2.3 Extensions of SIR Models

### 2.3.1 The SEIR model

The SEIR model (Susceptible-Exposed-Infected-Recovered), widely used today by statisticians and officials to analyze epidemic data in different stages. It is in fact adopted today by the most state of art models that are used to estimate the *Rt* values (reproduction number) to describe the epidemic dynamics and to predict its course. The SEIR model is a system of differential equations that considers the amount of the population susceptible to infection by being exposed to an infectious person and the individuals who either recover from infection or unfortunately die[9]

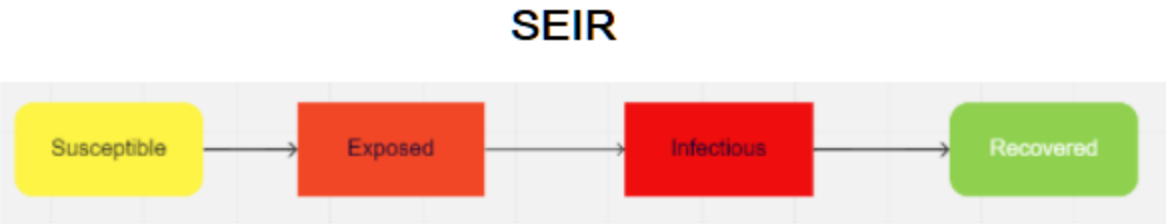


Figure 5: The picture shows the workflow of SEIR

The model in figure 5 is based on the assumption that there is a buffer period between being exposed and being infectious and represented by a parameter '$a$' which is a parameter for the average latency period. The second assumption made is based on the birth and death rates so that the total population remains constant. Therefore we have the model whereβis the contact rate, γis the recovery rate.

$$S' = -\beta IS/N$$

$$E' = \beta IS/N - aE$$

$$I' = aE - \gamma I \quad (8)$$

$$R' = \gamma I - R$$

## 2.3.2 SIR model with Birth and Deaths

The *SIR* model with birth and deaths considers including births in the susceptible class S and a death rate to every class in the model. Both the birth rate and death rates are proportional to the class they are associated with for instance since the births are associated with susceptible class S, therefore, the birth rate is proportional to the number of members in the susceptible class although in the case of no infections then it is directly proportional to the size of total population N[8].

The death rate, on the other hand, is associated with all the members of the class and is proportional to the size or number of members of its respective class it is associated with. However, if the birth and death rates are unbalanced it might either allow the total population size to grow or die out exponentially, therefore, the model can be applied to determine whether the disease will control the size of the population or increase exponentially[8].

To formulate an epidemic model with births and deaths one could consider the approach suggested by Hethcote [2] where birth and death rates are made to be equal and total population size $N$ is set to constant[8].

The model is represented as following,

$$S' = -\beta SI + \mu(N - S)$$

$$I' = \beta SI - \gamma I - \mu I \quad (9)$$

$$R' = \gamma I - \mu R$$

Here parameters $\beta, \gamma, \mu$ are the rates of infection, recovery, and mortality, respectively.

### 2.3.3 The MSEIR model

The MSEIR model is an extended model of the SEIR model where the model considers an additional compartment 'M' where a certain group of the population, especially a newborn infant, has something called passive immunity. M stands for maternally derived immunity.

### 2.3.4 The SIR model with Diffusion

The SIR model combined with a diffusion equation is useful to understand the distribution and density of individuals from all three categories susceptible/infectious/recovered. Spatial models combined with compartmental models may not be ideal for modelling the entire population, however, they are seen to leverage from modelling the distribution of infected persons in space, this is done by combining the SIR model with a diffusion equation.

## 2.4 Limitations of epidemic models

By definition, the deterministic epidemic models are the models whose output is fully determined by the parameter values and the initial conditions whereas the stochastic models may possess some randomness inherently with different outputs while the parameter values and initial conditions are the same. The deterministic models have long been used in the study of infectious diseases and one of the advantages of the deterministic models is that they are simpler to use than stochastic models. They are also useful because they have well developed numerical methods along with the benefits of the theory of dynamic systems. Although the deterministic models do sufficiently support disease analysis of large-sized populations they should be used cautiously. One of the main disadvantages would be their rigidity; these types of models aren't flexible enough to allow embedding new types of infectivity profiles[10].

Diffusion models on the other hand have been proposed to treat disease transmission as waves travelling among certain homogenous populations. However, these are known to be not very practical because of the nature of human contact through which diseases are very different from the spatial transmission. Although the spatial model has been used in a stochastic system rarely[10].

# 3. State-of-the-art Covid-19 models used for various countries

## 3.1 NIPH Norway

One of the primary models used to estimate the $Rt$ in Norway is the meta-population model. The model relies on many factors, one of which is the number of patients admitted into hospitals for medical attention or diagnosis in the entire country. The model also relies on the updated test data to update the R estimates regularly on their official website [12].

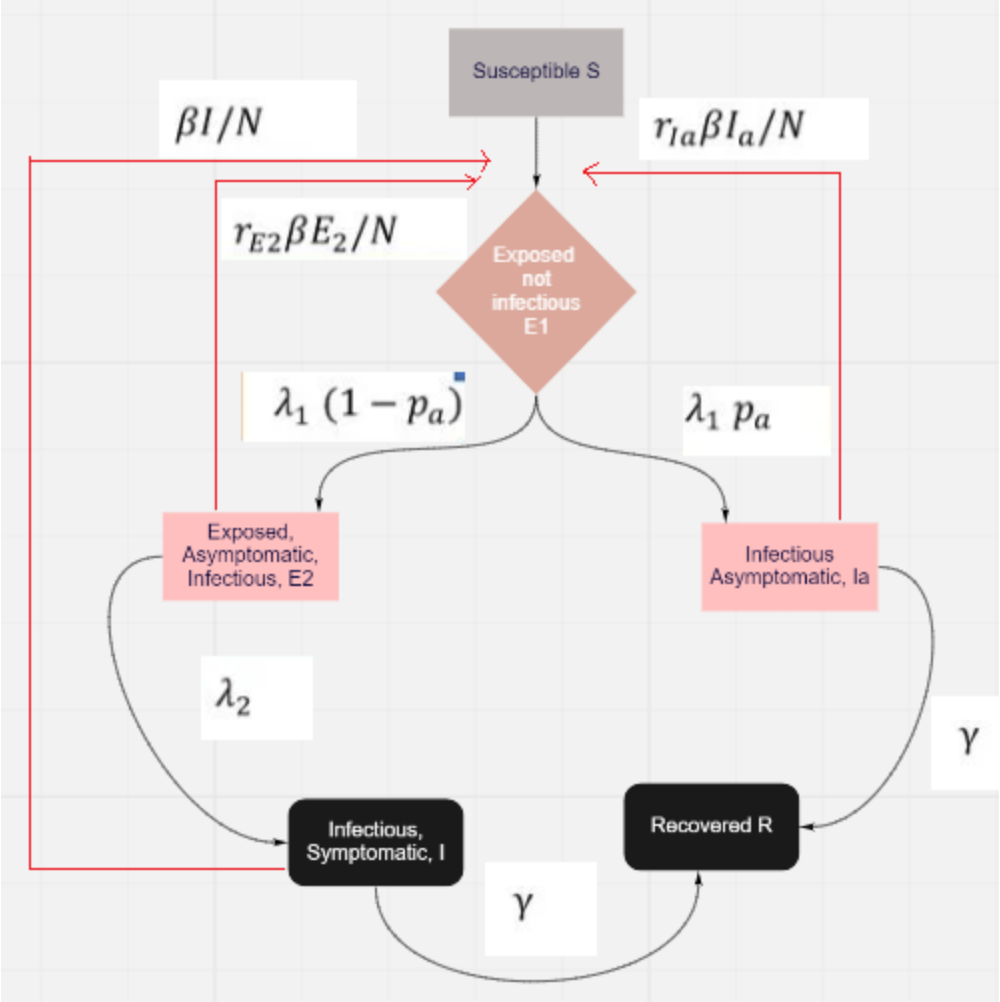


Figure 6. The meta-population model used by NIPH Norway to evaluate the national reproduction number estimates for Norway

The meta-population model in figure 6 is the SEIR type model that is stochastic in nature and an extension of the model by Engebretsen *et al.* (2019) and Engebretsen *et al.* (2020) [12][13][14]. The model used calculates disease transmission locally in each municipality separately based on people's mobility. To track the mobility of people between various municipalities the model relies on mobile data from Telenor.

