



Marcelo Henrique de Oliveira Mrtvi

Estimation of the Effective Reproduction Number for the COVID-19 Pandemic

Advisor: Prof. Isolde Previdelli
Co-Advisor: Prof. Anthony C. Davison

Maringá – Paraná
2021

Marcelo Henrique de Oliveira Mrtvi

Estimation of the Effective Reproduction Number for the COVID-19 Pandemic

Dissertação apresentada ao Programa de Pós-graduação em Bioestatística do centro de ciências exatas da Universidade Estadual de Maringá como requisito parcial para obtenção do título de mestre em Bioestatística.

Orientador: Prof. Isolde Previdelli

Coorientador: Prof. Anthony C. Davison

Universidade Estadual de Maringá - UEM

Departamento de Estatística - DES

Programa de Pós-Graduação em Bioestatística

Maringá – Paraná

2021

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá - PR, Brasil)

M939e

Mrtvi, Marcelo Henrique de Oliveira

Estimation of the effective reproduction number for the COVID-19 pandemic / Marcelo Henrique de Oliveira Mrtvi. -- Maringá, PR, 2022.
55 f.: il., figs.

Orientadora: Profa. Dra. Isolde Previdelli.

Coorientador: Prof. Dr. Anthony Davison.

Dissertação (Mestrado) - Universidade Estadual de Maringá, Centro de Ciências Exatas, Departamento de Estatística, Programa de Pós-Graduação em Bioestatística, 2022.

1. Bioestatística. 2. Covid-19. 3. Inferência Bayesiana - Método estatístico. 4. Número de reprodução - Covid-19. I. Previdelli, Isolde, orient. II. Davison, Anthony, coorient. III. Universidade Estadual de Maringá. Centro de Ciências Exatas. Departamento de Estatística. Programa de Pós-Graduação em Bioestatística. IV. Título.

CDD 23.ed. 570.15195

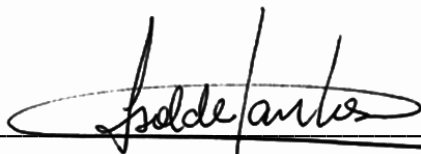
MARCELO HENRIQUE DE OLIVEIRA MRTVI

Estimation of the Effective Reproduction Number for the COVID-19

Pandemic

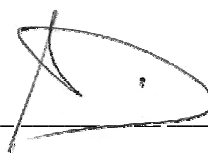
Dissertação apresentada ao Programa de Pós-Graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá, como requisito parcial para a obtenção do título de Mestre em Bioestatística.

BANCA EXAMINADORA



Prof.^a. Dra. Isolde Previdelli

Universidade Estadual de Maringá – PBE/UEM



Prof. Dr. Aluisio Jardim Dornellas Barros

Universidade Federal de Pelotas - UFPel-Pelotas-RS



Prof. Dr. Leonardo Soares Bastos

Programa de Computação Científica - PROCC-Fiocruz

Maringá, 01 de julho de 2021.

ACKNOWLEDGEMENTS

I would like to thank first my advisor, Professor Isolde Previdelli for the opportunity and her mother-like support and guidance during this project. It is very important to have an advisor that trusts in your potential but also knows how to push you to challenge yourself. Also just as important I would like to thank Professor Davison for the opportunity, his teaching and his infinite patience. I was very lucky to have the time and attention of such a prolific research. Working in his group and meeting such an attentive and caring professor was an remarkable experience. We from the department of Biostatistics of the State University of Maringá (UEM) want to thank the Chair of Statistics at EPFL for funding my stay in Switzerland during this project. I also want to thank my family and my wife Giovanna for their support during this period. Coming to another country is hard already, but the pandemic adds to the challenge. I am thankful for the people that I met in the academia, during the program in Biostatistics, André Ferreira and Breno Gabriel for their help during subjects and discussions. And also, for the people I met at EPFL: Jon, Mario, Servane, Sonia, Stefano and Tim that made this visit more special.

RESUMO

A pandemia de COVID-19 criou um dilema na sociedade, no qual a aplicação de medidas de saúde pública e restrições devem ser balanceadas com suas consequências econômicas. Para guiar as decisões durante a crise, um dos indicadores mais comuns usados por governos é o número reprodutivo (R). Na Suíça os efeitos da pandemia não foram diferentes do resto da Europa. Para ajudar durante a crise, o governo suíço criou a força tarefa de ciência suíça de COVID-19 (NCS-TF) cujo o grupo de modelagem e dados é responsável por produzir estimativas para R . Diversas abordagens para a estimação do número reprodutivo foram desenvolvidas e aplicadas à outras epidemias. O método da NCS-TF é baseado nos desenvolvimentos de Cori et. al. 2013 e o pacote do R "EpiEstim". Esse projeto utiliza uma abordagem Bayesiana para estimar o número reprodutivo na Suíça e em outros países. Ele se difere da abordagem atual pois estima a curva de incidência e os padrões semanais no mesmo algoritmo de Metropolis-Hastings. Apesar do maior tempo computacional em relação ao método da força tarefa Suíça, o uso de Splines como priori para R resultou em intervalos de confiança mais precisos em períodos de alta variação dos casos. Esse resultado foi expressivo se compararmos as estimativas no casos do Brasil na qual os intervalo de confiança do método utilizado trás uma maior segurança para as decisões governamentais.

Keywords: Algoritmo Metropolis-Hastings, COVID-19, Estimação Bayesiana, Número Reprodutivo.

ABSTRACT

The COVID-19 pandemic created a harsh dilemma for our society, in which the application of public health measures and restrictions has to be balanced with their economic consequences. To guide decisions during this crisis, one of the main indicators used by governments is the reproductive number (R). In Switzerland, the effects of the pandemic have not been different from the rest of Europe. To help during this crisis the Swiss government created the NCS-TF (Swiss National Covid-19 - Science Task Force) whose data and modelling group is responsible for producing estimates of R . Several approaches for the estimation of the reproductive number have been developed and applied in other epidemics. The NCS-TF method is based on the developments of Cori et. al. 2013 and the R package EpiEstim. This project uses a Bayesian approach to estimate the reproduction number in Switzerland and other countries. This differs from the current approach in the sense that estimates of the weekly patterns and the incidence curve are found using the same Metropolis–Hastings algorithm. Despite the longer computational effort compared to the NCS-TF, the use of splines as a prior for R resulted in narrower and more precise confidence intervals in periods of high variation on reported cases. This result is more evident in the estimates for Brazil, in which our method gives the decision–maker a narrower interval to decide on the implementation of public policies.

Keywords: Bayesian estimation, COVID-19, Metropolis–Hastings algorithm, Reproductive number.

CONTENTS

1	Introduction	7
	Introduction	7
1.1	A brief COVID-19 timeline: World, Switzerland and Brazil	7
2	Literature Review	12
2.1	Epidemiological Models	12
2.1.1	SIR Models	15
2.1.2	SEIR Models	18
2.1.3	Tailor-made models	19
2.2	Reproductive number	21
2.2.1	Instantaneous reproductive number	22
2.2.2	Case reproductive number	22
2.3	Serial interval and generation interval distributions	22
2.4	Reconstruction of the infection incidence curve	23
2.5	Swiss - COVID19 - Discussion	24
3	Methods	26
3.1	Data	26
3.1.1	Swiss data	27
3.1.2	Brazilian data	28
3.1.3	Other countries	30
R_t	- Estimation	30
3.2	Estimation Methods	30
3.3	Project Objectives	31
3.4	Project's Approach	32
3.4.1	Estimation of R_t	32
3.4.2	Splines	33
3.4.3	Weekly Pattern	35
3.5	Computational Methods	37
3.5.1	Markov Chain Monte Carlo (MCMC)	37
3.5.1.1	Definitions	37
3.5.2	Metropolis–Hastings Algorithm	39

4 Results	41
4.1 Simulated Data	41
4.2 Switzerland	44
5 Conclusion	47

Annex	49
ANNEX A ANNEX	50
Bibliography	52

CHAPTER 1

INTRODUCTION

1.1 A brief COVID-19 timeline: World, Switzerland and Brazil

- On 1 January 2020, the WHO requested information on the reported cluster of atypical pneumonia cases in Wuhan from the Chinese authorities and eight days later it was reported by Chinese authorities that the outbreak was caused by a novel coronavirus. The first mission to Wuha conducted by the WHO n occurred on 20-21st January 2020.
- Between 11th and 12th of February 2020 the WHO conducted a Global Research and Innovation Forum on the novel coronavirus, with participation by more than 300 experts and funders from 48 countries. Topics covered by the Forum included: the origin of the virus, its natural history, transmission and diagnosis; epidemiological studies; clinical characterisation and management; infection prevention and control; R&D for candidate therapeutics and vaccines; ethical considerations for research; and the integration of the social sciences into the response.
- Twelve days after the Forum, the WHO-China Joint Mission reported in a press conference that “much of the global community is not yet ready, in mindset and materially, to implement the measures that have been employed to contain COVID-19 in China” and that “to reduce COVID-19 illness and death, near-term readiness planning must embrace the large-scale implementation of high-quality, non-pharmaceutical public health measures”, such as case detection and isolation, contact tracing and monitoring/quarantining and community engagement.
- On 25 February 2020, Switzerland confirmed the first case of COVID-19 and on the next day the Brazilian Ministry of Health confirmed the first case in Brazil and Latin America. Four days later, the 100,000th case in the world was confirmed and the WHO

officially characterised COVID-19 as a pandemic. Unlike its Brazilian counterpart, the Swiss government response was fast, banning all events with more than 1000 participants early on.

- On 12 March 2020, the first death was registered in Brazil. On the next day Europe was declared by the WHO to have become the epicentre of the pandemic, with more reported cases and deaths than the rest of the world combined, apart from China.
- On 2 April 2020 the WHO reported evidence of transmission from pre-symptomatic and asymptomatic people infected with COVID-19, noting that transmission from a pre-symptomatic case can occur before symptom onset. This helps explain the fast dynamics of the spread of the disease. Two days later, the mark of 1 million cases worldwide was confirmed with more than a tenfold increase of cases in less than a month.
- On the 10th of April, Brazil reached 1,000 deaths from COVID-19. The next day the WHO published a draft landscape of COVID-19 candidate vaccines, on the basis of a systematic assessment of candidates from around the world.
- On the 19th of June 2020, Brazil reached one million COVID-19 cases. A diverse set of treatments was still under discussion. However, on 4th of July WHO announced that hydroxychloroquine and lopinavir/ritonavir were found to be ineffective regarding COVID-19 treatment and studies were discontinued. In the same month the 2020 edition of the UN's State of Food Security and Nutrition in the World was published, which forecasted that the COVID-19 pandemic could leave over 130 million more people in chronic hunger by the end of the year.
- On the 8th of August, Brazil reached three million cases and 100,000 deaths from COVID-19. On the 11th of November, the Brazilian Health Regulatory Agency authorized Sinovac to resume its vaccine trials less than 48 hours after halting the tests, which are being conducted by the Butantan Institute in the state of São Paulo.
- On the 23th of December of 2020 the Swiss vaccination campaign started.
- After more than a year of the pandemic, on the 5th of January 2021 the WHO's Strategic Advisory Group of Experts on Immunization (SAGE) reviewed the vaccine data for the Pfizer/BioNTech vaccine and formulated policy recommendations on how best to use it. The vaccine was the first to receive an emergency use validation from WHO for efficacy against COVID-19. Two days later Brazil reached 200,000 deaths from COVID-19, according to data from the state health secretariats.
- On the 14th of January the hospital system in Manaus, the capital of the state of Amazonas, started collapsing from the second wave of COVID-19 and ran out of oxygen.

On the same day the world surpassed two million COVID-19 deaths. Three days later, the Brazilian Health Regulatory Agency unanimously authorized the emergency use of the Corona Vac and Oxford vaccines. The state of São Paulo started vaccination against COVID-19 for health professionals at the University of São Paulo Faculty of Medicine Clinical Hospital. By the second week of February, Brazil reached five million people vaccinated in all 26 states and the Federal District, according to data from the state health secretariats.

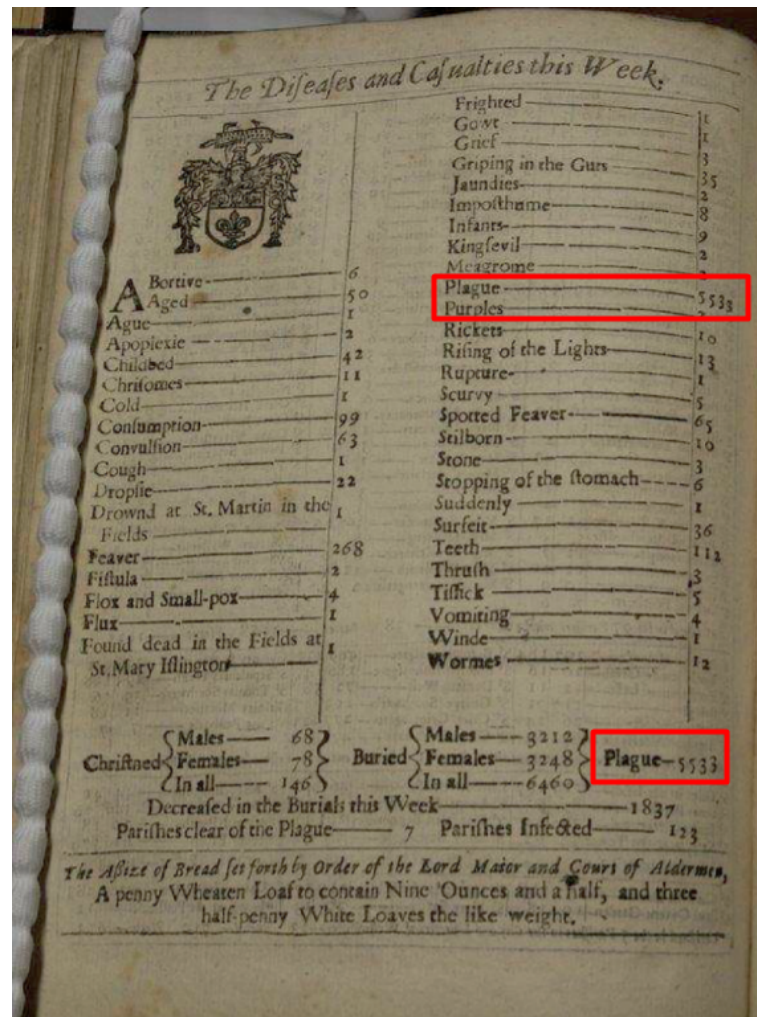
COVID-19 response

The COVID-19 pandemic will impact our society for years to come. Although we have had pandemics with similar characteristics before, the combination of different aspects of this pandemic makes it unique. For example, in the last 150 years we had the Spanish flu (1918–1919) and the Russian flu (1889–1900), which were greater or equal in size, but during those periods the world was less connected than it is today. Also, they were caused by a different virus (influenza), though it should be mentioned that studies have opened the discussion of the origin of the Russian flu from a coronavirus ([Vijgen et al., 2005](#)). Regarding recent events such as Ebola (2013), MERS(2012), H1N1 (2009) and Zika (2015), none had the same impact as COVID-19, even though our planet was intensely globalised.

