Linköping University | Department of Computer and Information Science
Master's thesis, 30 ECTS | Statistics and Machine Learning
2021 | LIU-IDA/STAT-A--21/014--SE

# HMM modelling for the spread of the SARS–CoV–2

**Martin Beneš**

Supervisor : Krzysztof Bartoszek
Examiner : Maryna Prus

Linköpings universitet
SE–581 83 Linköping
+46 13 28 10 00 , www.liu.se

## Upphovsrätt

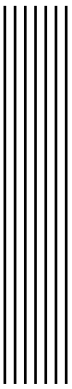


## Copyright

**Abstract**

The aim of the project is to develop an HMM for the current spread of the SARS–CoV–2 virus. The HMM could be coupled with a SIR+ based compartmental model for the different types of statistics—confirmed cases, hospitalizations, deaths. The confirmed cases should be treated as a random sample from the whole population of infected and the probability of sampling should try to take into account the different testing strategies.

The aim of the project would be to compare the spread of the virus in different countries (e.g. Czech Republic, Poland, Sweden, Italy, but other depending on the availability of data are possible) through regional (whenever possible) dynamics. For the thesis publicly available COVID–19 connected data will be used.

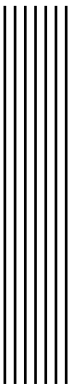


# Acknowledgments

During the writing of this thesis I got a great amount of support. My greatest gratitude goes to the thesis' supervisor Krzysztof Bartoszek, whose invaluable comments and piles of relevant research papers gave me much better insight into the field.

I would like to thank my girlfriend Kamila Nykiel, who encouraged me to work and gave me a helping hand whenever needed.

In addition, I would like to thank my parents and my family in our mother tongue. Mami a tati a celá rodino, jsem vám všem moc vděčný, že jste mě vždy podporovali nejen ve studiu. Mám veliké štěstí, že mými rodiči a rodinou jste zrovna vy.



# Contents

# List of Figures

# 1 Introduction

### Motivation

Currently there is an ongoing pandemic of Covid-19, one of the greatest challenges humans as a species had to face in last decades. Twentieth century introduced epidemiology as a research discipline and enabled spread of infections being viewed from mathematical rather than medical perspective. Public health, neglected just a few years ago [1, 2], got to the public eye now as the media chase its experts in these days.


Figure 1.1: Press Conference of Federal Chancellery of Austria, March 14, 2020, presenting new pandemic restrictions [3].

### Aim

This thesis is presenting a Hidden Markov model of Covid-19 spread and performs a simulation with it to approximately estimate the true situation about the infection for regions of several European countries: Czechia, Poland, Sweden and Italy.

### Research questions

To construct the model, the characteristics of the Covid-19 disease are investigated from relevant scientific literature and presented in form of probability distributions. The model definition requires answering of the question What are the distributions of parameters of Covid-19 - the incubation period, infection fatality ratio, reproduction number and duration of disease?

The Covid-19 statistics come from various sources and methods of measurement. Their correctness can be questioned not only in terms of accuracy, but in some cases even reliability of the source [4]. In other words, to what extent are the collected data used to fit the model reliable?

The result of a simulation with the HMM on various regional data can be evaluated using a similarity with the reported statistics. Apart from that, with a regional data one can answer Are there any patterns or similarities between the regions? Calendar of restriction can be analyzed such as Do the introduced restrictions influence the numbers?

The reliability of the simulation results is directly connected with the correspondence of the probabilistic definition of Covid-19 introduced in the thesis with reality. To what extent the results show that the drafted model of the disease is correct?

### Delimitations

The infection is modelled on a certain level of reality abstraction, so that many aspects of the infection are simplified. Those not included, but discussed are a multiple levels of infection severeness - asymptotic cases, a non-permanent immunity after recovery, mobility of population between regions/countries including incoming infectious population, vaccination and imperfect accuracy of the clinical tests.

# 2 Theory

## 2.1 Epidemics

Epidemic is commonly understood as an outbreak of a disease that freely spreads through the population. According to Encyclopedia Britannica, it is *an occurrence of disease that is temporarily of high prevalence* [5]. Epidemics and pandemics[1] are not only a matter of modern era, but occurred throughout the human history.

### Pre-modern epidemics

One of the oldest mentions in literature is an influenza epidemic in Persian Babylon in 1103 BC [6], however archaeological discoveries suggest even much older occurrences, such as the one in northeast China from $\sim$ 3000 BC [7].

By far the deadliest (in absolute numbers) [8] was a plague pandemic called Black Death from the $14^{\text{th}}$ century with a death toll around 25 mil. people [9], other epidemics caused by *Yersinia pestis* were Justinian's Plague ($540 - 750$ AD) [10], the Second Plague ($14^{\text{th}} - 19^{\text{th}}$ century) and the Third Plague ($1899 - 1940$'s) [11].

Another frequent epidemics were caused by influenza [12], cholera [13], tuberculosis, typhus or smallpox [14], the latter was eradicated in 1980 [15]. Some diseases are endemic such as yellow fever or malaria due to climate-dependent disease vector [16, 17], or Cocoliztli - a group of common diseases that decimated the Aztec population in mid $16^{\text{th}}$ century [18].[2]

### Modern epidemics

Regarding the pandemics, $20^{\text{th}}$ and $21^{\text{st}}$ centuries are dominated by the influenza - the Spanish flu ($1918 - 1920$), the Russian flu (1977) and the Swine flu (2019) caused by Influenza A/H1N1, the Hong Kong flu ($1968 - 1969$) and the Asian flu ($1957 - 1958$) caused by In-

[1] Epidemic is a general outbreak of disease. Pandemic is an epidemic that affects a significant portion of population of a continent, or worldwide [5].

[2] Civilizations on American continent developed isolated from the rest of the world for thousands of years and so did their immunity systems, adapted to the pathogens in the environment. Diseases brought by the first European colonizers (called *Cocoliztli*, in Nahuatl/Aztec meaning pest) were something absolutely novel for Americans' immunity systems and Cocoliztli wiped most of their population out.

fluenza A/H2N2 and its descendant A/H3N2 respectively [19]. The Spanish flu by itself directly caused 20 mil. deaths, far more than WWI [20].

Since early 1980's and still ongoing there has been a pandemic of a sexually-transmitted virus HIV[3], that causes AIDS[4], a disease that in the last 40 years killed more than 38 mil. people [21].

## 2.2 SARS-CoV-2

At the end of 2019 an outbreak of novel coronavirus occurred in Wuhan, China, later named *Severe acute respiratory syndrom coronavirus 2* (SARS-CoV-2). The virus shown in the figure 2.1 quickly spreaded around China and abroad and in just a matter of months, most of the world introduced epidemiological restrictions in order to stop the spread.

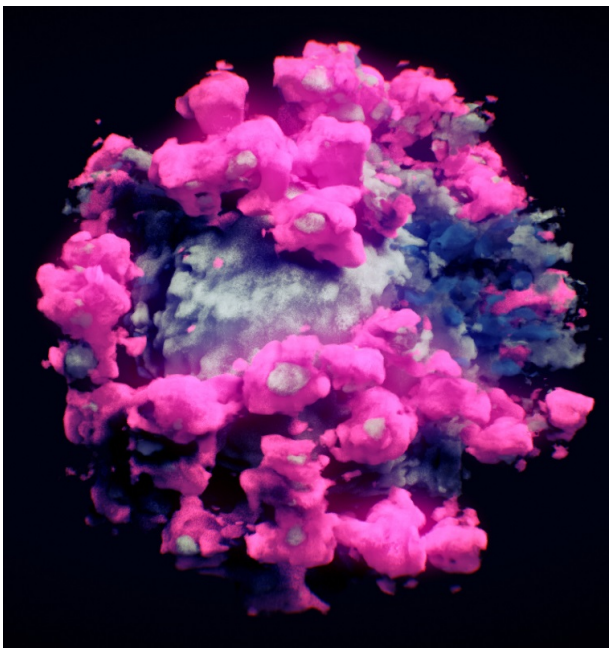


Figure 2.1: SARS-CoV-2 cryo-electron tomography scan [22].

The disease caused by SARS-CoV-2 is called Covid-19. The infected encounter respiratory illness with symptoms such as cough (67.6%), fever (62.2%), shortness of breath (32.4%), fatigue (24.3%), sore throat (21.6%) vomiting or diarrhea (21.6%), however their severity or absence vary significantly amongst patients. Fatality is significantly different (p-value $< 0.001$) for elderly population over 60 years, as shown in the figure 2.2 [23, 24, 25, 26, 27].

### Diagnostics

The method of collection of the data is an important factor for the correct evaluation of the analysis result. There are several broadly used diagnostic tests, lab-based and rapid, used for detecting of presence (past or present) of SARS-CoV-2 virus in the patient's organism [31]. There are two types with regard to what is being detected:

- *diagnostic tests* - virus itself or its parts (spike protein, RNA)
- *antibody tests* - antibodies produced by the host organism as a response to the virus

[3] Human Immunodeficiency Virus
[4] Acquired Immunodeficiency Syndrome

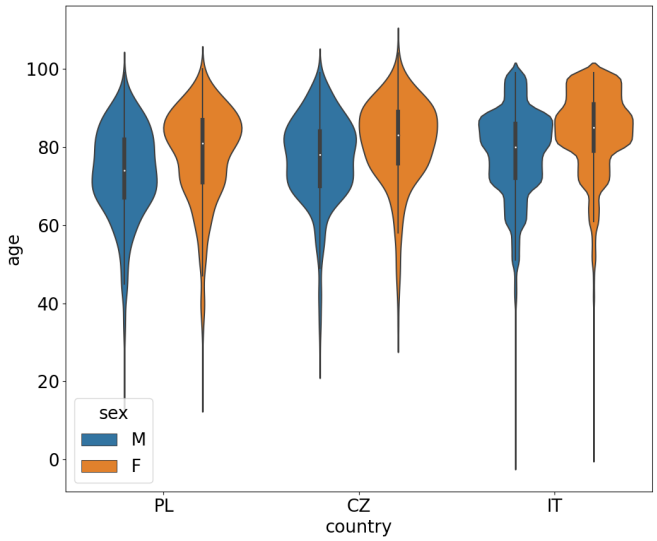



Figure 2.2: Covid-19 fatality, measured as deceased confirmed cases, per age and gender; Data from [28, 24, 29, 30].

**Diagnostic tests** Diagnostic tests detect an active Covid-19 infection. The sample for the test is a nasal or throat swab. After patient recovers, the test yields negative result again.

Currently the most commonly used is a RT-qPCR[5]. It consists of reverse transcription of viral RNA to DNA and amplification (replication) of the DNA, which happens only with the gene sequence of SARS-CoV-2.[6] Chain reaction activates fluorescent molecules, which indicate the presence of the DNA and hence the virus itself [32].

Different method of diagnosing an ongoing Covid-19 infection are antigen tests, that look for viral proteins specific for SARS-CoV-2. They are designed to have high sensitivity, but they have low specificity. The advantage is that they are rapid - they can be done by patient, do not require any special equipment and the result is known fast [33].

The accuracy of the tests depends on a sampling method and used kit, average performance from [34, 35] is listed in the table 2.3. Accuracy of the antigen tests is relative to RT-qPCR.

| Type of test | Specificity | Sensitivity |
|---|---|---|
| RT-qPCR | 98.787 % | 99.545 % |
| Antigen | 30.2 % | 100 % |

Figure 2.3: Accuracy of Covid-19 diagnostic tests [34, 35].

**Antibody tests** Antibody tests use blood serum and search for antibodies (IgM - early infection, IgG - long term immunity, ...) produced by the immune system as a response to encountered antigens - viral proteins. The testing makes only sense for person that already

[5]Reverse transcription quantitative polymerase chain reaction

[6]Replication of the DNA is selective thanks to customized primers, marking a start point of polymerase reaction. Polymerase is an enzyme capable of synthesis of DNA that duplicates each of the separated strains of DNA (only if matched by primers) in every reaction cycle. Amount of DNA grows exponentially.

had gone through the disease, as it measures a developed immunity. It is used for estimation of disease prevalence and infection fatality rate (IFR) [33].

## 2.3 Epidemiological Modeling

The epidemiology has experienced its first boom several years after so called Spanish flu[7], the first modern pandemic and at the beginning was described by medicine specialists, only later it was understood as an inter-discipline with mathematics and statistics [36].

### SIR* model

SIR* models, the most prominent class of epidemiological models, describe disease as a set of states with parameterized transitions between them. Each person in the modelled population has a state, that changes with certain probability according to the chosen model. Simple SIR model has three states and two connection: susceptible **S**, infected **I**, recovered **R**, connected such as S → I (getting sick) and I → R (recovering). Notation $X'$ denotes first derivation of $X$ w.r.t. $t$, i.e. $S' = \frac{dS}{dt}$, $I' = \frac{dI}{dt}$, etc.

$$\begin{aligned} S' &= -aSI \\ I' &= aSI - bI \\ R' &= bI \end{aligned} \tag{2.1}$$

SIR as a dynamic model can be expressed with differential equations (eq. 2.1). In this expression capital letters denote a number of people currently being in the state (as for time $t$). Each equation describes a change in the number of people for the state. Its terms correspond to the connections, e.g. number of people at time $t$ taking connection S → I is $aS_tI_t$.

SIR model describes diseases with fatality 0, no incubation period and permanent immunity. Since neither of this is true for SARS-CoV-2, for modelling we must extend SIR model by additional states: exposed **E** and dead **D**. All the states and connection of SEIRD model are shown in the figure 2.4.

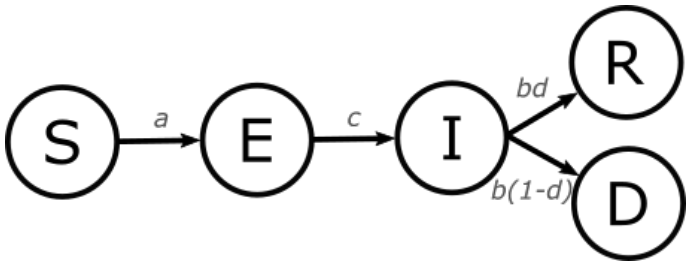



Figure 2.4: Example of SEIRD with permanent immunity.

As there is no connection R → S, this model assumes permanent immunity as well. Formally the SEIRD model can be defined as a set of differential equations (eq. 2.2).

The model parameters $a, c, b, d$ are directly connected (eq. 2.3) with the disease characteristics: basic reproduction number $R_0$, incubation period, symptom duration[8] and infection

[7]The name *Spanish flu* for the pandemic in late 1910's stems from the fact that media in Spain as one of the few neutral countries informed about the situation and casualties, while countries participating in WWI censored it not to cause a hysteria. Thus, situation in Spain looked much worse than elsewhere.

[8]Infectiousness and symptoms is in the SEIRD assumed to be equivalent terms. In reality they are not and they do not have to overlap.

$$\begin{aligned} S' &= -aSI \\ E' &= aSI - cE \\ I' &= cE - bI \\ R' &= b(1-d)I \\ D' &= bdD \end{aligned} \tag{2.2}$$

fatality rate IFR. These formulas are very often specified with the SIR* model definition as estimates for the model parameters [36, 37].

$$\begin{aligned} R_0 &= \frac{a}{b} \\ R_0(t) &= R_0 S(t) \\ \text{incubation period} &= c^{-1} \\ \text{symptom duration} &= b^{-1} \\ \text{IFR} &= d \end{aligned} \tag{2.3}$$

Basic reproduction number on whole population is computed as in equation 2.3. Time-dependent reproduction number $R_0(t)$ changes over time as individuals are leaving state S, becoming infected and recovered or dead. Thus $R_0(t)$ gets lower and lower with decrease of S.

The findings of epidemiology in a form of *theory of happenings* are applicable to wide range of different areas including marketing, malware, culture and others, making it a new and solid area of mathematics [38, 39].

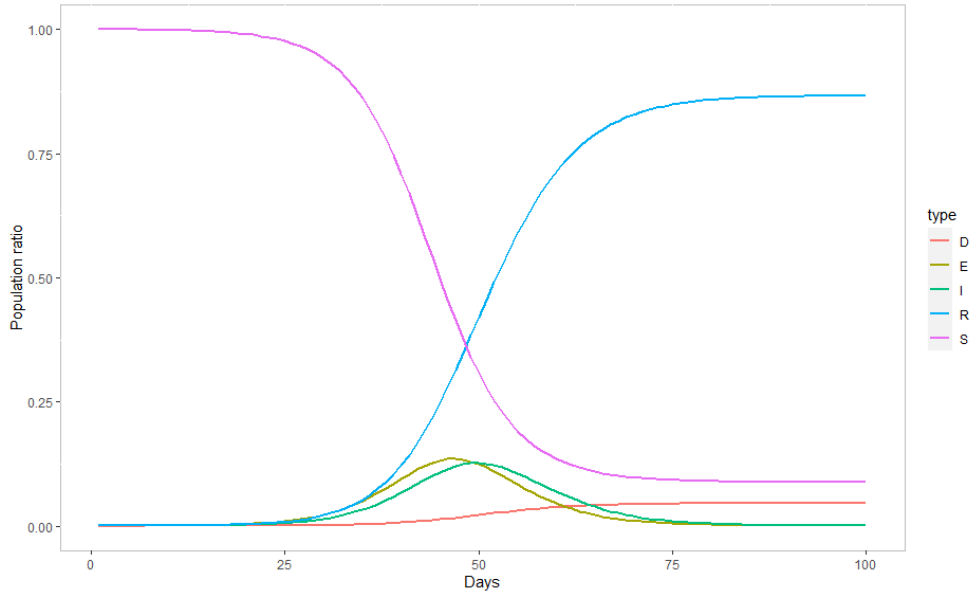



Figure 2.5: SEIRD dynamics, example 1
SEIRD dynamics for $(a, c, b, d) = (0.8, 0.3, 0.3, 0.05)$.

Figure 2.5 shows the dynamics for parameters $a = 0.8$, $c = 0.3$, $b = 0.3$ and $d = 0.05$ (constant over time) and initial values $I = 1$, $S = 9999$, $E = R = D = 0$. Incubation period is $\frac{1}{c} = \frac{1}{0.3} = 3.\overline{3}$ days and duration of symptoms $\frac{1}{b} = \frac{1}{0.3} = 3.\overline{3}$ days. Basic reproduction number with specified parameters is $R_0 = \frac{a}{b} = \frac{0.8}{0.3} = 2.\overline{6}$.

The disease is slow at the beginning. Then the exposed and infected exposed. Infected $I$ is delayed behind exposed $E$ and also covers larger AUC[9], because duration of symptoms is longer than incubation period. Death counts $D$ are also delayed behind infected $I$. At certain point infection starts slowing down and eventually stops. This moment of the epidemics is called *herd immunity*.

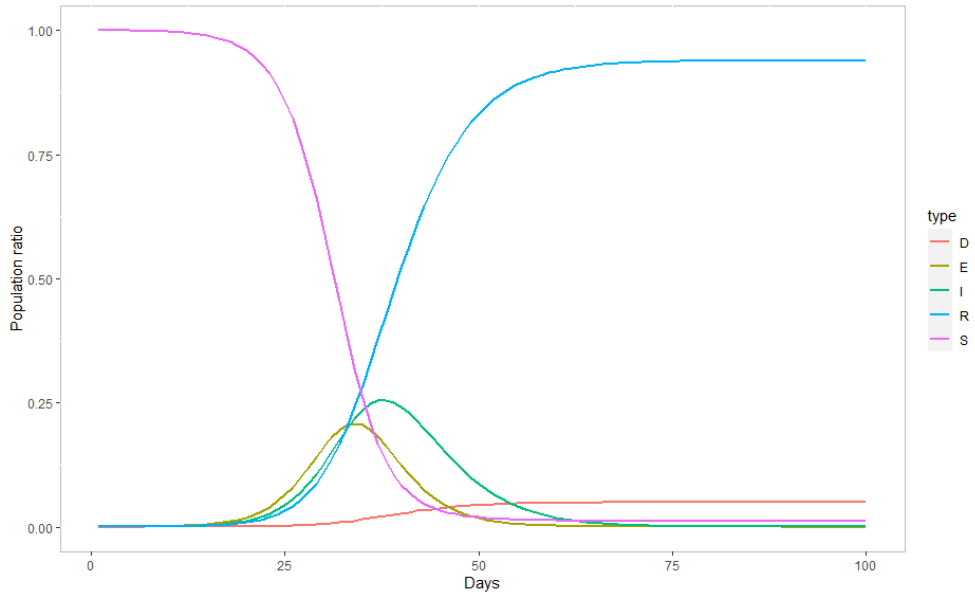



Figure 2.6: SEIRD dynamics for $(a, c, b, d) = (0.9, 0.3, 0.2, 0.05)$.

Figure 2.6 visualizes disease that spreads faster and whose symptoms last longer, parameters change such that $a = 0.5 \to 0.9$ and $b = 0.3 \to 0.2$. Basic reproduction number of this disease is $R_0 = \frac{a}{b} = \frac{0.9}{0.2} = 4.5$ and the symptoms' duration is $b^{-1} = 5$ days.

The herd immunity is reached with with almost all population being infected. This more contagious and longer lasting disease leaves more deaths behind and penetrates the population very fast. The curve of infected has much greater AUC, which is directly connected with occupancy of hospitals and potential collapse of hospitals.

Figure 2.7 visualizes SIR model with low value of parameter $a$ and y axis in logarithmic scale. Initial state assumes whole population of 10000 individuals susceptible (i.e. $S = 1$, as states are normalized by population), where single person becomes infectious. In the result, susceptibles are constant over time, recovered and deaths grow slightly at first as the initial infectious recover or die, later the differential system becomes stable.

To get the true numbers, real values are rounded: infected and exposed becomes 0, infected person recovers and the infection has no casulties. Basic reproduction number of this infection is $R_0 = \frac{a}{b} = \frac{0.5}{0.5} = 1$, time dependent reproduction number $R_0(t = 0) = R_0 S(t = 0) = \frac{9999}{1000} < 1$, so that each infection produces (nearly) one other infection, which means epidemic slowly dies out. If reproduction number is equal 1, number of infected is constant over time [40, 37].

[9] Area under curve

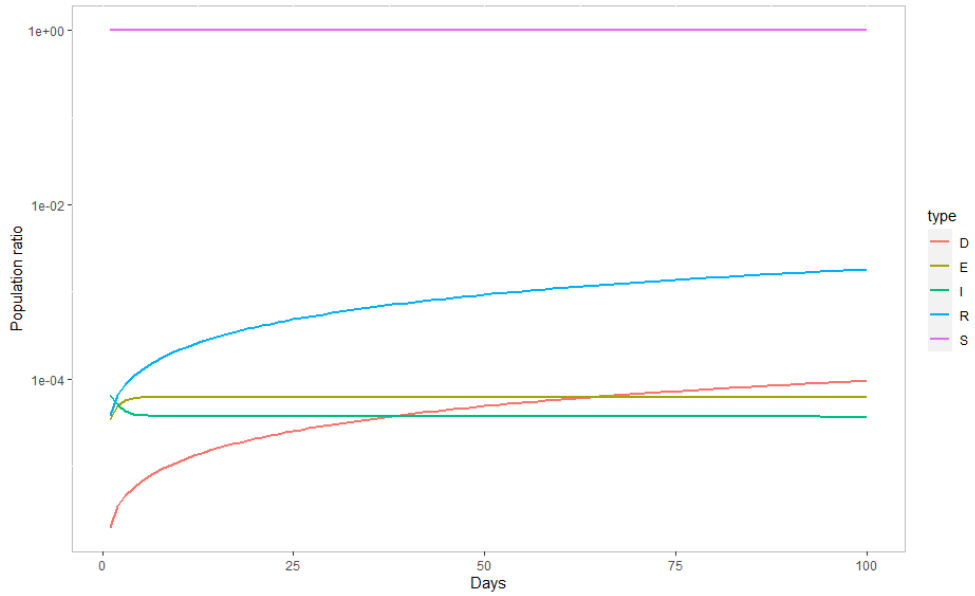



Figure 2.7: SEIRD dynamics for $(a, c, b, d) = (0.5, 0.3, 0.5, 0.05)$, $I = 1$.

### Bayesian SEIRD* model

Bayesian approach in epidemiological modelling brings uncertainty of parameter values. Instead of single value for the parameter, the model considers parameters $a, c, b, d$ to have a prior distribution representing our best guess supported by prior knowledge. Choice of prior is based on clinical measurements of the infected individuals.

Posterior probability contains both the prior and the information extracted from the data modelled by a likelihood distribution using a Bayes' theorem.

$$\begin{aligned} \theta &= (a, c, b, d) \\ P(\theta|D) &\propto P(D|\theta)P(\theta) \end{aligned} \tag{2.4}$$

Connection of parameters $a, c, b, d$ with the disease characteristics as specified in stochastic manner by the equation 2.4 is slightly different for Bayesian approach, the terms on both sides of equation sign are equal by distribution, denoted $\overset{d}{=}$. The alternation is shown in the equation 2.5.

$$\begin{aligned} R_0 &\overset{d}{=} \frac{a}{b} \\ R_0(t) &\overset{d}{=} R_0 S(t) \\ \text{incubation period} &\overset{d}{=} c^{-1} \\ \text{symptom duration} &\overset{d}{=} b^{-1} \\ \text{IFR} &\overset{d}{=} d \end{aligned} \tag{2.5}$$

## 2.4 Hidden Markov Models

Hidden Markov model (HMM) is a discrete stochastic model for time-series. It has two main components: latent (unobserved) variables called states and observations, that are to be modelled. Since the model is discrete, both the latent and the observed variables have predefined finite alphabet of symbols (e.g. finite set of numbers) they can contain.

The parameters of the model are transition and emission probability distributions. One of important assumptions of HMM is stationarity, the distributions are constant in time - what is changing are the states. The figure 2.8 shows the structure, transition probabilities are jointly denoted $\tau$ and emission probabilities $\varepsilon$.

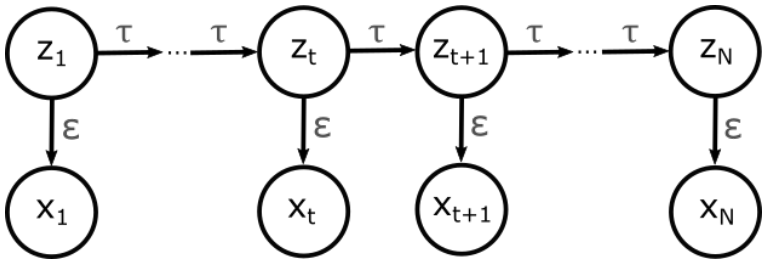



Figure 2.8: Hidden Markov model structure.

### Transitional probability

The transitional probabilities are defined between latent variables from time $t$ to time $t+1$ defining the distribution of the transition in each time step. These distributions can be aligned to a *transition matrix*.

### Emission probability

The emission probabilities define the observations $x_t$ based on the latent variables $z_t$. Similarly as the transitional matrix emission distributions can be aligned to *emission matrix*.

### Learning of HMM

Learning of HMM means an estimation of the distribution for $z(t)$ over $\forall t$. Analytical approach is Forward-Backward algorithm, producing distributions $\alpha(t)$ and $\beta(t)$, which are base for simulation using filtering and smoothing methods. Different way to predict the most probable path without learning the model is Viterbi algorithm.

**Forward-backward** The Forward-backward algorithm is used for filtering and smoothing. Filtering uses for estimating of distribution for $z_t$ (state at time $t$) only observations from the past $x^{0:t}$ and can be used for real time processing. Smoothing estimates distribution of $z_t$ all observations from the $0 : T$, where $t \in 0 : T$. Smoothing cannot be used for real time processing but is more accurate.

**Viterbi** Neither filtering nor smoothing do not guarantee that the output will be valid according to the transition and emission matrix, but they simply maximize the total score. The Viterbi algorithm does not learn the distribution for $z(t)$, but focuses on producing a valid output according to the transition and emission matrix, so its output "makes more sense" and it is widely used in some domains, such as natural language processing.