The model is used to simulate the stages of infection spread in a given population within each municipality separately. Each population within a certain municipality is classified into six infection categories depending on things like physical contact or symptoms.

According to the model, the Covid -19 virus spreads between people in the following stages through the six categories "S – Susceptible; E1 – Exposed, not infectious, no symptoms; E2 – Exposed,

infectious, pre-symptomatic; IA – Infectious, no symptoms; I – Infectious, with symptoms; R – Healthy/Immune"[12].

Out of six categories three of them are considered more infectious compared to others because they are symptomatic and have a tendency to spread the disease faster than the ones without the symptoms. Certainly, all other exposed and infectious categories whether symptomatic or not could infect more healthy people in the susceptible category[15].

Overall, the model seems to work well for predicting the course of the pandemic in Norway although there seems to be considerable uncertainty associated with it, the model seems to handle it by constantly updating its parameters and with more and more new data available about the disease, the final estimation seems to be more accurate to that of the real-time. Furthermore, the NIPH 's additional model for Covid-19 besides the metapopulation model is derived from the Di Ruscio et al. (2019) to estimate the effect of measures taken by the governments such as lockdowns by running the model to simulate the pattern of an outbreak of infection among people according to their respective age groups [12][16].

## 3.2 Yoyangs Covid-19 Death Forecasting Model (USA)

Another version of the compartmental model used for the evaluation of the course of pandemic and Covid-19 prediction that has been popularly used by many countries like Italy, the USA and the United Kingdom is the SEIR model. Although the epidemic models vary from country to country depending on the demographics and circumstances, this particular version was used as a simulator underneath their machine layer based on the SEIR (susceptible-exposed-infectious-recovered) model. Like the Norwegian model it also has a component called 'exposed' which is not infectious yet [17].

When it comes to implementation, this model also has a stochastic approach. This model has a probability distribution for each transmission between the compartments S-E-I-R. Although the model isn't much elaborated on, the simulator works through the formulation of probability distributions combined with actual cases to give daily estimates on new infections and deaths updated every day. For countries like Italy, similar stochastic SEIR models have been used along with the help of machine algorithms to get more accurate results [17].

## 3.3 Vaccination Model

As long as there are epidemic or pandemic diseases, having tools such as epidemic modelling is convenient and useful to describe and understand the extent of the effect of the outbreak of the disease, however, the most important question remains about how to prevent it or eradicate after all. Apart from masking, hand hygiene and social distancing which are not very sustainable or reliable for the long term, mass vaccination programs are seen to be promoted by all the governments around the world because of the clear effect it has on reducing infections. Taking the development of vaccines into consideration for any

communicable diseases, how does the availability of vaccines affect the SIR model? Especially when there has been enough discussion on the SIR model in the previous chapters what would happen if we introduce vaccines into the SIR model.

Consider that '*V*' stands for the number of individuals that are vaccinated. With the SIR Vaccination model, which is nothing but an extension of the SIR model with an added element of vaccine, we have three questions that may be crucial to understand the effects of vaccination on the system.

1. What is the condition for the epidemic or pandemic to stop?
2. How many individuals need to be vaccinated to put a stop to the epidemic?
3. How effective is the vaccine?

Addressing the first question which is the condition to stop the epidemic could be derived from the SIR model refer to equation (2) in chapter 2 which is $I' = I(\beta S - \gamma) < 0$.

That means when $R_0$value which stands for reproduction number: when $\beta s/\gamma < 1$, the epidemic slowly decreases which eventually dies out. By looking at the equation of the reproduction number there are many ways where the $R$ number can be brought down through different parameters such as contact rate and recovery rate $\beta$ $and$ $\gamma$. However, with the vaccine in the picture, the effective way of reducing the reproduction number is through targeting the susceptible population which is sustainable in the long term to keep the diseases at bay. But how do we know what proprotion of total population need to be vaccinated to reduce the susceptible population efficiently?

Since we have vaccination developed at a much later stage after the pandemic has started, we do have $R_0$values which are also known as a basic reproduction number readily available through the formula $\beta s_0/\gamma$. For an epidemic to end we need to ensure the number of susceptible individuals to be minimal to ensure $\beta s/\gamma$remains less than 1.

Consider that$S^*$as a population that is still susceptible after the vaccination and individuals who are vaccinated are denoted as V. To stop the pandemic, the proportion of population that needs to be vaccinated[22]

$$V > 1 - (1/R_0) \qquad (10)$$

Considering the fact the vaccines are not always 100 % effective, the way to estimate the efficiency of vaccination is $V_{eff} > 1 - (1/R_0)$ where $V_{eff}$= eV, and e stands for efficiency that ranges between 0 to 1. Therefore the proportion of the population that needs to be vaccinated taking into account that the efficiency of vaccine is not 100% effective[22]

$$V > 1/e(1 - \frac{1}{R_0}). \qquad (11)$$

## 3.4 Parameter Estimation model case study using simple SIR Model

For the estimation and analysis of parameters$\beta$values and the reproduction number$R_t$, this sliding window SIR epidemic model in python was originally derived from the research paper "Estimation of Time-Dependent Reproduction Number for Global COVID-19 Outbreak" by authors Petrova, T.; Soshnikov D.; Grunin A. The plots presented in the paper cover about 12 countries and the model runs for the first 90 days of the pandemic. The code used to run the model was retrieved from the Github link provided in the paper. For the study of the SIR model, the basic approach to estimating the effective and basic reproduction number *R0*, *Rt,* is done through optimization and fitting the SIR epidemic model to real data in a sliding time window. The model relies on an online open-source of datasets from John Hopkins University for COVID-19 for daily cases of infections, recovered and deaths. The paper further ventures into comparing with the mobility data however we chose not to delve into it due to lack of time and unavailability of relevant data[18].

### 3.4.3 Data Exploration

In this thesis, I will be using one of the most popular datasets among researchers today for the Covid-19 analysis published in Github. The dataset under the repository of Covid-19 used is maintained and published by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). The dataset contains worldwide records of confirmed, recovered and death cases and is updated every day.