During a pandemic epidemiologists and biostatisticians need to work with economists and other specialists to help guide public policy and discussions at the highest levels of governments about how to balance the necessity of harsh sanitary measures (e.g., lockdown) and the need to maintain the economy to avoid a financial crisis. The Disease Control Priorities Network comments on the balance between health and the economy, saying that high costs may occur as a result of interventions (such as quarantines and school closures) that lead to economic disruption. These interventions may be more cost-effective during a severe pandemic ([Madhav et al., 2017](#)).

As commented in the timeline above, right from the start of the outbreaks the WHO warned about the importance of non-pharmaceutical interventions (NPIs). These are actions, unlike getting vaccinated or taking medicine, that people and communities can take to help slow the spread of diseases. The efficacy of the implementation of these policies varies greatly between different cultures and governments they should be tailored to each situation and the progression of the disease. Some examples are closing schools and childcare facilities, requiring use of masks in public transport, banning events with a number of people, encouraging home-office, closing bars and cancelling social events. It is important that these policies are implemented early, mainly due to the efficiency of early actions, but also to avoid any catastrophe in the health systems that also have to deal with other issues that will continue to

The Diseases and Casualties this Week.



Disease/Casualty	Count
Frighted	1
Gout	1
Grief	1
Griping in the Guts	3
Jaundies	35
Impoethime	2
Infants	8
Kingevil	9
Meagrome	2
Plague	5538
Purples	1
Rickets	10
Rising of the Lights	13
Rupure	1
Scurvy	5
Spotted Fever	65
Stillborn	10
Stone	3
Stopping of the stomach	6
Suddenly	1
Surfeit	36
Teeth	112
Thrush	3
Tifick	5
Vomiting	4
Winde	1
Wormes	12

Category	Males	Females	In all
Christned	68	78	146
Buried	3212	3248	6460

Decreased in the Burial this Week 1837
Parishes clear of the Plague 7 Parishes Infected 123

*The Assize of Bread set forth by Order of the Lord Mayor and Court of Aldermen,
A penny Wheaten Loaf to contain Nine Ounces and a half, and three
half-penny White Loaves the like weight.*

Figure 1.1.1 – A bill of mortality for the City of London, England, for the week of 26 September to 3 October 1665. This photograph was taken by Claire Lees at the Guildhall in London, England, with the permission of the librarian

occur during the pandemic. Figure 1.1.1 is a bill of mortality for the City of London during the plague, which gives us a morbid reminder of this fact.

To assess if the situation can be considered a severe pandemic, one option is to look at how the numbers of cases are progressing. Even if the disease has a relatively low case fatality rate, e.g., according to Liang *et al.* (2020) that of COVID-19 is around 3.7% , it can have a significant impact in society when the whole population is affected. The main indicator of the progression of the disease is the reproductive number R , which measures how many people an infected person is going to affect during his or her infective cycle, and is a crucial input to policy-making.

This project discuss the current approach used by the Swiss National Covid-19 - Science Task force (NCS-TF) and develops a different method considering both the cases and the

infections, reconstructing the incidence curve in the same algorithm that estimates the reproduction number. The method is then applied to Switzerland and others countries to compare to the NCS-TF approach. Chapter 2 contains a brief literature review of epidemiological models and estimation of the reproduction number. Chapter 3 comments on the data used and discusses the methods applied to the data. The results are discussed in Chapter 4.

CHAPTER 2

LITERATURE REVIEW

2.1 Epidemiological Models

The first known application of a compartmental epidemic model was by Daniel Bernoulli in 1760. His model divided the population into susceptible and immune compartments and assumed an age-specific force of infection and case fatality rate, yielding a system of equations with an endemic equilibrium of susceptible and immune individuals. His motivation was similar to that of disease studies today, to predict the expected gain in life expectancy that would be brought about by applying smallpox control measures. Since variolation¹ was becoming widespread in Europe in the late 1700s, predicting the resulting increase in life expectancy would have been important for pricing annuities ([Allen et al., 2008](#)).

The transmission mechanism from an infective person to a susceptible person is understood for nearly all infectious diseases and the spread of diseases through a chain of infections is known. However, the transmission interactions in a population are very complex, so it is difficult to comprehend the large-scale dynamics of disease spread without the formal structure of a mathematical model. Instead of focusing on every interaction an epidemiological model tries to model the macroscopic behaviour of disease spread through a population ([Levin et al., 2012](#)). In certain communities it is possible to contain the spread of a disease by contact tracing and testing every individual, but this is logistically almost impossible after there is evidence of community transmission;

After Bernoulli, almost no significant work was done in epidemiological modelling for more

¹ Variolation was one of the first methods to intentionally create immunization in an individual. There were several types of procedure but the general idea was to take a scab of smallpox from a recently variolated patient, and by several methods such as exposing it to vapour or drying to create a milder infection that would result in the immunization of the patient. Luckily we live in a post-smallpox world where the last case was registered in 1978 in the United Kingdom.

than 100 years. However in the mid-nineteenth to the early twentieth centuries, the subject was studied by a number of authors who wrote papers on mathematical and statistical models for various types of infectious diseases. At this time it was already suspected that the density of susceptibles was an important quantity in the models. As Hirsch claimed in 1883, “the recurrence of the epidemics of measles at one particular place is connected neither with an unknown something (the mystical number of the Pythagoreans), nor with ‘general constitutional vicissitudes’, as Kostlin thinks; but it depends solely on two factors, the time of importation of the morbid poison, and the number of persons susceptible of it” ([Soper, 1929](#)).

The modelling of diseases during that period was mostly focused on an intriguing question that would arise to anyone that studied a little about disease during that period, “What causes the recurrences of disease?”. There were at least two competing hypotheses regarding the causes of recurrence. Scientists such as Brownlee hypothesized that seasonal recurrence in diseases such as measles was simply due to seasonal variation in pathogen virulence ([Brownlee, 1906](#)). By comparison, scientists such as Hamer and Davidson sought an endogenous explanation for recurrence. They suggested that it is unnecessary to invoke seasonal variation in host or pathogen properties and that this property would arise normally from appropriate modelling ([Hamer, 1906](#)). More specifically, Hamer hypothesized the concept that would be later known as the mass-action mixing assumption, in which the incidence is proportional to the product of the densities of susceptible and infected individuals. This concept is also used in chemical reactions, whose rates depend on the concentrations of the elements involved.

Unlike in chemistry and other sciences, it is impossible and unethical to conduct experiments regarding the spread of infectious diseases in human populations. Data are sometimes available from naturally occurring epidemics or from the natural incidence of endemic diseases; but it is often incomplete due to under-reporting, and the COVID-19 data are not different. This lack of reliable data makes accurate parameter estimation difficult. Since repeatable experiments and accurate data are usually not available in epidemiology, mathematical models and computer simulations are used to perform the needed theoretical experiments. ([Levin et al., 2012](#))

Compartmental models simplify the mathematical modelling of infectious diseases by putting individuals into categories. Every person is assigned to a compartment. In one of the most famous models, the SIR models, individuals receive the labels, S, I, or R:

- Susceptible (S): individuals who have no immunity to the infectious agent, so might become infected if exposed;
- Infectious (I): individuals who are currently infected and can transmit the infection to susceptible individuals whom they contact;

- Removed (R): individuals who are immune to the infection, and consequently do not affect the transmission dynamics in any way when they contact other individuals.

The total host population (N) size is the sum of all the compartments. If the disease is not deadly and ignoring demographics, the population is assumed to be constant, so

$$N = S + I + R.$$

People may progress between compartments, and the order of the labels usually shows the flow patterns between the compartments; for example SEIS means susceptible, exposed, infectious, then susceptible again. The models are most often run with ordinary differential equations (which are deterministic), but can also be used in a stochastic (random) framework. The numbers of individuals in each compartment must be integers, of course, but if the host population size N is sufficiently large we can treat S , I and R as continuous variables and express our model for how they change in terms of a system of differential equations.

An underrecognized value of epidemiological modeling is that it leads to a clear statement of the assumptions about the biological and sociological mechanisms which influence disease spread. The parameters used in an epidemiological model must have a clear interpretation such as a contact rate or a duration of infection. Models can be used to assess many quantitative conjectures. For example, one could check a conjecture that AIDS incidence would decrease if 90% of the sexually active heterosexual population started using condoms consistently. Epidemiological models can sometimes be used to predict the spread or incidence of a disease ([Levin et al., 2012](#)).

2.1.1 SIR Models

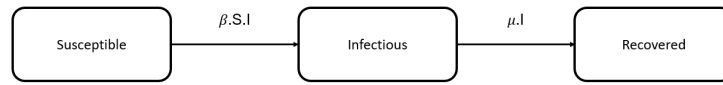


Figure 2.1.1 – Compartments of a SIR model

Having compartmentalised the host population, we need a set of equations that specify how the sizes of the compartments change over time. Solutions of these equations will give $I(t)$, which is the size of the infectious compartment at time t . The quality of the model can be judged by how well a plot of $I(t)$ resembles the real epidemic curve.

A common initial assumption is to not consider “vital dynamics”, i.e., births and deaths. SIR models can be defined by the following ordinary differential equations

$$\frac{dS}{dt} = -\beta SI,$$

$$\frac{dI}{dt} = \beta SI - \mu I,$$

$$\frac{dR}{dt} = \mu I,$$

where

t is a unit of time, commonly days;

S is the number of individuals who have no immunity to the infectious agent, so might become infected if exposed;

I is the number of individuals who are currently infected and can transmit the infection to susceptible individuals whom they contact;

R is the number of individuals who are immune to the infection, and consequently do not affect the transmission dynamics in any way when they contact other individuals;

β is the transmission rate per capita;

μ is the recovery rate.

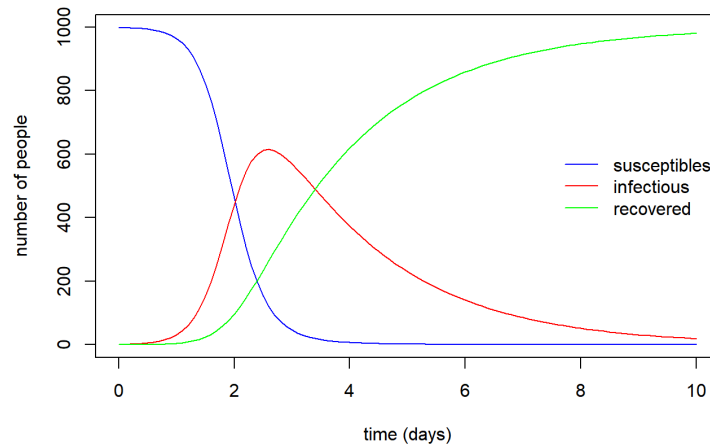


Figure 2.1.2 – Evolution of sizes of compartments in a SIR model in time, starting with 1000 susceptible individuals.

Figure 2.1.2 illustrates one example of how these equations can represent a epidemic with a initial population of 1000 susceptible individuals.

If we expand the SIR model to include B births per unit time and a natural mortality rate γ (per capita) then our equations become

$$\frac{dS}{dt} = B - \beta SI - \mu S,$$

$$\frac{dI}{dt} = \beta SI - \mu I - \gamma I,$$

$$\frac{dR}{dt} = \mu I - \gamma R, \quad t > 0.$$

The necessity to add demographic information to the models depends mostly on the incubation period of the disease and the size of the serial interval. Demographic aspects can be ignored for diseases with a short cycle such as the flu, but for a virus such as HIV that can have a survival time of 11 years on average without treatment it would be best to include this information in the model.

SIS model

Another common introductory model is the SIS epidemic model, in which a susceptible individual, after a contact with an infectious individual, becomes infected and infectious, but does not develop immunity. Thus, infected individuals return to the susceptible class after recovery. A obvious assumption is that there are no disease-related deaths. The compartmental diagram is also very simple; see Figure 2.1.3

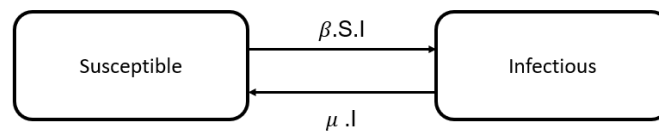


Figure 2.1.3 – Compartments of a SIS Model

This model can be applied to diseases such as the common cold, depending on the number of deaths in the season. If we want to include vital dynamics in the model, the equations become

$$\frac{dS}{dt} = B(N) - \beta(N)SI - \gamma S + f\alpha I,$$

$$\frac{dI}{dt} = \beta(N)SI - \gamma I - \mu I, \quad t > 0,$$

where

f is the fraction of infectives recovering with no immunity against reinfection;

α is the rate of recovery from infection;

B is births as a function of the number of individuals.