For estimation Viterbi uses all the observations from $0 : T$, same as smoothing, but due to the constraints it tends to be less accurate.

**Markov Chain Monte Carlo** Bayesian approach to HMM uses random simulations from the transition and emission distribution and minimizing of the log likelihood by searching for optimal parameters. There are frameworks that make the method easy to use, such as *Stan* [41].

## 2.5 Splines

Spline is a mathematical technique to interpolate a group of points with a piece-wise curve. Each segment or piece has is represented by a function, typically polynomial. Smoothness of the curves from two adjacent segments is ensured by a common (equal) derivative value, level of smoothness $C^k$ means equality in values of corresponding derivatives of orders $[0, \ldots, k]$. Well-known linear spline with level of smoothness $C^0$ is absolute value [42].

Figure 2.9 shows spline over interval $[-3, 3]$ with nodes in $[-3, 4], [-1, 0], [1, 1], [3, \frac{42}{18}]$. Level of smoothness in point $[-1, 0]$ is $C^2$, which means that both curves have the function values and the values of derivatives of first and second orders equal in this point. In point $[1, 1]$ level of smoothness is $C^0$, which means that the function values of both curves are equal, but the derivatives are different. Their difference is visible by bare eye.

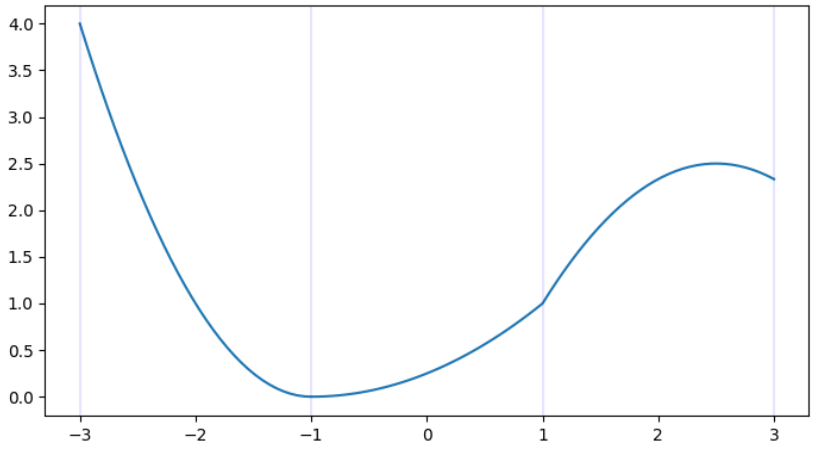



Figure 2.9: Spline with levels of smoothness $C^2$ in $-1$ and $C^0$ in 1.

Splines can be also used for approximation [43]. In this thesis, the piece components will be curves yielded as a result of numeric integration of SEIRD model.

The term spline is sometimes used for a range of different piece-wise functions, although it is its specific case, if we assume that spline must have level of smoothness at least 0. For this definition counterexamples are all non-continuous piece-wise functions, such as sign and indicator functions.

# 3 Data

The model is defined (that means its transition and emission components) using results of clinical measurements of patients' disease characteristics: incubation period and duration of symptoms, results of molecular and antibody tests, but also tracing information (for estimation of $R_0$) and others.

The input data for the model are the statistical information of the Covid-19 infection: counts of positively tested individuals, number of deceased on the disease etc.

## 3.1 Covid-19 statistics

The statistics are usually reported by national or regional authorities, responsible for publishing them - government institutions or public agencies - ministries of health, statistical offices or regional hygienic offices. The data can come in daily or weekly records with country-wise, regional, sub-regional (district-wise) or municipal administrative unit granularity.

### Data attributes

Although different authorities publish Covid-19 statistics in different formats, throughout those occur common attributes [44].

**Tests** First statistics is the number of performed tests. This statistic should ideally contain only number of diagnostic tests of individuals done to confirm their infection and repeated tests confirming recovery or antibody tests should be excluded. However some countries publish overall number of all tests performed regardless of type.

**Confirmed** The tests can be seen as a sample over a population and the confirmed cases is a number of positive tests per day. If the probabilities of infected and healthy getting tested are different, the test sample is biased.

The value of confirmed should ideally resemble the total infected, but since tests do not cover the whole population, it is influenced by number of tests and the sample bias.

**Deaths** Deaths is a death toll of Covid-19. This has been an intensive matter of dispute (dying with/on Covid-19), as a Covid-19 death can be understood on one hand solely as a direct

consequence of infection Covid-19, or dying while being Covid-19 positive regardless on the cause of death or comorbids on the other. In the latter case, Covid-19 positive passenger of a car dying in a car accident is reported in the statistics [45, 46, 47, 48].

**Hospitalized** Another reported number is number of hospitalized patients positive for Covid-19. Hospitalization can be in several modes coming with raising severeness of the infection, the data often contain additional current number of patients at intensive care unit and with connected ventilator, although often researches uses more detailed information of hospitalization.

Decision of hospitalization is often based on patient's state. Many researches are performed on sample of hospitalized people. Thus if asymptomatic and mild cases are eliminated from the sample, measurements on such sample might be skewed.

**Prevalence** Prevalence is the percentage of infected in population. The number can be either estimated real-time with molecular tests or backwards with antibody testing. The latter allows more careful sample selection and better results. Antibody testing can be under certain conditions performed even during active infection.

**Fatality ratio** Fatality ratio is a percentage of how deadly an infection is. It is derivated from the prevalence and deaths (eq. 3.1). Dependent on how the prevalence is estimated we distinguish *case fatality ratio* (CFR) and *infection fatality ratio* (IFR).

$$\begin{aligned} \text{CFR} &= \frac{\text{Number of deaths}}{\text{Confirmed by tests}} \\ \text{IFR} &= \frac{\text{Number of deaths}}{\text{Truly infected}} \end{aligned} \tag{3.1}$$

**Recovered** Recovered is the number of confirmed patients that did undergo the disease and on the given day received first negative test confirming their recovery.

### Data sources

**Czechia** The official data for Czechia are published by MZ ČR[1]. Most statistics cover the whole epidemic (since March 2020) and as for now contains following data attributes in *daily* time slots[23]:

- **Country**: RT-qPCR + antigen tests
- **District**: deaths, tests, hospital capacities and stock states
  - **Per age group**: incidence, prevalence, hospitalized, vaccinated
  - **Cases with age and gender**: confirmed, deaths
- **Municipality**: confirmed

The fetching of the data are implemented in the Python package `covid19czechia` [28]. The usage is shown in the listing 3.1.

[1] Ministerstvo Zdravotnictví České Republiky/The Ministry of Health of the Czech Republic

```
import covid19czechia as CZ
x = CZ.covid_deaths()
```

Listing 3.1: covid19czechia: usage example

**Poland** The responsible institution to publish the data for Poland is MZ RP[2]. Until October 10 2020, the regional data of confirmed and deaths were published via Twitter account @MZ_GOV_PL as daily updates. Deaths were reported as cases with gender and age. Via government webpage one could only acquire current counts in regions [49].

At the moment regional data between October 10 2020 to November 23 2020 are not published on either of the official sources mentioned. Since November 23 2020, MZ RP started to publish daily a CSV file on their webpage with regional counts (without gender or age information).

Currently data between January 20 2021 to February 28 2021 (today) are missing [50].

- **Country**: tests, recovered, hospitalized, quarantined
- **Region/municipality**: confirmed, deaths

The package `covid19poland` contains data collected both webscraped from Twitter and fetched from the MZ official webpages [24]. The sample code is in the list 3.2.

```
import covid19poland as PL
x = PL.covid_deaths()
```

Listing 3.2: covid19poland: usage example

In November 2020, Michał Rogalski pointed out issues with the official Polish statistics. The statistics reported by the government (MZ RP) should perfectly aggregate the data reported by the regional PSSE[3], but in fact in certain period of time they differed greatly (by 22000) [51]. Whatever reason for this phenomenon might have been, the statistics were heavily under-reported [4].

Additional source is a Michał Rogalski's public data collection COVID-19 w Polsce accessible on author's Google Drive, where he manually collects the statistics [52].

**Sweden** Sweden's official Covid-19 statistics are managed by FOHM[4], which publishes information about the current situation on weekly basis in PDF reports. FOHM also provides XLSX with more detailed data such as daily deaths, confirmed, intensive care unit cases and applied vaccines, and weekly confirmed and deaths per municipality [53].

- **Country**: deaths, icu, confirmed
- **Region**: icu (weekly), vaccines, tests - antibody
- **Municipality** (weekly): confirmed, deaths

The package `covid19sweden` contains data collected from the XLSX [54].

```
import covid19sweden as SE
x = SE.deaths()
```

Listing 3.3: covid19sweden: usage example

[2]Ministerstwo Zdrowia Rzeczypospolitej Polskiej/Ministry of Health of the Republic of Poland
[3]Powiatowe Stacje Sanitarno-Epidemiologiczne, Regional Sanitary-Epidemiological Stations, also *Sanepids*.
[4]Folkhälsomyndigheten/The Public Health Agency of Sweden

**Italy** In Italy, the data are published in a structured form on Github account of Dipartimento della Protezione, Presidenza del Consiglio dei Ministri - Civile[5] [55]. These data include

- **Country**: confirmed, deaths, recovered, tests (molecular, antigenic), quarantined, hospitalized (positive, positive with symptoms), suspected, ...
- **Region**: *Same as for country.*
- **Province**: *Same as for country.*

The implementation does not read the data directly from [55], but uses Python package `covid19dh` for it [30].

The dataset about the age distribution of death cases used in the plot 2.2 comes from the ISS[6].

**Covid-19 Data Hub** Since the Covid-19 data sources are publishing data in different formats, there are many projects collecting and unifying the data to make the access to them easy; *Covid-19 Data Hub* of Guidotti and Ardia used in this thesis for fetching Italian data is one of them [30].

### Data transformation

SIR model uses values in $[0; 1]$, so confirmed (daily incidence) are normalized by number of performed tests and cumulative recovered and deaths are normalized by cumulative tests. Data are unified by source-dependent data transformations as each source yields different format of the output.

### Data visual analysis

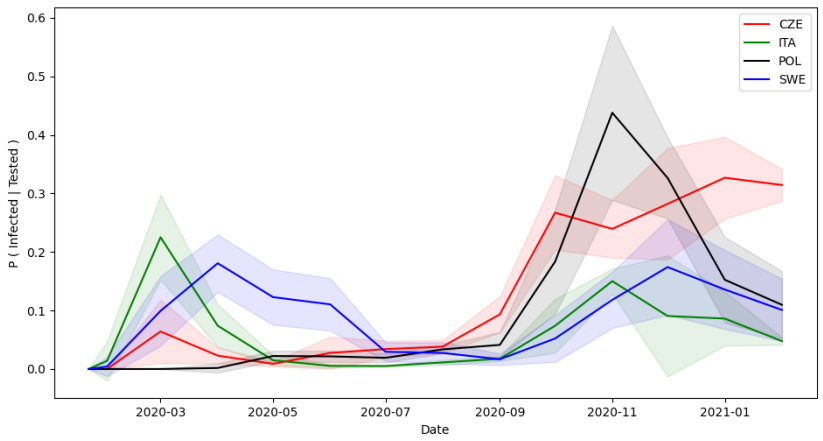



Figure 3.1: Ratio of positive tests in Poland, Sweden, Czechia and Italy.

In the ratio of positive tests to total tests (fig. 3.1), there are two peaks, separated by the summer 2020: in media and literature the period before July 2020 is called *the first wave*, while the period after July 2020 is called *the second wave* [56].

[5]Department of Civil Protection of Presidency of the Council of Ministers of Italy
[6]Istituto Superiore di Sanità, Higher Institute of Health of Italy.

The first wave is greater in Italy in March 2020 and in Sweden between April and June 2020. However the daily incidence by positive tests should not in the initial phase trusted without taking a look on the number of tests performed (fig. 3.2). Daily tested proportion of population is quite small before the summer ($0 - 0.15\%$) compared to second wave, so the numbers might not represent the true incidence.

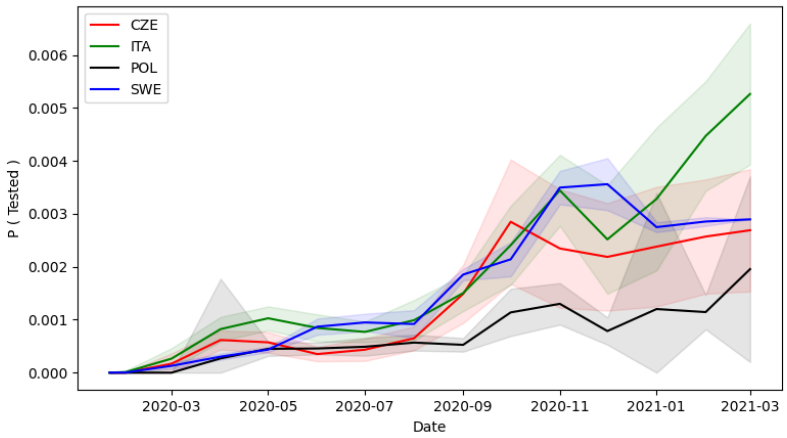



Figure 3.2: Ratio of tests over population in Poland, Sweden, Czechia and Italy.

During the second wave, Poland still performs low number of tests, other countries raise the test count almost 3 times.

The second wave grows first (and fastest) in September in Czechia and in October in Poland. Sweden and Italy has the second wave milder and delayed compared to Czechia and Poland. The curve peaks in November in Italy and Poland, then descends, in Sweden the peak comes in December. Positive tests' ratio stays equal over time in Czechia, but in Poland and Italy the curve abruptly falls down after November. In Sweden and Italy the second wave is more mild and delayed compared to Czechia and Poland.

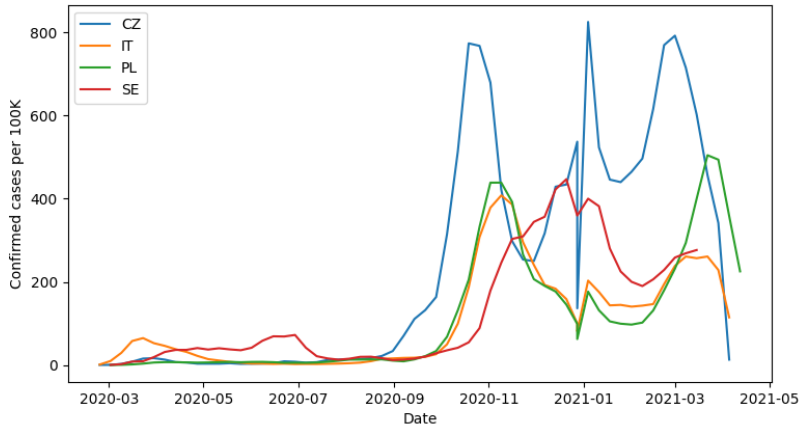



Figure 3.3: Weekly confirmed cases per 100 000 people in Poland, Sweden, Czechia and Italy.

Confirmed cases per 100000 people (fig. 3.3) look very similar to the positive tests' ratio.

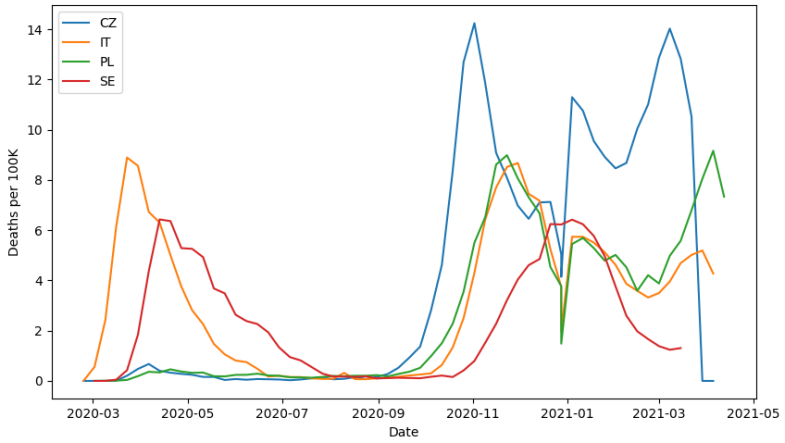



Figure 3.4: Weekly Covid-19 fatality per 100 000 people in Poland, Sweden, Czechia and Italy; Data from [28, 24, 54, 30, 57].

The curves of Covid-19 deaths (fig. 3.4) remind of the confirmed cases, but there is several interesting features. In Sweden, number of deaths goes down from May 2020, although the cases raise and peaks at the end June 2020. This might be connected with outbreaks of disease in retirement homes and national ban for visiting them by public in the second half of March [58].

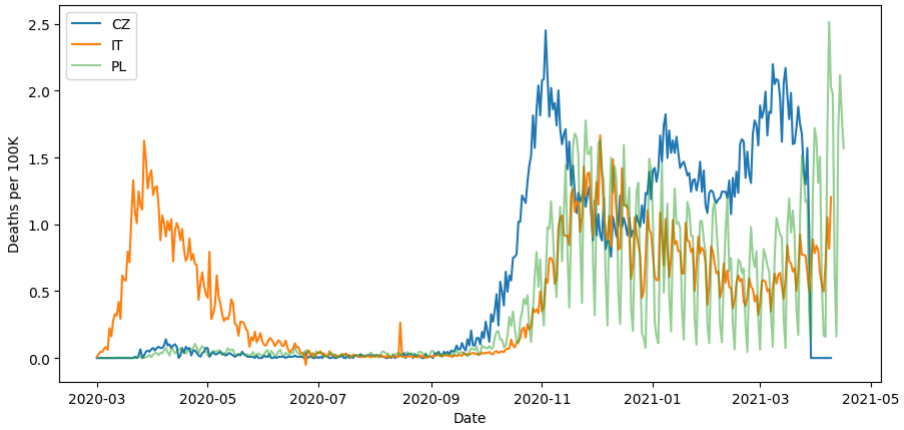



Figure 3.5: Daily Covid-19 fatality per 100 000 people in Poland, Czechia and Italy; Data from [28, 24, 54, 30, 57].

The greatest magnitude of weekly seasonality in daily deaths (fig. 3.5) is without any doubt present in Poland. According to my email correspondence with Bartosz Stawowski, a head of Departament Analiz i Strategii[7], Ministry of Health of Poland it is *a result of issuing the death certificate and reporting deaths to the Registy Office. As a consequence a certain percentage of fatal cases from weekends are reported on working days*.

[7]Department of Analyses and Strategy

In Italy and Czechia, this seasonality is oscillating much less and according to Veronika Hruba, the referent of the department of communication and PR of Ústav zdravotnických informací a statistiky (ÚZIS)[8], the reported deaths have the time of the actual death, not of issuing of the death certificate, as in case of Poland. Trend in Italy reminds of the one in Czechia. Sweden publishes only weekly data of deaths and so this trend can not be investigated.

**Administrative division** The administrative units for regional data are based on what division is used in the data of Covid-19 statistics, published by each of the countries. All 4 countries as EU members have regions with NUTS[9] codes for statistical purposes, table 3.6 specifies what level do the data have [59].

| Country | Division | Notes |
|---|---|---|
| *Czechia* | NUTS-3 | |
| *Italy* | NUTS-2 | ITH10 and ITH20 instead of ITH1 and ITH2. |
| *Poland* | NUTS-2 | PL91 and PL92 aggregated into PL9. |
| *Sweden* | NUTS-3 | |

Figure 3.6: Administrative divisions used in data.

**Timestep size** The minimal time step defined in the statistics is a day, although Sweden publishes some data only weekly, which is why the basic simulation step is a week, only in special cases simulation uses daily time step.

As the epidemic is changing slowly, it might be sufficient to estimate the time dependent parameters with fixed-sized windows. This accelerates the computation, but brings additional issues regarding alignment, because some changes might be on the edge of two windows. If the window is small enough, the problem is negligible.

## 3.2 Demographical Statistics

The demographical data of mortality and deaths have been acquired from the Eurostat.[10] The age distributions of mortality per country and gender are shown in the figure 3.7. The plot contains pure density over age groups of mortality, ignoring the population size in each age group.

Poland and Czechia, two countries with similar cuisine, lifestyle and historical context since the WWII remind each other with shape, although Poland has heavier tail towards younger age, especially in male population, which makes Polish life expectancy per 1000 people statistically lower than the Czech. The mode for both countries is between 80 and 85 years. There is a small bubble in Poland in the age group 0 - 4 years.

Mortalities of Sweden and Italy are similar, there is no statistical evidence for them differing in variance or mean, but they significantly differ from Czechia and Poland. There is a number of feasible explanations for the mutual similarities of the country mortalities Sweden-Italy and Czechia-Poland, they will be listed, but not further investigated. An important aspect for difference in life expectancy is balanced food, doing actively a sport on a regular basis, limited smoking, drugs and consumption of alcohol, mental health, positive attitude to life but also for example air pollution or political freedom. [61, 62, 63]

In Italy, Poland and Czechia, women live significantly longer than men ($\alpha = 5\%$), in Sweden there is no statistical evidence for that.

[8] Institute of medical information and statistics of the Czech Republic

[9] Nomenclature of territorial units for statistics (NUTS) is a European standard encoding of regions.

[10] European Statistical Office (Eurostat) is an EU institution responsible for data managing and publishing.

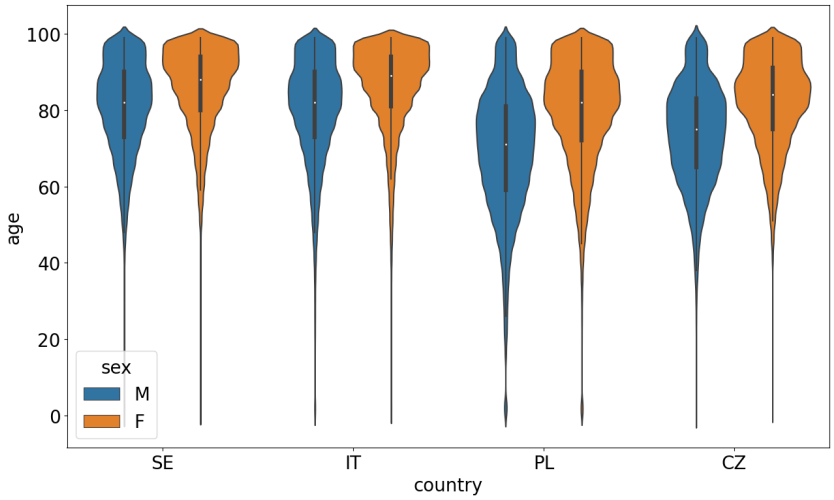



Figure 3.7: Country mortality in 2020 [57, 60].

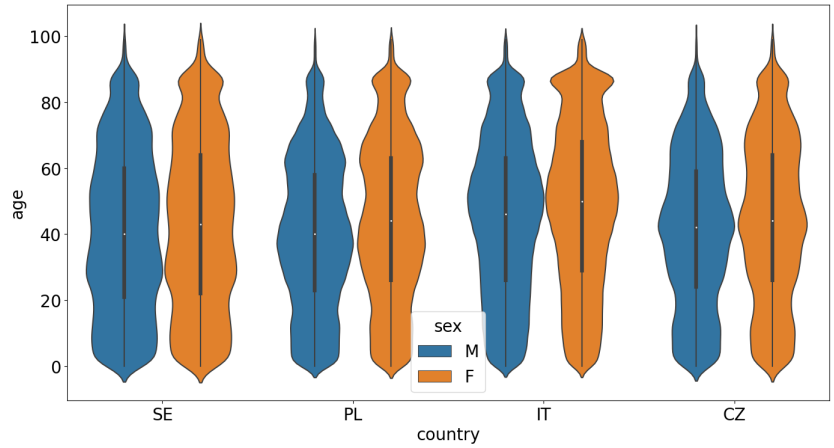



Figure 3.8: Country populations in 2020 [57, 64].

Figure 3.8 contains demographic distribution of population in year 2020. Italy has a peak in age of $45 - 55$, i.e. year of birth $1965 - 1975$, most likely a consequence of *il boom economico* (Italian economic boom) that transformed economy from agriculture into industrial and had a great economical and sociological impact on Italian society, which definitely might have caused great changes in trends of natality and immigration [65].

The effects of WWII are visible in the plot of all countries as there is much less people in the age group $80 - 85$, born $1940 - 1945$. As the supply of food and basic needs were limited during the war and the overall economy was influenced not only in the countries directly participating in the war, but also neutral countries, such as Sweden. Italy has a great population peak in the population born right before the WWII in age group $85+$, which could be a result of Battle of Births, Benito Mussolini's pro-natal politics to boost the birth rate up [66].

Poland contains two major peaks - $60-65$ years and $35-45$ years, born $1955-1960$ and $1985-1995$. Poland was devastated after the war and population was significantly reduced. After the war there occurred an effect of *demographic compensation*, that is common after wars or other significant population reduction. Second wave is the baby boom echo of the first wave [67].

Czechia has a great peak at age group $45-55$, which is caused by family-supportive normalization politics, which produced strong generation known as *Husákovy děti*. Fall of communistic regime opened new opportunities for realizations in professional and personal life and as a consequence even the average age of mothers grew up. The backside of this was a lower birth rate between $1990-2000$. This weak generation is sometimes called *Havlovy děti* [68].

| | $H_0 : \sigma_1 = \sigma_2$ (F-test) | | | | $H_0 : \mu_1 = \mu_2$ (T-test) | | | |
|---|---|---|---|---|---|---|---|---|
| | **CZ** | **IT** | **PL** | **SE** | **CZ** | **IT** | **PL** | **SE** |
| **CZ** | | 0.0862 | 0.3783 | 0.0504 | | 0.0000 | 0.2557 | 0.1973 |
| **IT** | 0.0862 | | 0.0033 | 0.2193 | 0.0000 | | 0.0000 | 0.0000 |
| **PL** | 0.3783 | 0.0033 | | 0.0088 | 0.2557 | 0.0000 | | 0.3356 |
| **SE** | 0.0504 | 0.2193 | 0.0088 | | 0.1973 | 0.0000 | 0.3356 | |

Figure 3.9: Hypotheses' tests for populations comparison (p-values), population in ten thousands.

According to results of tests (table 3.9), Italian mean age differs from the rest of the countries. Population is divided by $10^4$ as a sensitivity setting. By variance, Poland differs from Italy and Sweden on usual level of significance ($\alpha = 0.95$), using $\alpha = 0.9$ Czechia differs from Italy and Sweden too.

### Mortality in age group 0-4 years

As mentioned in the visual analysis of figure 3.7, there is a small bubble in Poland in the age group $0-4$ years. The mortality is compared over years $2014-2020$ with other age groups of young age (fig. 3.10).

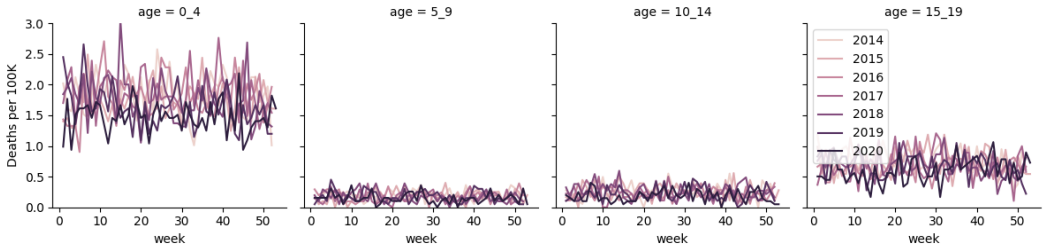



Figure 3.10: Mortality in Poland over age groups $0-4$, $5-9$, $10-14$ and $15-19$ in $2014-2020$.

