Mathematical and Statistical Modelling During a Pandemic Event

by

Adriana-Stefania Ciupeanu

A thesis submitted to the Faculty of Graduate and Postdoctoral Studies of
The University of Manitoba
in partial fulfillment of the requirements of the degree of

Doctor of Philosophy

Individual Interdisciplinary Studies
Department of Mathematics
University of Manitoba
Winnipeg

Abstract

This thesis addresses challenges in epidemiological research through innovative methodological approaches across multiple domains. It begins by presenting a novel approach to conducting a comprehensive literature review on COVID-19 epidemiological modelling. Given the vast volume of research, traditional manual methods proved insufficient. To address this, we employed a data-driven methodology using natural language processing (NLP) techniques to systematically analyze and synthesize key findings in the field.

One focus is on COVID-19 transmission in Alberta, where the transition from time-dependent to constant parameters improved the clarity of intervention scenario analyses. Findings highlight the critical role of social distancing and strategic testing in controlling disease spread, offering practical insights for public health decision-making.

Another key focus is infectious disease variant interactions, introducing the concept of "practical coexistence" to examine how initial infection levels and reproduction numbers influence variant dynamics. The analysis of Alpha and Gamma variants in Alberta and British Columbia reveals the impacts of containment strategies on variant spread.

Additionally, the thesis develops a Python package for temporal network analysis, applying innovative techniques to transportation networks. By examining air travel and urban bike-sharing systems, the research illuminates complex network adaptability and resilience which has implications for pandemic preparedness.

Ultimately, the work emphasizes the critical role of interdisciplinary collaboration, modelling flexibility, and comprehensive intervention strategies in addressing future public health challenges.

Acknowledgements

I offer my sincerest gratitude to my thesis advisors, Dr. Julien Arino of the Department of Mathematics at the University of Manitoba and Dr. Saman Muthukumarana of the Department of Statistics at the University of Manitoba. Dr. Arino's office door was always open whenever I encountered a problem or had a question about my research or writing. His patience, especially with my sarcasm over the years, has been unwavering. I am deeply grateful for his meticulous review of my thesis and his constructive criticism, which has significantly improved my work.

I also wish to acknowledge Dr. Stephanie Portet of the Department of Mathematics at the University of Manitoba and Dr. Aleeza Gerstein of the Departments of Statistics and Microbiology for their willingness to serve on my PhD committee. Their careful reading of my thesis and their valuable corrections and suggestions have been indispensable.

My deepest gratitude goes to my parents, whose constant encouragement throughout my study years has been a cornerstone of my success. They instilled in me a love of science and supported all my pursuits, making this achievement possible.

I extend my thanks to Dr. Michael Li from the Department of Mathematical and Statistical Sciences at the University of Alberta and Dr. Betsy Varughese of the Institute of Health Economics. Their invitation to join their research groups during the COVID-19 pandemic was invaluable. This experience broadened my research perspective, particularly by allowing me to learn and apply time-dependent parameters modelling, which has been instrumental in my work.

Additionally, I am grateful to my colleagues and friends for their support and camaraderie throughout this journey.

Lastly, I wish to thank the Visual and Automated Disease Analytics (VADA) graduate training program that supported part of my research.

Nothing of me is original. I am the combined effort of everyone I've ever known.

Chuck Palahniuk, Invisible Monsters

Thank you all for your unwavering support and contributions to my academic journey.

Contributions of Authors

Chapter 3: Assisted literature review

Adrian-Stefania Ciupeanu: conceptualization, data collection, codes analysis, methodology, writing

Julien Arino: conceptualization, supervision, writing (review and editing)

Chapter 4: Quantifying the effects of public health interventions in Alberta during the first wave of COVID-19

Adrian-Stefania Ciupeanu: conceptualization (from Section 4.2.4 onwards), codes, investigation, analysis, methodology, simulations, writing

Weston C Roda: initial codes (which ended up not being used), fitting of the 1st wave results, conceptualization (Section 4.2.7)

Donglin Han: initial codes (which ended up not being used), fitting of the 1st wave results

Marie Betsy Varughese: supervision

Michael Li: conceptualization, supervision

Chapter 5: Mathematical modelling of the dynamics of COVID-19 variants of concern: asymptotic and finite-time perspectives

Adrian-Stefania Ciupeanu: conceptualization, modification of codes, investigation, analysis, methodology, simulations, writing (original draft)

Weston C Roda: initial codes

Donglin Han: initial codes

Qun Cheng:initial codes

Marie Betsy Varughese: conceptualization, supervision, writing (review and editing)

Michael Li: supervision, writing (review and editing)

Chapter 6: Quantifying Change in a Network

Adrian-Stefania Ciupeanu: conceptualization, data collection, codes, analysis, methodology, writing

Julien Arino: conceptualization, supervision, writing (review and editing)

Contents

List of Figures

List of Tables

1

Introduction

1.1 How this thesis came to be

I started working on my Individual Interdisciplinary Studies (IIS) PhD in September 2019. The proposed research was titled "Addressing the Scarcity of Data during Epidemic Events". The plan was to utilise several significant infection events, such as the spread of SARS in 2003 and the Ebola Virus Disease outbreak in Western Africa in 2014, along with other international spread events for which data was available, to constitute a library of cases for investigation. The main idea was that, as an epidemic or pandemic event unfolds, there is initially very limited data, making statistical models not reliable. This scarcity of data can be compensated by incorporating expert knowledge through the use of mathematical models and our aim was to investigate such aspects.