2.1.2 SEIR Models

The SIR model was a initial step for the development of epidemiological models. After that initial idea several other models were derived to be more specific to certain diseases. For example, some diseases have a significant time between contact with the virus and the infection itself. With this observation in the real world a model with a period considering the individual as exposed (E) between susceptible and infectious would be appropriate. Such models are called SEIR models. Figure 2.1.4 illustrates the different stages.

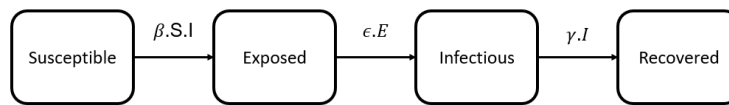


Figure 2.1.4 – Compartments of a SEIR Model

For the SEIR compartmental model the set of differential equations is

$$\frac{dS}{dt} = -\beta SI,$$

$$\frac{dE}{dt} = \beta SI - \epsilon E,$$

$$\frac{dI}{dt} = \epsilon E - \mu I,$$

$$\frac{dR}{dt} = \mu I, \quad t > 0.$$

2.1.3 Tailor-made models

Considering how diversely a disease can manifest itself and spread, it is not uncommon to develop tailor-made models for a specific important case. One example of those models for COVID-19 can be seen in the paper “Extensions of the SEIR model for the analysis of tailored social distancing and tracing approaches to cope with COVID-19” from [Grimm et al. \(2021\)](#). In this paper they add compartments to the SEIR model in order to better differentiate the individuals in the population. These compartments allow them to incorporate different parameters that represent important aspects about COVID-19 in a individual immunological response, such as young, old, vulnerable, non-vulnerable, recovered, dead, asymptomatic, symptomatic and severe cases. Thus, transforming the SEIR model into what they named SEI³RD, resulted in the equations

$$\begin{aligned}\frac{dS_k}{dt} &= - \sum_{l=1}^k (\beta_{lk}^{asym} I_l^{asym} + \beta_{lk}^{sym} I_l^{sym} + \beta_{lk}^{sev} I_l^{sev}) S_k, \\ \frac{dE_k}{dt} &= - \sum_{l=1}^k (\beta_{lk}^{asym} I_l^{asym} + \beta_{lk}^{sym} I_l^{sym} + \beta_{lk}^{sev} I_l^{sev}) S_k - \epsilon_k E_k, \\ \frac{dI_k^{asym}}{dt} &= \eta_k \epsilon_k E_k - \gamma^{asym} I_k^{asym}, \\ \frac{dI_k^{sym}}{dt} &= (1 - \eta_k)(1 - \nu_k) \epsilon_k E_k - \gamma^{sym} I_k^{sym}, \\ \frac{dI_k^{sev}}{dt} &= (1 - \eta_k) \nu_k \epsilon_k E_k - \left((1 - \sigma_k(t)) \gamma_k^{sev-r} + \sigma_k(t) \gamma_k^{sev-d} \right) I_k^{sev}, \\ \frac{dR_k}{dt} &= \gamma^{asym} I_k^{asym} + \gamma^{sym} I_k^{sym} + (1 - \sigma_k(t)) \gamma_k^{sev-r} I_k^{sev}, \\ \frac{dD_k}{dt} &= \sigma_k(t) \gamma_k^{sev-d} I_k^{sev}, \quad k = 1, \dots, K,\end{aligned}$$

where

K = number of groups;

N_k = total number of individuals in group k ;

S_k = susceptible individuals in group k ;

E_k = exposed individuals in the latent period in group k ;

I_k^{asym} = asymptomatic infectious individuals in group k ;

I_k^{sym} = symptomatic infectious individuals in group k ;

I_k^{sev} = severely symptomatic infectious individuals in group k ;

R_k = recovered individuals with immunity in group k ;

D_k = dead individuals in group k ;

η_k = fraction of asymptomatic infectious individuals;

ν_k = fraction (of I_{symk}) of severely symptomatic infectious individuals;

σ_k = lethality rate conditional on severe infection;

$\beta_{kj}^{asym}, \beta_{kj}^{sym}, \beta_{kj}^{sev}$ = group-specific infection rates.

At first glance these ordinary differential equations look complex, but they do not differ from the initial concept of compartmentalising the population into groups and estimate the infectious period, recovery time and other parameters that are present in the original SEIR model. This model allowed the group to simulate different scenarios of intervention measurements and to estimate important information for policymakers, for example, the probable number of deaths and required ICU capacity.

2.2 Reproductive number

The *reproduction number* R_0 is a key epidemiological variable to guide decisions during a pandemic. If a compartmental model is used it can be obtained from the set of ordinary equations. Considering a SIR compartmental model and supposing that at $t = 0$ the number of susceptibles is equal to the total population $S = N$, then the first infected person introduced to the system will be expected to infect other individual at the rate βN during the expected infectious period, $1/\gamma$. Thus, we obtain

$$R_{0_{SIR}} = \frac{\beta N}{\gamma}.$$

When referring to the reproduction number is important to make a distinction between the basic reproduction number and the effective reproduction number. The latter can be defined as the instantaneous reproductive number or as the case reproductive number (Gostic *et al.*, 2020). Those different quantities may look similar but conceptually they are significantly different.

The basic reproduction number, R_0 , is defined as the expected number of secondary cases produced by a typical primary case in an entirely susceptible population (Dietz, 1993). This is an epidemiological metric used to describe the contagiousness or transmissibility of infectious agents. It is affected by numerous biological, sociobehavioral and environmental factors that govern pathogen transmission and, therefore, is usually estimated with various types of complex mathematical models, which make R_0 easily misrepresented, misinterpreted, and misapplied (Delamater *et al.*, 2019).

When infection is spreading through a population, it is often more convenient to work with the effective reproduction number R , which is defined as the actual average number of secondary cases per primary case. This is typically smaller than R_0 , and it reflects the impact of control measures and depletion of susceptible persons during the epidemic. If R exceeds 1, the number of cases will inevitably increase over time, and a large epidemic is possible. To stop an epidemic, R needs to be persistently below 1 (Wallinga and Teunis, 2004).

The case reproductive number is useful for retrospective analyses of how individuals infected at different time points contributed to spreading the disease. The instantaneous reproductive number is more appropriate for estimating the reproductive number of the infected population on specific dates, especially when aiming to study how interventions or other extrinsic factors have affected transmission (Gostic *et al.*, 2020). One useful analogy is to think about the instantaneous reproductive number as the life expectancy when a person is born, and the case reproductive number as the number of years this person will eventually live.

2.2.1 Instantaneous reproductive number

The instantaneous reproductive number involves fewer assumptions about the future than does the case reproduction number, making it more appropriate to real-time estimation. The method from [Cori et al. \(2013\)](#) has been applied in several epidemics in which R_t is estimated as

$$R_t = \frac{I_t}{\sum_{s=1}^t I_{t-s} w_s},$$

where I_t is the number of new infections on day t and w_s is the generation interval. This estimator describes the ratio between the number of new infections on a day relative to the number of infections and the infectivity profile of the cases on previous days. The usual assumption is that w_s is given by a discretized gamma distribution.

2.2.2 Case reproductive number

The case reproductive number, sometimes called the cohort reproductive number, is the expected number of secondary infections that an individual will eventually cause. The proposed method by [Wallinga and Teunis \(2004\)](#) is to calculate the likelihood that a case j was infected by case i relative to the likelihood that i was infected by other cases. Using pairs of cases was the innovative idea on their approach. They obtained the ratio

$$p_{ij} = \frac{w(t_i - t_j)}{\sum_{i \neq k} w(t_i - t_k)},$$

so the individual reproductive number of the case j is

$$R_j = \sum_i p_{ij}.$$

They also assume that the generation interval follows a discretised gamma distribution.

2.3 Serial interval and generation interval distributions

In most methods for estimating R , one necessary input is information regarding the infectiousness of the disease, normally in the form of the generation interval or serial interval. The generation interval is the time between infection of the host and that of a second case. This applies to both clinical cases and unidentified infections. With person-to-person transmission of infection, the interval between cases is determined by the generation time. The serial interval is the period of time between analogous phases of an infectious illness, in successive cases of a chain of infection that is spread from person to person ([Porta, 2014](#)).

As [Svensson \(2007\)](#) points out, other terms are also used, such as transmission time or transmission interval. Although the term generation interval is frequently used, it is more common to observe the serial interval, since it is easier to identify symptom onset and hospitalisation than the time of infection ([Kenah et al., 2008](#)).

Misspecification of the generation interval is a large potential source of over- or under-estimation, and estimates of R_t are most prone to this kind of bias when the true value is substantially greater or less than one ([Gostic et al., 2020](#)).

2.4 Reconstruction of the infection incidence curve

One problem in dealing with real-world data is the reconstruction of the incidence curve. Methods for R_t estimation are based on the knowledge of the epidemic curve, which in general is unobserved.

The data generally available concern the symptom, report, hospitalisation or death curves, which do not represent our input of interest (the incidence curve) because the infections are blurred in time owing to variation amongst individuals. For example, supposing symptoms of a certain disease takes 5 days on average to manifest, not all individuals infected at time t will be identifiable by their symptoms at time $t + 5$ and thus our symptom curve will be a smoothed representation of the incidence curve. For some infections (e.g., HIV), diagnostic symptoms (i.e., AIDS-defining illness) may occur years after infection, so the symptom curve is a poor reflection of the evolution of the epidemic ([Goldstein et al., 2009](#)).

Naive approaches for dealing with observation delays, such as subtracting delays sampled from a distribution, can introduce bias ([Gostic et al., 2020](#)). The most recommended method to obtain the epidemic curve is that of [Goldstein et al. \(2009\)](#), who applied the Richardson–Lucy deconvolution to the influenza epidemic of 1918. In real-world data it is important to understand that the existence of super-spreaders can generate anomalies in the initial days of the epidemic ([Wallinga and Teunis, 2004](#)).

2.5 Swiss - COVID19 - Discussion

Switzerland is a unique country in many aspects. It is located in the centre of Europe but is not part of the European Union, has three official languages, has 25% foreigners in its population and several other characteristics that make it unique. It has borders with Liechtenstein, France, Germany, Austria and Italy, which was one of the European countries most impacted by the first wave of COVID-19. Due to cross-border commuters and its dependence on workers from France and Italy, a total closure of its borders was not an option. But nevertheless, several public health measures and restrictions were implemented in the first wave. These measures and their timing can be seen in Figure 2.5.1 .

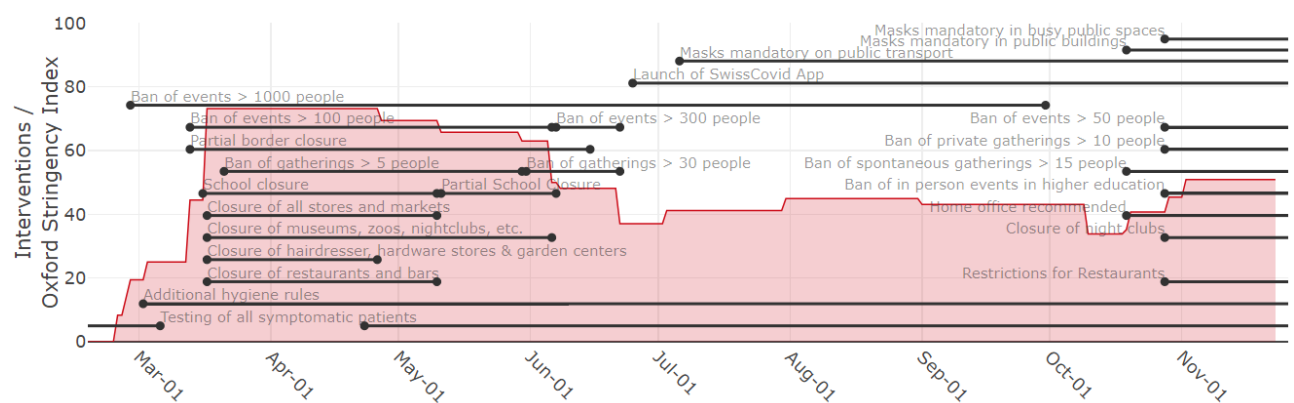


Figure 2.5.1 – Implementation and removal of COVID-19 restrictions in Switzerland by health authorities. Source:Swiss National Covid-19 - Science Task Force (<https://ncs-tf.ch/en/situation-report>). Accessed:15/11/2020

Even with restrictions and campaigns to inform the public, Switzerland was not spared from the first wave of the COVID-19, though it had a better pandemic response than most of its neighbours. The second wave in October was more severe, having days with more than 10,000 new COVID-19 cases, as can be seen in Figure 2.5.2.