For comparison in international context, similar plots are produced for Czechia (fig. 3.11), Italy (fig. 3.12) and Sweden (fig. 3.13)

Polish mortality in the age group $0-4$ years actually appears to be higher than in age groups $5-9$, $10-14$ and $15-19$. If we take only the data from 2020, mortality of the age group $0-4$ in Poland is significantly greater than in Czechia, Italy and Sweden ($H_0 : \mu_{PL} < \mu_X$, all p-values 1). Running both-sided t-test between these three countries ($H_0 : \mu_X = \mu_Y$), all of them have comparable mortalities in the age group $0-4$ (p-values $> 0.8$)

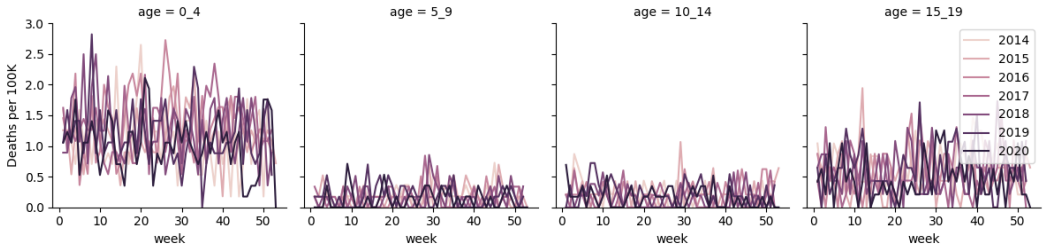



Figure 3.11: Mortality in Czechia over age groups $0-4$, $5-9$, $10-14$ and $15-19$ in $2014-2020$.

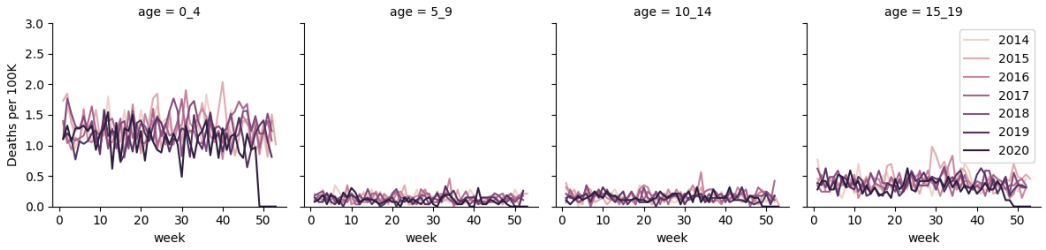



Figure 3.12: Mortality in Italy over age groups $0-4$, $5-9$, $10-14$ and $15-19$ in $2014-2020$.

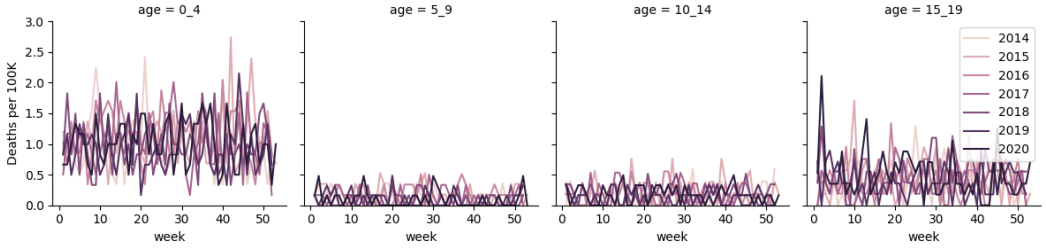



Figure 3.13: Mortality in Sweden over age groups $0-4$, $5-9$, $10-14$ and $15-19$ in $2014-2020$.

## 3.3 Calendar

To stop the spread of the Covid-19, national and regional governments imposed various restrictions or recommendations about behavior. These could have change the parameters of the disease and thus cause a change of progress, shown as a certain feature in the statistics.

To interpret these features in terms of possible restrictions that could have caused them, these events were collected. Amongst items of interest are dates of imposing or releasing of restrictions that influenced a behavior of population, but also special events, which cause that people move, gather and meet each other, such as national holidays, elections or demonstrations. Other important events are those related to change in statistics, such as change in strategy of testing or statistical corrections.

Later in the thesis it is discussed, what event could have caused various peas or changes in trend and whether it seems that a certain restriction does in general imply improvement of the pandemic situation.

# 4 Method

## 4.1 Model

HMM consisting of **transition** and **emission** models. Transition model connects latent states of time step $t$ with time step $t + 1$. Emission model connects latent state with observed variable of time step $t$. Structures for both transition and emission models are shown in the figures 4.1 and 4.2 respectively.

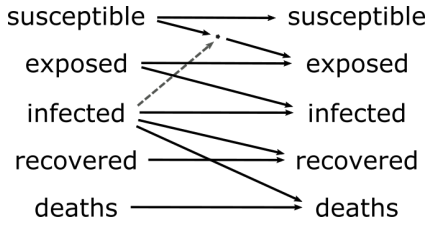



Figure 4.1: HMM transition structure.

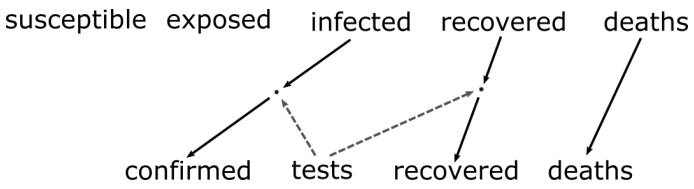



Figure 4.2: HMM emission structure.

Precise constructing of the models means seeking distribution of the parameters corresponding with the characteristics of the Covid-19 disease.

### Covid-19 characteristics

The characteristics of Covid-19 infection are needed to be able to model the outbreak. As SEIDR is used, the objectives are distributions of following variables

- Duration of incubation period
- Duration of disease since symptoms
- Reproduction number $R_0$
- Infection fatality rate (IFR) - investigated in age groups separately

There is several methods to acquire these characteristics

- Clinical measurements = (anonymized) information about hospitalized patients - *incubation period*, *duration of symptoms*
- Antibody tests = presence of antibodies in organism, signs that person had the disease - *prevalence*, *IFR*
- Tracing = reconstruction of the infection transmission graph in the population by detecting contacts of positively tested - *serial interval*, *reproduction number*

**Incubation period** A research measuring the incubation period length cited even by WHO[1] in precausion recommendation [69] estimates the median incubation to be 5.1 days, although 95% of all cases experiencing $2.2 - 11.5$ days and 50% of all cases experiencing $3.8 - 6.7$ days.

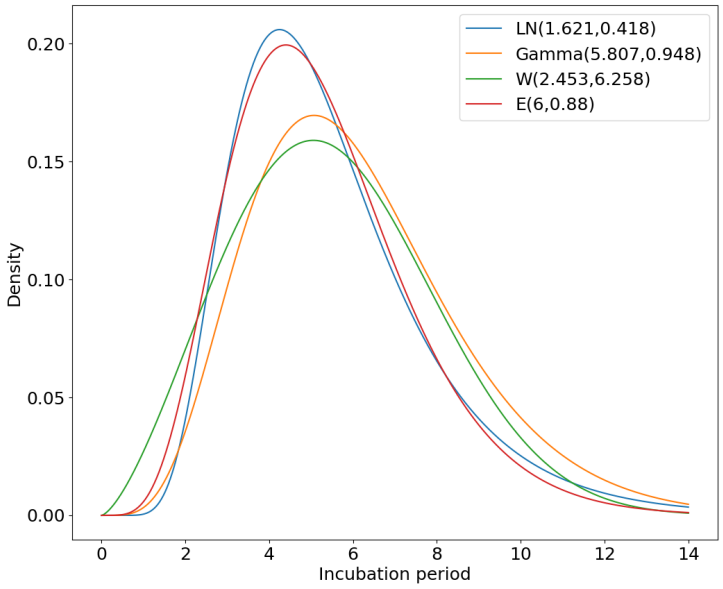



Figure 4.3: Estimated distributions of incubation period duration [70].

The paper also estimated several parametric distributions to the data, shown in the figure 4.3. The best one fitting to the data selected using lowest MSE of its quantiles to the data quantiles is $\Gamma(5.807, 0.948)$, the results are shown in the table 4.4 [70].

[1] World Health Organization

| **Distribution** | LN(1.621, 0.418) | Γ(5.807, 0.948) | W(2.453, 6.258) | E(6, 0.88) |
|---|---|---|---|---|
| **MSE** | 0.438798 | 0.427651 | 0.666146 | 1.022750 |

Figure 4.4: Incubation period distributions' goodness-of-fit by quantile MSE.

$$P_{X_d}(i) = \int_i^{i+1} f_X(t)\,dt = F_X(i+1) - F_X(i), i = 0, 1, 2, \ldots \tag{4.1}$$

If the distribution is to be modelled in using transition matrix, we need to discretize the distribution to get probability of symptom onset per day since exposure. Using equation 4.1 we get distribution from the figure 4.5. The probability density function is denoted $f_X(t)$, the distribution function $F_X(t)$.

The domain of the random variable $x$ is limited to $i \in \{0, 1, \ldots, 20\}$, as less than 0.01% of cases had incubation period longer than 20 days. The probabilities can be found in `data/distr/incubation.csv`. Similarly looking distribution was reported by [71] too.

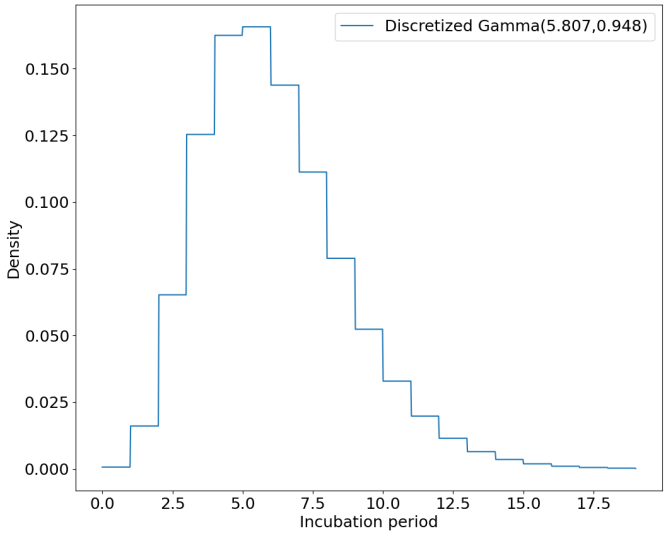



Figure 4.5: Discretized incubation period duration distribution.

**Disease duration** A proper description of disease is far more complicated than just its duration, usually to evaluate disease dynamics, epidemiological research estimates serial interval, attack rate, reproduction number, incubation period and branches the disease based on symptomatic and asymptomatic patients into scenarios.

The model designed for this thesis will simplify the disease dynamic into disease duration, only using two scenarios:

- Symptomatic - infectiousness and symptoms occur at the same time
- Asymptomatic - symptoms do not occur, but infectiousness does. The individuals in this scenario have lower probability to go and get themselves tested for coronavirus.

The scenarios probabilities for a patient depends on age of the patient, symptoms are more likely with older patients and from the literature [72] was created the table 4.6.

| **Age group** | **Asymptomatic** | 95% **CI** |
|---|---|---|
| *Total* | 0.308 | $0.077 - 0.538$ |
| $0 - 15$ | 0.6 | 0.4 |
| $16 - 64$ | 0.45 | 0.55 |
| $65+$ | 0.3 | 0.7 |

Figure 4.6: Asymptomatic scenario probability per age group [72].

The dataset for the duration of symptoms (fig. 4.7) consists of 129 samples of hospitalized patients diagnosed with COVID-19. Of those 69% were also at ICU[2] and out of them 91% had to be connected to mechanical ventilation. Immunosuppressed was 23% of the patients. The data contains only hospitalized patients, thus the sample is biased [73].

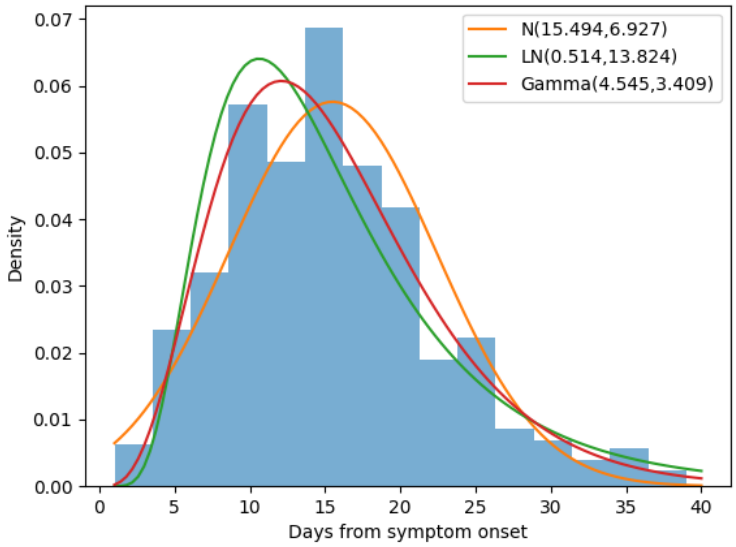



Figure 4.7: Estimated distributions of duration of symptoms [73].

The data shows that the mean of symptoms duration is 15.5 days and 95% of all samples lay within $4.225 - 32.775$ days and 50% of all samples in $11 - 19$ days. The distribution fitting to the data the best seems to be lognormal or gamma. Using AIC from the equation 4.2 estimated for each distribution $m$ (with likelihood $Pr(x|\overrightarrow{\theta_m})$ denoted $L_m(\cdot)$ and $\mathrm{df}_m$ degrees of freedom) from all the distributions $M$ it is analytically determined that the best fitting distribution is $\Gamma(4.545, 0.293)$ (table 4.8).

$$\begin{aligned} \text{best model} &\equiv \underset{m \in M}{\operatorname{argmin}}\ \mathrm{AIC}(m) \\ \mathrm{AIC}(m) &= 2\mathrm{df}_m - 2\ln\Big[L_m(\cdot)\Big] \end{aligned} \tag{4.2}$$

[2]Intensive care unit

| **Distribution** | $\mathcal{N}(15.4942, 6.9272^2)$ | $log\mathcal{N}(0.5142, 13.8266^2)$ | $Gamma(4.545, \frac{1}{3.409})$ |
|---|---|---|---|
| **AIC** | 4635.0654 | 4670.95 | 4594.3844 |

Figure 4.8: Goodness-of-fit of distributions for duration of symptoms by AIC.

As before, for modelling using transition matrix we discretize the distribution to get daily probabilities using equation 4.1, the result is shown in the figure 4.5 and in the file `data/symptoms.csv`. Similar to ours are also the results of [74].

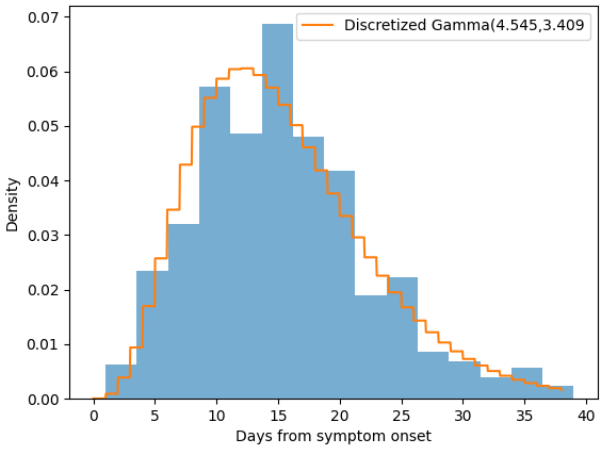



Figure 4.9: Discretized duration of symptoms distribution [73].

Hospitalized cases are either risky patients or patients with severe symptoms of the disease, measuring characteristics only on hospitalized people is biased, as in the case of estimate 4.9) - different publications state duration of symptoms for patients with milder Covid-19 within 10 days [75, 76].

An estimate of the disease duration is dependent on a sample type - usually a tissue from upper respiratory specimens is used, but it can be measured from various samples: there are studies measuring SARS-CoV-2 presence of Covid-19 positive patients from rectal swabs, where it turns out the viral persistence is longer (than usual nasopharyngeal swab) [77]. In addition, symptoms negatively affecting digestive system has also occurred in some cases [78].

**Generation time / serial interval** Generation period $w(t)$ is an experimentally measured characteristic of the disease - time between infection of two successive cases. The serial interval on the other hand is the time between symptoms onset of two successive cases [79].

Paper [80] estimates serial interval to be $\Gamma(\alpha, \beta)$ distribution with mean $\mu = 4.55$ and standard deviation $\sigma = 3.3$. When using formula for expected value and variance of gamma distributed random variable, the serial interval has distribution $\Gamma(1.901, 0.41781)$, as shown in the equation 4.3.

The distribution both continuous and per-day discretized is visualized in the figure 4.10.

**Reproduction number** Basic reproduction number, average count of new infection generated by single infected individual, is a tricky statistic to estimate, as it is computed from incidence, which is also unobserved. There are several methods used in literature that ap-

$$\begin{aligned} \text{Serial interval} &\sim \Gamma(\alpha, \beta) \\ E[\text{Serial interval}] &= \frac{\alpha}{\beta} = 4.55 \\ V[\text{Serial interval}] &= \frac{\alpha}{\beta^2} = 3.3^2 = 10.89 \\ \alpha &= 1.901,\ \beta = 0.41781 \end{aligned} \tag{4.3}$$

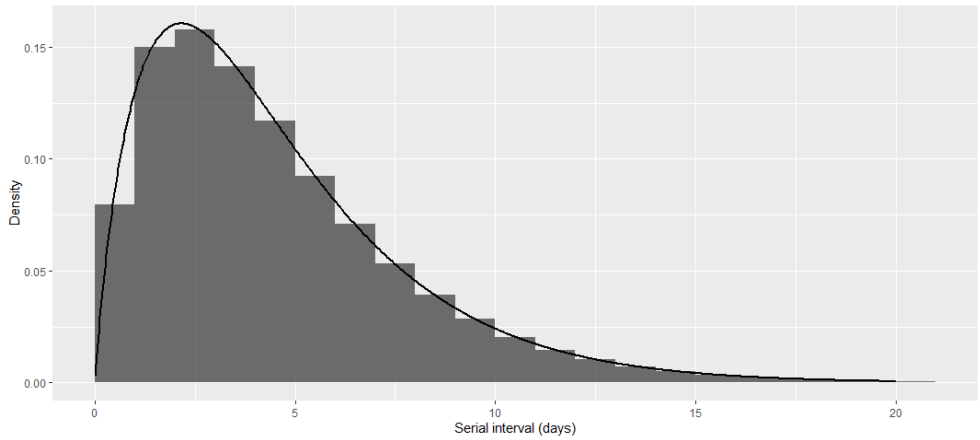



Figure 4.10: Serial interval distribution [80].

proximates reproduction number, both time varying $R_0(t)$ and basic reproduction number $R_0$ [81].

Their strategy is estimation using clinical measurements of disease characteristics - incubation period, serial interval and infectious period. Most current research papers and WHO estimates basic reproduction number $R_0$ of SARS-CoV-2 to be 2 - 4 [82, 83, 80, 84], simulation result is shown in the figure 4.11. However the conditions of the environment (e.g. sufficient precaution of people) change the $R_0(t)$ significantly.

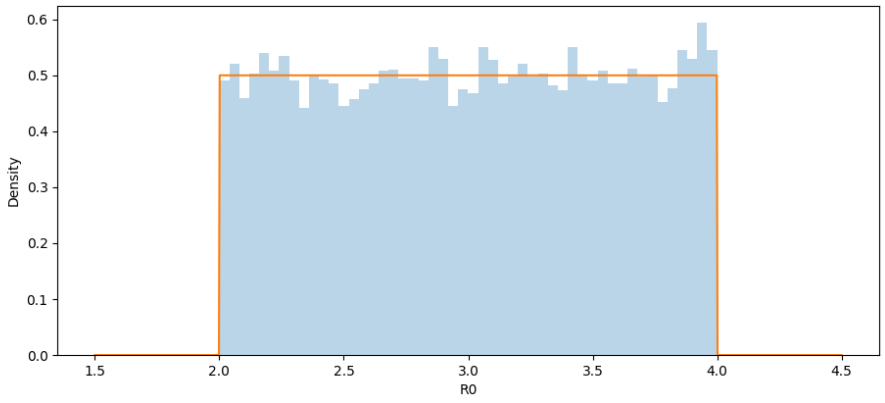



Figure 4.11: $R_0$ estimate using PCR incidence.

The method introduced by [85] is minimizing objective function from the equation 4.4. Equation describes case $j$ infected at time $t_j$ (days), number of total infected is $K$. Any

subsequent case $i$ from time $t_i$ has been potentially directly infected by $j$, the probability $P(i \text{ is caused by } j) = p_{ij}$ is dependent on the generation period $w(t)$ (fig. 4.10). Marginalizing $i$ from $p_{ij}$ gives the number of cases infected by $j$ independently on time of the infection $t_i$. Effective reproduction number is acquired by averaging all the cases with the same symptom onset $t_j$, $R_t = N^{-1} \sum_{t_j=t} R_j$. This method is described in detail in [86].

$$\begin{aligned} p_{ij} &= \frac{w(t_i - t_j)}{\sum_{k=1}^{i-1} w(t_i - t_k) + \sum_{k=i+1}^{K} w(y_i - t_k)} \\ R_0(j) &= \sum_{i=1}^{K} p_{ij} \end{aligned} \tag{4.4}$$

The above mentioned algorithm is implemented by R package `EpiEstim` and was used to estimate $R_0(t)$ over the incidence confirmed by tests in the data. The result aggregated per months is shown in the figure 4.12.

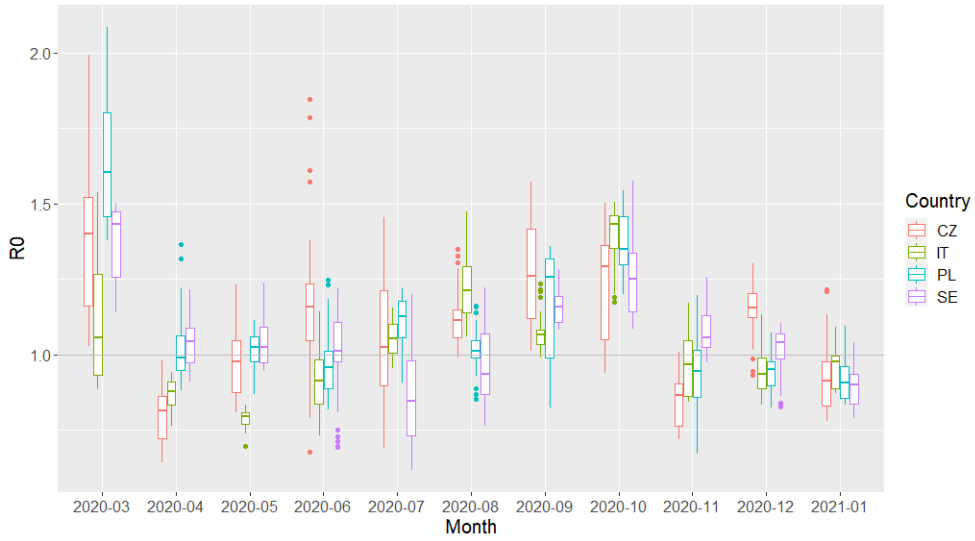



Figure 4.12: $R_0(t)$ monthly estimates using confirmed cases.

Reproduction number greater than 1 means epidemic being started. $R_0$ estimates from the figure 4.12 are tested by a one-sided t-test such as null hypothesis $H_0 : R_0(t) \geqslant 1$, the results are presented as p-values in table 4.13.

**Infection fatality rate** Fatality rates are estimated from prevalence and death counts. While case fatality ratio (CFR) uses molecular test results and thus can be estimated in real time with the disease, infection fatality ratio uses true prevalence, measured with antibody testing, that is in case of Covid-19 mostly performed several weeks after the patient's recovery to be reliable. CFR estimate is dependent on the sample collected during the pandemic, which can be biased as infected would be more likely to go and get tested that the healthy.

It was shown that AgM test can turn positive already during the infection [34, 87]. The values estimated by [88] are shown in the table 4.14. Very similar results are presented by [89] and [90].

As no other information of Covid-19 IFR was found except the table 4.14 from [88], IFR is modelled by uninformative Uniform distribution over the credible interval as shown in the equation 4.5.

| Date | CZ | IT | PL | SE |
|---|---|---|---|---|
| Mar 2020 | $9.25 \cdot 10^{-6}$ | 0.0212 | $1.16 \cdot 10^{-9}$ | $1.25 \cdot 10^{-9}$ |
| Apr 2020 | 1 | 1 | 0.1777 | $6.89 \cdot 10^{-3}$ |
| May 2020 | 0.8404 | 1 | 0.1497 | $4.13 \cdot 10^{-3}$ |
| Jun 2020 | $4.25 \cdot 10^{-4}$ | 1 | 0.8241 | 0.4894 |
| Jul 2020 | 0.0639 | $9.71 \cdot 10^{-6}$ | $2.42 \cdot 10^{-7}$ | 0.9999 |
| Aug 2020 | $3.8 \cdot 10^{-8}$ | 0 | 0.1826 | 0.9503 |
| Sep 2020 | 0 | $3.41 \cdot 10^{-7}$ | $4.46 \cdot 10^{-5}$ | 0 |
| Oct 2020 | $1.36 \cdot 10^{-8}$ | 0 | 0 | 0 |
| Nov 2020 | 1 | 0.9305 | 0.9754 | $2.25 \cdot 10^{-6}$ |
| Dec 2020 | 0 | 0.9986 | 0.9999 | 0.1934 |
| Jan 2021 | 0.9919 | 0.9928 | 1 | 1 |

Figure 4.13: P-values of epidemic hypothesis $H_0 : R_0(t) \leqslant 1$, $H_A : R_0(t) > 1$.

| Age group | IFR estimate [%] | Credible interval [%] |
|---|---|---|
| $5-9$ | 0.0016 | $[0; 0.019]$ |
| $10-19$ | 0.00032 | $[0; 0.0033]$ |
| $20-49$ | 0.0092 | $[0.0042; 0.016]$ |
| $50-64$ | 0.14 | $[0.096; 0.19]$ |
| $>= 65$ | 5.6 | $[4.3; 7.4]$ |
| *Total* | 0.64 | $[0.38; 0.98]$ |

Figure 4.14: Infection fatality rate estimates [88].

$$\text{IFR} \sim \text{Uniform}(0.004, 0.01) \tag{4.5}$$

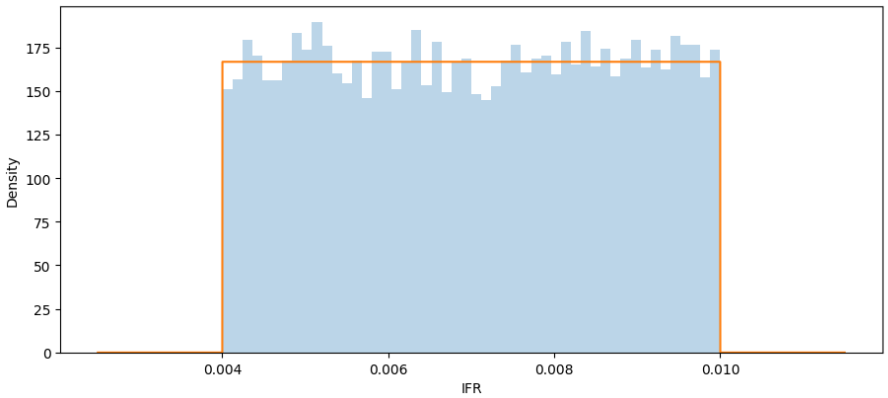



Figure 4.15: Simulated IFR.

### Transition model

The transition model denotes the transition of latent variables between times $d$ and $d+1$. Model parameters can be either defined as a scalar (meaning single value for given time point) or a vector if the parameter value differ significantly for different groups - either duration of infectiousness for both symptomatic and asymptomatic progress of disease.