The data is freely available for public use on the Github link provided below:
https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series

The time-series datasets for infected recovered and deaths cases are located separately. To import, the dataset needs to be in the raw format to be called by your python program to read and analyse further. During the exploratory analysis, I found that all the datasets seem to have the same number of columns (403), however, the number of rows slightly differ for the recovered dataset i.e 258 although the number of rows of the remaining two datasets i.e confirmed and deaths are the same (273). All three of the time-series datasets cover data from about 177 countries. While working with data I also discovered that the number of cases under each date is cumulative of both old and new cases. Each data set also has its fair share of null values and contains the same number of columns which keep increasing every day as the data gets updated. A typical dataset from John Hopkins something like in figure 7.

| | Province/State | Country/Region | Lat | Long | 1/22/20 | 1/23/20 | 1/24/20 | 1/25/20 | 1/26/20 | 1/27/20 | 1/28/20 | 1/29/20 | 1/30/20 | 1/31/20 | 2/1/20 | 2/2/20 | 2/3/20 | 2/4/20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | NaN | Afghanistan | 33.93911 | 67.709953 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | NaN | Albania | 41.15330 | 20.168300 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | NaN | Algeria | 28.03390 | 1.659600 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | NaN | Andorra | 42.50630 | 1.521800 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | NaN | Angola | -11.20270 | 17.873900 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

5 rows × 393 columns

Figure 7: The picture displays how a typical structure of a dataset looks like from a John Hopkins university database on Github.

### 3.4.4 Data Preparation

To start with, the data is loaded from the GitHub repository. The data has been filtered according to the countries for further evaluation. The dataset contains records since 21 January 2020. While exploring the data, it seems like not all countries have enough infections starting from January, so the start dates have been chosen when the infected cases are greater than 100. The daily data are stored in a cumulative sum format.

The table contains the cumulative sum for all cases. During the analysis, however, new daily cases of infection have been derived by using functions that return subtracted values between columns.

### 3.4.5 Code Workflow

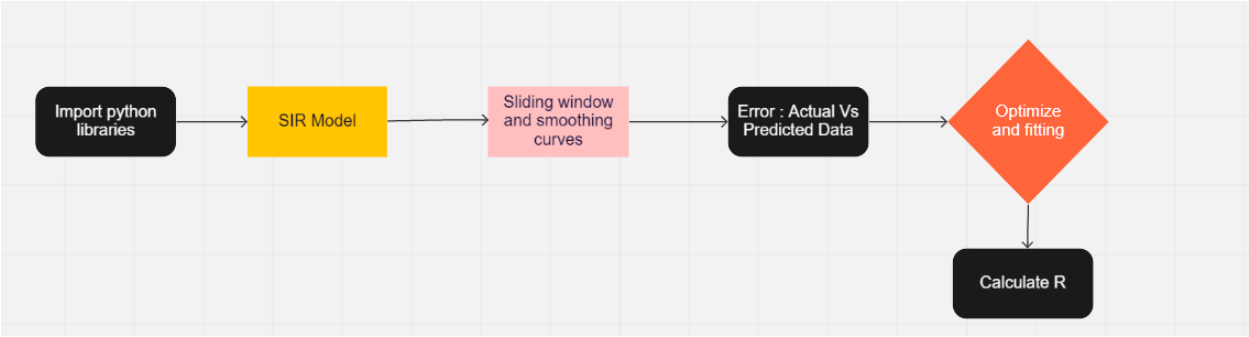


Figure 8: Code workflow for implemented Sliding SIR from the paper "Estimation of Time-Dependent Reproduction Number for Global COVID-19 Outbreak" by authors Petrova, T.; Soshnikov D.; Grunin A

The code originally begins with importing necessary libraries as shown in Figure 8.

*Load Data:*

To start with, the data is directly loaded from the John Hopkins University Covid-19 repository in Github so that we have updated estimates, every time the code runs.

The datasets for infected, recovered and country populations are available in separate CSV files. In order to extract the information we need, these datasets are coupled together for further use[18].

### *SIR Model :*

Next, the SIR model function is defined in the code with initial values and parameters to solve differential equations that contain three variables *S(t), I(t), R(t)* numerically. The authors derived their example code from a chapter from the Scipybook[7].

### *Smoothing Curves and Sliding Window:*

The sliding window feature is implemented in the SIR model through a method called "compute_params` The method basically performs a 7-day fitting of the SIR model using the `fit` function for all the days[18].

Since we need a smoothing function in order to remove the outliers from the data for all days beginning from the day when the number of infected people is above 1000, i.e. $t0 = \min\{t|I(t) > 1000\}$.

It is done so through a method called 'make_frame'[18].

### *Optimization:*

In this part of the code, the value of the parameter γ is fixed. However, since we have access to the actual data, to determine the parameter value β,the code performs a fitting function where the results from the SIR model are minimized to fit the real data through optimization. In this case, it is done by solving optimization problems using a Python built-in function called 'minimize' using Powell's method[18]. The math behind the optimization looks something as follows.

Considering variables *V'* and *I'* that represent daily new infected cases from actual data and computed data by the model. The optimization function for calculating minimal discrepancy between *V'* and *I'* that corresponds to $\beta^*$ value is:

$$\beta^* = argmin_{\beta} \sum_{t=0}^{t=n} (V'(t) - I'(t))^2$$

The process of finding argmin is however a quite complex task to perform a numerical solution at each step of ODE equation (1), (2) and (3) from the SIR model, so the authors chose to go for an optimization algorithm called Powell's method (Powell's conjugate direction method) because it works fast and performs efficiently[18].

Now that we have the $\beta^*$ estimates, everything is put together in a function called 'analyze' to perform sliding SIR fitting and Rt calculations.

# 4. Results

## 4.1 SIR model with different γvalues

Some modifications were made to the original code to experiment for this thesis. The code originally was designed to cover only a 90 days period, however, it has been modified to cover all days starting from the date the infected cases are more than 1000 until the most recent date the data is updated. Second, the sliding interval considered to calculate the R-value has also been modified to 15 days because it seems to be a more popular time duration used to calculate in most countries. The Figure 9(a),9(b), 9(c) are the Rt estimates run for different γ *values* i.e. 1/45, 1/60, 1/30 for 12 countries. The γvalues considered are based on assumptions of average time taken for recovery.