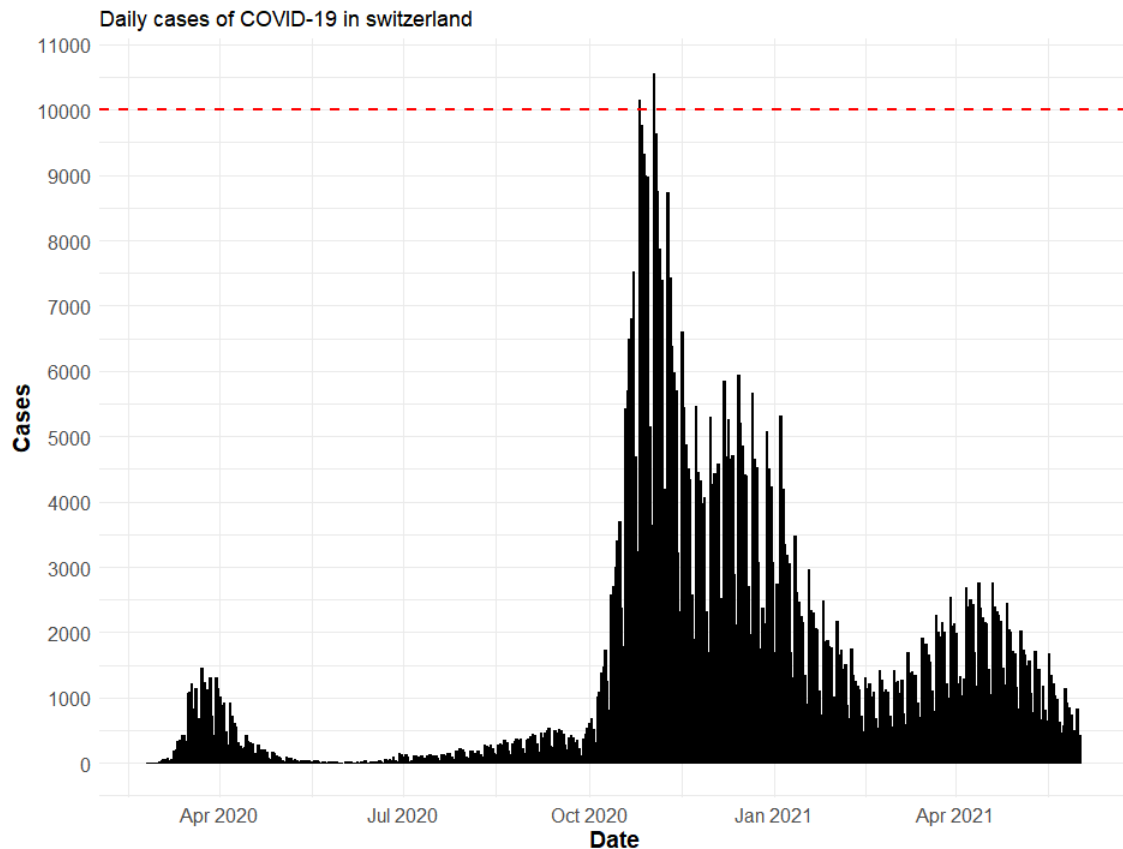


Figure 2.5.2 – Number of daily confirmed cases of COVID-19 in Switzerland

The NCS-TF is responsible for guiding the government in several aspects of the pandemic, such as testing and the progression of the disease. The data modelling group generates daily updates of the R_t estimates which can be seen in Figure 2.5.3. Their updates have been a fundamental part of the government's response and decisions about the implementation of restrictions. However, one thing that catches the eye at first glance of their R_t estimation is the huge confidence intervals, which will be one of the topics discussed in this project.

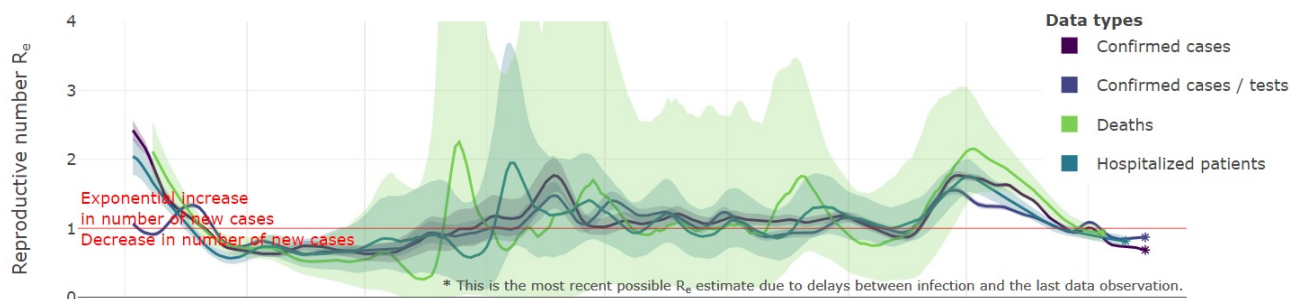


Figure 2.5.3 – R estimates for Switzerland by the NCS-TF for 2020 based on reported cases, deaths and hospitalizations. Source: Swiss National Covid-19 - Science Task Force (<https://ncs-tf.ch/en/situation-report>) Accessed: 15/11/2020

CHAPTER 3

METHODS

3.1 Data

When a new disease starts, the research and medical community has to put in place infrastructure that allows epidemiologists to estimate the reproduction number. For example, at the moment there are two tests that are used to identify cases, the rapid diagnostic test (RDT) which is colloquially called "the antibody test" and the PCR test that detects the presence of the RNA of the virus. Having such tests widely available in a small time frame is already a tremendous achievement for the scientific community. To achieve this different aspects of the disease are studied in order to develop each test with its specific characteristics.

RT-PCR test

The RT-PCR test (Reverse transcription polymerase chain reaction) detects the presence of viral RNA. It is known to have be very accurate and efficient. Although the this test can be costly compared to the antigen test, it is able to detect a COVID-19 infection even before the person becomes infectious and will allow early isolation. Thus, this method is able to prevent the transmission of the virus to other hosts. There is also the discriminant PCR which is performed on positive samples and allows to determine whether it is a specific mutation for which the discriminant test is designed.

Rapid antigen test

The rapid antigen test is cheaper than the PCR test but it is also not as accurate because it need high concentrations of the proteins during the infection to detect it. This test does not require specialised staff and can give results within 30 minutes. Another disadvantage is that a significant percentage of those infected pass the test as a false negative. In a few days, these people will spread the virus among others, thinking they are healthy.

Antibody test

This test measures antibodies to the SARS-CoV-2 virus in the bodies of people who have already had COVID-19 or are successfully recovering from the disease. Antibodies are not present at the onset of the disease.

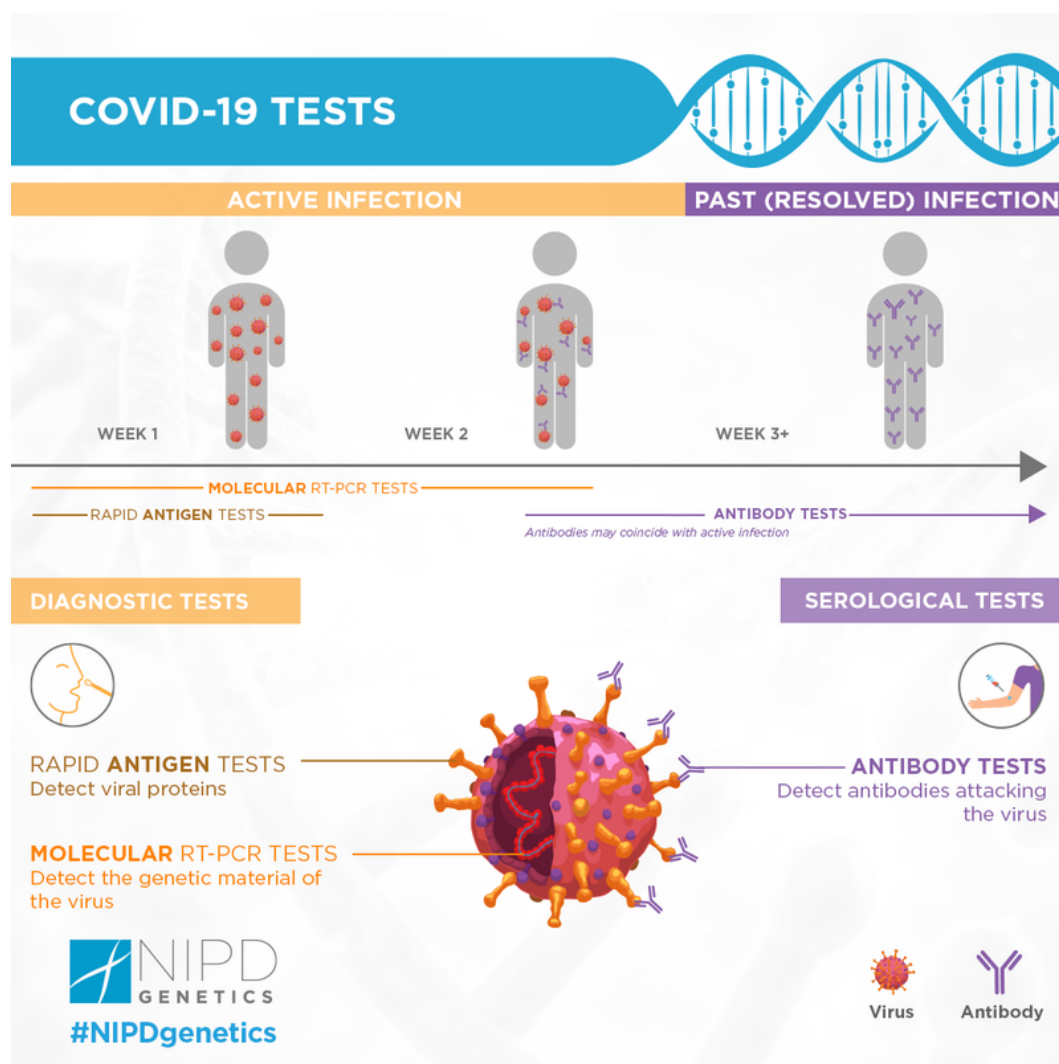


Figure 3.1.1 – Different types of COVID-9 tests Source: NIPD Genetics

After a reliable method for identification of the disease has been put in place it is possible to have a time series of confirmed infections. In some diseases if the symptoms are evident enough it is possible to create this time series for confirmed infections, hospitalisations and deaths using only medical observation.

3.1.1 Swiss data

The data for Switzerland are obtained from the Federal Office of Public Health (FOPH), which collects the data from several cantonal authorities and compiles them to know the total

cases in Switzerland which will be later used by the Swiss Covid-19 task force. The time series of cases and deaths can be seen in Figure 3.1.2. Before doing any estimation or statistical analysis there is information we can obtain just by looking at the graph.

We can see that Switzerland had a first wave, and after public authorities took action the cases got under control. There is also the possibility that testing was not so widespread in the first wave as in the second wave, which caused the first wave to be considerably smaller than the first, but this is just a initial guess from the time series. Countries may have different patterns in the time series of infections, and in Switzerland it is very clear that the country has a weekend pattern.

During the first wave doctors had to learn to treat patients with COVID-19, so the fatality rate in later months of the pandemic should be lower if everything else stays the same, and this may explain why the peaks for deaths are similar in Figure 3.1.2 but the cases are very different. However, we also have to consider that in the first wave testing was not as widespread as in the second and there is a possibility that Switzerland was not identifying as many cases as it was during the second wave. This illustrates how hard it is to understand the underlying process that are resulting in these variations of cases and deaths during the pandemic only looking at cases and deaths. The same results can be caused by different phenomena.

3.1.2 Brazilian data

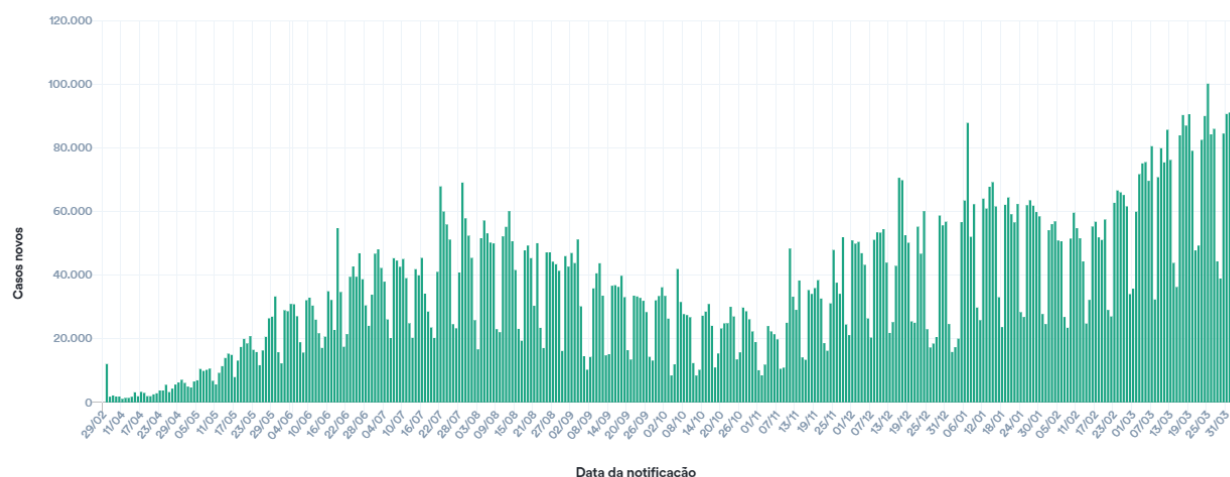


Figure 3.1.3 – – Number of cases and death for the 1st and 2nd wave of COVID-19 in Brazil
Source: <https://covid.saude.gov.br/>. Accessed:06/04/2021