The disease progress is projected in the transition model using an SEIRD model. If a Bayesian definition is used, vector parameters, e.g. incubation, infection or immunity periods, are expressed as random variables with appropriate prior distributions. The structure of the model including the transition parameters $a, c, b, d$ is shown in the figure 2.4.

Parameters can be also time-dependent, which takes into account that pandemic characteristics change over time. To lower the computational costs, a time unit for parameter values might differ from the data time unit, for $n = 7$ shown in the figure 4.16.


Figure 4.16: Illustration of parameter time slots, window size $n = 7$.

Priors for parameters $a, b, c, d$ can be estimated using the clinically measured characteristics of Covid-19. The formulas for parameters come from the SEIRD model.

**Incubation period** Parameter $c_t$ contains the information about incubation period, also represented as a probability of transition $E \to I$. Incubation period is derived from $c_t$ by the equation 4.6.

$$\text{Incubation period} \overset{d}{=} c_t^{-1} \implies c_t \overset{d}{=} \text{Incubation period}^{-1} \tag{4.6}$$

Samples of $c_t$ are acquired using a simulation from the incubation period distribution and following transformation defined by the equation 4.6. The prior distribution is acquired as a fit to the simulated draws.

For parameter $c$ it results in in the distribution specified in the equation 4.7. The samples and the fitted distribution are shown in the figure 4.17.

$$c \sim Beta(3.478, 51.059) \tag{4.7}$$

**Duration of symptoms** Parameter $b_t$ is associated to duration of individual leaving in the state $I$, thus duration of symptoms. In the current definition of the compartment model, the deceased and surviving cases are assumed to have the same symptom duration. Together with $d_t$, both parameters control the connections $I \to R$ and $I \to D$. Distribution of $b_t$ is related to symptom duration as defined in the equation 4.8.

**Infection fatality rate** Parameter $d_t$ is related to IFR, ratio of people dying in total infected cases. This ratio is projected in number of individuals taking transition $I \to D$ to those taking transition $I \to R$. Distributions of $b_t$ and $d_t$ are related to known disease characteristics as specified in the equation 4.9.

Similarly as before, samples of $b_t$ and $d_t$ are produced by simulation from the distributions in the equation 4.9. Then the distribution is fitted to the draws as shown in the figures 4.18 and 4.19, the final distributions are specified by the equation 4.10.

**Reproduction number** Parameter $a_t$ represents the infection rate, the transition $S \to E$ and is directly connected to reproduction number as specified by the equation 4.11.

Together equations 4.11 and 4.9 imply that the distribution of $a_t$ defined according to equation 4.12 and shown in the figure 4.20.

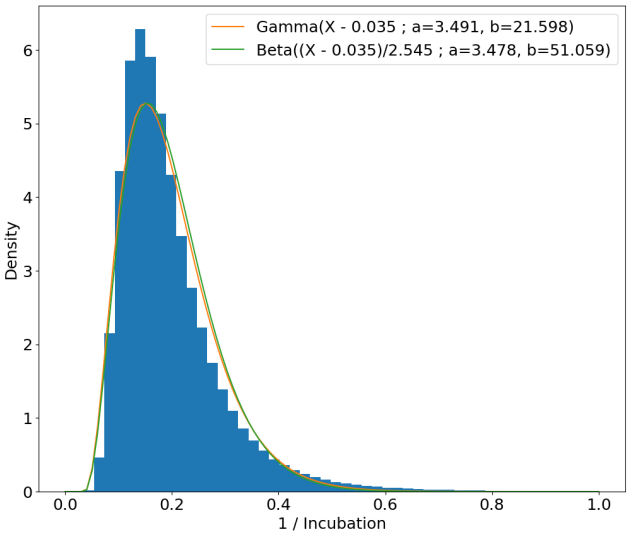



Figure 4.17: Estimate for parameter $c$.

$$\text{Symptom duration} \overset{d}{=} b_t^{-1} \tag{4.8}$$

$$\begin{aligned} b_t &\overset{d}{=} \text{Symptom duration}^{-1} \\ d_t &\overset{d}{=} \text{IFR} \end{aligned} \tag{4.9}$$

$$\begin{aligned} b &\sim Beta(2.585, 1.58 \cdot 10^6) \\ d &\sim Uniform(0.004, 0.01) \end{aligned} \tag{4.10}$$

### Emission model

Active infection is measured by tests, certain percentage of whose turns out to be positive. Simplest distribution to model percentage of positively tested individual is Bernoulli, as shown in equation 4.13 with parameter $p$ interpreted as ratio of positive tests, $\frac{\text{\#Positive Tests}}{\text{\#Tests}}$. The monthly positive tests' ratio per over time per each of the countries is shown in the figure 3.1.

Prior distribution for Tested (fig. 4.14) is represented with Beta distribution with parameters $\alpha$ and $\beta$. Relevant for choice of their values is the ratio of performed tests in the population, $\frac{\text{\#Tests}}{\text{\#Population}}$ (fig. 3.2).

**Posterior** On given day $t$ in given administrative unit with population $N$ there are $I_t$ infected people. Sample of $T_t$ perfect tests is taken, yielding $(x_1, \ldots, x_T)$, where $x_i \in \{0, 1\}$. The tests' results, the positive tests' ratio $\frac{|\{x_i|\ \forall\ i \in 1{:}T \text{ such that } x_i = 1\}|}{T}$ denoted as $\bar{x}$. Distribution

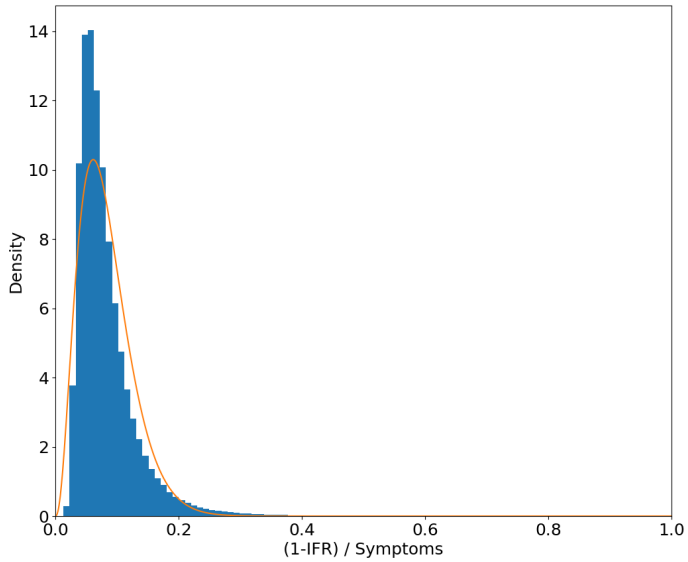



Figure 4.18: Estimate for parameter $b$.

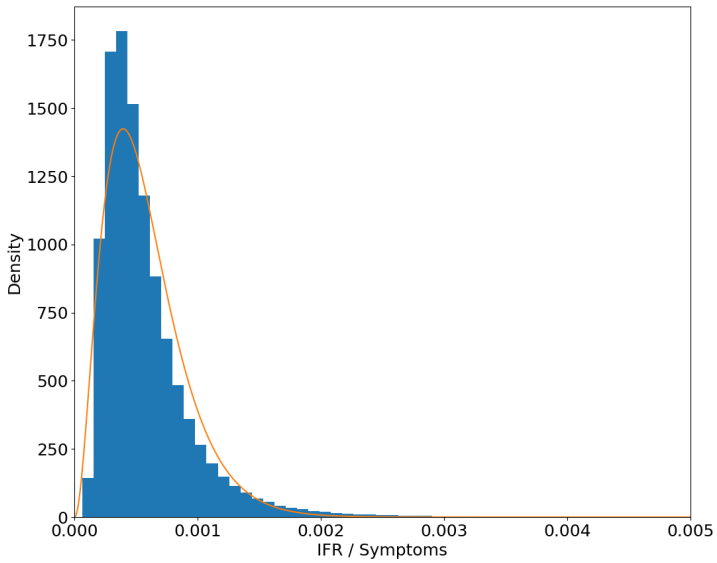



Figure 4.19: Estimate for parameter $d$.

$$R_0(t) \stackrel{d}{=} \frac{a_t}{b_t} S(t) \tag{4.11}$$

$$\begin{aligned} a_t &\overset{d}{=} \frac{R_0(t)}{\text{Symptom duration}_t} \\ a &\sim Weibull(1.836352, 0.365743) \end{aligned} \tag{4.12}$$

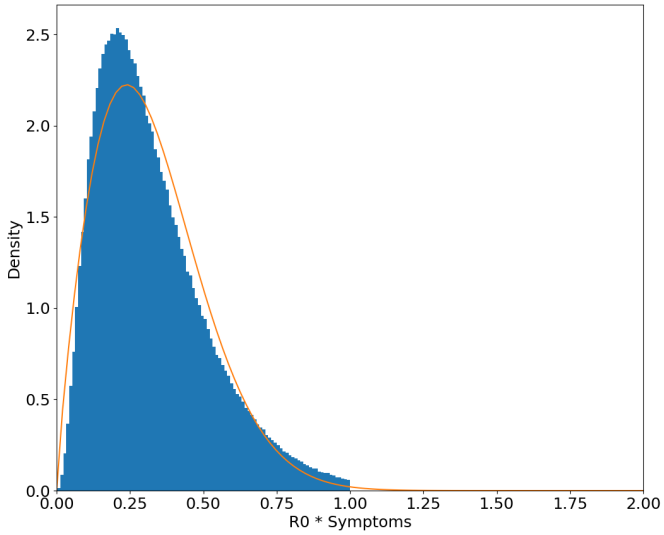



Figure 4.20: Estimate for parameter $a$.

$$\begin{aligned} \text{Infected} \mid \text{Tested} &\sim Bernoulli(p) \\ P(\text{Infected} = x_i \mid \text{Tested} = p_i) &= p_i^{x_i}(1-p_i)^{1-x_i} \end{aligned} \tag{4.13}$$

$$\text{Tested} \sim Beta(\alpha, \beta),\ \alpha, \beta > 0 \tag{4.14}$$

of posterior Infected | Tested is derived using Bernoulli$(p)$ model and conjugate Beta$(\alpha, \beta)$ prior (eq. 4.15).

$$\begin{aligned} &P(Tested = p \mid Infected = \vec{x}) \overset{\perp}{\propto} P(Tested) \prod_{i=1}^{T} P(Infected = x_i \mid Tested) = \\ &= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} p^{\alpha-1}(1-p)^{\beta-1} p^{\sum_{i=1}^{T} x_i}(1-p)^{T-\sum_{i=1}^{T} x_i} \propto \\ &\propto p^{\alpha-1}(1-p)^{\beta-1} p^{T\bar{x}}(1-p)^{T(1-\bar{x})} = \\ &= p^{(\alpha+T\bar{x})-1}(1-p)^{(\beta+T-T\bar{x})-1} \propto Beta(\alpha' = \alpha + T\bar{x}, \beta' = \beta + T - T\bar{x}) \end{aligned} \tag{4.15}$$

Identical derivation is used for statistics of recovered cases and true number of people that recovered from Covid-19, as only the confirmed cases are contained in the statistics,

parameters $\alpha$,$\beta$ for the posterior distribution, defined in the equation 4.16 for recovered, can be chosen differently from the infected.

$$P(Tested = p \mid Recovered = \overrightarrow{x}) \propto Beta(\alpha' = \alpha + T\bar{x}, \beta' = \beta + T - T\bar{x}) \tag{4.16}$$

Deaths use same derivation as well, although the choice of prior parameters should make the result of the simulation closer to the true numbers. For the following postulation we assume healthy population with natural immune systems reacting to Covid-19 antigens with usual response as symptoms mentioned in Section 2.2.

Infected individuals that die due to the disease will at some point get severe symptoms. People with symptoms are more likely to get tested as well as being hospitalized - and in the hospital patients with respiratory symptoms do get tested for Covid-19. There will be generally only few cases where the person dies without being tested or hospitalized at all. Post-mortem Covid-19 diagnostic testing is not being done.

Given this postulation, $\alpha$ and $\beta$ parameters in the posterior Tested | Deaths (eq. 4.17) should be set so that result of simulation is close to the reported statistics.

$$P(Tested = p \mid Deaths = \overrightarrow{x}) \propto Beta(\alpha' = \alpha + T\bar{x}, \beta' = \beta + T - T\bar{x}) \tag{4.17}$$

## 4.2 Model training

Training the model means seeking the values of parameters $a$,$c$,$b$,$d$ such that with specified emission prior parameters $\alpha_{\mathrm{I}}$, $\beta_{\mathrm{I}}$, $\alpha_{\mathrm{R}}$, $\beta_{\mathrm{R}}$, $\alpha_{\mathrm{D}}$, $\beta_{\mathrm{D}}$ the HMM simulation reminds the training data (eq. 4.18). There are several analytical methods used for HMM fitting, briefly described in the subsection 2.4, such as forward-backward or Viterbi algorithms. However they assume transition and emission models to be probabilistic models, while the HMM from the section 4.1 has its transition model defined as ODE[3].

The training can be done numerically and the best fitting parameters found with optimization. Objective function is the negative log-likelihood of confirmed, cumulative recovered and cumulative deaths on I, R and D respectively,

$$\begin{aligned}
(\overrightarrow{S}, \overrightarrow{E}, \overrightarrow{I}, \overrightarrow{R}, \overrightarrow{D}) &= \text{simulate from SEIRD}(a, c, b, d) \\
\overrightarrow{\text{Confirmed}} &\sim \text{Emission model}(\overrightarrow{I}, \overrightarrow{\text{Tests}}, \alpha_{\mathrm{I}}, \beta_{\mathrm{I}}) \\
\overrightarrow{\text{Recovered}} &\sim \text{Emission model}(\overrightarrow{R}, \overrightarrow{\text{Tests}}, \alpha_{\mathrm{R}}, \beta_{\mathrm{R}}) \\
\overrightarrow{\text{Deaths}} &\sim \text{Emission model}(\overrightarrow{D}, \overrightarrow{\text{Tests}}, \alpha_{\mathrm{D}}, \beta_{\mathrm{D}})
\end{aligned} \tag{4.18}$$

Alternative approach is assuming single-valued SEIRD parameters and numerically optimize with negative log-likelihood from emission model used as objective value.

## 4.3 Implementation

Most of the program that produces the results in the chapter 5 is implemented in Python 3. Some parts such as an estimation of $R_0(t)$ from incidence used for the figures 4.12 and 6.10 is written in R.

[3]Ordinary differential equations

In Python except for the standard library of Python, the implementation uses external libraries NumPy [91], Matplotlib [92], SciPy[93], Pandas [94], Seaborn [95], Scikit-learn [96], OpenPyXl [97], Requests [98], Statsmodels [99] and geneticalgorithm [100]. R code uses external packages EpiEstim [101], ggplot2 [102], mosaicCalc [103], rstan [104], bayesplot [105] and dplyr [106].

The code for SEIRD as the execution of HMM transition model with parameters $a, c, b, d$ and initial values $S_0, E_0, I_0, R_0, D_0$ as an input is shown in the listing 4.1 [107].

```
def seird(y, t, POP, a, c, b, d):
    """SEIRD step.

    Args:
        y (tuple): Values (S,E,I,R,D) at time t.
        t (float): Time.
        POP (int): Population size.
        a,c,b,d (float): Parameters.
    Returns:
        (tuple): Values (dS,dE,dI,dR,dD) between t and t+1.
    """
    S, E, I, R, D = y
    dSdt = - a*S*I
    dEdt = a*S*I - c*E
    dIdt = c*E - b*I - d*I
    dRdt = b*(1-d)*I
    dDdt = b*d*I
    return dSdt, dEdt, dIdt, dRdt, dDdt

# parameters
initial_values = np.array([POP-1,0,1,0,0]) / POP
D = 100 # days
a,c,b,d = get_params() # stochastic or constant, predefined or optimized
# numerical integration
from scipy.integrate import odeint
r = odeint(seird, initial_values, np.linspace(0, D, D+1), args=(POP, a, c, b, d))
# r is of size |D x 5|
```

Listing 4.1: SEIRD: usage example.

Objective function of the HMM implemented in the listing 4.2 first simulates S,E,I,R,D values from transition model and then computes negative log likelihood of emission model score from confirmed, recovered and deaths.

```
def posterior_objective(params, dates, pars):
    """Score of HMM with given parameters.

    Args:
        params (tuple): Parameters (a,c,b,d) of SEIRD.
        dates (): Dates of simulation.
        pars (): Emission prior parameters for I, R and D.
    """
    # data and parameters
    x = _posterior_data(region, dates)
    # run transition model
    latent = transition(params=params, D=dates.days())
    T,Tc = x.tests,x.tests.cumsum()
    xbarI,xbarR,xbarD = x.confirmed/T,x.recovered.cumsum()/Tc,latent.D/Tc
    I,R,D = latent.I,latent.R,latent.D
    # emission model score
    score = 0
    score += beta.logpdf(I, pars.I.alpha + T*xbarI, pars.I.beta + T*(1 - xbarI))
    score += beta.logpdf(R, pars.R.alpha + Tc*xbarR, pars.R.beta + Tc*(1 - xbarR))
    score += beta.logpdf(D, pars.D.alpha + Tc*xbarD, pars.D.beta + Tc*(1 - xbarD))
    return - score / D
```

Listing 4.2: Objective of HMM to optimize.

Function `posterior_objective()` from the listing 4.2 is minimized using optimization. Local search does not perform well, so search with mutation (genetic algorithm) is used for optimization, as shown in the listing 4.3. The variable boundaries are the domains of the parameters, but they can be more specific (e.g. `[[0,.25],[0,.25],[0,.1],[0,.1]]`), so that the optimization converges faster.

```
# objective function
dates = ("2020-08-01","2021-03-13")
emissionParameters = {"I": [1,10], "R": [1,10], "D": [1,1]}
def objective(pars):
    return posterior_objective(pars, dates, emissionParameters)
# optimize
from geneticalgorithm import geneticalgorithm as ga
model = ga(objective, dimension=4, variable_type='real',
           variable_boundaries=[[0,1],[0,1],[0,1],[0,1]])
model.run()
# best params
return model.output_dict['variable']
```

Listing 4.3: Optimization of HMM.

# 5 Results

## 5.1 Transition model

Transition model is based on compartment model SEIRD. Formally the SIR* models are introduced in the subsection 2.3 and how the SIR* models are used in the HMM presented by this thesis is described in the subsection 4.1. To demonstrate the functionality of the model by segments, a sample epidemic was generated with parameters from the table 5.1 and shown in the figure 5.2.

| | | Parameters | | | | |
|---|---|---|---|---|---|---|
| **From** | **To** | **a** | **c** | **b** | **d** | **R0** |
| 1 March 2020 | 14 April 2020 | 0.4 | 0.4 | 0.2 | 0.05 | 2 |
| 15 April 2020 | 31 May 2020 | 0.15 | 0.4 | 0.2 | 0.05 | 0.75 |
| 1 June 2020 | 31 August 2020 | 0.6 | 0.4 | 0.2 | 0.05 | 3 |

Figure 5.1: Parameters for transition model example.

The model does reflect the changes of the parameters, but it still follows the property, that the population is getting closer to the herd immunity effect, when compartment models flatten in I. In the example epidemic from the figure 5.2, the first segment contains epidemic with R0 2 and the third segment contains more aggresive epidemic with R0 3. The second segment represents e.g. restrictions, so that R0 gets less than 1 (in this case 0.75). At the beginning of the epidemic segments, there is a slow period and at sudden point, the epidemic goes faster. The second segment contains mild slowdown in susceptibles' descent and recovered' ascent, but abrupt fall in infected and exposed.

Finally, the epidemic starts to slowing down in the third segment at the end of July as a result of herd immunity effect and at the end of August there are almost no cases of active infection in the population.

Segments are connected, because the last value is used as the initial value for the next segment in the implementation, but the linkage is not smooth as the derivation of the first order is different in the edge point for each of the segments.

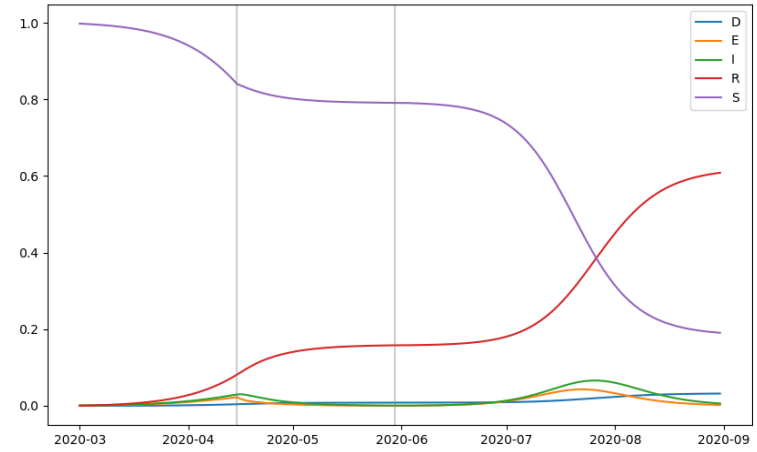



Figure 5.2: Transition model example.

## 5.2 Emission model

Emission model is introduced in the subsection 4.1. It does transformation of latent unobserved series $z[t]$ into observation $x[t]$, ergo $z[t] \rightarrow x[t]$. Its behavior is demonstrated in the figure 5.3, where series $z[t]$ is produced by moving average process (denoted MA), defined by equation 5.1.

$$z[t] = \left[\mathrm{MA}[t] + 3\sin\frac{2\pi t}{T}\right]_{[0,1] \text{ normalized}} \tag{5.1}$$

The parameters for the plot 5.3 are $(\alpha, \beta) = (1, 50)$, number of iterations is $N = 1000$ and constant daily number of tests $T[t] = 100$. The time axis $t \in \{0, \ldots, T\}$ uses $T = 365$.

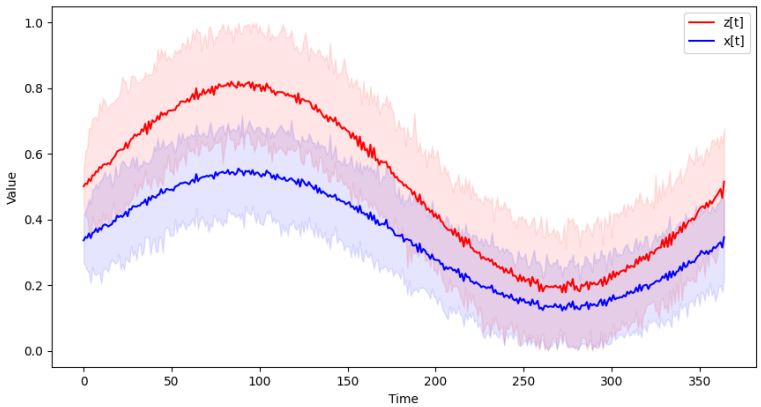



Figure 5.3: Emission model example.

## 5.3 Results of simulations

This section contains sample of various results yielded by the model. They are analysed later in the chapter 6.

### Parameters from Covid-19 characteristics

Parameter priors of parameters specified in subsection 4.1 are used to produce results in the figures 5.4 (confirmed) and 5.5 (recovered and deaths). Simulation is ran on Polish country data between March and the end of September 2020.

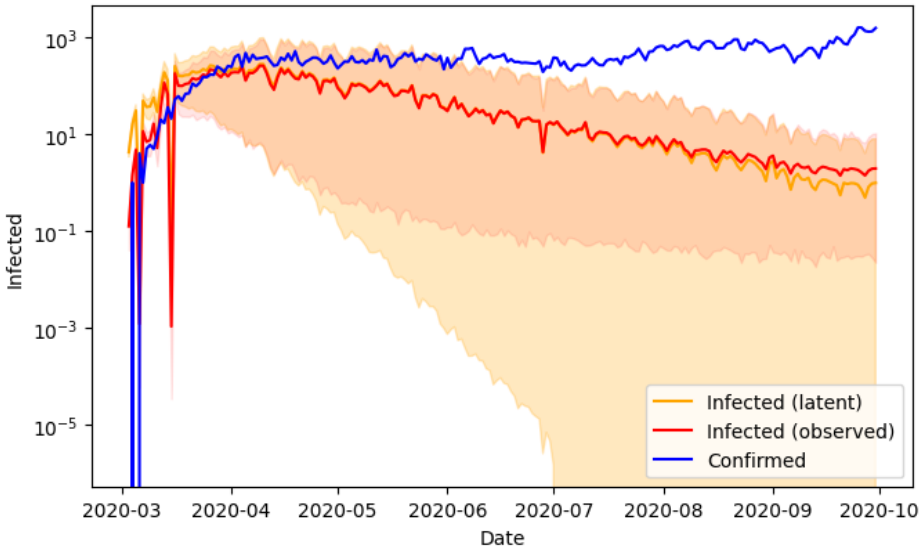



Figure 5.4: PL (*Poland* country), daily incidence, parameters from literature.

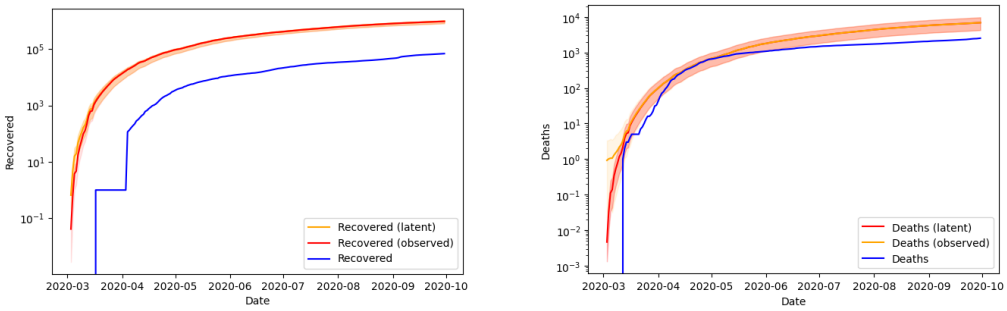



Figure 5.5: PL (*Poland* country), daily recovered and deaths, parameters from literature.

### Optimized parameters

**Daily data** Figures 5.6 and 5.7 show incidence, recovered and deaths' predictions produced by the model with optimized parameters and data with daily time step. The simulation time range goes from 1 August 2020 (as preceding data are not published) till the end of March 2021.

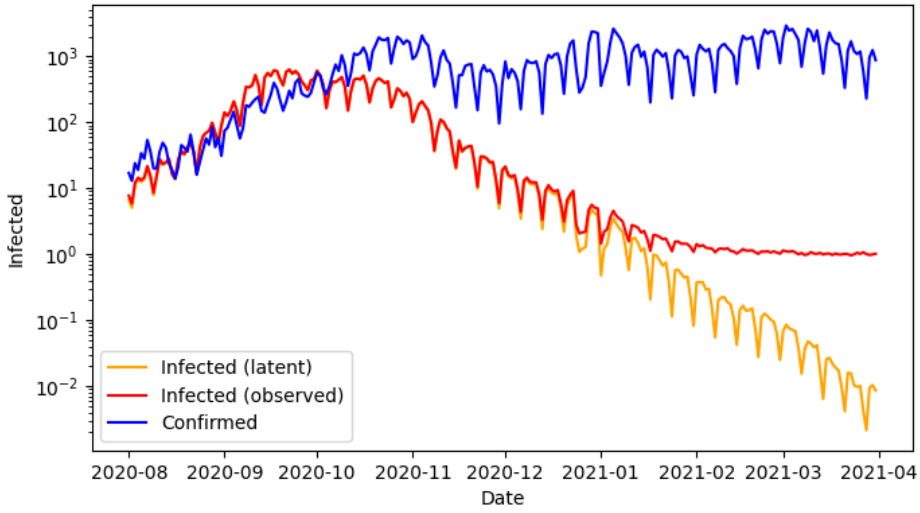



Figure 5.6: CZ020 (*Central Bohemian* region), daily incidence, optimized parameters.