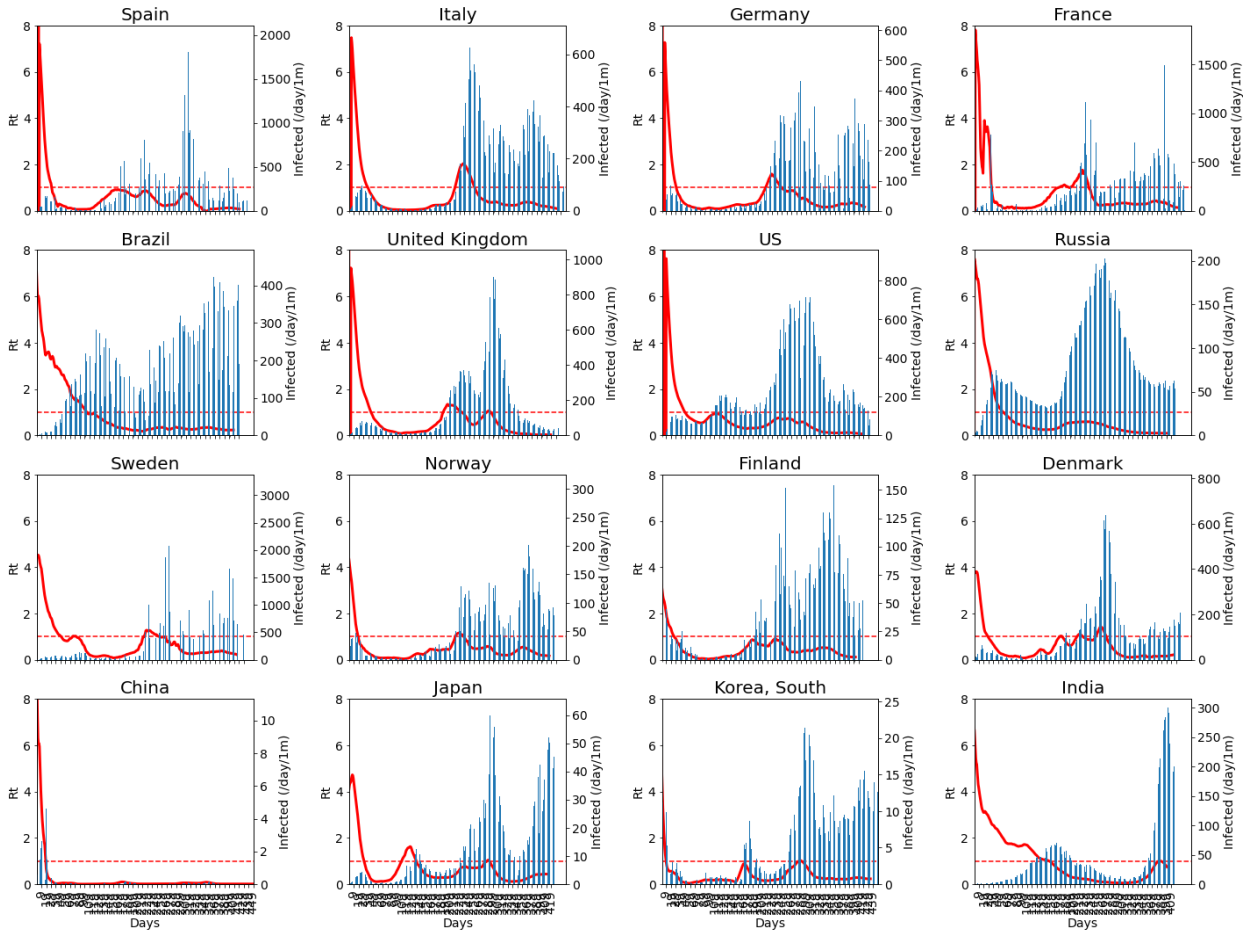


Figure 9(a) : $R_t$ Estimation plot run for 12 different countries for γ= 1/45 with $R_t$ in red and normalised number of new infected by their respective population in blue bars on the y-axis, whereas x-axis is for the number of days starting from the day cases of infection is greater than 1000. The infected cases in blue are normalised to fit the scales [18]

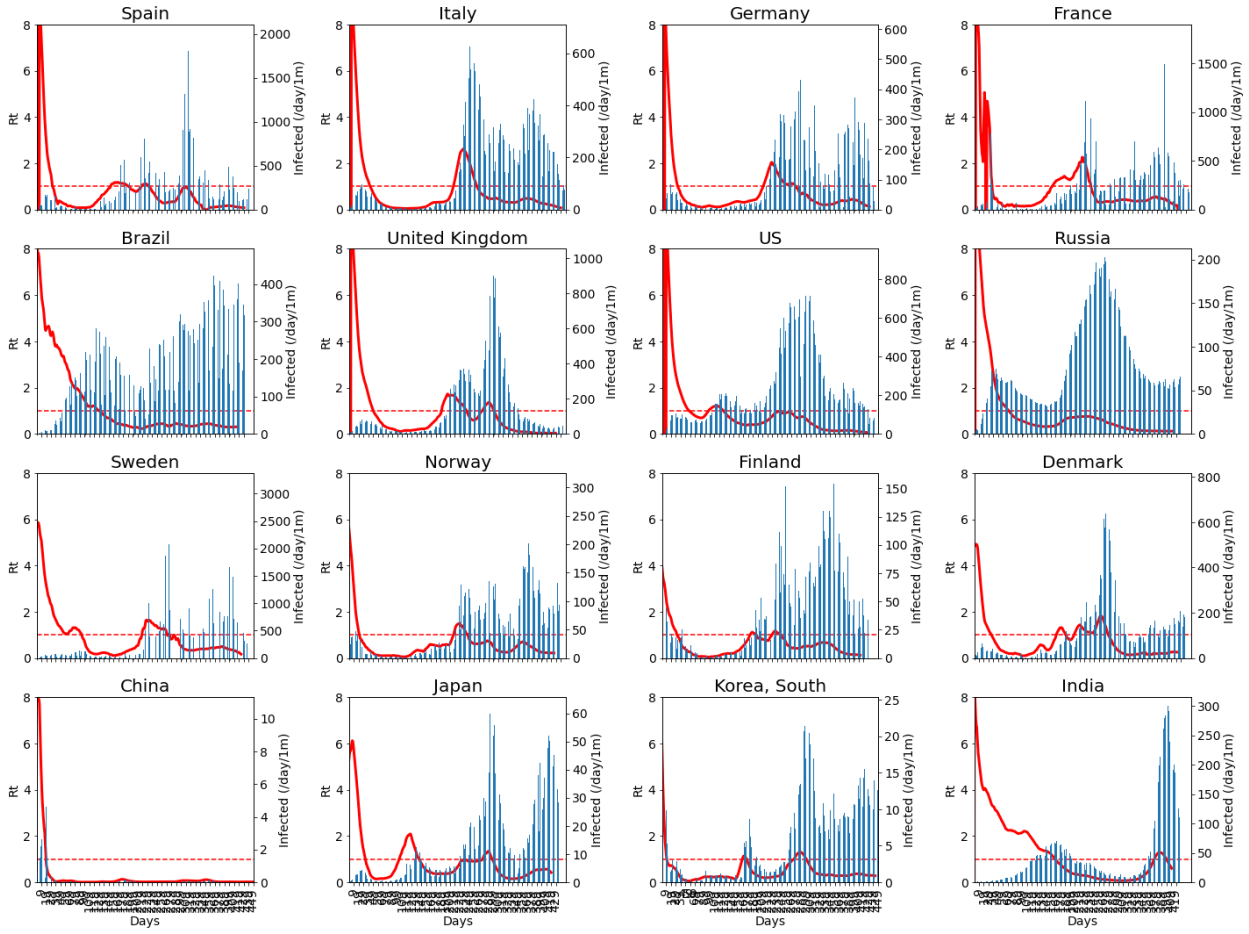


Figure 9(b) $R_t$ Estimation plot run for 12 different countries for γ= 1/60 with $R_t$ in red and normalised number of new infected by their respective population in blue bars on the y-axis, whereas x-axis is for the number of days starting from the day cases of infection is greater than 1000. The infected cases in blue are normalised to fit the scales [18]