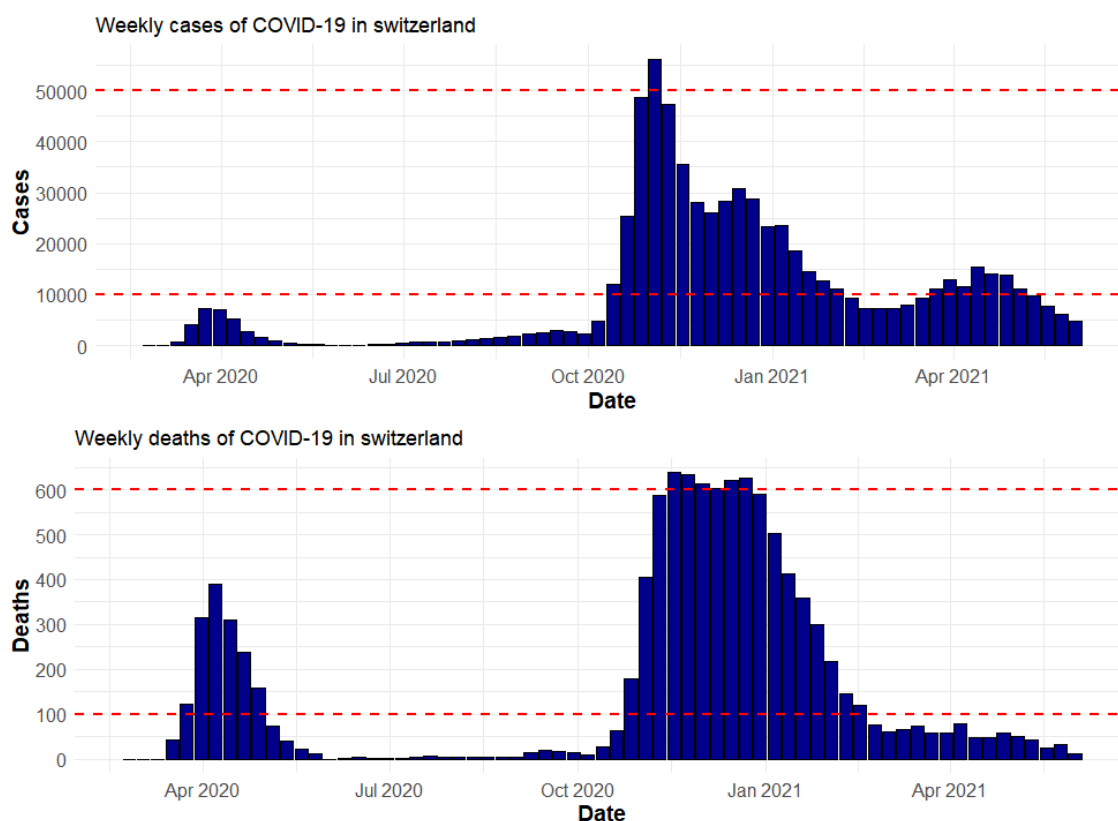


Figure 3.1.2 – Number of cases and deaths for COVID-19 in Switzerland based on data reported from the Federal Office of Public Health

Brazil is a particular country when we refer to the pandemic; it has had four health ministers. The initial handling by Luiz Henrique Mandetta saw national cohesion of the fight against COVID-19, but after he was fired most of the work was left to the states and the country lacked a centralised organ to deal with the epidemic. The data that we are using is from the website "<https://covid.saude.gov.br/>" which gathers the information from the 26 states and compiles it every day, updating it at 7PM (GMT-3). This is the current official source, however due to the lack of centralised governance during the handling of the pandemic the media created a group to gather the data from different states and publish it every day at 8PM. The mainstream media groups participating are the G1, O Globo, Extra, Estadão, Folha and UOL, who cite the constant attacks on the media from the current president Jair Bolsonaro as the main reason for the necessity of an independent count of the COVID-19 cases.

Unlike Switzerland, in Brazil the difference between the first and the second wave is not so clear. This can be due to several reasons: one possibility is how the virus dispersed in the country, so that when some states were starting to reduce the number of cases others were just being introduced to the virus. Another point is that the case time series definitely have a weekend pattern but this is not specific to Brazil.

3.1.3 Other countries

The method developed here was also applied to other countries, but we did not use the data directly from official state sources, but from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). In the beginning of the pandemic their dashboard was used widely, but it is be useful to check the data source if you want to analyse specific countries. For example, there is a clear difference in the data from JHU and the FOPH if we look at Switzerland: the first has no cases reported at the weekends.

3.2 Estimation Methods

Several approaches for the estimation of R_t have been developed, and due to the complexity surrounding this epidemiological problem it is important to understand the underlying assumptions when applying those different methods. For example, [Bettencourt and Ribeiro \(2008\)](#) derived a method based on a SIR model which assumes that the generation interval follows a exponential distribution, but this is not the case with COVID-19 and several other diseases. This assumption was one of the main reasons reported by [Gostic et al. \(2020\)](#) for bias in their estimation of R_t , though they point out that if adapted this method can produce smoother estimates than Cori's method due to the penalisation of jumps in R_t .

Another method frequently used is from [Wallinga and Teunis \(2004\)](#), who showed that the relation between the reproduction number and the epidemic curve is determined by the generation interval. However, this method is not recommended for real-time estimation since it requires incidence data from times later than the moment it is trying to estimate.

The method most applied for real-time estimation of the effective reproduction number, comes from the paper "A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics". [Gostic et al. \(2020\)](#) concluded in their analyses that this is the best method for near real-time estimation.

This method starts from the assumption that the distribution of infectiousness is independent of calendar time and models transmission with a Poisson process. The number of infections on day t depends on the infections in previous days, the reproductive number and the infectiousness profile, which is generally assume to be gamma distributed. So the number of new cases x_{t+1} on day $t + 1$ given the past is distributed as

$$x_{t+1} \sim \text{Pois}\left(R_t \sum_{s=0}^t w_s x_{t-s}\right).$$

Also, the conditional distribution of the incidence I_t at time t is

$$P(I_t \mid I_0, \dots, I_{t-1}, w, R_t) = \frac{\left(R_t \sum_{s=0}^t w_s x_{t-s}\right)^{I_t} e^{-R_t \sum_{s=0}^t w_s x_{t-s}}}{I_t!}.$$

The next step is to assume that the transmissibility is constant over a period $[t - \tau + 1; t]$, so the likelihood of incidence during this period is

$$P(I_{t-\tau+1}, \dots, I_t \mid I_0, \dots, I_{t-\tau}, w, R_{t,\tau}) = \prod_{s=t-\tau+1}^t \frac{\left(R_t \sum_{s=0}^t w_s x_{t-s}\right)^{I_t} \exp\left(-R_t \sum_{s=0}^t w_s x_{t-s}\right)}{I_t!}.$$

With the equation above and using a Bayesian framework we obtain the following posterior distribution, assuming for the prior $P(R)$ a gamma distribution,

$$P(I_{t-\tau+1}, \dots, I_t, R_{t,\tau} \mid I_0, \dots, I_{t-\tau}, w) = S_1 \exp(S_2) S_3,$$

in which to ease the notation we used the terms S_1, S_2, S_3 that replaced

$$S_1 = R_{t,\tau}^{\left(a + \sum_{s=t-\tau+1}^t I_s - 1\right)},$$

$$S_2 = -R_{t,\tau} \left(\sum_{s=t-\tau+1}^t \Lambda_s + \frac{1}{b} \right),$$

$$S_3 = \prod_{s=t-\tau+1}^t \frac{\Lambda_s^{I_s}}{I_s! \Gamma(a) b^a}.$$

The method above is used by the Swiss National Covid-19 Task Force to estimate the values of R_t for Switzerland, using the standard values for the prior gamma distribution in the package EpiEstim (Mean = 5, Standard deviation = 5). Their results can be found in [Sciré et al. \(2020\)](#).

3.3 Project Objectives

Considering the impact of the COVID-19 and the possibility of new pandemics in the future, works to improve the estimation of the effective reproduction to orient governmental decisions are in high demand. Aligned with this necessity it may be possible to improve the confidence intervals used by the Swiss task force. Our proposed method differ from the Task force because it considers the weekly patterns and the reconstruction of the incidence curve in

the same Metropolis-Hastings algorithm. Thus, it may be possible to have a better estimation, considering that we are estimating everything at once (R_t , weekly patterns and infections) while the Task Force does these steps separately.

The goal of this project is to estimate the effective reproduction number of COVID-19 in Switzerland and compare to the method from the NCS-TF. For this to be achieved we present some of the methodologies in use and apply our method that will be discussed further to simulated and real data. This project is part of a larger group project at the EPFL. The project included six master's students and was supervised by Professor Anthony C. Davison and Hélène Ruffieux. My studies were funded for 10 months by EPFL and I stayed in Lausanne working in the Department of Mathematics with the Chair of Statistics. There were weekly meetings on Mondays in which the students presented important papers and the results of different approaches for the estimation of the reproductive number. These approaches included a frequentist method using Generalized Additive Models (GAMs) and other Bayesian methods using the software STAN.

3.4 Project's Approach

3.4.1 Estimation of R_t

Considering Y_t as the number of cases at day t , which is a fraction of the number of infections X_t on the same day and assuming a Poisson distribution for both, the number of cases and infections on day $t + 1$ depends on the number of cases in the previous days and is

$$x_{t+1} \mid x_1, \dots, x_t \sim \text{Pois}(\Delta_t),$$

$$y_{t+1} \mid y_1, \dots, y_t \sim \text{Pois}(\mu_t),$$

where

R_t is the reproduction number;

w_s is a discretized gamma distribution with mean = 5.3 and sd = 3.2;

$$\Delta_t = R_{t-1} \sum_{s=0}^{t-1} w_s x_{t-s-1};$$

$$\mu_t = \sum_{s=0}^{t-1} a_s x_{t-s-1}.$$

The values for the distribution were taken from [Linton et al. \(2020\)](#), who studied the distribution for COVID-19 in the early pandemic.

The joint distribution of x, y for a fixed R is

$$f(x, y | R) = \prod_{t=2}^n \left(\frac{(\Delta_{t-1})^{x_t} e^{-\Delta_{t-1}}}{x_t!} \frac{(\mu_{t-1})^{y_t} e^{-\mu_{t-1}}}{y_t!} \right).$$

The joint distribution for Y and X is used as part of Bayes Theorem to obtain the distribution for $P(R | y, x)$,

$$P(R|y, x) = \frac{P(y, x|R)P(R)}{P(y, x)} \propto P(y, x|R)P(R).$$

In this Bayesian framework we still lack a prior distribution for R . Considering the nature of the problem at hand we do not expect that the reproduction number will vary wildly in a brief period of time nor do we expect any discontinuity. Thus, one of the desired characteristics of our distribution for R is dependence between neighbouring days, which we will impose with our prior. This will be achieved by writing R_t as a combination of splines and coefficients to be estimated. Thus, a brief discussion of splines is necessary. For more details on splines the source material for this subsection are the books “Generalized Additive Models: an introduction with R” from [Wood \(2017\)](#) and “Splines and PDEs: From approximation theory to numerical linear algebra” from [Kunoth et al. \(2018\)](#).

3.4.2 Splines

A spline is a function defined piecewise by polynomials. In interpolating problems, spline interpolation is often preferred to polynomial interpolation because it yields similar results, even when using low degree polynomials. Also the ease with which splines can be stored and evaluated on a computer makes them powerful for a variety of applications. In general, a function defined on an interval $[a, b]$ is defined as a polynomial spline of degree k , having knots x_1, \dots, x_n , if the following three conditions hold:

1. $a < x_1 < \dots < x_n < b$, so the knots x_1, \dots, x_n partition the interval $[a, b]$ into $n + 1$ smaller subintervals;
2. in each subinterval $[x_i, x_{i+1}]$, the spline is given by a polynomial function of at most degree k ; and
3. the spline and its derivatives up to order $k - 1$ are all continuous on $[a, b]$.

The points at which the sections join are known as the knots of the spline. Typically the knots would either be evenly spaced through the range of observed x values, or placed at quantiles of the distribution of unique x values.

Cubic Splines

Consider a set of points $\{x_i, y_i : i = 1, \dots, n\}$ where $x_i < x_{i+1}$. The cubic spline, $g(x)$, interpolating these points, is a function made up of sections of cubic polynomial, one for each $[x_i, x_{i+1}]$, which are joined together so that the whole spline is continuous to its second derivative, while $g(x_i) = y_i$ and $f''(x_1) = f''(x_n) = 0$. The spline that has zero second derivatives at the end knots is a “natural spline”.

B- Splines

To efficiently deal with splines, one needs a suitable basis for their representation. B-splines stands out as one of the most useful spline basis functions. Any spline $g(x)$ of degree k can be written as a linear combination of B-splines $B_i(x)$:

$$g(x) = \sum_i a_i B_i(x),$$

where each B-spline $B_i(x)$ is a spline of degree k . B-splines permit the efficient evaluation of a spline and its derivatives because they have local support, in other words, outside a small range, they take the value of zero. To define a k -parameter B-spline basis, we need to define $k + n + 1$ knots, $x_1 < x_2 < \dots < x_{k+n+1}$, where the interval over which the spline is to be evaluated lies within $[x_{n+2}, x_k]$ (so that the first and last $n + 1$ knot locations are essentially arbitrary). An $(n + 1)$ th order spline can then be represented as

$$g(x) = \sum_{i=1}^k B_i^m(x) \beta_i,$$

where the B-spline basis functions are most conveniently defined recursively as follows,

$$B_i^m(x) = \frac{x - x_i}{x_{i+m+1} - x_i} B_i^{m-1}(x) + \frac{x_i + m + 2 - x}{x_{i+m+2} - x_{i+1}} B_{i+1}^{m-1}(x), \quad i = 1, \dots, k,$$

and

$$B_i^{-1}(x) = \begin{cases} 1, & x_i \leq x < x_{i+1}, \\ 0, & \text{otherwise.} \end{cases}$$

With this we can return to our problem of creating a dependence structure in the reproductive number. We can then use a B-spline basis and estimate the coefficients associated with this basis. Let M be a joint matrix with the elements of the first column equal to one and the other columns filled by the spline basis with degree $q - 1$,

$$M_{n,q} = \begin{pmatrix} 1 & B_1 & \dots & B_q \\ 1 & B_1 & \dots & B_q \\ \vdots & \vdots & \ddots & \vdots \\ 1 & B_1 & \dots & B_q \end{pmatrix}.$$

Also, we define γ as a single column matrix with length equals to the number of days in our data, in which its element are the β coefficients to be estimated

$$\gamma_{q,1} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_q \end{pmatrix}.$$

Thus, we write the reproduction number as the product $R = M\gamma$. For the splines basis we used equally spaced splines and varied the number of them to analyse how well they can capture the variation between days.