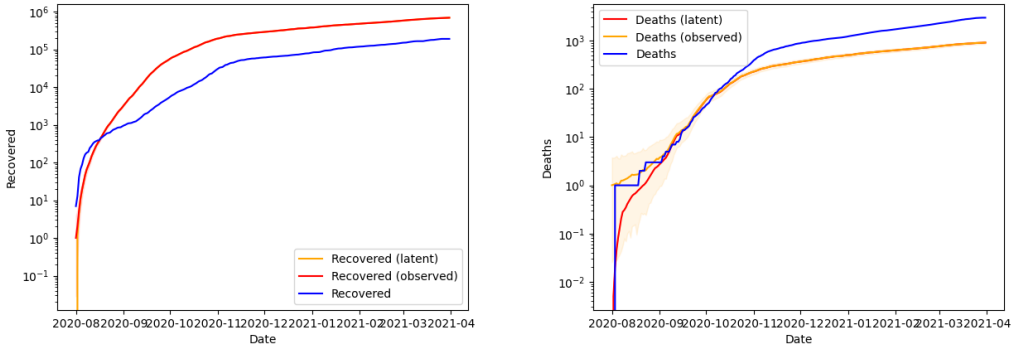



Figure 5.7: CZ020 (*Central Bohemian* region), daily recovered and deaths, optimized parameters.

**Weekly data** Result with optimized parameters using data aggregated per week is shown in the figures 5.8 and 5.7. The time range goes from the beginning of March 2020 to the end of September 2020. Swedish authorities do not publish recovered, so the optimization is done only using confirmed and deaths, so the recovered are predicted based on the other statistics.

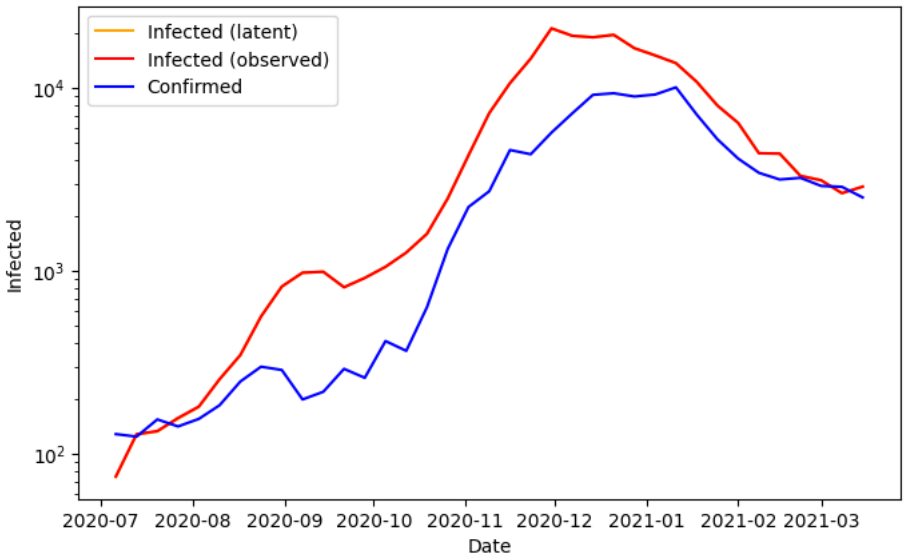



Figure 5.8: SE224 (*Skåne* region), weekly incidence, optimized parameters.

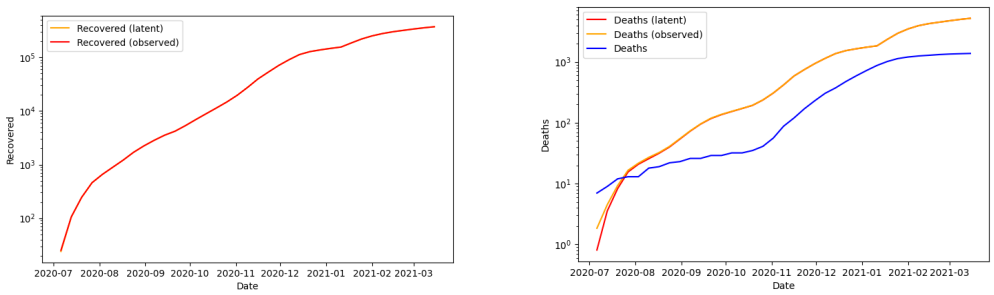



Figure 5.9: SE224 (*Skåne* region), weekly deaths, optimized parameters.

**Changing the prior** All the results has used prior parameters $\alpha = \beta = 1$. With greater $\beta$ we believe that there is more infected, than what was reported in the statistics (fig. 5.10). In the beginning before the curve drops down, the prediction $z_t$ is actually greater than the prediction $x_t$ as intended.

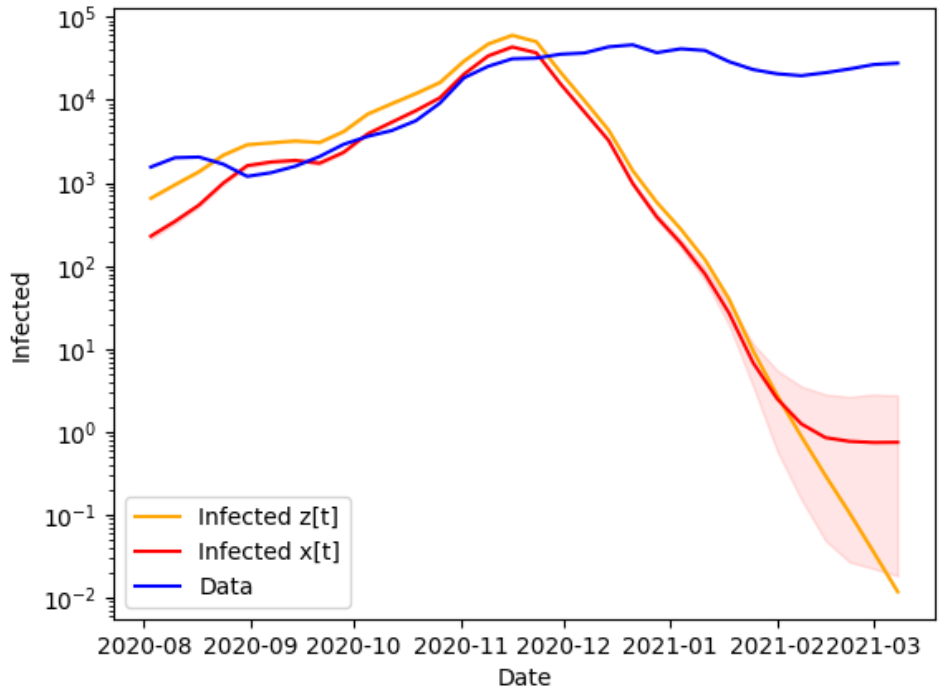



Figure 5.10: SE (*Sweden* country), weekly incidence, optimized parameters, prior Beta($\alpha = 1, \beta = 10^5$).

## 5.4 Restrictions

The discussed countries had following development of restrictions during the Covid-19 pandemic up to week 11, 2021. All the events mentioned below are cited from the calendar, a supplementary material to this thesis, where all the events contain references to newspaper articles, government decrees or other relevant sources.

- Czechia

National lockdown with closed schools and businesses (hotels, restaurants, etc.) was imposed after a few first people tested positive in week 10, 2020. Obligatory mask wearing, compulsory quarantine for citizens/residents and ban for entering the country was introduced in week 11, 2020. This state lasted over Easter holidays, which are typically connected with visiting of family and friends as part of *pomlázka* tradition.

Gradual releasing started at the end of April, from the end of April, during May more and more types of businesses were allowed to reopen, mostly with a certain degree of restrictions and public events up to 500 and later 1000 people were allowed in June, from weeks 23 and 25 respectively. From the July, 1, all the remaining restrictions were released including covering of face with mask. At the end of July, a statistical correction was done.

At the end of August, the restrictions were regionalized (set using a score for each region separately). Pupils returned to school only for September, as second wave of the epidemic raised, government imposed new series of measures centrally, including closed schools, businesses and night lockdown. On October 2-3, the regional elections took place with turnout of almost 38%.

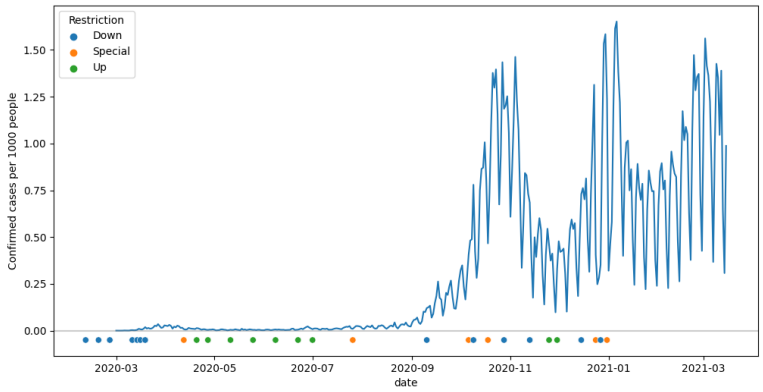



Figure 5.11: Restrictions in Czechia vs. the daily positive confirmed cases.

At the beginning of November, the situation got better and restrictions got relaxed, even schools reopened for 3 weeks from November 25 to December 15, 2020. During Christmas, the Czech government imposed restrictions on all the businesses. As the situation was not improving for some time, the national lockdown for all municipalities was imposed on February 26 and stayed until April.

- Italy

Italy was the first European epicenter of the disease and Covid-19 struck it hard in the first wave. Regions fought with the situation by issuing various regional restrictions about gathering and mobility. Government reacted with a full national lockdown in week 10, ban of gatherings of all kind (#IoRestoaCasa decree) and leaving home without reason. All schools were closed as well.

During Easter the restrictions stayed in force and the first relaxations started coming after week 18 at the beginning of May. At the end of May and in June sports (week 21) and cultural activities (week 24) were allowed and a mobile app Immuni was introduced for easier tracing.

Although situation got better, Italy stayed alerted and started reintroducing restrictions on sport, culture (week 28) and masks (week 32) again during July and August already, while e.g. Czechia or Poland did so 2 months later. September 20-21, Italy held a public referendum with turnout over 51%.

Second Italian lockdown with closed businesses and public areas was ordered in the second half of October and stays in effect when this thesis is written as of April 2021.

- Poland

Restrictions in Poland were being gradually tightened from week 10 to 13, amongst introduced measures were closing of all schools and businesses, closing of borders with quarantine for arriving citizens and ban for non-residents to enter, ban for entering public areas and mobility limitations, i.e. through limiting of public transport connections. There restrictions were also in place over Easter, which in Poland is usually connected with visiting of family and friends and attending of religious services.

First releasing of the restrictions started after Easter in week 16 by reopening of public places and churches. After that hotels, museums, libraries and restaurants reopened as well. From week 21, gatherings for up to 150 people could take place, including weddings, theaters and gyms.

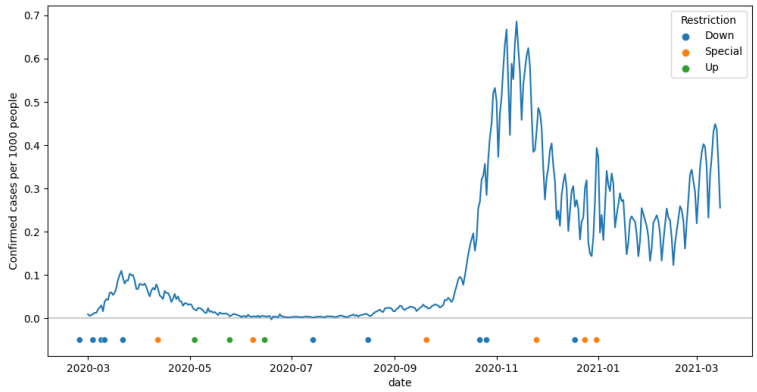



Figure 5.12: Restrictions in Italy vs. the daily positive confirmed cases.

During the summer, two rounds of presidential elections were held, on June 28 and July 12. At the end of August, regional restrictions were used instead of the central ones and schools reopened with the new school year. First centralized restrictions came in the middle of October, closing most of the businesses. On October 22 the Constitutional Tribunal passed an anti-abortion verdict, which immediately caused demonstrations all over the Poland, where the Varsovian being by far the largest one with more than 100000 participants [108]. On 11 November the annual March of Independence was held despite being explicitly banned and turned into riots in the streets.

Before All-Hallows Eve on November 1, all cemeteries got closed. All the remaining opened businesses and schools were shut down in weeks 44 and 45. Restrictions stayed active over the Christmas and got partially released after the new year. Later in March, Polish government reacted on the third wave with new restrictions closing all the businesses.

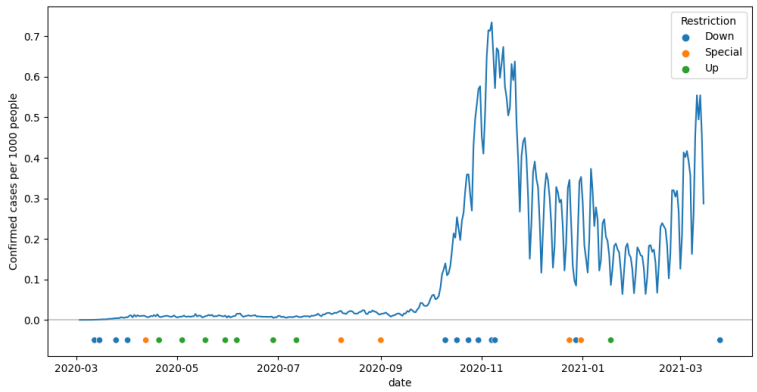



Figure 5.13: Restrictions in Poland vs. the daily positive confirmed cases.

- Sweden

From the beginning Sweden chose very different approach from most of other countries, as the authorities only published recommendations and limited gatherings, but never imposed a full lockdown as in other countries. People were recommended to keep social dis-

tancing and schools were closed based on decision of the directors, not government, so some schools kept open. From week 10, gatherings were limited to 500 people and from week 12 to 50 people.

There was no releasing during spring or summer and so the second wave in October did not mean any significant change from the government. Gatherings were eased to 300 people on November 1.

Government imposed restrictions during Christmas to avoid gatherings such as serving of alcohol is limited by time in restaurants. The recommendations stayed pretty much the same even after the New Year.

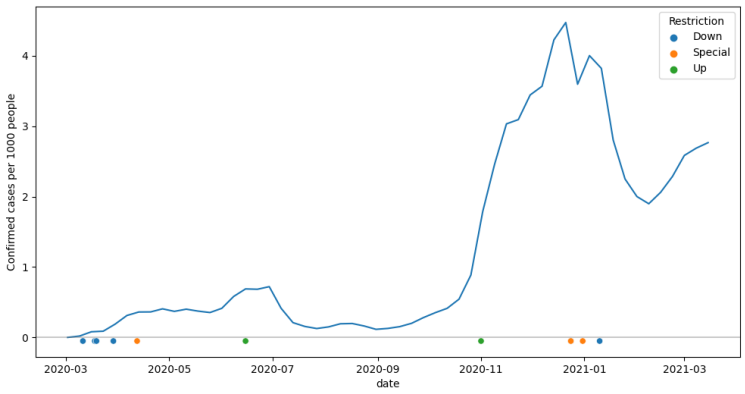



Figure 5.14: Restrictions in Sweden vs. the weekly positive confirmed cases.

# 6 Discussion

## 6.1 Results

### Equilibria of differential equations

Differential equation can have equilibria, an unchanging state of the model. Equation of a form $\frac{dx}{dt} = f(x)$ has asymptotic equilibria in zero points $\frac{dx}{dt} \overset{t\to\infty}{\longrightarrow} 0$, i.e. $f(x) \overset{t\to\infty}{\longrightarrow} 0$. These can be either stable or unstable, After a small change the state always goes towards the stable equilibrium, or always away from the unstable equilibrium.

SIR* models without feedback connections generate a single peak of the epidemic and then converges into a stable equilibrium, as shown in the figures 2.5, 2.6 and 2.7.

### Analysis of results

**Parameters from Covid-19 characteristics** Figure 5.4 shows model with parameter priors from literature used. In the first month the prediction $x[t]$ follows the data, but at the beginning of April 2020 the model reaches herd immunity threshold and incidence starts to descend. The credible interval is very wide, especially for the latent infection $z[t]$. From August 2020, the observed infections $x[t]$ overgrow the latent infections $z[t]$ as latent incidence $z[t]$ descends too low and emission model's Beta$(\alpha, \beta)$ prior becomes dominant over the insufficient number of confirmed discovered by tests in the data.

Deaths and recovered for the same simulation are shown in the figure 5.5. Predicted recovered are high above the reported statistics, while predicted deaths follow the reported deaths much closer and they also have wider credible interval.

Model output with latent infections $z[t]$ below observed infections $x[t]$ is invalid. It could be surmounted by detecting $z[t] < x[t]$ and either producing `NA` or 0 on both $x[t]$ and $z[t]$.

**Optimized parameters** If an SEIRD spline is used, such as in the figures 5.6 and 5.7, similar behavior is still present, but the latent predicted infections fits the reported cases much better in the initial phase and for longer time before it starts to descend, which is in October 2020.

At the same moment, predictions of deaths becomes lower than the reported deaths. The predicted recovered overestimates the recovered from the statistics until October 2020, from November 2020 it follows the trend as the lines go parallelly.

Weekly time step in the figures 5.8 and 5.9 causes that the compartment model will not reach the herd immunity at all and thus the predicted latent infections follow the reported cases pretty well and the same do the deaths. Recovered is not reported in statistics published by Folkhälsomyndigheten, so the objective function contains only sum of negative log-likelihood of confirmed cases and deaths, but recovered are produced by the model too.

Possible explanation for the predictions per week performing better is that too many steps makes the model *saturate* too soon. It is likely that even weekly steps will get saturated at some point and thus the model is not suitable for long-term epidemics with multiple waves or seasonal diseases, but rather for single peak outbreaks.

**Comparison of results and the reported statistics** Figures 6.1, 6.2 show the distribution of correlation of predicted latent mean $z[t]$ with the reported statistics in regions of all four countries for infected (I) and deaths (D). For the figure 6.1, only the first 60 days of simulation are used (August 1, 2020 to September 30, 2020), i.e. the days before the prediction drops and flattens, the figure 6.2 shows the results of the simulation in its whole length (August 1, 2020 to March 13, 2021).

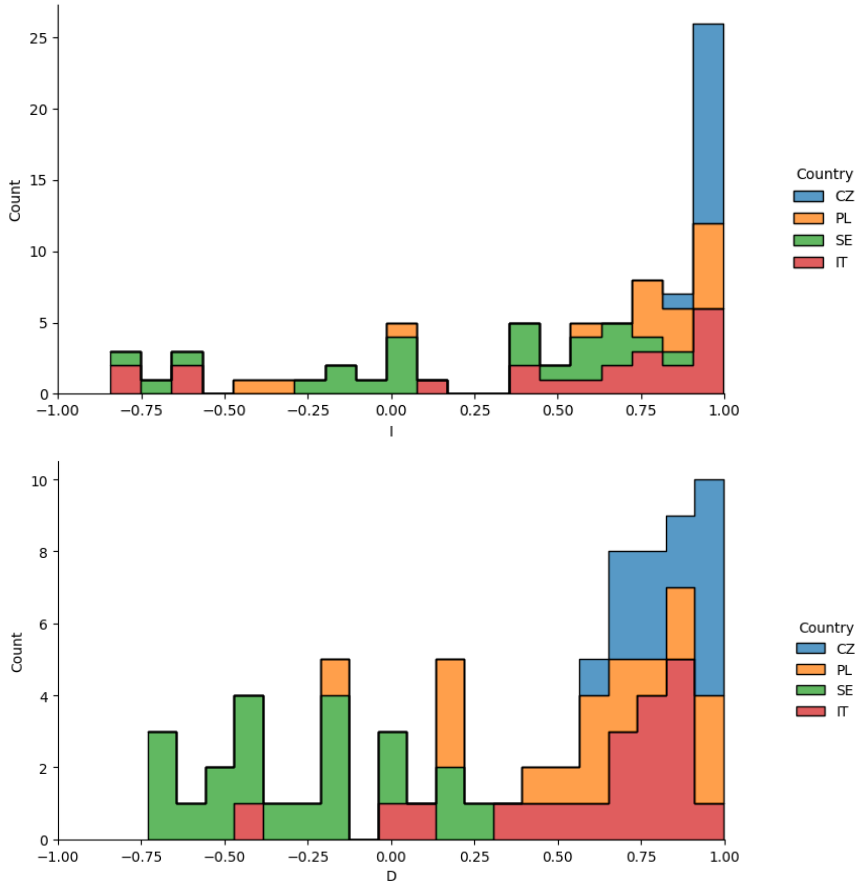



Figure 6.1: Correlation of prediction of infected to reported confirmed cases (left) and deaths to reported deaths (right), August 1 to September 30, 2020.

Regarding accuracy of infected in the first 60 days, visualized in the figure 6.1, the most accurate are the regions of Czechia, which all except 1 have correlation close to 1.. Poland has 3 regions with low positive or negative correlations. Sweden has 7 regions with negative correlation and about 8 with correlation positive but less than 0.5. Italy has 4 regions of

negative correlation, most of the regions has correlation above 0.3. The correlation of deaths with the predicted number is the highest for Czechia. Poland and Italy perform comparably, where each has about 1 region of negative correlation and some more regions of low positive correlation. Sweden has most of the regions negatively correlated. On the first 60 days, prediction seems to be the most accurate for Czechia and the least accurate for Sweden.

Now we shall compare the prediction vs. reported statistics on the whole modelled period from August 1, 2020 to March 13, 2021. The distributions of correlation for both infected vs. reported confirmed cases and deaths vs. reported deaths are presented in the figure 6.2.

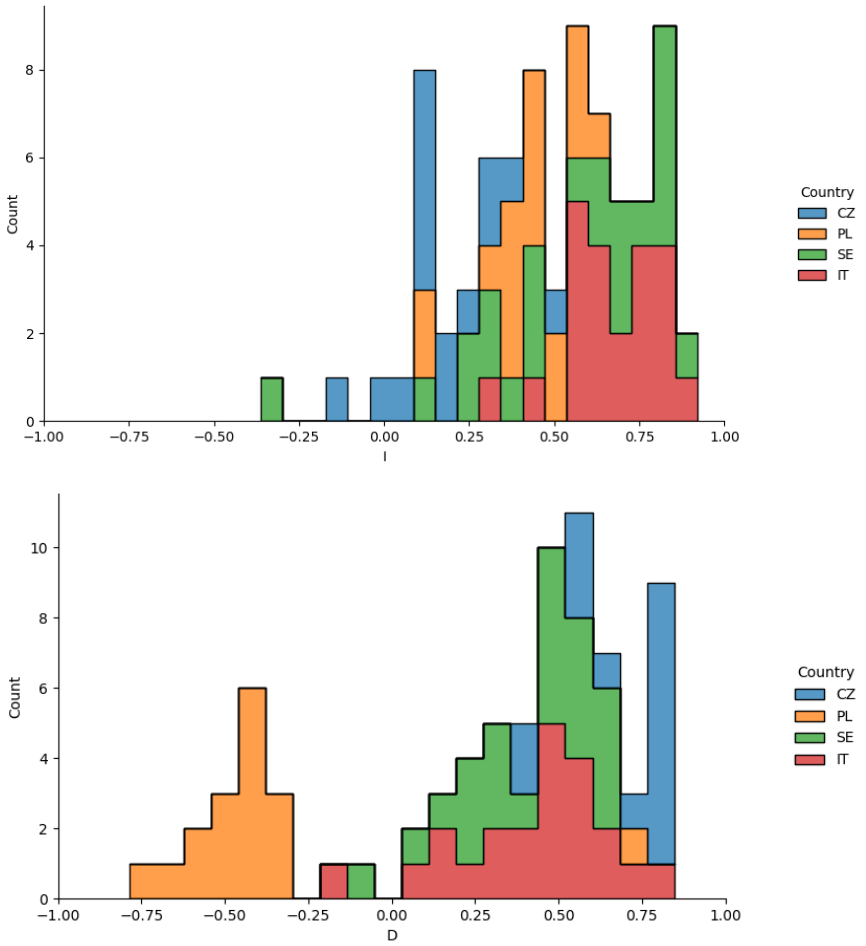



Figure 6.2: Correlation of prediction of infected to reported confirmed cases (left) and deaths to reported deaths (right), August 1, 2020 to March 13, 2021.

In the figure 6.2 the highest correlation seem to have Italy, as for all but two regions it is greater than 0.5. The lowest overall correlation is of Czechia this time. Poland has all the regions with positive correlation, but Sweden seems to be a bit more correlated with the statistics, despite the fact that it has one region with correlation lower than $-0.25$. In predictions of deaths, Czechia seems to perform the highest in similarity to statistics. Sweden and Italy have the correlation about 0.5. Poland's prediction of deaths on the whole period is for all regions but one anti-correlated.

In general the method does not produce prediction that would be very similar to the training set, especially not for long-term modelling.

**Covid-19 characteristics' estimates** Parameters optimized per week $a, c, b, d$ are used to estimate characteristics of Covid-19 characteristics - reproduction number $R_0$, infection fatality ratio IFR and infectiousness duration, their relations are defined by the equation 2.5. Experimentally acquired estimates should in the best case remind of the parameter estimates' from other literature, defined in the subsection 4.1.

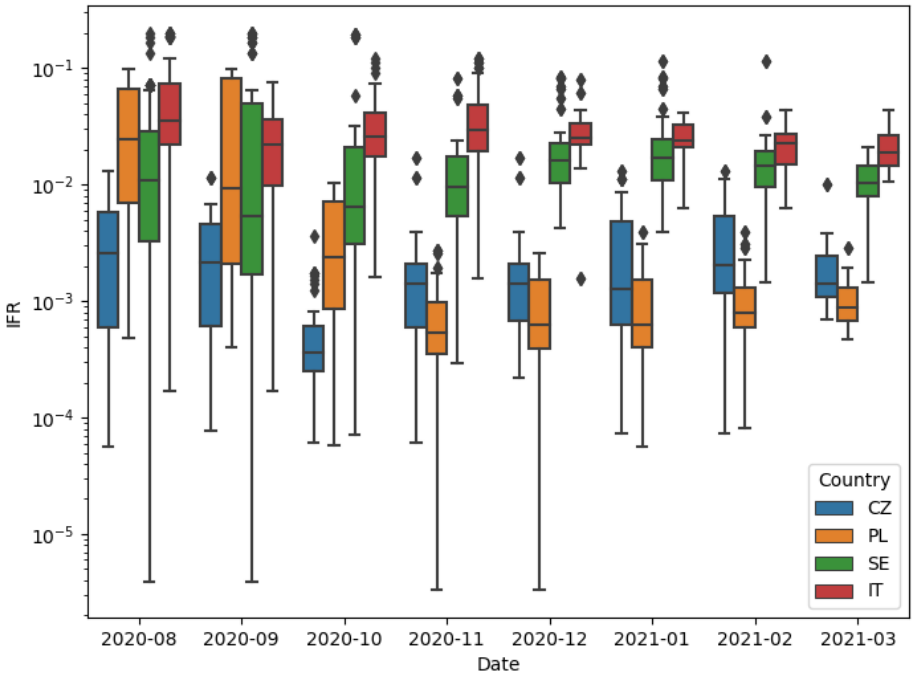



Figure 6.3: Boxplot series of IFR regional estimates per country.