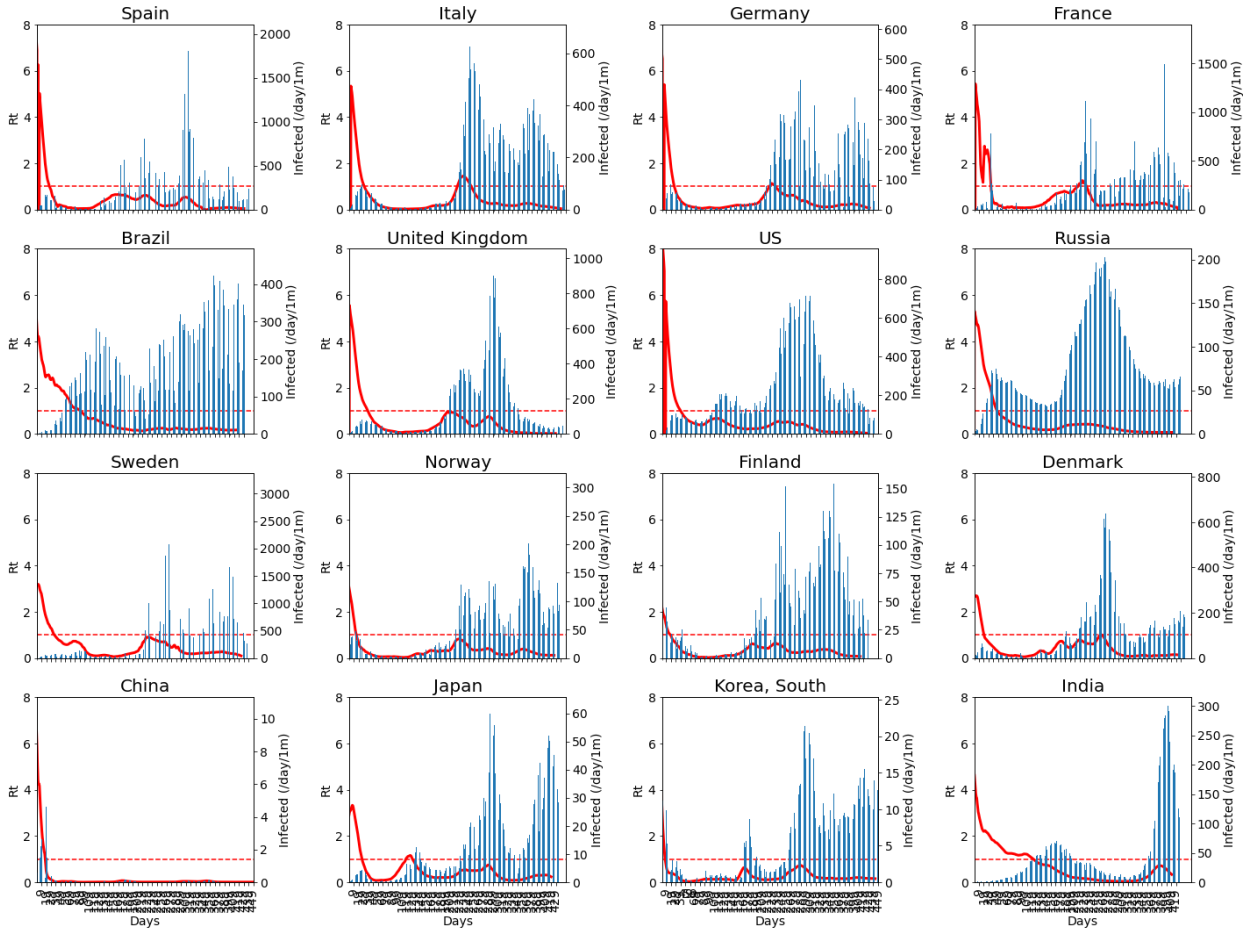


Figure 9(c) : $R_t$ Estimation plot run for different countries for γ= 1/30 with $R_t$ in the red and normalised number new infected by their respective population in blue bars on the y-axis, whereas on the x-axis for time i.e number of days starting from the day cases of infection is greater than 1000 cases. The infected cases in blue are normalised to fit the scales [18]

## 4.1.0 Results Interpretation

One can observe from Figures 9(a), 9(b) and 9(c) that the γvalues influence the reproduction numbers$R_0 = \beta/\gamma$.The higher the recovery period is, the higher the reproduction number$R_t$.When making comparisons of the real estimates that are sourced from the official websites to the *Rt* estimates produced by the model withγvalues 1/45 referring to fig 9(a), the model seems to give the closest *Rt* as compared to other γvalues 1/30, 1/60. Although for some countries there are no official reports for their estimated $R_t$to be able to make comparisons with the given figures we have been able to get out hands on official estimates for Norway and UK which is discussed in the next section 4.2 'Comparing Estimated $R_t$ from the Model and Official $R_t$values'.

## 4.2 Comparing Estimated $R_t$ from the Model and Official $R_t$ values

*Norway*

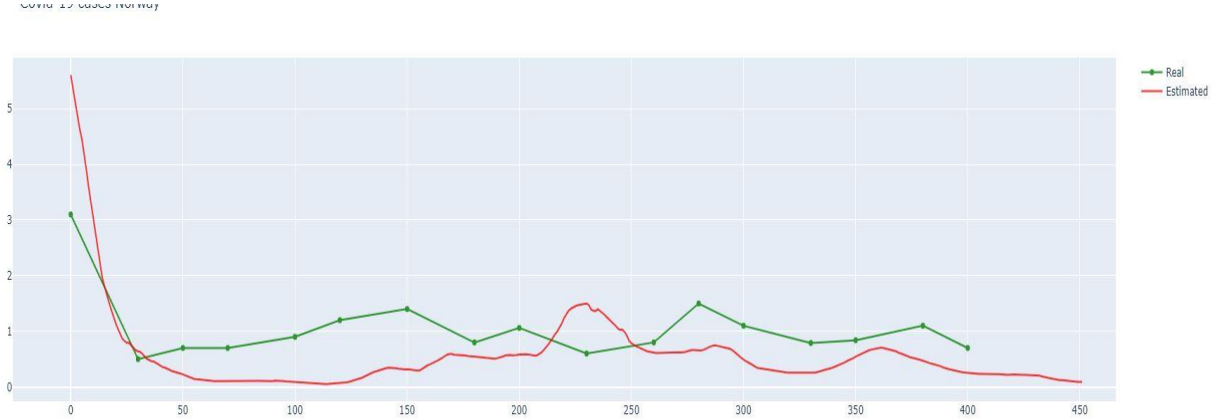


Figure 10(a): Model Comparison: Norway $R_t$ estimated in red versus $R_t$ values from the official estimates in green in y-axis and x-axis corresponds to the number of days with MSE (mean square error) = 63.79480382778067

According to the official sources, the first confirmed case of coronavirus infection was found in Norway on 26th February 2020. Since then there has been a quick increase in the number of confirmed cases of infection in the month of March 2020 which led the government to start introducing social distancing measures by mid of march 2020. The early wave of the virus seems to have first been traced back to some Norwegian tourists returning from Austria and Italy. So according to Table(1) which are the official *Rt* numbers released by the NIPH Norway, the first estimate for the *R0* reproduction number started at 3.1 average and then numbers went down due to lockdown. The *Rt* numbers, however, went back as the lock measures were eased out [19].

The results in figure 10(a) represent the comparisons of *Rt* values estimated by the model for Norway in red and *Rt* estimates released by the official departments in Norway in green. The trends in figure 10(a) do not exactly match up with the estimated values of the model versus an official, although they seem to intersect at certain points, the official estimates seem to be overestimating at many places compared to model *Rt* value estimates, which is beneficial in a way because it prevents from being underprepared for essential things like arranging beds and respirators in health care services as well as the policy-making for a pandemic like Covid-19 disease that has already claimed many lives.