3.4.3 Weekly Pattern

There is still one important aspect to add to the model, which is how to deal with the weekly patterns that frequently appear. For this we use the fact that the cases y_t reported on day t are actually a combination of cases from previous days. Furthermore, not all cases on day t will be reported on this day. Thus, for each day of the week we have a vector of delay probabilities P^{day} that will distribute the cases of this day to the following days. For example, assuming that 50% of the cases occur on Sunday will be reported on Monday, 20% on Tuesday, 20% on Wednesday and the rest 10% on Sunday it self, the vector of delay probabilities for Sunday would be

$$P^{Sunday} = (0.1, 0.5, 0.2, 0.2, 0.0, 0.0, 0.0).$$

We then have a 7×7 matrix that maps the reporting delays from every day of the week to another. Some elements of the matrix are expected to be 0, because it would be unlikely that a case on Monday will be reported on Sunday. To illustrate this the following matrix is an example of a 1-day delay pattern.

$$\begin{pmatrix} M & Tu & W & Th & F & Sa & Su \\ \begin{matrix} 1 - P^{M \rightarrow Tu} & P^{M \rightarrow Tu} & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 - P^{Tu \rightarrow W} & P^{Tu \rightarrow W} & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 - P^{W \rightarrow Th} & P^{W \rightarrow Th} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 - P^{Th \rightarrow F} & P^{Th \rightarrow F} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 - P^{F \rightarrow Sa} & P^{F \rightarrow Sa} & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 - P^{Sa \rightarrow Su} & P^{Sa \rightarrow Su} \\ P^{Su \rightarrow M} & 0 & 0 & 0 & 0 & 0 & 1 - P^{Su \rightarrow M} \end{matrix} \end{pmatrix}$$

Using this we can rewrite the average of the Poisson distribution for Y_t as a linear combination of previous days,

$$\mu_{t+1} = \sum_{d=1}^D \mu'_{t+1-d} P_d^{t+1-d}.$$

Thus, we use this weighted averaged of reported cases in our algorithm accounting for reporting delays.

3.5 Computational Methods

3.5.1 Markov Chain Monte Carlo (MCMC)

Markov Chain Monte Carlo (MCMC) methods are a class of algorithms that allow you to sample from a probability distribution. The idea is that you can construct a Markov chain that has the target sampled distribution as its equilibrium. Ideally you can start from a point in the chain and take random steps that can be accepted or not depending on the target distribution and as you take more steps you get closer to the target distribution. Next we present a few important concepts to know before working with MCMC methods, this topic was based on the books “Monte Carlo Statistical Methods” by [Robert and Casella \(2013\)](#) and “Markov Chain Monte Carlo in Practice” by [Gilks et al. \(1996\)](#), which can be consulted for more explanation.

3.5.1.1 Definitions

To talk about MCMC it is important to have some definitions regarding a few properties of the Markov Chain X_n .

Transition Kernel - A transition kernel is a function $K(x, A)$ that for $x \in \chi$ and the state-space S belonging to the Borel set of χ in other words, $S \in \mathcal{B}(\chi)$, satisfies

1. For all $x \in \chi$, $K(x, \cdot)$ is a probability measure;
2. For all $A \in \mathcal{B}(\chi)$, $K(\cdot, A)$ is measurable,

where $\mathcal{B}(\chi)$ represents the Borel sets of χ . In the cases where χ is discrete the kernel is a transition matrix K with elements

$$P_{xy} = P(X_n = y \mid X_{n-1} = x), \quad x, y \in \chi.$$

Markov chain - Given a transition kernel K , a sequence X_0, X_1, \dots, X_n of random variables are a Markov chain, denoted by X_n , if, for any n , the conditional distribution of X_n given $X_{n-1}, X_{n-2}, \dots, X_0$ is the same as the distribution of X_n given X_{n-1} . This means that the probability of the next element on the sequence depends only on the current point of the sequence of random variables, or in a more formal statement,

$$P(X_{n+1} \in A \mid x_0, x_1, \dots, x_n) = P(X_{n+1} \in A \mid x_n) = \int_A K(x_n, dx).$$

The chain is time-homogeneous if the distribution of $(X_{n_1}, \dots, X_{n_k})$ is the same as the distribution of $(X_{n_1-n_0}, X_{n_2-n_0}, \dots, X_{n_k-n_0})$ for any n_0 .

Irreducibility - In the discrete case, the chain is irreducible if all states communicate, namely if

$$P_x(\tau_y < \infty) > 0, \quad x, y \in \mathcal{X},$$

τ_y being the first time y is visited. Irreducibility is important because it tells us if the Markov chain is sensitive to the initial conditions and it guarantees convergence.

Transience and recurrence

Although irreducibility guarantees that every set A will be visited by the Markov chain X_n , this property is too weak to tell us how often the trajectory of X_n will enter A . In a finite state-space S , a state $\omega \in S$ is transient if the average number of visits to ω , $\mathbb{E}[\nu_\omega]$ is finite, and recurrent if $\mathbb{E}[\nu_\omega] = \infty$.

In the discrete case, the recurrence of a state guarantees its return. For irreducible chains, recurrence and transience are properties of the chain, not of a particular state.

Reversibility - A Markov chain is reversible if the direction of time has no effect on its dynamics. In other words X_{t+1} conditional on $X_{t+2} = x$ has the same distribution as X_{t+1} conditional on $X_t = x$.

Ergodic - A state ω is said to be ergodic if it is aperiodic and positive recurrent. This means that we are certain to revisit this state in the future, which guarantees convergence to the desired distribution. If all states in an irreducible Markov chain are ergodic, then the chain is said to be ergodic.

Markov Chain Monte Carlo - A Markov chain Monte Carlo (MCMC) method for the simulation of a distribution f is any method producing an ergodic Markov chain X_n whose stationary distribution is f .

3.5.2 Metropolis–Hastings Algorithm

There are several Markov Chain Monte Carlo methods commonly used in statistics, such as, Hamiltonian Monte Carlo (HMC), slice sampling and the Metropolis–Hastings (MH) algorithm that was used in this project. The Metropolis–Hasting algorithm is a type of Markov Chain Monte Carlo method that is used to sample from distributions that are hard to sample. The MH algorithm is normally used for multi-dimensional distributions, while for single dimensions there are simpler algorithms.

The algorithm starts with the target density f . A conditional density $q(y | x)$, defined with respect to the dominating measure for the model, is then chosen. The Metropolis–Hastings algorithm can be implemented in practice when $q(u | x)$ is easy to simulate from and is either explicitly available (up to a multiplicative constant independent of x) or symmetric; that is, such that $q(x | y) = q(y | x)$. The target density f must be available to some extent: a general requirement is that the ratio $f(y)/q(y | x)$ is known up to a constant independent of x (Robert and Casella, 2013).

The general concept of this algorithm is that the probability of acceptance of a distribution depends on the position we are in the sample space and where we want to move to next. The idea is that we write this acceptance in order to be more likely for us to move to a more densely populated area of the sample space; even though the steps are random, for example using a random walk, the acceptance of our movements is not.

Random Walk - A random walk is a sequence that starting from a initial point X_0 can be constructed by

$$X_{n+1} = X_n + \epsilon_n,$$

where ϵ_n is a random value.

The Metropolis-Hasting algorithm can be divided in the following steps (using R_t as an example):

1. start with R_t and choose a initial state for $t = 0$;
2. generate a candidate for R_0^* from $g(R_0^* | R_t)$;
3. calculate the ratio

$$\frac{P(R_t)g(R_0^* | R_t)}{P(R_0^*)g(R_t | R_0^*)};$$

4. compute the acceptance probability

$$A(R_0^*, R_t) = \min \left(1, \frac{P(R_t^*)g(R_0 | R_t^*)}{P(R_0^*)g(R_t^* | R_0)} \right);$$

5. generate u from the uniform distribution and replace or update R_t based on

$$R_{t+1} = \begin{cases} R_1^*, & u \leq A(R_0^*, R_t), \\ R_t, & u > A(R_0^*, R_t). \end{cases}$$

It is common to ignore the first steps of the simulation since they are most likely not in equilibrium and thus do not represent the target distribution. In our approach the proposed distributions are obtained by random walks. To run our method and the algorithm we used the software R, in which we used only the basic R tools and the package "Splines" version 4.0.1. For a new value to be accepted in our method three proposals had to be accepted, the time series for infections, the reproductive number and the weekly pattern. This creates difficulties for convergence of the method and the necessity of several simulations to get a representative mapping of the space of the distribution. This problem can be overcome by tuning the parameters of the random walk so that the proposals are more likely to be accepted in order to have a higher acceptance ratio.

CHAPTER 4

RESULTS

4.1 Simulated Data

First we applied the method from Section 3.4 to simulated data. We simulated with a constant R_t equal to 3 starting from one case until 21 days of the progression of the disease. In Figure 4.1.1 we can see a time series of the cases y , remember, this is not the true incidence curve and just the reported cases.

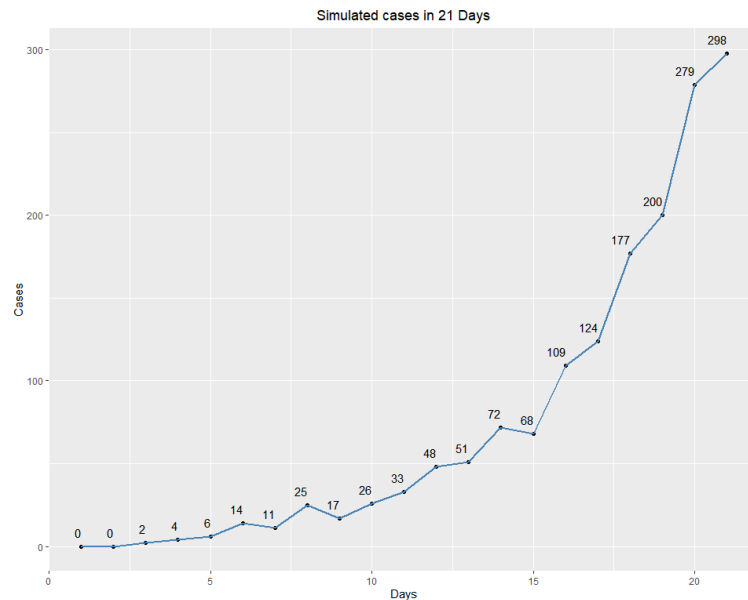


Figure 4.1.1 – Time series of infections (y) resulted from simulated data with $R_t = 3$ during a period of 21 days.

Figure 4.1.2 show the results for R_t in boxplots for the 21 days of simulations. The first thing that is noticeable is the overestimation in the beginning and the underestimation in the

end. This is common to all methods since the data in this type of problem always suffer from right and left truncation. This happens because there are no more cases after the 21st day so the model compensates by reducing the value of R_t in the final estimates, thus with a smaller value of R_t the model justifies the data available, i.e, the absence of cases on the 22th day. The opposite occurs in the initial days, for example, considering the second day of the time series all the cases should come from the first day and thus the model overestimates the values of R_t in the first day.

Although the value of R_t in the first and last five days are far from the true value $R_t = 3$ shown in the red horizontal line, this is a important indication that the method is working properly due to the fact that the over and under estimation happens about one serial interval from the beginning and the end of the time series, indicated by the vertical red lines..

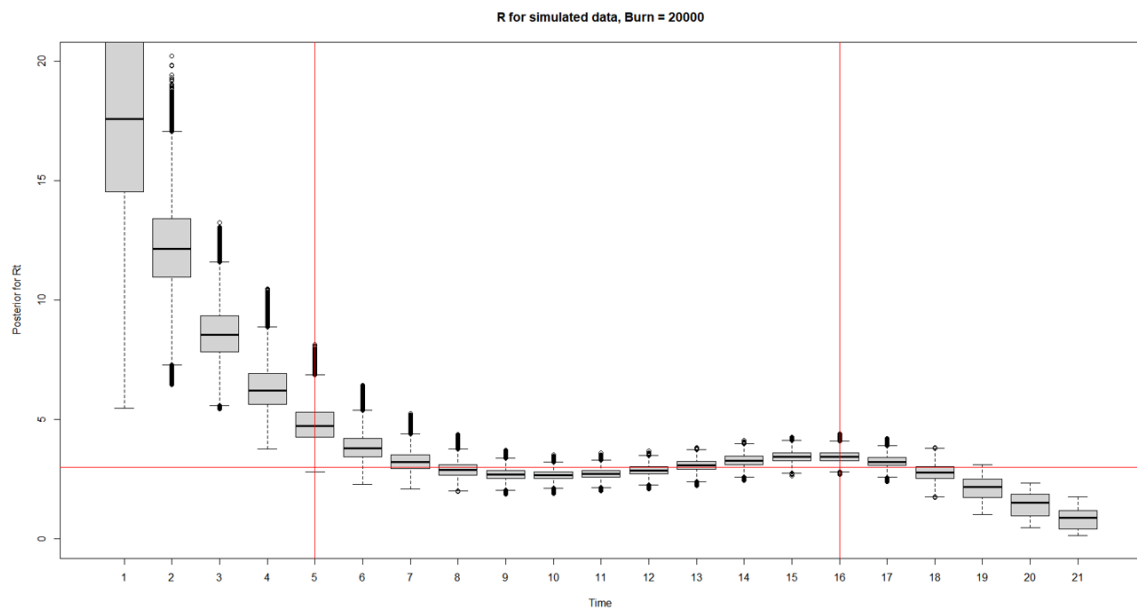


Figure 4.1.2 – Estimation of R_t from the simulated data with 200.000 iterations of the Metropolis-Hastings algorithm. Vertical red lines indicate one serial interval from the beginning and the end of the time series, while the horizontal red line indicates the true value of R_t

Figure 4.1.3 illustrates another way to check if we are correctly estimating the reproductive number. We used the average of the Poisson distribution obtained to reconstruct the cases and as we can see it followed the simulated data (red dots) well, giving confidence that the method is correctly estimating R_t .