IFR estimate from literature is shown in the table 4.14, the experimentally measured IFR from the SEIRD model is shown in the figure 6.3. As shown in the results of the modelling, later in the modelling the curve falls down and does not follow the data curve well.

| IFR estimates [%] | Mean of regional | | | | Country-wise | | | |
|---|---|---|---|---|---|---|---|---|
| **Month** | **CZ** | **IT** | **PL** | **SE** | **CZ** | **IT** | **PL** | **SE** |
| Aug 2020 | 0.351 | 5.777 | 3.516 | 2.311 | 0.700 | 6.581 | 0.300 | 0.986 |
| Sep 2020 | 0.301 | 2.548 | 3.328 | 3.928 | 0.139 | 0.581 | 0.192 | 0.224 |
| Oct 2020 | 0.058 | 3.324 | 0.412 | 2.257 | 0.029 | 1.243 | 0.124 | 0.406 |
| Nov 2020 | 0.217 | 3.909 | 0.078 | 1.407 | 0.066 | 1.287 | 0.132 | 1.097 |
| Dec 2020 | 0.221 | 2.787 | 0.085 | 2.279 | 0.268 | 2.252 | 0.056 | 3.203 |
| Jan 2021 | 0.316 | 2.526 | 0.100 | 2.428 | 0.268 | 2.400 | 0.132 | 3.124 |
| Feb 2021 | 0.376 | 2.237 | 0.108 | 1.671 | 0.383 | 2.075 | 0.132 | 1.398 |
| Mar 2021 | 0.227 | 2.106 | 0.109 | 1.148 | 0.383 | 1.823 | 0.043 | 0.995 |

Figure 6.4: IFR estimates of the model.

The table 6.4 presents the estimates of IFR on country-level data as well as the mean estimates over regions of the country from the figure 6.3. The estimates on data from Italy, Sweden and the two first months in Poland are higher ($2 - 3.5\%$) than the respective estimate

from the table 4.14. Czechia estimates the IFR lower, around $0.2\% - 0.3\%$. Given that the estimate in the table 4.14 is correct, Sweden, Italy and Poland are not testing sufficiently, however this is not supported by the ratio of performed test in the figure 3.2.

Another hypothesis for this is wrong initial value for deaths. Both E and I are initialized with $0.01 \cdot$ #Tests, D and R are 0, which makes difference as they are used as cumulative, this explains IFR in Italy around 5% in August 2020. September IFR could be more trusted. Drop in Polish IFR is explained by missing deaths' statistics after October 10, 2020.

To sum up, using the simulation results I estimate the IFR of Covid-19 for $2 - 3\%$.

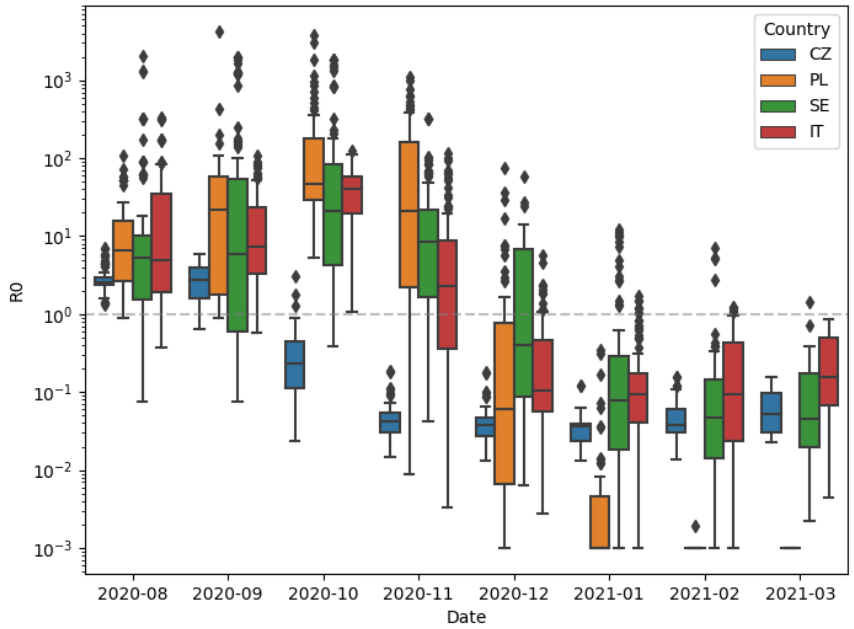



Figure 6.5: Boxplot series of $R_0(t)$ regional estimates per country.

The basic reproduction number $R_0$ of Covid-19 from the literature can be found in the subsection 4.1. According to WHO it is about $2 - 4$.

| $R_0$ estimates | Mean over regional | | | | Country-wise | | | |
|---|---|---|---|---|---|---|---|---|
| **Month** | **CZ** | **IT** | **PL** | **SE** | **CZ** | **IT** | **PL** | **SE** |
| Aug 2020 | 2.921 | 34.541 | 14.109 | 65.224 | 5.936 | 32.631 | 5.794 | 3.866 |
| Sep 2020 | 2.819 | 18.736 | 105.372 | 220.478 | 6.014 | 6.286 | 2.090 | 1.822 |
| Oct 2020 | 0.372 | 42.699 | 284.019 | 179.284 | 0.389 | 19.297 | 1.037 | 4.879 |
| Nov 2020 | 0.005 | 10.989 | 136.721 | 25.842 | 0.029 | 3.994 | 0.134 | 3.431 |
| Dec 2020 | 0.004 | 0.438 | 2.653 | 4.155 | 0.026 | 0.028 | 0.101 | 0.303 |
| Jan 2021 | 0.004 | 0.248 | 0.002 | 1.127 | 0.019 | 0.007 | 0.099 | 0.055 |
| Feb 2021 | 0.005 | 0.258 | 0.000 | 0.321 | 0.027 | 0.006 | 0.098 | 0.017 |
| Mar 2021 | 0.006 | 0.284 | 0.000 | 0.144 | 0.033 | 0.008 | 0.104 | 0.019 |

Figure 6.6: $R_0$ estimates of the model.

The figure 6.5 shows the box plot of $R_0(t)$ estimated over all the regions and aggregated as a month. These mean values as well as the estimates on the country data are presented in

the table 6.6. Italy, Poland and Sweden specify numbers far away from what reproduction number should be according to WHO. Mean regional estimates for Czechia are within the range for the first two months. Country $R_0(t)$ estimates are lower than the mean estimates over regions.

$R_0(t)$ estimates visibly show the observation that the simulation flattens after some time. This is property of SIR* models.

Given the results it is hard to say, what the reproduction number of the Covid-19 could be. According to only the results on regions of the Czech Republic, the WHO estimate $2-4$ seems reasonable.

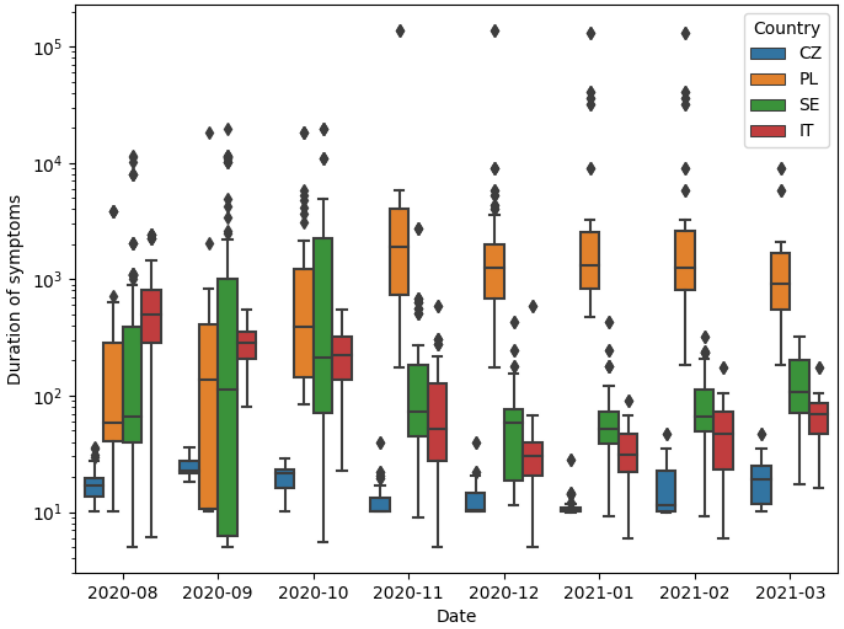



Figure 6.7: Symptoms' duration boxplot series per country.

In the subsection 4.1 symptoms duration is estimated for 15.5 days, while 50% of all cases encounter symptoms for $11-19$ days and 95% of all cases for $4.225-32.775$ days.

If parameter $b$ is optimized over $[0;.25]$, this effectively means that symptoms' duration has a domain $[4;\infty]$. Optimization yields the country-wise estimates (tab. 6.8) of Poland to be within this range, although as described before, only the first few months (Aug-Oct 2020) are taken, as in the later months the prediction of the model flattens. Considering the symptoms' duration estimate being the mean over predictions of the models of all the regions, Poland is yielding very high and unrealistic prediction, the results most closest to the literature estimate are the ones for the Czech Republic.

Bands for estimates for Poland and Sweden are very wide during the August and September and then get narrower. Italian estimates are the highest at first. Polish estimates later raise in value, while Swedish and Italian descend and are fairly similar for most of the time series.

The closest to the literature are Poland country-wise estimate and the Czech mean over regions in the first 3 months. These two yield symptoms' duration to be in interval $12-24$ days.

| | Mean over regional | | | | Country-wise | | | |
|---|---|---|---|---|---|---|---|---|
| **Month** | **CZ** | **IT** | **PL** | **SE** | **CZ** | **IT** | **PL** | **SE** |
| Aug 2020 | 17.748 | 688.366 | 324.689 | 857.886 | 62.580 | 408.194 | 23.724 | 76.063 |
| Sep 2020 | 24.848 | 282.625 | 519.648 | 1881.616 | 47.320 | 57.401 | 12.362 | 23.770 |
| Oct 2020 | 19.482 | 251.896 | 1726.343 | 1931.998 | 22.601 | 125.195 | 17.176 | 63.248 |
| Nov 2020 | 12.797 | 93.907 | 5726.890 | 217.930 | 15.864 | 88.769 | 10.404 | 46.238 |
| Dec 2020 | 13.170 | 42.462 | 5254.635 | 69.959 | 16.704 | 44.332 | 10.186 | 89.646 |
| Jan 2021 | 11.475 | 35.002 | 8900.052 | 70.362 | 12.364 | 40.814 | 10.083 | 83.504 |
| Feb 2021 | 16.507 | 51.721 | 8952.778 | 97.350 | 17.526 | 54.466 | 10.292 | 71.867 |
| Mar 2021 | 20.808 | 69.062 | 1727.617 | 136.657 | 21.761 | 67.907 | 10.565 | 96.878 |

Figure 6.8: Symptoms' duration estimates of the model.

### Regional comparison

Regional data can be clustered and thus find regions with similar epidemiological progress. Figure 6.9 shows the weekly incidence (from tests) in Czechia, Italy, Poland and Sweden, normalized by the region population. The histogram is seriated using hierarchical clustering, which is shown on the left side with a dendrogram. Distance matrix was constructed using cosine metric, which assumes each week as an orthogonal dimension, and thus time lag makes a great difference in the comparison.

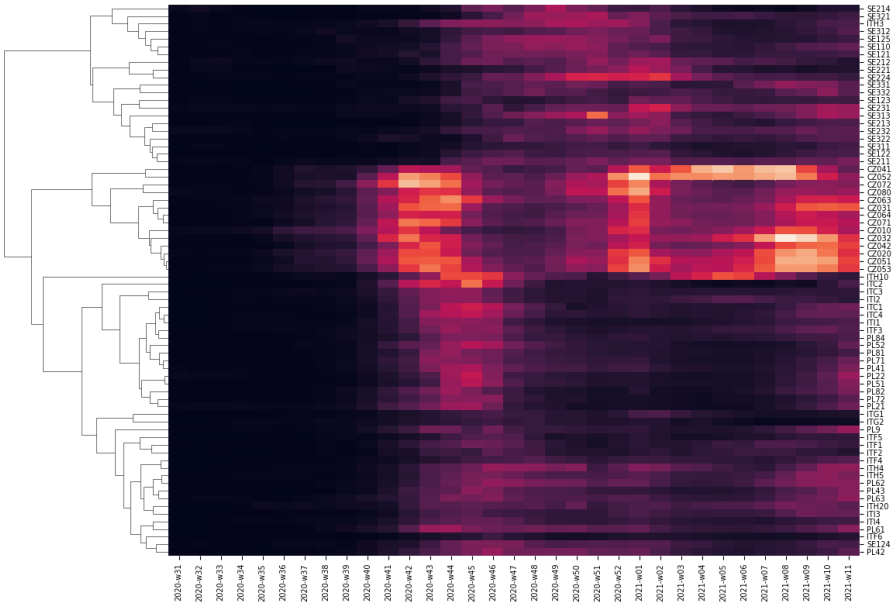


Figure 6.9: Clustering of regions based on weekly confirmed cases per 1000 people.

By countries, the second wave starts in Czechia (week 36), then in Poland and Italy (week 40) and latest it came to Sweden (week 42). There is an outlier ITH10 (*South Tyrol*). The normalized incidence is by far the highest in the regions of Czechia.

Clustering (fig. 6.9) follows the countries quite well, which is most likely a result of country lockdowns and closed borders, reducing the international mobility, and restrictions on national level, which caused different environments for the virus to spread. Present are following clusters The histogram shows several clusters of the regions shows several groups of regions, where the epidemic developed differently or had different timing.

- Sweden + ITH3 (*Veneto*)
  - Cluster 1: SE214 (*Gotland*), SE321 (*Västernorrland*), ITH3 (*Veneto*), SE312 (*Dalarna*), SE125 (*Västmanland*), SE110 (*Stockholm*), SE121 (*Uppsala*) SE211 (*Jönköping*)

  Cluster 1 mostly consists of several Swedish regions, especially Stockholm area and Italian region Veneto around Venezia city. These regions have increase in incidence from week 44, 2020 and decrease in week 2, 2021, after which the incidence stays low. These regions are densely populated, rather urban or even metropolitan, with a strong industry.

  - Cluster 2: SE212 (*Kronoberg*), SE221 (*Blekinge*), SE224 (*Skåne*)

  Cluster 2 covers regions at the most southern part of Sweden. They had similar epidemic development as cluster 3 with slightly delayed raise of the second wave against the cluster 1, but after the January 2021, there is no raise in incidence, as do have the regions of cluster 3. South of the Sweden is densely populated with mild climate and the virus might have similar conditions for spreading in the regions and due to the geographical closeness there is a room for intensive infection interchange between the regions.

  - Cluster 3: SE331 (*Västerbotten*), SE332 (*Norrbotten*), SE123 (*Östergötland*), SE231 (*Halland*), SE313 (*Gävleborg*), SE213 (*Kalmar*), SE232 (*Västra Götaland*), SE322 (*Jämtland*), SE311 (*Värmland*), SE122 (*Södermanland*), SE211 (*Jönköping*)

  Cluster 3 contains all other regions of Sweden, namely most of Norrland, the west coast and Götaland without the regions from cluster 2. These regions differ greatly in density of population, lifestyle and intensity of mobility within and between regions, as there are densely populated regions from Stockholm metropolitan area, such as Södermanland as well as the sparsely populated Norrbotten or Västerbotten.

  The incidence raise comes a week or two later than for regions in cluster 1. After Christmas the incidence starts to descend, but in February, the numbers start slightly raising again.

- Czechia
  - Cluster 4: CZ041 (*Karlovarský*), CZ052 (*Královehradecký*)

  Cluster 4 contains two regions of Czechia, where Covid-19 struck harder than in the other regions of Czechia. These two regions had a problem with lack of capacities in hospital and were locked down from the rest of the Czechia. In the peak between week 41-46, cluster 4 does not differ from the cluster 5. During the second peak after Christmas, the incidence raises and does not fall back as in other regions, but stays high until March.

  - Cluster 5: CZ072 (*Zlínský*), CZ080 (*Moravskoslezský*), CZ063 (*Vysočina*), CZ031 (*Jihočeský*), CZ064 (*Jihomoravský*), CZ071 (*Olomoucký*), CZ010 (*Praha*), CZ032 (*Plzeňský*), CZ042 (*Ústecký*), CZ020 (*Středočeský*), CZ051 (*Královehradecký*), CZ053 (*Pardubický*)

  Cluster 5 are all the regions of Czechia except the two from cluster 4. After summer 2020, there are 3 peaks of Covid-19, in October 2020, January and February/March 2021.

- Italy and Poland + SE124 (*Örebrö*)

– Cluster 6: ITC2 (*Valle d'Aosta*), ITC3 (*Liguria*), ITI2 (*Umbria*), ITC4 (*Lombardia*), ITI1 (*Toscana*), ITF3 (*Campania*), PL84 (*Podlaskie*), PL52 (*Opolskie*), PL81 (*Lubelskie*), PL71 (*Łódzskie*), PL41 (*Wielkopolskie*), PL22 (*Śląskie*), PL51 (*Dolnośląskie*), PL82 (*Podkarpackie*), PL72 (*Świętokrzyskie*), PL21 (*Małopolskie*)

Cluster 6 are most of Polish regions without the coast, some of the northern Italian regions and Campania (Napoli region). These had a peak in October, then a long flat region with low incidence. In the March 2021 at the edge of the date range of the data, there is another raise of the numbers.

– Cluster 7: ITG1 (*Sicilia*), ITG2 (*Sardegna*), PL9 (*Mazowieckie*), ITF5 (*Basilicata*), ITF1 (*Abruzzo*), ITF2 (*Molise*), ITF4 (*Puglia*), ITH4 (*Friuli-Venezia Giulia*), ITH5 (*Emilia-Romagna*), PL62 (*Warmińsko-Mazurskie*), PL43 (*Lubuskie*), PL63 (*Pomorskie*), ITH20 (*Trento*), ITI3 (*Marche*), ITI4 (*Lazio*), PL61 (*Kujawsko-pomorskie*), ITF6 (*Calabria*), SE124 (*Örebrö*), PL42 (*Zachodniopomorskie*)

Regions of cluster 7 are the Polish coast and the regions of Italy, mostly southern and central Italy, the islands and alpine regions of Trento and Fruili-Venezia Giulia. It also contains a Swedish region of Örebrö. All of these regions had a mild raise in October 2020, which lasted till January 2021, in Feburary the incidence was low and in March it went up again.

**Calendar** The section 5.4 describes the time series of events influencing the pandemic, especially the introduced and released restrictions, including plots over time together with incidence to visualize their potential effect on the curve for each of the countries, Czechia (fig. 5.11), Italy (fig. 5.12), Poland (fig. 5.13) and Sweden (fig. 5.14)

- Czechia

March restrictions preceded the raise of the incidence and so the peak was low and flattened fast. Incidence went up a few weeks after public events are allowed in June. Masks stopped being required July 1, but was not followed by any raise in the incidence on country level. This could be a partial evidence of masks' required not being an influential factor for the disease spreading, however one must take into consideration that during summer when the temperature is higher, the disease might have different conditions for spread, as it is in the case of influenza.

First round of the presidential election was held when the incidence had been already raising for more than a month, two weeks after the second round the numbers stopped to grow. Hence, it does not seem that the election would cause the virus to spread more. Restriction to wear a mask was reintroduced at the beginning of September, but the raise in incidence continued for more than a month after that. Two weeks after ban of leaving homes during night there happened an abrupt fall in incidence present in statistics.

Reopening of schools was followed by the incidence growing again up. The incidence raised and during Christmas, a full lockdown was ordered again. Two weeks after that, at the beginning of January, the numbers fell abruptly again. In February incidence raised and restrictions, such as municipalities' lockdown and compulsory FFP2 masks came in effect in February.

- Italy

The top peak of the first wave in Italy was preceded by closing of schools, a ban of gathering and a regional lockdown. Curve flattened just a few days after the lockdown restricting leaving home was introduced on March 22. As restrictions seem to make an effect about 2 weeks after they were imposed, the lockdown on March 22 might have caused the descend

in April, but the flattening at the second half of March had to have a had its impulse in early March.

Easter and releasing of the restrictions in May and July does not show to cause any change in the progress of the epidemic. In late August incidence raised, although slowly than exponential, certainly affected by a restrictions such as masks, as in other countries the curve appears to be much steeper.

However a steep raise came after the public referendum was held. Lockdown and closing of businesses imposed in the late October were followed by 2 more weeks of raise and then the raise flattens.

Incidence was descending in December, the peak after Christmas is most likely its effect. In March positively tested cases raised again.

- Poland

The first wave in Poland does not fade out and the numbers are the same from April to July. Elections do not seem to make an effect on the incidence. Masks stayed compulsory in the interiors and since August regionalization, some regions introduced masks everywhere. When the incidence grew in October, there was a series of country-wise measures: masks inside and outside, limitation of gatherings, closing of businesses and cemeteries, and the peak eventually drops, but it is hard to postulate which one of the restrictions had the main effect.

During the December the incidence abruptly drops and then it stays the same and at the beginning of January 2021 even raises, most likely as a result of Christmas and New Year. In January schools are reopened, which does not have any immediate effect, but at the end on February numbers starts to raise again anyway.

- Sweden

The Swedish response to the first wave might have prevent the incidence from the exponential growth, however the incidence does not drop and the main peak came in early July and then drops.

Moving the limit for gathering up is imposed during a rapid growth of incidence in October and November and the turnabout comes right after Christmas, during which certain restrictions were in force in fact. From January the incidence declines till late February, when it raises again.

Sweden is a rare example of a country with low pandemic restrictions (and hence quite natural development of the disease) and well reported data.

Czechia, Italy and Poland made wearing of masks compulsory, but the incidence was raising even after that, so the effect of the restriction to wear a mask do not seem to be as significant from the curves. On the other hand gathering seems to be the key factor for the pace of the spread and lockdown or limiting of people to meet and gather the key mechanism to control the disease. In Czechia after releasing of gatherings in June 2020 number of confirmed cases grew, in Poland and Italy after the Christmas, that are connected with higher mobility and gatherings, we also observe a peak in incidence.

Poland reopened schools in September 2020 and in January 2021, which did not cause any immediate raise of incidence, it does not seem that opened schools would make a significant difference in spreading, however after opening of schools in Czechia at the end of November, the numbers went quickly up.

## 6.2 Method

### Estimation of reproduction number

Figure 4.12 contains boxplots of incidence reproduction number, aggregated over months and shown as a time series. Table 4.13 tests, whether the monthly $R_0(t)$ are greater or lower than

1, i.e. epidemic or not. As for Czechia, Italy and Poland, the reproduction number is less than 1 from April to June (with exception of Czechia in June). Sweden is delayed against the other three countries and the epidemic is active until May, during the summer the reproduction number is significantly lower than 1. From November, the epidemic is inactive in Italy and Poland and from December in Sweden. In Czechia, November and January the reproduction number is less than 1, but greater in December.

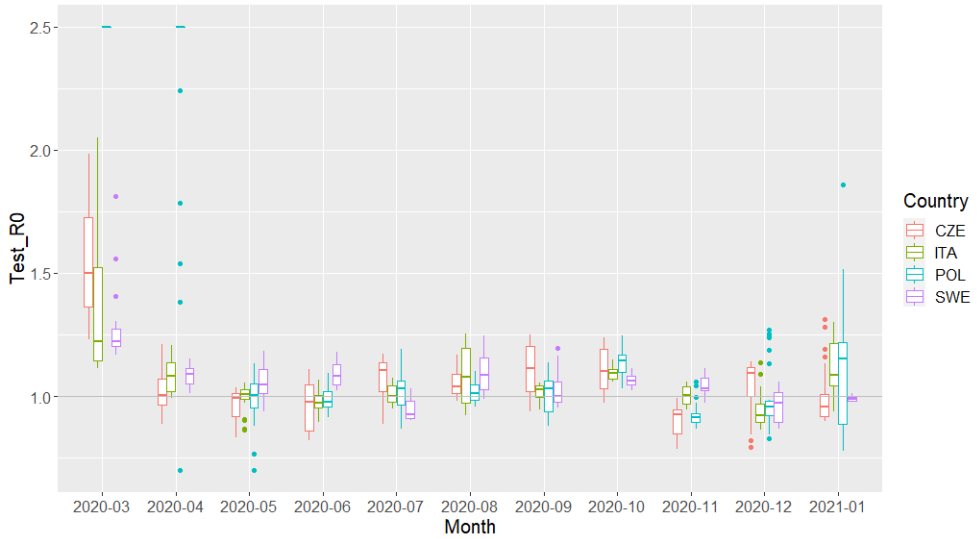



Figure 6.10: $R_0(t)$ estimated on tests (treated as incidence).

Ratio of positive tests in test sample as already said does not reflect the true prevalence of the infection of the population, as it depends on the number of performed tests and testing strategy. To evaluate the reliability of the $R_0(t)$ estimated, we might use the same method to estimate $R_0(t)$ using daily performed tests treated as incidence input. A simple correlation of the output with the $R_0(t)$ estimate gives a good picture about dependence of the estimate on the number of tests. The reproduction number of tests is shown in the figure 6.10.

Comparing the $R_0(t)$ estimated using a performed tests (fig. 6.10) to the $R_0(t)$ estimated using incidence from confirmed tests (fig. 4.12), the boxes are much more narrow and closer to 1 in most of the months - this means that the number of tests do not raise abruptly. The estimates of the first month are very high, which reminds of the situation in the March 2020.

| | **Reference** | | **EpiEstim** | | **HMM** | |
|---|---|---|---|---|---|---|
| **Period** | **CZ** | **SE** | **CZ** | **SE** | **CZ** | **SE** |
| Before Mar 7 | 2.64 | $1-7$ | 3.73 | 3.26 | $<1$ | $<1$ |
| Mar 7 - Mar 12 | 1.84 | $<1$ | 2.57 | 1.64 | $<1$ | $<1$ |
| Mar 12 - Mar 16 | 1.28 | $<1$ | 2.18 | 1.10 | $<1$ | $<1$ |
| Mar 16 - Apr 1 | 1.00 | $1-2$ | 1.34 | 1.39 | $<1$ | $\sim 15$ |
| Apr 1 - May 1 | 0.72 | $\sim 1$ | 0.80 | 1.04 | $0.1-1.6$ | $\sim 30$ |
| After May 1 | 1.08 | $\sim 1$ | 0.98 | 1.01 | $<1$ | $\sim 10$ |

Figure 6.11: $R_0(t)$ estimates together with reference from official sources [109, 110].

Folkhälsomyndigheten (Sweden) published plot of $R_0(t)$ estimates between Feb 17 and Jul 9, 2020 [110], Ministerstvo zdravotnictví (Czechia) published $R_0(t)$ for March - May 2020 [109], both are shown in the table above (fig. 6.11). For Poland and Italy, official estimates of $R_0(t)$ in this period together were not found.

The table in the figure 6.11 also shows the estimates of described HMM and method from R package `EpiEstim`, described in the equation 4.4. While official $R_0(t)$ is to some extent similar to the `EpiEstim` estimates, HMM estimates are a way off, mostly not making sense and being outside of meaningful domain for $R_0$.

### Parameter priors

Figure 5.4 presents results of simulation using data of Poland. The parameters of model have priors described in the subsection 4.1 that were created using relevant clinical research. Model uses total number of performed tests, but not the number of positive cases. Curves of recovered and deaths remind of the true data, although both are higher, which means that some recoveries and deaths were not captured by the tests. Infected curve does follow well the first wave from March to May, but then stops following the infected curve and stays close to 0.

Emission prior parameters are all $(1, 1)$, in all three cases the mean latent and observed curves for all three attributes are overlapping, however their credible intervals differ a lot.

Transition model can have the parameters either constant or stochastic, as used in the figure 5.4. Emission model is always stochastic. If transition parameters are stochastic, i.e. have prior distribution, the HMM simulation's results' credible intervals are wider and dependent on informativeness (width) of the prior distribution. With constant parameters the SIR* equations are ODE[1], with random parameters SDE[2]. Another option for the transition component are discrete models from SARIMA[3] families, such as AR, MA, ARMA, etc.

### Alternative modelling

Another possible solution how to apply analytical HMM fitting algorithms such as forward-backward or Viterbi is to approximate the SIR-based transition model with probabilistic model [111].