Table(1) Official Reproduction number dated from 17 February 2020 - 13 th June 2021 released by FHI, Norway
https://www.fhi.no/en/publ/2020/weekly-reports-for-coronavirus-og-covid-19/

| Reproduction Figures | Average |
|---|---|
| R0(Starting from Outbreak- 15th march) | 3.1 (2.5 - 3.9) |
| R1(15th March - 20th April) | 0.5(0.4- 0.6) |
| R2 (20th April - 11 may) | 0.7(0.3 - 1.0) |
| R3(11 May - 30 jun) | 0.7(0.3 - 1.0) |
| R4(1st July - 31st July) | 0.9(0.2- 1.5) |
| R5(1st August - 30 th August) | 1.1(0.8 - 0.14) |
| R6(1st Sept - 31st September) | 0.9(0.7 - 1.1) |
| R7(1st Oct - 25 th October) | 1.2(1.0 - 1.04) |
| R8(26th Oct - 4 November) | 1.4(1.1 - 1.6) |
| R9(5 th Nov - 30 th November) | 0.8(0.7 - 0.9) |
| R10(1st Dec - 4th January) | 1.06(1.0 - 1.11) |
| R11(4 th January - 21 st January) | 0.6(0.5 - 0.7) |
| R12(22 nd January - 7 th February) | 0.8(0.7 - 1.0) |
| R13 (8 Feb - 1st March) | 1.5(1.3 - 1.6) |
| R14( 1st March - 24th March ) | 1.1(1.0 -1.2) |
| R15(25 March - 15 th April 2021) | 0.79(0.74 - 0.84) |
| R16(16 th April - 5 May 2021) | 0.84(0.77- 0.93) |
| R17(from 6th May to 19 May 2021) | 1.1(0.9- 1.2) |
| R18 (from May 19 2021) | 0.7(0.5 - 0.8) |

*United Kingdom*

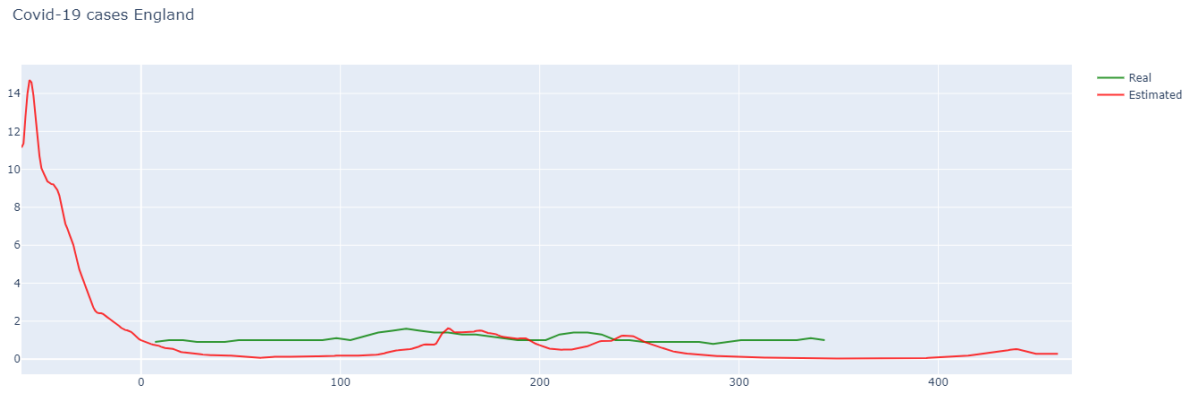


Figure 10(b) Model Comparison: United Kingdom $R_t$ estimated in red versus $R_t$ values from the official estimates in green in y-axis and x-axis corresponds to the number of days with MSE (mean square error) = 43.876002474772825

The first confirmed case in the UK was reported on 31 January 2020, The UK government started introducing social distancing and other measures to the public in February 2020 and more stringent regulations such as lockdown for England were introduced sometime in March[20].

Official *Rt* estimates from the UK are available on their official website. From what you can observe trends in figure 10(b) the official estimates known as the "Real" estimates in the figure did not seem to cover the period entirely as compared to the model due to insufficient data. The time series provided on the official website does not seem to be updated since the end of March 2021. The real *Rt* estimates in the 10(b) do not seem to have many highs and lows compared to the Norwegian figure 10(a), although they are known to have handled the pandemic more poorly compared to the rest of Europe. In my experience, the official estimates sourced from the time series dataset do not seem to be reliable for performing comparisons of *Rt* values. One of the reasons being that the *Rt* estimates that the report provided seem to be around 0.8 dated 29 May 2020 which seems not correct *R0* value considering the pandemic started sometime in February and the other being due to lack of complete data[20].

## 4.3 Actual cases of Infection versus fitted data by the model

The results for the daily actual cases versus cases fitted by the model are presented in figures 11(a) and 11(b) to represent the countries Norway and the United Kingdom. The trend that is apparent in both the figures shows a similar kind of overlap of actual cases of infection which is denoted by V' in green versus daily cases fitted by the model denoted by I' in red.
Although the red lines seem to be more prominent in both the figures due to the overlap, once zoomed in, you clearly see separated lines like in figures 12(a) and 12(b) which means there is a discrepancy between actual cases versus fitted although it seems very minor.

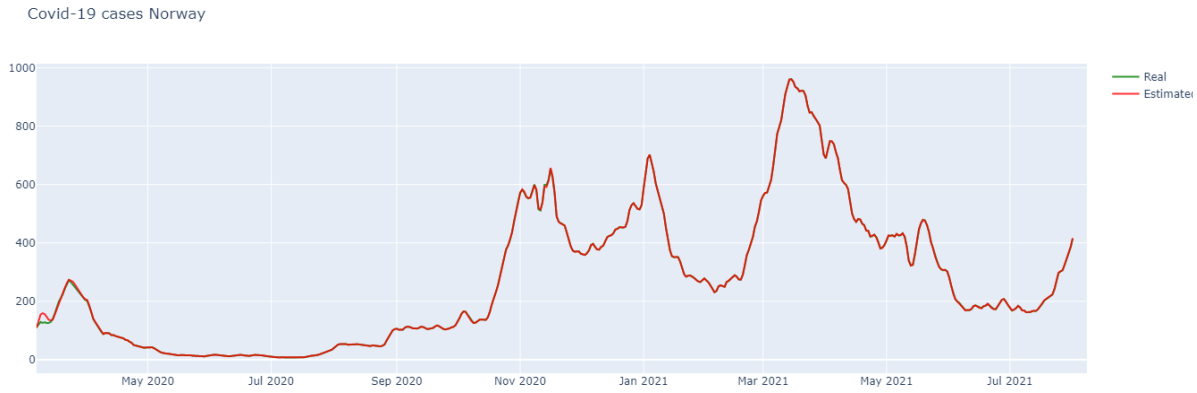


Figure 11(a): The plot corresponds to the difference between V'(actual number of daily infected cases) and I' (daily infected cases fitted by the model) for the country Norway. The green line here represents the actual number of cases V' and the red line represents the I' fitted daily cases by the model. In this figure both the red and green seem to overlap each other hence red looks more prominent it could because of minimal discrepancy. Here, the MSE(Mean Square Error) calculated for Norway is 8.161696788800855.