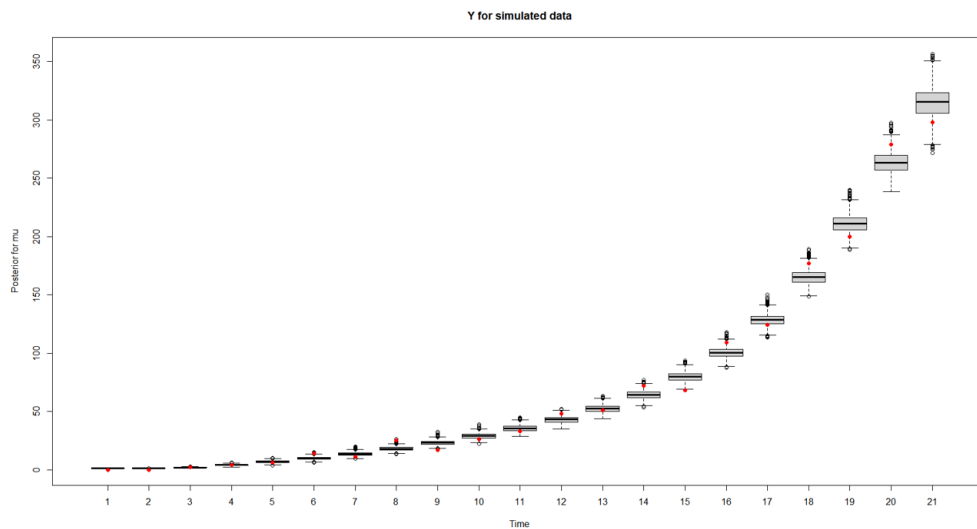


Figure 4.1.3 – Reconstruction of infected series based on the average of the Poisson distribution calculated during the Metropolis-Hastings algorithm. Red dots represent the true infected time series and the boxplots contain the values estimated from 200,000 iterations

Another performance check during the test with the simulated data was based on checking the convergence of the Metropolis-Hastings to a distribution. This can be seen in the annex in Figures A.0.1 and A.0.2, which show the values of the reconstructed cases and R_t converging to a distribution from its initial starting point.

4.2 Switzerland

Using the Swiss data we estimated the reproductive number varying the number of splines (q) in Figure 4.2.1. With $q = 15$ it is noticeable that we were not able to capture most of the variations during a short period. The figure shows that with $q = 30$ and $q = 60$ we capture the significant increase of cases during October 2020.

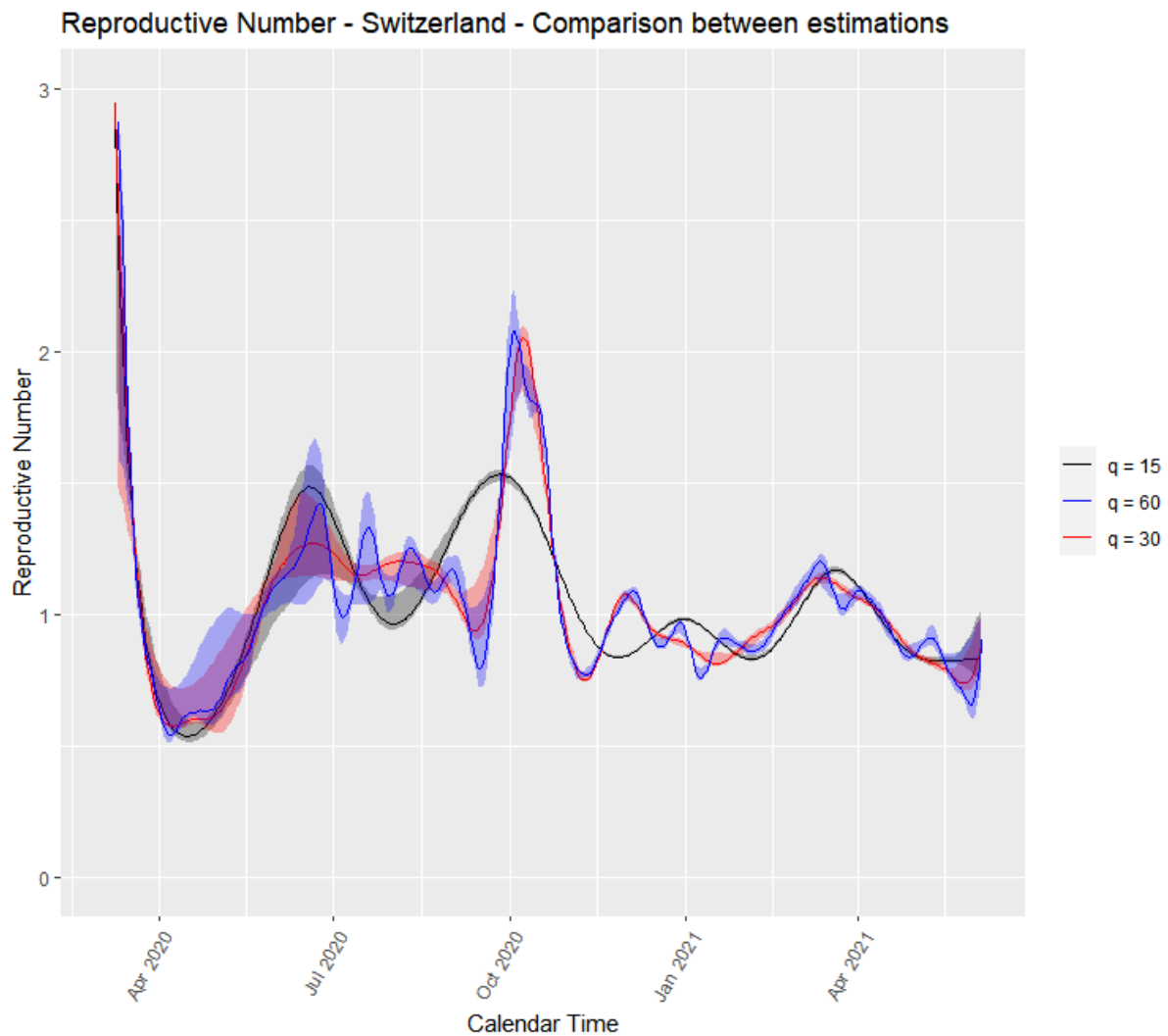


Figure 4.2.1 – Estimates of the reproductive number using the data from Switzerland. Bold line indicates the median and the translucent lines indicates the 95% confidence interval, q is equal to the number of splines used

We then use the values of the Swiss Task-Force to compare to our results. As Figure 4.2.2 shows our results were very similar, with the main differences occurring in the initial period of the pandemic, but after July both are very close, although the NCS-TF method captures more short-term variations.

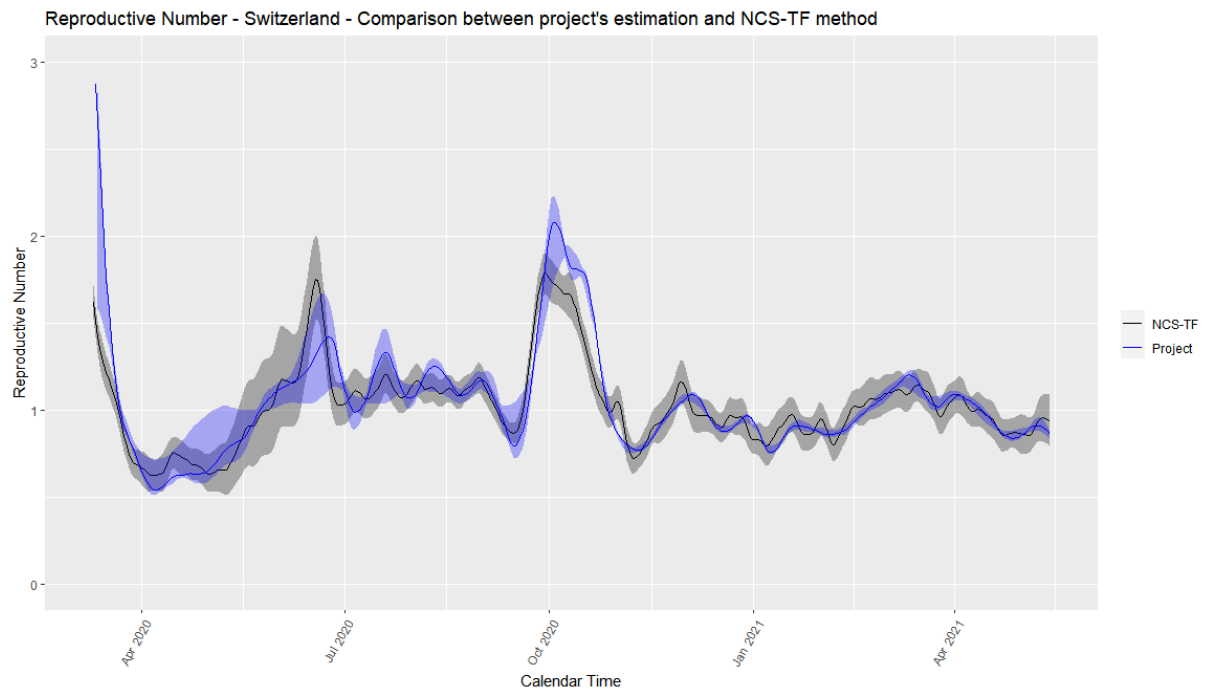


Figure 4.2.2 – Estimated reproductive number for Switzerland comparing the project's and the NCS-TF method. Bold line indicates the median and the trans-lucid lines indicates the 95% confidence interval

Figure 4.2.3 uses the method for the data from Brazil. We see that the task force has very different results from the project. The first thing we notice is the large confidence interval from the task force method. Also, the task force method capture more short-term variations, but this could be achieved by increasing the number of splines in our prior distribution.

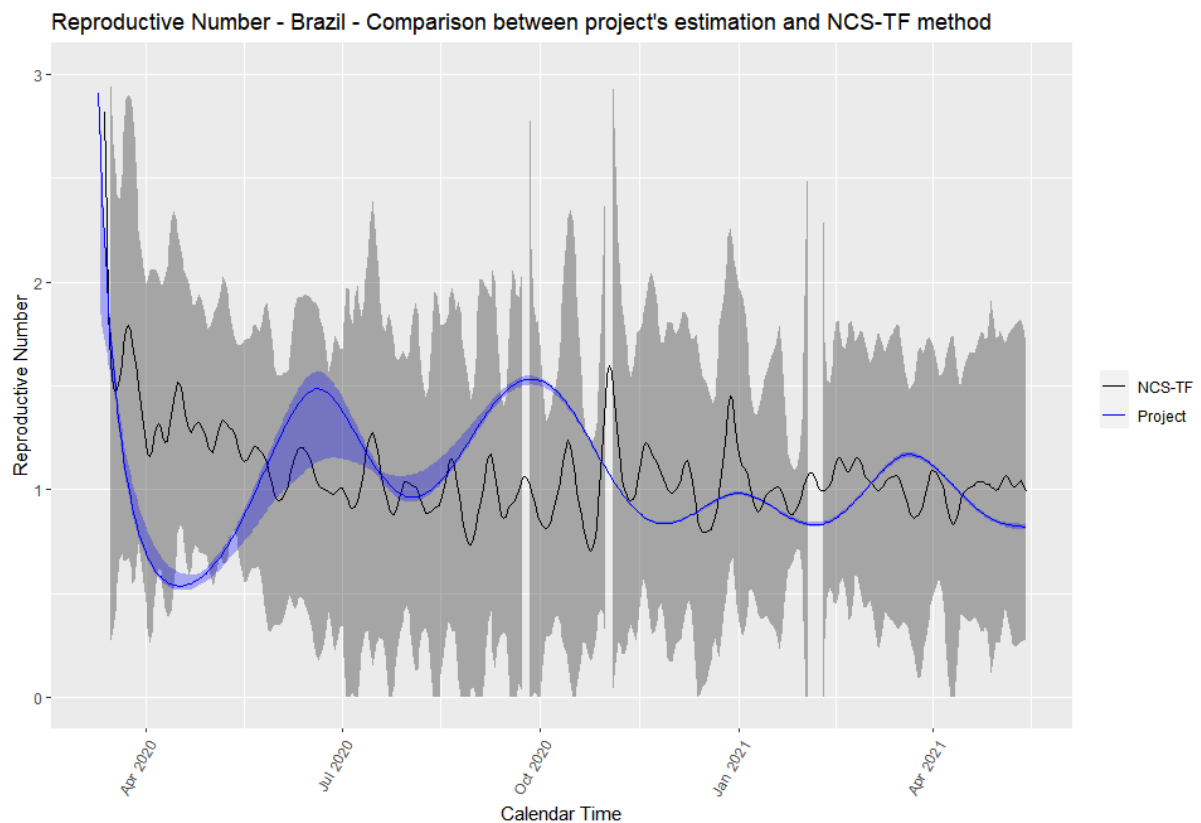


Figure 4.2.3 – Estimated reproductive number for Brazil comparing the project's and the NCS-TF method. Bold line indicates the median and the translucent lines indicates the 95% confidence interval

CHAPTER 5

CONCLUSION

The method developed worked well in the simulated data as our estimates were close to the true values and our algorithm converged to the target distribution. Also, comparing to the results of the NCS-TF we could see similar results reassuring that we can apply this method to real world data. Despite the longer computation time in our method (up to 10 minutes) we saw a significant decrease in the confidence interval in the Brazil time series, giving more confidence for public officials to make decisions. Like all methods developed so far, our method suffers from left truncation, which generates an over-estimate of the reproductive number in the beginning and also from right truncation, which causes underestimation at the end of the time series. The most important problem to deal with the right truncation, since this affects how fast the government can respond to changes in the reproductive number. Also the left truncation issue occurs in a fixed point in time, at the beginning of the pandemic when the disease has not spread throughout the country and society is just starting to notice that there is something different, while the right truncation is a moving set of days during the ongoing pandemic. When applying this approach to other data sets it should be noted that, depending on the serial interval of the disease, the official reporting of values of the reproductive number R_t should start at least one serial interval from the extreme points of the time series. For example, the NCS-TF starts reporting the value of R_t after 10 days of the first case and stops reporting 10 days before the last day. This would also be appropriate for our method due to what we analysed in both the simulated and real world data. Data augmentation methods could be used to increase the incidence time series with artificial cases which will reduce the issue with left and right truncation. There are several other aspects of the pandemic that we did not introduce in the model but may be considered, such as, how to deal with more infectious variants, the inclusion of vaccinated people in the population, sensitivity of the test and several others that our method would probably benefit from if added.