One of classical ecological systems is environment with populations of predators and preys, where predators reduce the population of prey and are naturally reduced when there is too few prey for too many of predators. A common model for such systems is *Lotka-Volterra*, sometimes called as *predator-prey*, defined in the equation 6.1.

$$\begin{aligned} S' &= -aSI + bS \\ I' &= aSI - bS \end{aligned} \tag{6.1}$$

With certain changes, Lotka-Volterra can be used for epidemiological modelling, so that susceptibles are prey and infected are predators. Such model does not have *Exposed* state from SEIRD used in this thesis, but automatically contains non-permanent immunity, as Lotka-Volterra does reduction of both prey and predators and with appropriate parameters one can achieve periodicity [112].

Unlike compartment models, basic Lotka-Volterra does not assume constant population size, but it depends on parameters $a$,$b$. Figure 6.12 shows the dynamics of the model with parameters $a = b = 0.1$.

Different alternative for transitional part is using spatial models, such as cellular automaton (CA). They separate the environment, population in this case, into homogenous grid of cells. Each cell has a simple dynamics based on a state and affecting and being affected by

[1] Ordinary differential equation
[2] Stochastic differential equation
[3] Seasonal auto-regressive integrated moving average

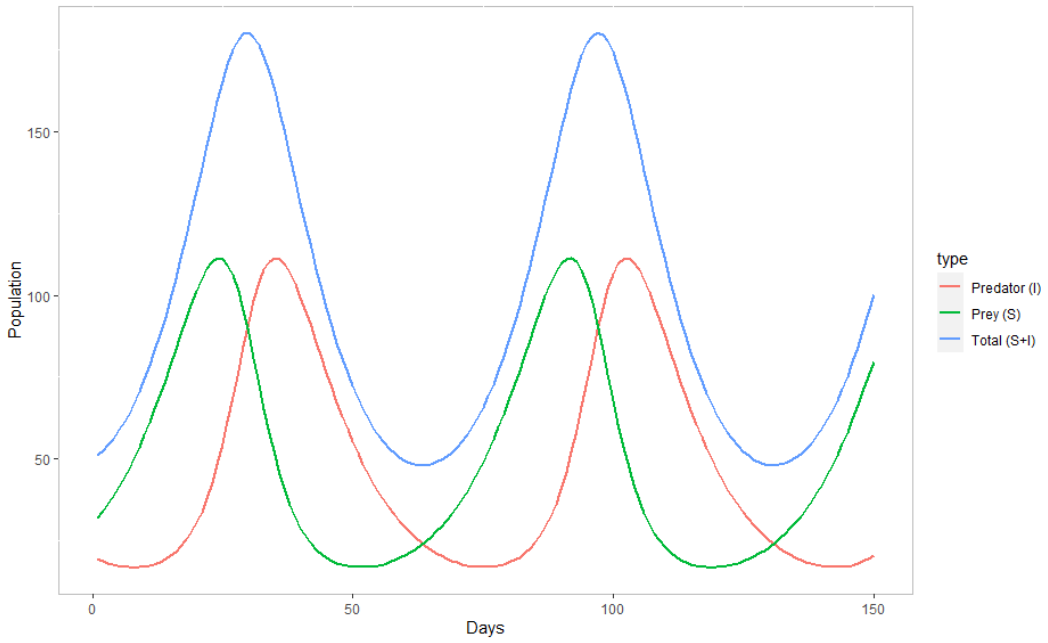



Figure 6.12: Lotka-Volterra dynamics, $a = b = 0.1$.

the adjacent cells. CA can be used for wide variety of problems, e.g. biology, fluid dynamics, but also epidemiology to model the infection in population [113].

### Potential extensions for the model

**Non-permanent immunity** Used compartment model assumes permanent immunity. Non-permanent immunity can be introduced to the model with additional connection from **R**ecovered and **S**usceptible with parameter $e$, connected with immunity period, such as $e \overset{d}{=} (\text{Immunity period})^{-1}$. Formally this model is defined by the equation 6.2.

$$\begin{aligned} S' &= -aSI + eR \\ E' &= aSI - cE \\ I' &= cE - bI \\ R' &= b(1-d)I - eR \\ D' &= bdD \end{aligned} \tag{6.2}$$

Behavior of such model is shown in the figure 6.13, identical to the figure 2.5 only with non-permanent immunity and parameter $e = 0.0\overline{3}$, which stands for immunity period of 30 days. In the plot there are two peaks visible, around days 50 and 120 and it is notable to say, that here the number of susceptibles can grow and recovered can descend. After day 120 the epidemic slowly dies out, because population is reduced as number of deceased is quite high already then.

**Vaccination** Pandemic of Covid-19 is going on for over a year and many people are hopeless about the situation. Vaccination is thus often the last hope people look up to [114]. Some

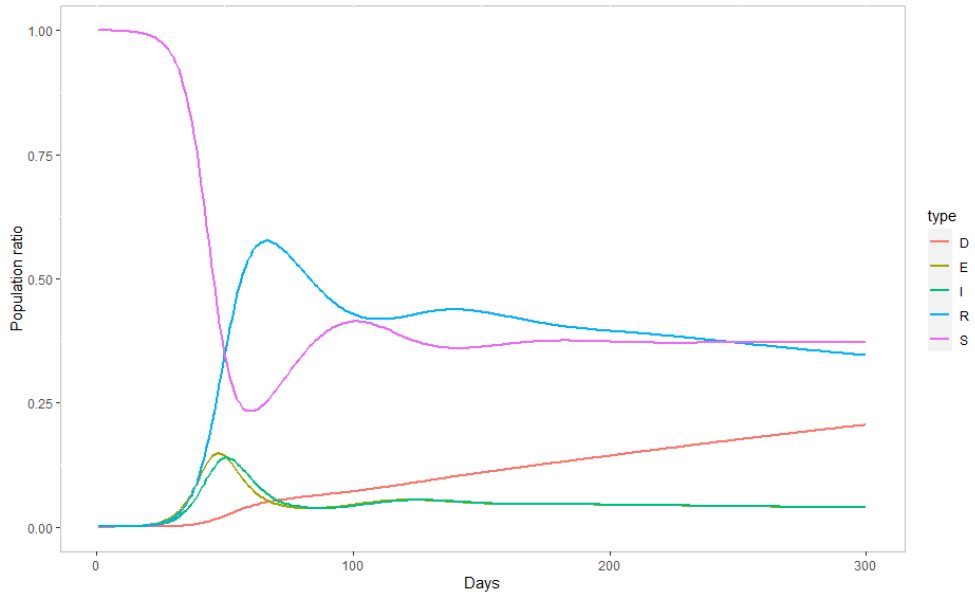



Figure 6.13: SEIRD dynamics for $(a, c, b, d) = (0.8, 0.3, 0.3, 0.05)$, $e = 0.0\overline{3}$, $I = 1$.

countries have already vaccinated a great portion of population [115] and report a decrease in epidemic [116]. However the demand for the vaccines is high and so the vaccination goes slow [114]. In case of the SEIRD, vaccination can be implemented as an additional state $V$ (vaccinated) connected from the susceptibles with time dependent parameter $v_t = \frac{V[t]}{N}$, where $V[t]$ is number of people vaccinated at day $t$ (in discrete time) and $N$ stands for the population size. SEIRD with vaccination is defined in the equation 6.3.

$$\begin{aligned} S' &= -aSI - (1 - v_t)N \\ E' &= aSI - cE \\ I' &= cE - bI \\ R' &= b(1 - d)I \\ D' &= bdD \\ V' &= (1 - v_t)S \end{aligned} \tag{6.3}$$

Using SEIRD with parameters $(a, c, b, d) = (0.8, 0.3, 0.3, 0.05)$ and no vaccination, after 100 days and population of size 10000, there is 456 deaths. Vaccination requires 1 dose and gives perfect immunity once applied, every day 0.25 % of population (i.e. 25 people) is vaccinated. As an effect after 100 days there are only 367 deaths. The effect of the vaccination is shown in the plots 6.14 (without vaccination) and 6.15 (with vaccination).

**Temperature** Although Covid-19 pandemic lasts for over a year, with stronger transmission activity during colder months there is a partial evidence that Covid-19 is capable to become a seasonal disease [117], similarly to influenza or malaria[4] [118]. Even though various feed-

[4]Malaria is seasonal because its vector, mosquito of genus *Anopheles*, is in the adult phase only at certain time of year to bite humans and transmit the malaria-causing parasite *Plasmodium falciparum*.

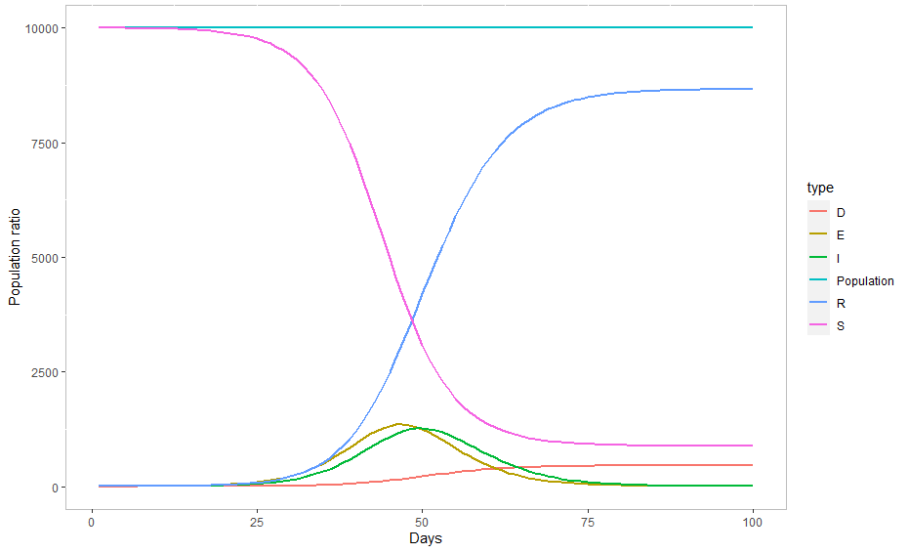



Figure 6.14: SEIRD dynamics for $(a, c, b, d) = (0.8, 0.3, 0.3, 0.05)$, $v = 0$, $I = 1$.

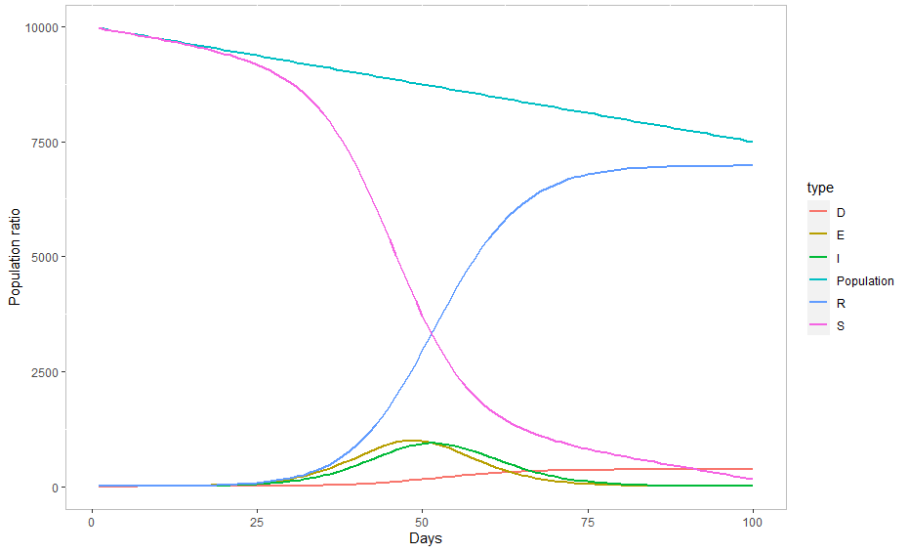



Figure 6.15: SEIRD dynamics for $(a, c, b, d) = (0.8, 0.3, 0.3, 0.05)$, $v = 0.0025$, $I = 1$.

back connections might cause oscillation in SIR* models as shown in the figure 6.15, yearly seasonality can be projected onto model by adding an input with meteorological information, such as data of temperature, precipitation and humidity [119].

**Asymptomatic cases** There have occurred cases of Covid-19 that were asymptomatic, that is no or very mild symptoms during the period of infection, [72] suggests 30.8% on sample of 565 positive cases. Asymptotic cases might differ from the infections with symptoms by time of recovery (or quarantine/isolation as the infected does not generate new infections anymore) and less careful behavior - fever causes that infected rather stays at home and have fewer contacts to transmit disease.

SIR* model can be extended with a new state A to capture asymptomatic cases, the structure including the transitions is shown in the figure 6.16 and defined formally by the equation

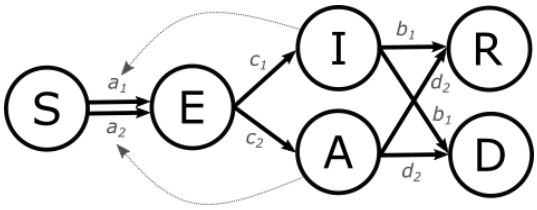



Figure 6.16: SEIARD model schema.

6.4. Longer infectious period is represented by lower parameter $c$ value. Less careful behavior producing more infections is represented by greater parameter $a$ value.

$$\begin{aligned} S' &= -a_1SI - a_2SA \\ E' &= a_1SI + a_2SA - c_1E - c_2E \\ I' &= c_1E - b_1I - d_1I \\ A' &= c_2E - b_2A - d_2A \\ R' &= b_1I + b_2A \\ D' &= d_1I + d_2A \end{aligned} \tag{6.4}$$

SEIARD behavior is shown using simulation with parameters $c_1 = c_2 = 0.3$, $d_1 = d_2 = 0.05$ and parameters specific for asymptotic behavior $a_1 = 0.5$, $a_2 = 0.8$, $b_1 = 0.5$, $b_2 = 0.2$ in the figure 6.17. In the plot, $I$ and $A$ do behave the same, although there is more asymptotic cases, because they have longer infectious period and so symptomatic cases sooner become R or D.

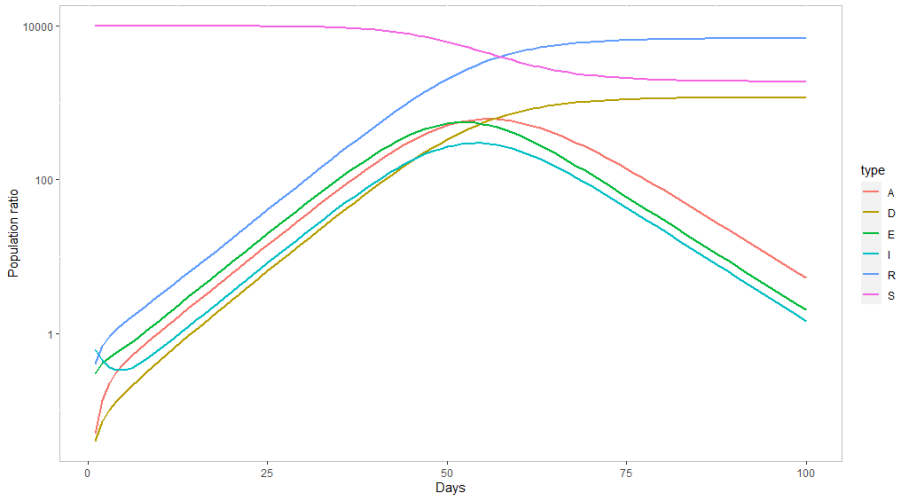



Figure 6.17: SEIARD dynamics for $(a, c, b, d) = ((0.5, 0.8), 0.3, (0.5, 0.2), 0.05)$, $I = 1$.

Period of symptoms and infectiousness do not necessarily overlap and one can be delayed or hastened to the other, or symptoms might not occur at all. These cases can be solved by scenarios [120].

**Test errors** One of the limitations of the method is assumption of perfect clinical tests. However in reality, PCR, antigen and antibody tests are not flawless and their accuracy is specified as sensitivity (true positive rate) and specificity (true negative rate) in the table 2.3. Dropping this assumption is possible by adding noise to the result.

**Mobility** Compartment models assume fixed-sized population that is homogeneously shuffled with regard to model parameters, e.g. no clusters of retired in retirement homes, children in schools etc. In the thesis, each region is modelled in a mode of hard lockdown when no people are moving between the regions.

Mobility between the regions can be modelled, so that each region has its separate SIR model and there are connections between there regions, or using spacial models, such as cellular automaton, for the mobility. Behavior of the model could remind of fluid dynamics.

### Evaluation of the method

The model performs well until the SIR starts flattening, then the prediction descends to 0 and is a lot off from the data. If the model works with daily time steps, it also reaches herd immunity sooner. Model with weekly time steps performs the same, but it takes longer to reach the herd immunity.

Observations above imply that the model might perform well for a single peak short-term outbreaks, such as local outbreaks of measles or dengue.

For long-term epidemics with multiple waves or seasonal diseases it saturates and descends after some time. This could be overcome by restarting the model for each peak, such as single season of influenza (autumn, winter and spring).

## 6.3 The work in a wider context

Reported statistics became a political topic in some countries. There are voices suggesting that some countries might be over-reporting or on the other hand under-reporting their Covid-19 deaths and cases for various reasons [121, 122]. Whether these claims are true or not is far beyond the extent of this thesis, however their presence itself brings a question, whether some political representations would be capable of changing the image of pandemic state in their country.

If so, the extent of their influence on the statistics would be a matter of local political culture as well as presence of power control mechanisms, such as independence of the authority responsible for collecting and publishing the data or freedom of press. Except of direct statistics' fabricating, politicians could change the strategy of testing, such as making Covid-19 testing free of charge, which would raise a number of performed tests and thus number of confirmed cases. Less testing has the opposite effect. Other ways would be post-mortem testing to raise number of deaths.

Although political representations might have various motivation for statistics' fabrication, there seems to be ways to influence the Covid-19 statistics even in the countries with separation of power and system of checks and balances.

# 7 Conclusion

- *What are the distributions of characteristics of Covid-19 - the incubation period, reproduction number, infection fatality ratio and duration of disease?*

Literature suggests incubation period 5.1 days (CI 95% $2.2 - 11.5$ days and CI 50% $3.8 - 6.7$ days), reproduction number $2 - 4$ and infection fatality ratio 0.0064 (credible interval 95% $0.0038 - 0.0098$). The latter depends on age, where age group $20 - 49$ years has an IFR estimate of 0.000092, while for $65+$ it is 0.056. Mean duration of disease (infectiousness duration, excluding incubation period) has been estimated from data of hospitalized patients as 15.5, with 95% band $4.225 - 32.775$.

Experimentally IFR estimate is $2 - 3\%$, duration of symptoms $12 - 24$ days and the basic reproduction number $2 - 3$, although estimated on the cases confirmed by tests using `EpiEstim` package, the $R_0$ was mostly less than 2. Simulation sometimes yielded unrealistic estimates of the characteristics, which have been removed.

- *To what extent are the collected data used to fit the model reliable?*

During the first wave of Covid-19 (first half of 2020) lower number of tests was performed in all four countries. Since August 2020, the ratio of tests starts growing and the sample is more reliable, it is about $0.2\% - 0.5\%$ of population tested daily.

Data have defects, such as significantly lower deaths in Poland connected with wrong method of reporting the time of deaths. In Poland, Michał Rogalski has indicated that the statistics published by the government have been under-reported [4]. In addition, there is still an ongoing discussion about when to label a death with cause Covid-19, i.e. to die on Covid-19 or with Covid-19 if the patient has a serious comorbidity [45].

- *How much are the reported statistics projected in the results of simulation?*

The model fits well to the first few samples and follows them closely. After some time, the prediction deflects from the data, drops down and flattens to 0. It is possible to derive the disease characteristics from the model parameters, but experiments have shown, that in some cases they were a lot off compared to the estimates in other research.

- *Are there visible patterns or similarities between regions?*

Similarities between all the regions strongly correspond with the countries regions belong to. There is not a lot of outlying regions, although countries do differ from each other quite a lot in time of the outbreak, intensity or duration of the high incidence.

Assuming the data from the second and the incomplete third wave, i.e. between August 2020 and March 2021, Italy and Poland have a very similar pandemic development during the second and the third wave. Sweden had a delayed start and Czechia encountered the greatest intensity regarding per-population normalized incidence.

- *Do the introduced restrictions influence the numbers?*

Effect of restrictions comes delayed caused by the serial interval. Although it is impossible to infer a direct causality just from the incidence, some events seem to cause a changes in numbers such as peaks during Christmas in Italy and Poland.

Regarding potential power of the restrictions, full lockdown is very often followed by a descend of the incidence and thus seems to be the most efficient for slowing the spread down. Example is Czechia, where lockdowns were repeatedly imposed and released (fig. 7.1, 7.2 and 7.3).

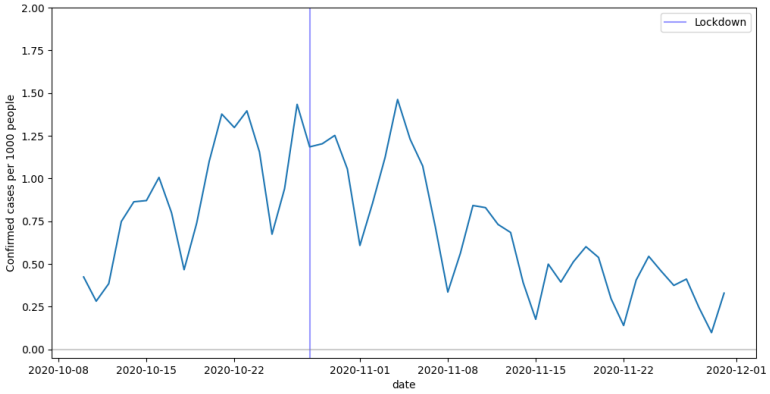



Figure 7.1: Second lockdown in Czechia, 28 Oct 2020.

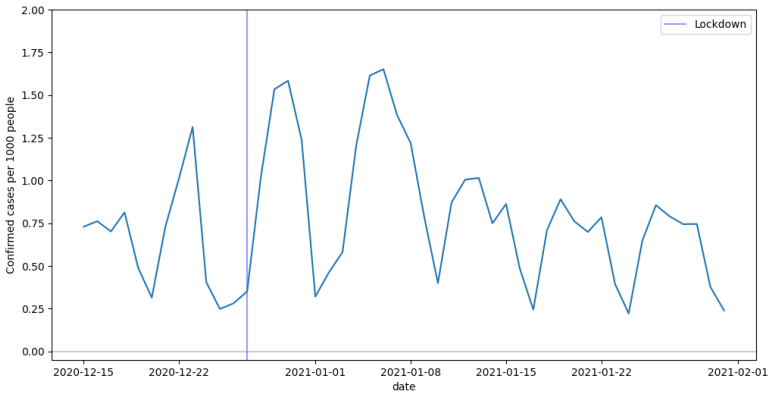



Figure 7.2: Third (nightly) lockdown in Czechia, 27 Dec 2020.

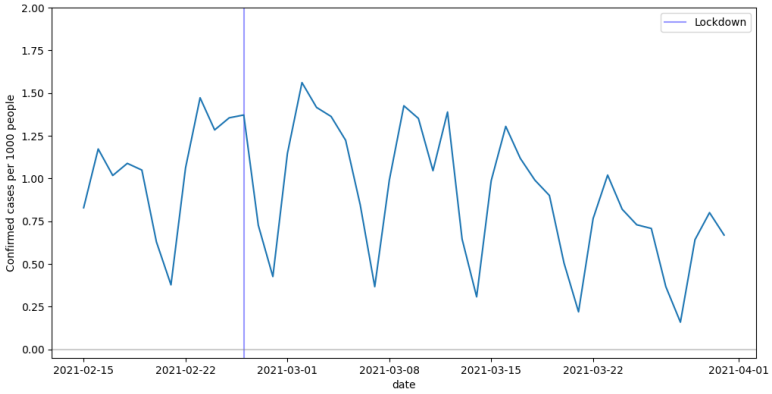
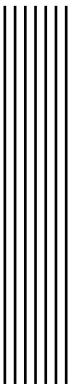



Figure 7.3: Fourth (district) lockdown in Czechia, 26 Feb 2021.

For compulsory mask wearing there has been no observation where imposing the restriction would precede a drop in incidence. For closing the schools there is not clear evidence as the data contains examples of both reopening of schools without any change afterwards and with an abrupt turn.

- *To what extent do the results show that the drafted model of the disease is correct?*

Compartment models are able to fit to a certain segment of the data, but they do not seem to be appropriate for modelling outbreaks with a long duration (lasting months or years) as they flatten after some time.



# Bibliography


[1] Janetta Němcová. *Hygienické stanice se potýkají s nedostatkem lidí. Odborníky se nedaří najít i kvůli nízkým platům*. June 2020. URL: https://www.irozhlas.cz/zpravy-domov/zdravotnictvi-hygienicke-stanice-lekari-personalni-krize-koronavirus_2006110714_ada.

[2] Kamil Turecki. *Pandemia obnażyła braki, choć budżet na sanepidy jest największy od lat*. Oct. 2020. URL: https://wiadomosci.onet.pl/tylko-w-onecie/koronawirus-sanepid-spoznione-inwestycje-w-stacjach-sanitarno-epidemiologicznych/dv1yynr.

[3] Bundesministerium für Finanzen. *Pressekonferenz am 14.03.2020*. Mar. 2020. URL: https://www.flickr.com/photos/159530260@N03/49657522158/.

[4] Radek Kosarzycki. *Teraz już nikt nieprawidłości nie znajdzie. Rząd zmienił sposób raportowania zakażeń koronawirusem*. Nov. 2020. URL: https://spidersweb.pl/2020/11/mz-zmiana-raportowania-nowych-zakazen.html.

[5] Encyclopaedia Britannica. *Epidemic*. URL: https://www.britannica.com/science/epidemic.

[6] Arthur Albert St. M. Mouritz. *"The Flu" A brief history of influenza in U.S. America, Europe, Hawaii*. 1921. URL: http://resource.nlm.nih.gov/101283076.

[7] Owen Jarus. *Gruesome Find: 100 Bodies Stuffed into Ancient House*. July 2015. URL: https://www.livescience.com/51662-100-bodies-found-prehistoric-house.html.

[8] Nicholas LePan. *Visualizing the History of Pandemics*. Mar. 2020. URL: https://www.visualcapitalist.com/history-of-pandemics-deadliest/.

[9] Jenny Howard. *Plague was one of history's deadliest diseases—then we found a cure*. July 2020. URL: https://www.nationalgeographic.com/science/article/the-plague.

[10] John Horgar. *Justinian's Plague (541-542 CE)*. Dec. 2014. URL: https://www.ancient.eu/article/782/justinians-plague-541-542-ce/.

[11] Barbara Bramanti, Katharine R. Dean, Lars Walløe, and Nils Chr. Stenseth. *The Third Plague Pandemic in Europe*. Apr. 2019. DOI: 10.1098/rspb.2018.2429.

[12] Chris W. Potter. "A history of influenza". In: *Journal of Applied Microbiology* 91.4 (2001), pp. 572–579. DOI: 10.1046/j.1365-2672.2001.01492.x.

[13] Paul A. Blake. *Historical Perspectives on Pandemic Cholera*. Mar. 1994. DOI: 10.1128/9781555818364.ch18.

[14] Didier Raoult, Theodore Woodward, and J. Stephen Dumler. "The history of epidemic typhus". In: *Infectious Disease Clinics of North America* 18.1 (2004). Historical Aspects of Infectious Diseases, Part I, pp. 127–140. ISSN: 0891-5520. DOI: 10.1016/S0891-5520(03)00093-X.

[15] Joel G. Breman and Isao Arita. *The Confirmation and Maintance of Smallpox Eradication*. 1980.

[16] Alan D. T. Barrett and Stephen Higgs. "Yellow Fever: A Disease that Has Yet to be Conquered". In: 52 (Jan. 2007), pp. 209–229. DOI: 10.1146/annurev.ento.52.110405.091454.