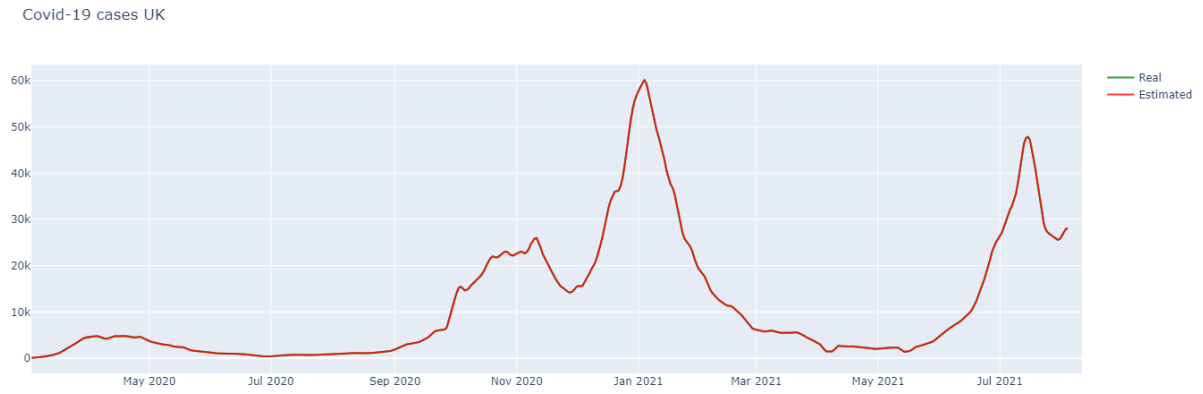


Figure 11(b): The plot corresponds to the difference between V'(actual number of daily infected cases) and I' (daily infected cases fitted by the model) for Norway. The green line here represents the actual number of cases V' and the red line represents the I' fitted daily cases by the model. In this figure both the red and green seem to overlap each other hence red looks more prominent it could because of minimal discrepancy. Here, the MSE(Mean Square Error) calculated for the UK is 1729.156815617208.

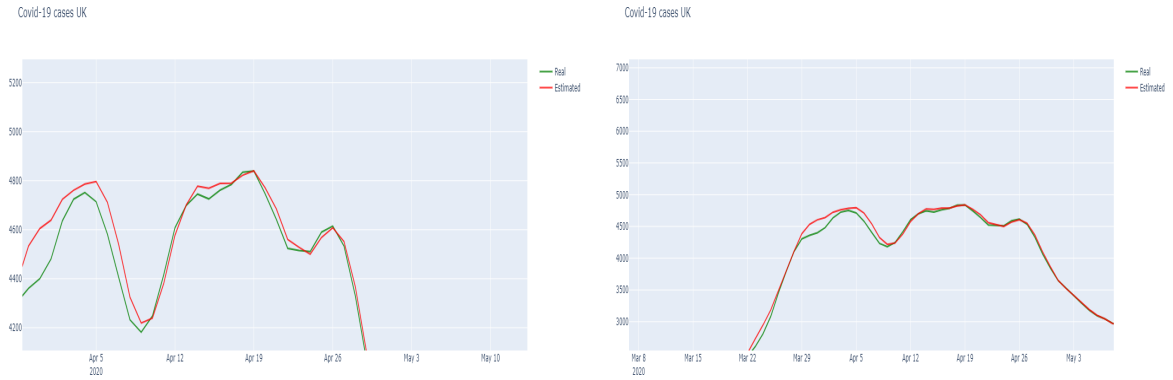


Figure 12(a): This is the zoomed-in version of Figure 11(b) where you see the line for the daily actual infection cases V' in green and I' is the daily infection cases fitted by the model in red.

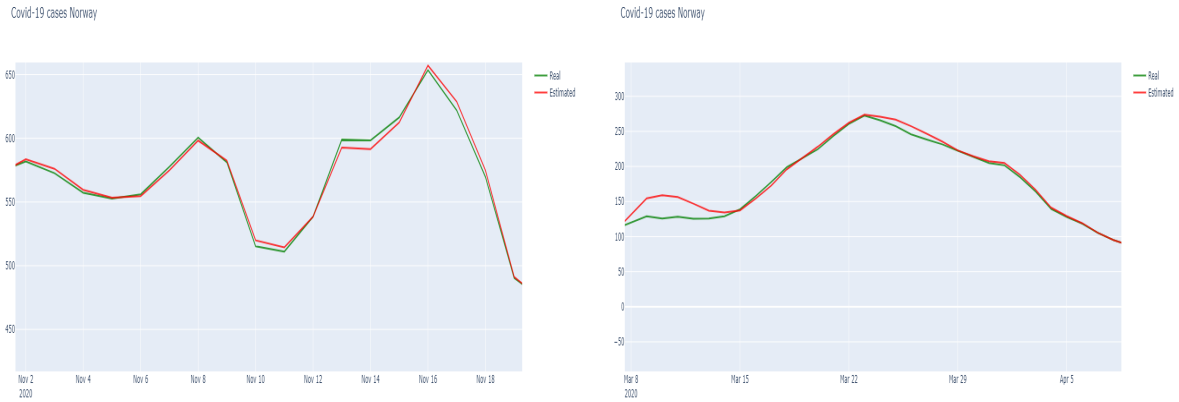


Figure 12(b): This is a zoomed-in version of Figure 11(a), where you see the line for the daily actual infection cases V' in green and I' is the daily infection cases fitted by the model in red.

# 5. Discussions and Conclusions

In this thesis we delved into the history of epidemics to learn about why and how the modelling of epidemics started, describing the different models from gathering information to building the most advanced models of today that are being used to analyse the corona pandemic. I follow it up with briefing the associated limitations of these models as well. I further discussed the concepts that were first introduced by the various experts that shaped today's epidemiological models and the detailed timeline of its evolution. In the next chapter about 'Compartmental Models,' I studied the SIR models and it's extensions such as SIS, SEIR, MSEIR and SIR-D along with some brief analysis on SIR and SIS model, and then progressed to discuss the highly relevant state-of-the-art epidemic models in Chapter 3.

For analysis, I go ahead and apply a specific SIR model proposed in "Estimation of Time-Dependent Reproduction Number for Global COVID-19 Outbreak" authors Petrova, T.; Soshnikov D.; Grunin A [18] to derive estimates for the number of infected people and the associated *Rt* values. In this implementation, the recovery rate is kept constant. Furthermore, I tried to explore the effect of the recovery rate, by varying its values and performing a comparative analysis. The main findings when varying the recovery rate were:

- The higher the recovery period the higher the *Rt* values.
- A recovery rate of 1/45 seems to yield more accurate results compared to 1/60 and 1/30.

And the findings from the comparative analysis between *Rt* values that are official versus estimated by the model were :

- The *Rt* value trends by model do not exactly match with real official estimates.
- The official *Rt* values always seem to be overestimating compared to *Rt* model estimates.
- The calculated MSE(Mean Square Error) for Norway is higher around 63.79 and the UK is nearly 44.

Finally, important findings from actual cases of infection versus fitted data by the model tell us that:

- The results for daily new cases from actual data and the one fitted by the model seem to be an overall good fit except for a few places.
- The more accurately both the results fit together the more accurate βvalues are derived from the model.

Putting these main results into perspective, several remarks should be pointed out.