Even though the project focuses on estimating the reproduction number to orient public policies, using it as the only guidance is not wise and the countries that we selected to our analysis illustrate this issue.

Switzerland had a small percentage of its population infected in comparison to Brazil. But the reproduction number of Switzerland shows high variability during the year. On the other hand in Brazil's case the country had a significant percentage of its population infected and the reproduction number does not rise above 2. If we direct public policies during the worst period of the pandemic in both countries in order to both to have a reproduction number around one, in Brazil's case it would maintain the ICUs full like it was in the peak of the pandemic while in Switzerland which faced much less during its peak the country would have a significant lower burden on its health systems. This illustrates that even when ignoring economical/political factors and looking only based on an epidemiological point of view, the reproduction number is not a stand-alone statistic and requires more information to grasp the entirety of each country's situation. Other epidemiological statistics must be consider with the reproductive number, such as the numbers of cases and deaths, the availability of ICU and others.

Annex

ANNEX A

ANNEX

Figures [A.0.1](#) and [A.0.2](#) show a form of checking the convergence of a Metropolis-Hastings algorithm to a target distribution. Figure [A.0.1](#) shows that the time series starts far from the rest of the time series but after a few iterations it reaches a new level and oscillates around it. Figure [A.0.2](#) the initial values of R_t were closer to the target value, but we still see the same phenomenon occurring in the beginning of the time series (panels “Time 2” and “Time 4”) since their target values are far from the initial values proposed due to the under-estimation that occurs in the beginning of the period as commented before.

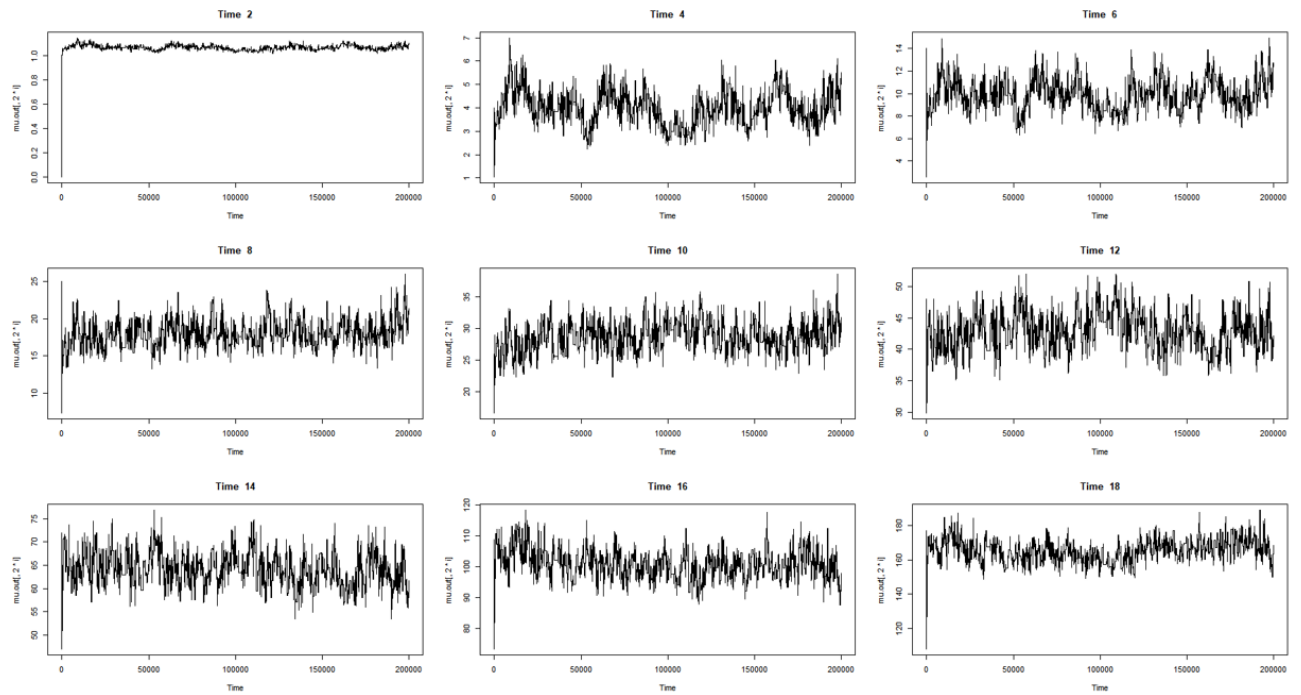


Figure A.0.1 – Time series of the number iterations from the Metropolis-Hastings algorithm in the x-axis and the value of the number of reported cases on the y-axis based on the simulated data

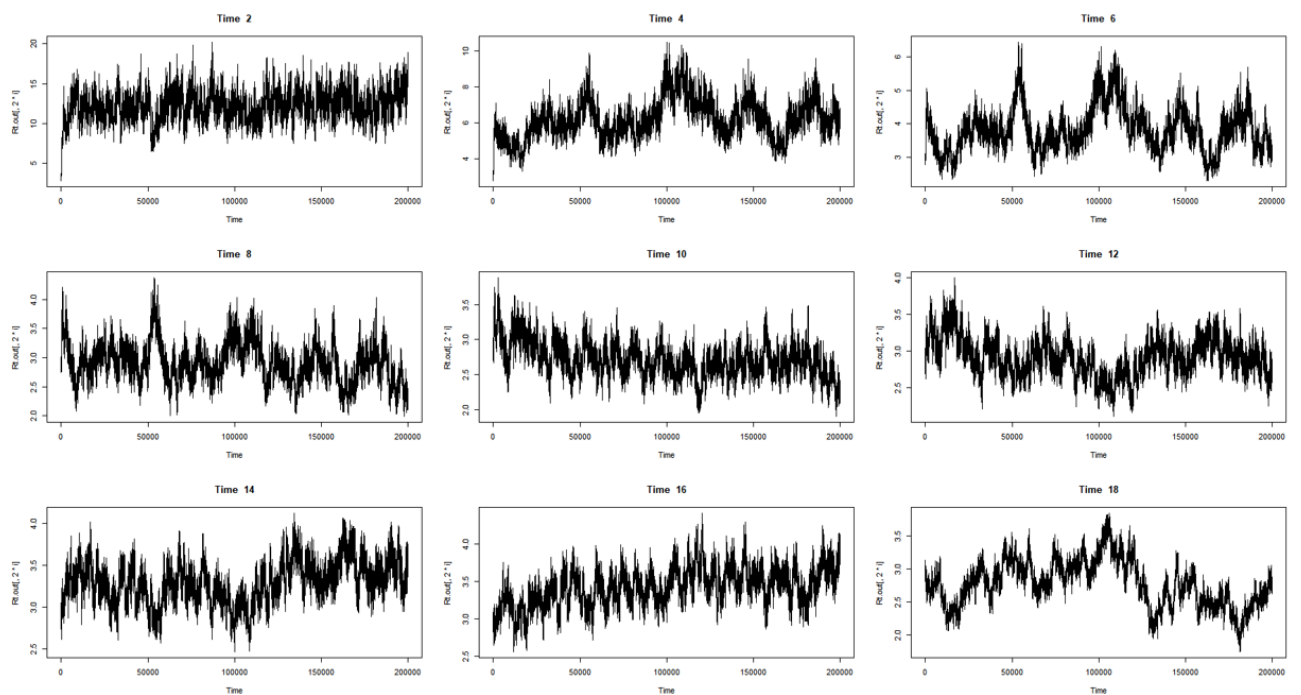


Figure A.0.2 – Time series of the number iterations from the Metropolis-Hastings algorithm in the x-axis and the value of the reproduction number on the y-axis based on the simulated data

BIBLIOGRAPHY

- Allen, L. J., Brauer, F., Van den Driessche, P. and Wu, J. (2008) *Mathematical Epidemiology*. Berlin, Germany: Springer.
- Bettencourt, L. M. and Ribeiro, R. M. (2008) Real time Bayesian estimation of the epidemic potential of emerging infectious diseases. *PLoS One* **3**(5), e2185.
- Brownlee, J. (1906) Statistical studies in immunity: the theory of an epidemic. *Proceedings of the Royal Society of Edinburgh* **26**(1), 484–521.
- Cori, A., Ferguson, N. M., Fraser, C. and Cauchemez, S. (2013) A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology* **178**(9), 1505–1512.
- Delamater, P. L., Street, E. J., Leslie, T. F., Yang, Y. T. and Jacobsen, K. H. (2019) Complexity of the basic reproduction number (r_0). *Emerging Infectious Diseases* **25**(1), 1.
- Dietz, K. (1993) The estimation of the basic reproduction number for infectious diseases. *Statistical Methods in Medical Research* **2**(1), 23–41.
- Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996) *Markov Chain Monte Carlo in Practice*. CRC Press.
- Goldstein, E., Dushoff, J., Ma, J., Plotkin, J. B., Earn, D. J. and Lipsitch, M. (2009) Reconstructing influenza incidence by deconvolution of daily mortality time series. *Proceedings of the National Academy of Sciences* **106**(51), 21825–21829.
- Gostic, K. M., McGough, L., Baskerville, E. B., Abbott, S., Joshi, K., Tedijanto, C., Kahn, R., Niehus, R., Hay, J. A., De Salazar, P. M. *et al.* (2020) Practical considerations for measuring the effective reproductive number, R_t . *PLOS Computational Biology* **16**(12), e1008409.

- Grimm, V., Mengel, F. and Schmidt, M. (2021) Extensions of the SEIR model for the analysis of tailored social distancing and tracing approaches to cope with Covid-19. *Scientific Reports* **11**(1), 1–16.
- Hamer, W. H. (1906) *Epidemic Disease in England: the Evidence of Variability and of Persistence of Type*. London, UK: Bedford Press.
- Kenah, E., Lipsitch, M. and Robins, J. M. (2008) Generation interval contraction and epidemic data analysis. *Mathematical Biosciences* **213**(1), 71–79.
- Kunoth, A., Lyche, T., Sangalli, G. and Serra-Capizzano, S. (2018) *Splines and PDEs: From approximation theory to numerical linear algebra*. Springer.
- Levin, S. A., Hallam, T. G. and Gross, L. J. (2012) *Applied Mathematical Ecology*. Berlin, Germany: Springer.
- Liang, L.-L., Tseng, C.-H., Ho, H. J. and Wu, C.-Y. (2020) Covid-19 mortality is negatively associated with test number and government effectiveness. *Scientific Reports* **10**(1), 1–7.
- Linton, N. M., Kobayashi, T., Yang, Y., Hayashi, K., Akhmetzhanov, A. R., Jung, S.-M., Yuan, B., Kinoshita, R. and Nishiura, H. (2020) Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: a statistical analysis of publicly available case data. *Journal of Clinical Medicine* **9**(2), 538.
- Madhav, N., Oppenheim, B., Gallivan, M., Mulembakani, P., Rubin, E. and Wolfe, N. (2017) *Pandemics: Risks, impacts, and mitigation*. Washington, DC: Elsevier.
- Porta, M. (2014) *A Dictionary of Epidemiology*. New York, NY, USA: Oxford University Press.
- Robert, C. and Casella, G. (2013) *Monte Carlo Statistical Methods*. Springer.
- Sciré, J., Nadeau, S. A., Vaughan, T. G., Gavin, B., Fuchs, S., Sommer, J., Koch, K. N., Misteli, R., Mundorff, L., Götz, T. et al. (2020) Reproductive number of the Covid-19 epidemic in Switzerland with a focus on the cantons of Basel-Stadt and Basel-Landschaft. *Swiss Medical Weekly* **150**(19-20), w20271.
- Soper, H. E. (1929) The interpretation of periodicity in disease prevalence. *Journal of the Royal Statistical Society* **92**(1), 34–73.
- Svensson, Å. (2007) A note on generation times in epidemic models. *Mathematical Biosciences* **208**(1), 300–311.

- Vijgen, L., Keyaerts, E., Moës, E., Thoelen, I., Wollants, E., Lemey, P., Vandamme, A.-M. and Van Ranst, M. (2005) Complete genomic sequence of human coronavirus oc43: molecular clock analysis suggests a relatively recent zoonotic coronavirus transmission event. *Journal of Virology* **79**(3), 1595–1604.
- Wallinga, J. and Teunis, P. (2004) Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of Epidemiology* **160**(6), 509–516.
- Wood, S. N. (2017) *Generalized Additive Models: an Introduction with R*. CRC Press.