[17] Kenneth J. Arrow, Claire B. Panosian, and Hellen Gelband. *Saving Lives, Buying Time: Economics of Malaria Drugs in an Age of Resistance*. 2004. DOI: 10.17226/11017.

[18] Rodolfo Acuna-Soto, David W. Stahle, Malcolm K. Cleaveland, and Matthew D. Therrell. "Megadrought and Megadeath in 16th Century Mexico". In: (2004). DOI: 10.3201/eid0804.010175.

[19] Edwin D. Kilbourne. "Influenza Pandemics of the 20th Century". In: 12 (Jan. 2006), pp. 9–14. DOI: 10.3201/eid1201.051254.

[20] Krzysztof Kuszewski and Lidia Brydak. "The epidemiology and history of influenza". In: *Biomedicine Pharmacotherapy* 54.4 (2000), pp. 188–195. ISSN: 0753-3322. DOI: 10.1016/S0753-3322(00)89025-3.

[21] World Health Organization. "HIV/AIDS". In: (Nov. 2020). URL: https://www.who.int/news-room/fact-sheets/detail/hiv-aids.

[22] Nanographics GmbH. *High resolution renderings of SARS-CoV-2 Cryo-ET*. 2021. URL: https://nanographics.at.

[23] Ministry of Health of Czechia. *Onemocnění aktuálně*. 2021. URL: https://onemocneni-aktualne.mzcr.cz/covid-19.

[24] Martin Beneš. *Covid19Poland: Web Scraper of COVID-19 data for Poland*. 2021. URL: https://pypi.org/project/covid19poland.

[25] Instituto Nazionale di Statistica. *Impact of Covid-19 Epidemic on Mortality: Causes of Death in Covid-19 Laboratory Confirmed Cases*. July 2020. URL: https://dc-covid.site.ined.fr/en/data/italy/.

[26] European Centre for Disease Prevention and Control. *Clinical characteristics of COVID-19*. Aug. 2020. URL: https://www.ecdc.europa.eu/en/covid-19/latest-evidence/clinical.

[27] Marco Cascella, Michael Rajnik, Abdul Aleem, Scott C. Dulebohn, and Raffaela Di Napoli. *Features, Evaluation, and Treatment of Coronavirus*. 2020. URL: https://www.ncbi.nlm.nih.gov/books/NBK554776/.

[28] Martin Beneš. *Covid19Czechia: Web Scraper of COVID-19 data for Czechia*. 2021. URL: https://pypi.org/project/covid19czechia.

[29] Instituto Superiore di Sanità. *Sorveglianza integrata COVID-19: i principali dati nazionali*. 2021. URL: https://www.epicentro.iss.it/coronavirus/sars-cov-2-sorveglianza-dati.

[30] Emanuele Guidotti and David Ardia. "COVID-19 Data Hub". In: *Journal of Open Source Software* 5.51 (2020), p. 2376. DOI: 10.21105/joss.02376.

[31] Marco Cascella, Michael Rajnik, Arturo Cuomo, Scott Dulebohn, and Raffaela Napoli. "Features, Evaluation, and Treatment of Coronavirus (COVID-19)". In: (Jan. 2021). URL: https://www.ncbi.nlm.nih.gov/books/NBK554776/.

[32] Waidi Folounso Sule and Daniel Oladimeji Oluwayelu. "Real-time RT-PCR for COVID-19 diagnosis: challenges and prospects". In: *The Pan African Medical Journal* (35 July 2020). DOI: 10.11604/pamj.supp.2020.35.24258.

[33] Georgia Guglielmi. *Fast coronavirus tests: what they can and can't do*. Sept. 2020. URL: https://www.nature.com/articles/d41586-020-02661-2.

[34] Robert Kubina and Arkadiusz Dziedzic. "Molecular and Serological Tests for COVID-19. A Comparative Review of SARS-CoV-2 Coronavirus Laboratory and Point-of-Care Diagnostics". In: *Diagnostics* 10.6 (2020), p. 434. DOI: 10.3390/diagnostics10060434.

[35] Anaïs Scohy, Ahalieyah Anantharajah, Monique Bodéus, Benoît Kabamba-Mukadi, Alexia Verroken, and Hector Rodriguez-Villalobos. "Low performance of rapid antigen detection test as frontline testing for COVID-19 diagnosis". In: *Journal of Clinical Virology* 129 (2020), p. 104455. DOI: 10.1016/j.jcv.2020.104455.

[36] Yutaka Okabe and Akira Shudo. "A Mathematical Model of Epidemics—A Tutorial for Students". In: *Mathematics* 8.7 (2020). DOI: 10.3390/math8071174.

[37] Dan Connelly. *SEIRD model of COVID-19*. 2020. URL: https://medium.com/@djconnel_14663/seird-model-of-covid-19-596e6754c2c4.

[38] Adam Kucharski. *The Rules of Contagion*. 2020, p. 352. ISBN: 978-1-78816-019-3.

[39] Changguo Li, Yongzhen Pei, Meixia Zhu, and Yue Deng. "Parameter Estimation on a Stochastic SIR Model with Media Coverage". In: 2018 (2018). DOI: 10.1155/2018/3187807.

[40] Agnieszka Bartłomiejczyk and Marcin Wata. "Analizy epidemiologiczne w środowisku Matlab/Octave". In: *VI Konferencja e-Technologie w Kształceniu Inżynierów eTEE'2019*. 2019, pp. 11–16. DOI: 10.32016/1.65.01.

[41] Stan Development Team. *Stan*. Feb. 2021. URL: https://mc-stan.org/.

[42] Guerino Mazola, Gérard Milmeister, and Jody Weissmann. *Comprehensive Mathematics for Computer Scientists 2*. Springer, 2005. ISBN: 3-540-20861-5.

[43] Nikolaj Ezhov, Frank Neitzel, and Svetozar Petrovic. "Spline approximation, Part 1: Basic methodology." In: 12.2 (2018), pp. 139–155. DOI: 10.1515/jag-2017-0029.

[44] Bernd Kamps and Christian Hoffmann. *COVID Reference*. Steinhäuser Verlag, Jan. 2021. ISBN: 978-3-942687-53-9. URL: https://amedeo.com/CovidReference06.pdf.

[45] Thomas A. Slater, Sam Straw, Michael Drozd, Stephe Kamalathasan, Alice Cowley, and Klaus K. Witte. "Dying 'due to' or 'with' COVID-19: a cause of death analysis in hospitalised patients". In: *Clinical Medicine* 20.5 (2020), e189–e190. ISSN: 1470-2118. DOI: 10.7861/clinmed.2020-0440.

[46] AFP. *UK 'overestimates' coronavirus death toll: Study*. July 2020. URL: https://health.economictimes.indiatimes.com/news/diagnostics/uk-overestimates-coronavirus-death-toll-study/77034966.

[47] Nina Schwalbe. *We could be vastly overestimating the death rate for COVID-19. Here's why*. Apr. 2020. URL: https://www.weforum.org/agenda/2020/04/we-could-be-vastly-overestimating-the-death-rate-for-covid-19-heres-why/.

[48] Laurie Davis. "Excess deaths, baselines, Z-scores, P-scores and peaks". In: (Oct. 2020). URL: https://arxiv.org/abs/2010.10320.

[49] Ministry of Health of the Republic of Poland. *Oficjalny profil Ministerstwa Zdrowia*. 2020. URL: https://twitter.com/MZ_GOV_PL.

[50] Ministry of Health of the Republic of Poland. *Koronawirus: informacje i zalecenia*. 2021. URL: https://www.gov.pl/web/koronawirus/pliki-archiwalne-powiaty.

[51] Philipp Fritz. *Polens Pandemie-Prophet*. Mar. 2021. URL: https://www.welt.de/politik/ausland/plus229336649/Corona-Analyst-Michal-Rogalski-Polens-Pandemie-Prophet.html.

[52] Michał Rogalski. *COVID-19 w Polsce*. 2021. URL: https://docs.google.com/spreadsheets/u/1/d/1ierEhD6gcq51HAm433knjnVwey4ZE5DCnu1bW7PRG3E.

[53] The Public Health Agency of Sweden. *Statistics and analyses*. 2021. URL: https://www.folkhalsomyndigheten.se/smittskydd-beredskap/utbrott/aktuella-utbrott/covid-19/statistik-och-analyser/.

[54] Martin Beneš. *Covid19Sweden: Web Scraper of COVID-19 data for Sweden*. 2021. URL: https://pypi.org/project/covid19sweden.

[55] Presidenza del Consiglio dei Ministri. *Dati COVID-19 Italia*. 2021. URL: https://github.com/pcm-dpc/COVID-19.

[56] Simona Iftimie, Ana F. López-Azcona, Immaculada Vallverdú, Salvador Hernàndez-Flix, Gabriel de Febrer, Sandra Parra, Anna Hernández-Aguilera, Francesc Riu, Jorge Joven, Jordi Camps, Antoni Castro, and REUSCOVID Study Group. "First and second waves of coronavirus disease-19: A comparative study in hospitalized patients in Reus, Spain". In: *medRxiv* (2020). DOI: 10.1101/2020.12.10.20246959.

[57] Martin Beneš. *A simple interface for parsing data from Eurostat*. 2021. URL: https://pypi.org/project/eurostat-deaths/.

[58] The Local. *Coronavirus: What went wrong in Sweden's care homes?* May 2020. URL: https://www.thelocal.se/20200506/coronavirus-what-went-wrong-in-swedens-care-homes/.

[59] Eurostat. *NUTS Maps*. URL: https://ec.europa.eu/eurostat/web/nuts/nuts-maps.

[60] Eurostat. *Deaths by week, sex, 5-year age group and NUTS 3 region*. 2021. URL: https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=demo_r_mweek3.

[61] Keiko Miyamoto, Fumiya Kawase, Tomoko Imai, Ayako Sezaki, and Hiroshi Shimokata. "Dietary diversity and healthy life expectancy—an international comparative study". In: *European Journal of Clinical Nutrition* 73 (2019), pp. 395–400. DOI: 10.1038/s41430-018-0270-3.

[62] Bert Brunekreef. "Air pollution and life expectancy: is there a relation?" In: *Occupational and Environmental Medicine* 54 (11 1997), pp. 781–784. DOI: 10.1136/oem.54.11.781.

[63] Johan P. Mackenbach, Yannan Hu, and Caspar W. N. Looman. "Democratization and life expectancy in Europe, 1960–2008". In: *Social Science Medicine* 93 (2013), pp. 166–175. ISSN: 0277-9536. DOI: 10.1016/j.socscimed.2013.05.010.

[64] Eurostat. *Population on 1 January by age group, sex and NUTS 3 region*. 2021. URL: http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=demo_r_pjangrp3.

[65] Michele Tommasi. *Il Miracolo economico italiano (1958-1963): storia, origini e conseguenze*. URL: https://www.studenti.it/miracolo-economico-italiano-1958-1963-storia-origini-conseguenze.html.

[66] Richard J. B. Bosworth. *Mussolini's italy : life under the fascist dictatorship, 1915-1945*. Penguin Group US, 2014. ISBN: 978-1-101-07857-0.

[67] Sławomir Dmowski. *Struktura demograficzna ludności Polski*. 2017. URL: https://www.geografia24.eu/geo_prezentacje_rozsz_3/383_2_ludnosc_urbanizacja/r3_2_03a.pdf?fbclid=IwAR1Obd_vKnuWJ-APcBg41hcMW4UUoKrbQ1XnG6S6SKURIL46NYI-WjsK0MQ.

[68] Český statistický úřad. *"Husákovy" versus „Havlovy děti"*. 2014. URL: https://www.czso.cz/csu/czso/52002e2055.

[69] World Health Organization. *Transmission of SARS-CoV-2: implications for infection prevention precautions*. July 2020. URL: https://www.who.int/news-room/commentaries/detail/transmission-of-sars-cov-2-implications-for-infection-prevention-precautions.

[70] Stephen A. Lauer, Qifang Grantz Kyra H. amd Bi, Forrest K. Jones, Qulu Zheng, Hannah R. Meredith, Andrew S. Azman, Nicholas G. Reich, and Justin Lessler. "The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application". In: *Annals of Internal Medicine* 172.9 (2020), pp. 577–582. DOI: 10.7326/M20-0504.

[71] Xuan Jiang, Simon Rayner, and Min-Hua Luo. "Does SARS-CoV-2 has a longer incubation period than SARS and MERS?" In: *Journal of Medical Virology* 92.5 (2020), pp. 476–478. DOI: 10.1002/jmv.25708.

[72] Hiroshi Nishiura, Tetsuro Kobayashi, Takeshi Miyama, Ayako Suzuki, Sung-mok Jung, Katsuma Hayashi, Ryo Kinoshita, Yichi Yang, Baoyin Yuan, Andrei R. Akhmetzhanov, and Natalie M. Linton. "Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19)". In: *International Journal of Infectious Diseases* 94 (May 2020), pp. 154–155. DOI: 10.1016/j.ijid.2020.03.020.

[73] Jeroen J. A. van Kampen, David A. M. C. van de Vijver, Pieter L. A. Fraaij, Bart L. Haagmans, Mart M. Lamers, Nisreen Okba, Johannes P. C. van den Akker, Henrik Endeman, Diederik A. M. P. J. Gommers, Jan J. Cornelissen, Rogier A. S. Hoek, Menno M. van der Eerden, Dennis A. Hesselink, Herold J. Metselaar, Annelies Verbon, Jurriaan E. M. de Steenwinkel, Georgina I. Aron, Eric C. M. van Gorp, Sander van Boheemen, Jolanda C. Voermans, Charles A. B. Boucher, Richard Molenkamp, Marion P. G. Koopmans, Corine Geurtsvankessel, and Annemiek A. van der Eijk. "Duration and key determinants of infectious virus shedding in hospitalized patients with coronavirus disease-2019 (COVID-19)". In: *Nature Communications* 12.1 (Jan. 2021). DOI: 10.1038/s41467-020-20568-4.

[74] Qingxian Cai, Deliang Huang, Pengcheng Ou, Hong Yu, Zhibin Zhu, Zhang Xia, Yinan Su, Zhenghua Ma, Yiming Zhang, Zhiwei Li, Qing He, Lei Liu, Yang Fu, and Jun Chen. "COVID-19 in a designated infectious diseases hospital outside Hubei Province, China". In: *Allergy* 75.7 (2020), pp. 1742–1752. DOI: 10.1111/all.14309.

[75] Roman Wölfel, Victor M. Corman, Wolfgang Guggemos, Michael Seilmaier, Sabine Zange, Marcel A. Müller, Daniela Niemeyer, Terry C. Jones, Patrick Vollmar, Camilla Rothe, Michael Hoelscher, Tobias Bleicker, Sebastian Brünink, Julia Schneider, Rosina Ehmann, Katrin Zwirglmaier, Christian Drosten, and Clemens Wendtner. "Virological assessment of hospitalized patients with COVID-2019". In: *Nature* 581 (May 2020), pp. 465–469. DOI: 10.1038/s41586-020-2196-x.

[76] Centers for Disease Control and Prevention. "Duration of Isolation and Precautions for Adults with COVID-19". In: (Oct. 2020). URL: https://www.cdc.gov/coronavirus/2019-ncov/hcp/duration-isolation.html.

[77] Yi Xu, Xufang Li, Bing Zhu, Huiying Liang, Chunxiao Fang, Yu Gong, Qiaozhi Guo, Xin Sun, Danyang Zhao, Jun Shen, Huayan Zhang, Hongsheng Liu, Huimin Xia, Jinling Tang, Kang Zhang, and Sitang Gong. "Characteristics of pediatric SARS-CoV-2 infection and potential evidence for persistent fecal viral shedding". In: *Nature* 26 (2020), pp. 502–505. DOI: 10.1038/s41591-020-0817-4.

[78] Chaoqun Han, Caihan Duan, Shengyan Zhang, Brennan Spiegel, Huiying Shi, Weijun Wang, Lei Zhang, Rong Lin, Jun Liu, Zhen Ding, and Xiaohua Hou. "Digestive Symptoms in COVID-19 Patients With Mild Disease Severity: Clinical Presentation, Stool Viral RNA Testing, and Outcomes". In: *The American journal of gastroenterology* 115 (2020), pp. 916–923. DOI: 10.14309/ajg.0000000000000664.

[79] Sonja Lehtinen, Peter Ashcroft, and Sebastian Bonhoeffer. "On the relationship between serial interval, infectiousness profile and generation time". In: (Jan. 2021). ISSN: 1742-5662. DOI: 10.1098/rsif.2020.0756.

[80] Mohammad Aghaali, Goodarz Kolifarhood, Roya Nikbakht, Hossein Mozafar Saadati, and Seyed Saeed Hashemi Nazari. *Estimation of the serial interval and basic reproduction number of COVID-19 in Qom, Iran, and three other countries: A data-driven analysis in the early phase of the outbreak*. June 2020. DOI: 10.1111/tbed.13656.

[81] Chong You, Yuhao Deng, Wenjie Hu, Jiarui Sun, Qiushi Lin, Feng Zhou, Cheng Heng Pang, Yuan Zhang, Zhengchao Chen, and Xiao-Hua Zhou. "Estimation of the time-varying reproduction number of COVID-19 outbreak in China". In: *International Journal of Hygiene and Environmental Health* 228 (2020), p. 113555. ISSN: 1438-4639. DOI: 10.1016/j.ijheh.2020.113555.

[82] World Health Organization. *Coronavirus disease 2019 (COVID-19): Situation Report – 46*. Mar. 2020. URL: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200306-sitrep-46-covid-19.pdf.

[83] Kevin Linka, Mathias Peirlinck, and Ellen Kuhl. *The reproduction number of COVID-19 and its correlation with public health interventions*. 2020. DOI: 10.1007/s00466-020-01880-8.

[84] Farid Najafi, Nazanin Izadi, Seyed Saeed Hashemi-Nazari, Fatemeh Khosravi-Shedmani, Roya Nikbakht, and Ebrahim Shakiba. *Serial interval and time-varying reproduction number estimation for COVID-19 in western Iran*. June 2020. DOI: 10.1016/j.nmni.2020.100715.

[85] Jacco Wallinga and Peter Teunis. "Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures". In: *American Journal of Epidemiology* 160.6 (Sept. 2004), pp. 509–516. ISSN: 0002-9262. DOI: 10.1093/aje/kwh255.

[86] Jacco Wallinga and Peter Teunis. "Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures". In: (Sept. 2004). DOI: 10.1093/aje/kwh255.

[87] World Health Organization. *Estimating mortality from COVID-19*. Aug. 2020. URL: https://www.who.int/news-room/commentaries/detail/estimating-mortality-from-covid-19.

[88] Javier Perez-Saez, Stephen A. Lauer, Laurent Kaiser, Simon Regard, Elisabeth Delaporte, Idris Guessous, Silvia Stringhini, Andrew S. Azman, and Serocov-POP Study Group. "Serology-informed estimates of SARS-CoV-2 infection fatality risk in Geneva, Switzerland". In: 21 (4 July 2020). DOI: 10.1016/S1473-3099(20)30584-3.

[89] John P. A. Ioannidis. "Infection fatality rate of COVID-19 inferred from seroprevalence data". In: *Bulletin of the World Health Organization* 99.1 (Oct. 2020), pp. 19–33. DOI: 10.2471/blt.20.265892.

[90] Gideon Meyerowitz-Katz and Lea Merone. "A systematic review and meta-analysis of published research data on COVID-19 infection fatality rates". In: *Intenational Journal of Infectious Diseases* 101 (Dec. 2020), pp. 138–148. DOI: 10.1016/j.ijid.2020.09.1464.

[91] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. "Array programming with NumPy". In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2.

[92] John D. Hunter. "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[93] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[94] Wes McKinney. "Data structures for statistical computing in python". In: *Proceedings of the 9th Python in Science Conference*. Vol. 445. Austin, TX. 2010, pp. 51–56.

[95] Michael Waskom, Olga Botvinnik, Drew O'Kane, Paul Hobson, Saulius Lukauskas, David C. Gemperline, Tom Augspurger, Yaroslav Halchenko, John B. Cole, Jordi Warmenhoven, Julian de Ruiter, Cameron Pye, Stephan Hoyer, Jake Vanderplas, Santi Villalba, Gero Kunter, Eric Quintero, Pete Bachant, Marcel Martin, Kyle Meyer, Alistair Miles, Yoav Ram, Tal Yarkoni, Mike Lee Williams, Constantine Evans, Clark Fitzgerald, Brian, Chris Fonnesbeck, Antony Lee, and Adel Qalieh. *mwaskom/seaborn: v0.8.1 (September 2017)*. Version v0.8.1. Sept. 2017. DOI: 10.5281/zenodo.883859.

[96] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830. URL: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

[97] Charlie Clark and Ericand Gazoni. *A Python library to read/write Excel 2010 xlsx/xlsm files*. Version 3.0.7. 2021. URL: https://openpyxl.readthedocs.io/en/stable/index.html.

[98] Kenneth Reitz. *Requests: Python HTTP for Humans*. Version v2.25.1. URL: https://docs.python-requests.org/.

[99] Skipper Seabold and Josef Perktold. "statsmodels: Econometric and statistical modeling with python". In: *9th Python in Science Conference*. 2010. DOI: 10.25080/Majora-92bf1922-011.

[100] Ryan Solgi. *Genetic Algorithm: An easy implementation of genetic-algorithm (GA) to solve continuous and combinatorial optimization problems*. Version v1.0.2. URL: https://pypi.org/project/geneticalgorithm/.

[101] Anne Cori, Zhian Kamvar, Jake Stockwin, Thibaut Jombart, Elisabeth Dahlqwist, Rich FitzJohn, and Robin Thompson. *EpiEstim v2.2-3: A tool to estimate time varying instantaneous reproduction number during epidemics*. 2021. URL: https://github.com/mrc-ide/EpiEstim.

[102] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

[103] Randall Pruim, Daniel T. Kaplan, and Nicholas J. Horton. "The mosaic Package: Helping Students to 'Think with Data' Using R". In: *The R Journal* 9.1 (2017), pp. 77–102. URL: https://journal.r-project.org/archive/2017/RJ-2017-024/index.html.

[104] Stan Development Team. *RStan: the R interface to Stan*. R package version 2.21.2. 2020. URL: http://mc-stan.org/.

[105] Jonah Gabry and Tristan Mahr. *bayesplot: Plotting for Bayesian Models*. R package version 1.8.0. 2021. URL: https://mc-stan.org/bayesplot/.

[106] Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*. R package version 0.7.6. 2018. URL: https://CRAN.R-project.org/package=dplyr.

[107] Christian Hill. "The SIR epidemic model". In: *Learning Scientific Programming with Python*. Cambridge University Press, 2015. ISBN: 978-1-107-07541-2. URL: https://scipython.com/book/chapter-8-scipy/additional-examples/the-sir-epidemic-model/.

[108] Magdalena Kulej. *To pokazuje skalę protestu. "Marsz na Warszawę" na nieprawdopodobnym nagraniu z drona*. Oct. 2020. URL: https://wiadomosci.radiozet.pl/Polska/Warszawa/Protest-w-Warszawie-30.10.-Nieprawdopodobne-nagranie-i-zdjecia-z-drona.

[109] Ondřej Májek, Ondřej Ngo, Jiří Jarkovský, Monika Ambrožová, Barbora Budíková, Jan Kouřil, Ladislav Dušek, and Tomáš Pavlík. *Dokumentace k epidemiologickému modelu ÚZIS ČR pro krátkodobé predikce*. June 2020. URL: https://onemocneni-aktualne.mzcr.cz/doc/dokumentace-modely.pdf.

[110] Folkhälsomyndigheten. *Riket: skattning av det momentana reproduktionstalet*. June 2020. URL: https://web.archive.org/web/20200720001846/https://www.folkhalsomyndigheten.se/contentassets/4b4dd8c7e15d48d2be744248794d1438/riket-skattning-av-effektiva-reproduktionsnumret-2020-07-13.pdf.

[111] Howida Slama, Abdullah Hussein, Nabila A. El-Bedwhey, and Mustafa M. Selim. "An approximate probabilistic solution of a random SIR-type epidemiological model using RVT technique". In: *Applied Mathematics and Computation* 361 (2019), pp. 144–156. ISSN: 0096-3003. DOI: 10.1016/j.amc.2019.05.019.

[112] Sayan Nag. *A mathematical model in the time of Covid-19*. Mar. 2020. DOI: 10.31219/osf.io/8n92h.

[113] Günter Schneckenreither, Nikolas Popper, Günther Zauner, and Felix Breitenecker. "Modelling SIR-type epidemics by ODEs, PDEs, difference equations and cellular automata – A comparative study". In: *Simulation Modelling Practice and Theory* 16.8 (2008), pp. 1014–1023. ISSN: 1569-190X. DOI: 10.1016/j.simpat.2008.05.015.

[114] BBC. *Covid-19 vaccine: 'Hopefully next year we'll be living a normal life'*. Dec. 2020. URL: https://www.bbc.com/news/uk-55230423.

[115] Rachel Treisman. *The Country That Vaccinated 93% Of Adults In Under 2 Weeks*. Apr. 2021. URL: https://www.npr.org/sections/coronavirus-live-updates/2021/04/12/986450910/the-advantage-of-our-smallness-bhutan-vaccinates-93-of-adults-in-under-2-weeks.

[116] Jacqui Wise. "Covid-19: Pfizer BioNTech vaccine reduced cases by 94% in Israel, shows peer reviewed study". In: *BMJ* 372 (2021). DOI: 10.1136/bmj.n567.

[117] Robert Preidt. *Like Flu, COVID-19 May Turn Out to Be Seasonal*. Feb. 2021. URL: https://www.webmd.com/lung/news/20210202/like-flu-covid-19-may-turn-out-to-be-seasonal.

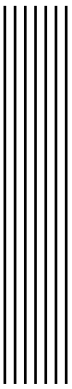
[118] Behailu Taye, Kidane Lelisa, Daniel Emana, Abebe Asale, and Delenasaw Yewhalaw. "Seasonal Dynamics, Longevity, and Biting Activity of Anopheline Mosquitoes in Southwestern Ethiopia". In: *Journal of Insect Science* 16.1 (Jan. 2016). 6. ISSN: 1536-2442. DOI: 10.1093/jisesa/iev150.

[119] Ahmad Naserpor, Sharareh Kalhori, Marjan Ghazisaeedi, Rasoul Azizi, Hosseini Ravandi, and Sajad Sharafie. "Modification of the Conventional Influenza Epidemic Models Using Environmental Parameters in Iran". In: *Healthcare Informatics Research* 25.1 (2019), pp. 27–32. DOI: 10.4258/hir.2019.25.1.27.

[120] Michael A. Johansson, Talia M. Quandelacy, Sarah Kada, Pragati Venkata Prasad, Molly Steele, John T. Brooks, Rachel B. Slayton, Matthew Biggerstaff, and Jay C. Butler. "SARS-CoV-2 Transmission From People Without COVID-19 Symptoms". In: *JAMA Network Open* 4.1 (Jan. 2021), e2035057–e2035057. ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2020.35057.

[121] Amy Walker, Lisa Jones, and Lazaro Gamio. *Is the Coronavirus Death Tally Inflated? Here's Why Experts Say No*. July 2020. URL: https://www.nytimes.com/interactive/2020/06/19/us/us-coronavirus-covid-death-toll.html.

[122] David Trilling. *Statistics show Central Asia underreporting COVID deaths*. Feb. 2021. URL: https://eurasianet.org/statistics-show-central-asia-underreporting-covid-deaths.



# Appendices

## Github code repository

The code implementing the model and producing the results, including the plots in this thesis is publicly accessible at `https://github.com/martinbenes1996/732A64`.

## Calendar

The calendar is a collected list of events, potentially influential on the pandemic situation in the analyzed countries. It is a supplement material of the thesis in a file `calendar.csv` or it can be found in the code repository as a file `data/calendar.csv`.