First, acquiring the right data set *Rt* values for several countries has been challenging. Due to compatibility issues and time constraints, I have been able to choose only two countries namely, Norway and the United Kingdom to do the comparison of *Rt* values (official versus model) in this project. Comparisons between estimated *Rt* values by the model and the official estimates show different trends at different timelines, however, they do intersect at certain points. Overall, they don't

seem to be an exact match which is a given considering the models being used to produce the official estimates for Norway and UK are advanced and complex stochastic models and data relied upon is constantly updated and so are the parameters. The model used in the paper [18] requires more updated model formulation to keep with accuracy and changing disease dynamics such as SEIR. To get better results, one could adapt the SEIR model approach for accurate results in the future instead of SIR as it captures epidemic better and is closer to the real-world scenarios. Also, a standard SIR model still lags in describing external factors that influence pandemic diseases.

Second, there is the effect of heterogeneity of all the parameters such as$\beta$and $\gamma$and *Rt* that varies in different parts of the geographical location and demographics of a population. If you need to estimate the *R-value* at a national level, it is usually done so by taking an average of all *R* values of the respective provinces/municipalities just the way mentioned in the meta-population model in Norway[12]. Although it could be a challenge estimating *R* in smaller municipalities where the population and infected cases are scarce. So the national estimate for *R-value* may not correspond to disease trends in all of its subregions, especially the ones with very low numbers in population or infections [20].

Finally, it is also important to consider the fact that the mathematical models used by different organisations to simulate the spread of the diseases may vary from one another and the data is also sourced differently. Although the models are built to replicate closest to real-life situations even with the stochastic models there is always still a bit of uncertainty involved. As far as policymaking is concerned, some experts say that a single epidemic model is not a very reliable source to answer questions about whether they exactly fit the real-world scenarios but I believe that they are somewhat a good start[20].

For a more accurate approach, results from several models should be taken into account to accurately conclude about the disease. Therefore for future work consideration, I think it would be interesting to experiment with models such as Marchov chains or spatial distribution to evaluate parameters that could define the disease model in a new context.

*Acknowledgement: I would like to thank my supervisors Leiv Øyehaug and Pedro G. Lind for their unwavering support throughout my thesis. Working on this thesis has indeed opened my mind more towards understanding mathematical models and has inspired me to pursue them more in future.*

# References


1. Brauer F. (2017). Mathematical epidemiology: Past, present, and future. *Infectious Disease Modelling*, *2*(2), 113–127. https://doi.org/10.1016/j.idm.2017.02.001

2. Hethcote HW (2000). "The mathematics of infectious diseases". Society for Industrial and Applied Mathematics. 42: 599–653.

3. Kermack, W; McKendrick, A (1991). "Contributions to the mathematical theory of epidemics – II. The problem of endemicity". *Bulletin of Mathematical Biology*. 53 (1–2): 57–87.doi:10.1007/BF02464424. PMID 2059742.

4. Maziarz M, Zach M. Agent-based modelling for SARS-CoV-2 epidemic prediction and intervention assessment: A methodological appraisal. *J Eval Clin Pract*. 2020;26(5):1352-1360. doi:10.1111/jep.13459

5. Kermack, W; McKendrick, A (1991). "Contributions to the mathematical theory of epidemics I". *Bulletin of Mathematical Biology*. 53 (1–2): 33–55. doi:10.1007/BF02464423. PMID 2059741.

6. Kermack, W; McKendrick, A (1991). "Contributions to the mathematical theory of epidemics – III. Further studies of the problem of endemicity". *Bulletin of Mathematical Biology*. 53 (1–2): 89–118. doi:10.1007/BF02464425. PMID 2059743

7. Scipy Book: https://scipython.com/book/chapter-8-scipy/additional-examples/the-sir-epidemic-model/

8. Brauer, F., Castillo-Chavez, C., & Feng, Z. (2019). Simple Compartmental Models for Disease Transmission. *Mathematical Models in Epidemiology*, *69*, 21–61. https://doi.org/10.1007/978-1-4939-9828-9_2

9. Godio, A., Pace, F., & Vergnano, A. (2020). SEIR Modeling of the Italian Epidemic of SARS-CoV-2 Using Computational Swarm Intelligence. *International journal of environmental research and public health*, *17*(10), 3535. https://doi.org/10.3390/ijerph17103535

10. Mick Roberts, Viggo Andreasen, Alun Lloyd, Lorenzo Pellis, Nine challenges for deterministic epidemic models,Epidemics,Volume 10, 2015, Pages 49-53,ISSN 1755-4365, https://doi.org/10.1016/j.epidem.2014.09.006(https://www.sciencedirect.com/science/article/pii/S1755436514000553)

11. Newman, M. E. J. (2002-07-26). "Spread of epidemic disease on networks". *Physical Review E.* **66** (1): 016128. arXiv:cond-mat/0205009. Bibcode:2002PhRvE..66a6128N. doi:10.1103/PhysRevE.66.016128. PMID 12241447. S2CID 15291065.

12. Coronavirus modelling at the NIPH https://www.fhi.no/en/id/infectious-diseases/coronavirus/coronavirus-modelling-at-the-niph-fhi/

13. Engebretsen S, Engø-Monsen K, Frigessi A, Freiesleben de Blasio B (2019) A theoretical single-parameter model for urbanisation to study infectious disease spread and interventions. PLoS Comput Biol 15(3): e1006879. https://doi.org/10.1371/journal.pcbi.1006879

14. Time-aggregated mobile phone mobility data are sufficient for modelling influenza spread: the case of Bangladesh Solveig Engebretsen, Kenth Engø-Monsen, Mohammad Abdul Aleem, Emily Suzanne Gurley, Arnoldo Frigessi, Birgitte Freiesleben de Blasio medRxiv 2020.03.11.20033555; doi: https://doi.org/10.1101/2020.03.11.20033555.

15. Data on how Norwegians move around allow for a fine tuned model of calculating the spread of coronavirus https://sciencenorway.no/covid19-epidemic-information-technology/data-on-how-norwegians-move-around-allow-for-a-finetuned-model-of-calculating-the-spread-of-coronavirus/1689767

16. Quantifying the transmission dynamics of MRSA in the community and healthcare settings in a low-prevalence country by Francesco Di Ruscio, Giorgio Guzzetta, Jørgen Vildershøj Bjørnholt, Truls Michael Leegaard, Aina Elisabeth Fossum Moen, Stefano Merler, Birgitte Freiesleben de Blasio, Proceedings of the National Academy of Sciences Jul2019, 116 (29) 14599-14605; DOI: https://www.pnas.org/content/116/29/14599.abstract

17. Youyang Gu Death forecasting Model https://covid19-projections.com/model-details/

18. Petrova, T.; Soshnikov, D.; Grunin, A. Estimation of Time-Dependent Reproduction Number for Global COVID-19 Outbreak. Preprints 2020, 2020060289 (doi: **10.20944/preprints202006.0289.v1**)

19. Weekly reports for coronavirus and COVID-19: https://www.fhi.no/en/publ/2020/weekly-reports-for-coronavirus-og-covid-19/

20. The R value and growth rate: https://www.gov.uk/guidance/the-r-value-and-growth-rate#:~:text=The%20R%20range%20for%20England,agreed%20by%20SAGE%20this%20week.&text=The%20R%20range%20for%20the,as%20of%2026%20March%202021.

21. Notes on R0 https://web.stanford.edu/~jhj1/teachingdocs/Jones-on-R0.pdf

22. Almut Scherer, Angela McLean, Mathematical models of vaccination, *British Medical Bulletin*, Volume 62, Issue 1, July 2002, Pages 187–199, https://doi.org/10.1093/bmb/62.1.187