I had also been admitted to the NSERC CREATE Visual and Automatic Disease Analytics graduate training program hosted by the University of Manitoba and run jointly with the University of Victoria. Besides taking a course on the *Foundations of disease analytics*, the program involved a four-month internship, which we had arranged for me to spend in May–August 2020 in Lyon (France) working with a collaborator of Dr. Julien Arino (one of my supervisors) who heads the vaccine modelling team at Sanofi.

However, on December 31, 2019, the WHO China Country Office was informed of cases of pneumonia of unknown aetiology in Wuhan [1]. The pathogen was later identified as SARS-CoV-2, responsible for COVID-19, which evolved from a coronavirus infecting wild bats and then spread to humans. Consequently, the thesis took a different direction than initially envisioned but still addresses some of the originally proposed methodological problems.

First of all, Julien Arino became directly involved in the Public Health Agency of Canada (PHAC) modelling response to the pandemic, which in turn means that I became involved. Secondly, I could not leave Canada (as I would not have been able to return), so my VADA internship was changed to (remote) work with Dr. Michael Li's lab at the University of Alberta, alongside Dr.

Marie Betsy Varughese from Alberta Health Services (now at the Institute of Health Economics). The work with Dr. Li evolved into a collaboration that is still ongoing.

This thesis documents my personal journey through the pandemic, providing a collection of some of the work conducted during this period. It aims to offer insights into the practical applications of mathematical modelling in infectious disease outbreaks. The chapters of the thesis present snapshots of the activities of a mathematical modeller of infectious diseases during such events. This work spanned various stages of the pandemic, using WHO and CDC terminology [4] from the "Alert phase" (when a new pathogen has been identified in humans) to the "Transition phase" (preparation for future pandemic waves).

In detailing this journey, the thesis highlights several key lessons learned during the pandemic. These include the importance of preparedness, the need for adaptable models that can be quickly adjusted to new data and scenarios, and the benefits of interdisciplinary collaboration in managing public health crises. By sharing these experiences and insights, I hope to contribute valuable lessons to the community, emphasising that the question is not if a pandemic will occur in the future, but rather "when" and "in what form".

1.2 Epidemics and pandemics

Epidemics, defined as sudden outbreaks of disease, and pandemics, which are epidemics occurring across various regions of the globe, have happened throughout human history, significantly impacting societies, economies, and public health systems. These phenomena have shaped human history by causing widespread mortality and morbidity, disrupting daily life, and prompting advancements in medical and public health practices. The bubonic plague, also known as the Black Death, is an example of an epidemic's destructive potential. Sweeping across Europe in the 14th century, the plague was caused by the bacterium *Yersinia pestis*, transmitted by infected fleas living on rodents. It resulted in the deaths of an estimated one-third of the European population [148]. This catastrophic event led to the development of early public health measures, such as quarantine and sanitation practices, which laid the groundwork for modern infectious disease control. Quarantine stations, or "lazarettos", were established in port cities to isolate and monitor incoming ships and travellers, a practice that is still employed today to control the spread of infectious diseases [180].

Similarly, the 19th-century smallpox epidemic, caused by the variola virus, was a global catastrophe. Highly contagious and often fatal, smallpox ravaged communities worldwide. The development of a vaccine in the 18th century eventually led to the eradication of smallpox in 1980, marking a significant victory in public health history [174].

The Spanish flu pandemic of 1918 infected approximately 500 million people and caused the deaths of between 20 and 100 million people, illustrating the devastating potential of pandemics

[89]. During 1918–1919 the limitations of social distancing measures in controlling a pandemic became evident. While public-gathering bans, school closures, and transportation restrictions were implemented, these measures were often difficult to enforce and faced societal resistance [181]. This historical example underscores the enduring challenges of implementing social distancing interventions during public health emergencies. The work [180] found that despite understanding the risks of infection, adherence to social distancing measures was difficult for Hispanic communities in New York City during COVID-19 due to factors including crowded multi-generational households, essential jobs requiring close contact with others, and providing unpaid care to family members. This emphasises the need for targeted public health approaches that consider the unique social and economic realities of different populations during pandemics.

More recently, HIV has claimed about 40 million lives worldwide since the beginning of the epidemic in the early 1980s, with around 85 million people infected [150]. In Eswatini, the HIV/AIDS epidemic led to the highest prevalence rates globally, causing life expectancy to drop to 32 years in 2009 [3].

Certain diseases, such as plague, hemorrhagic fevers, measles, and poliovirus, continue to erupt occasionally, while others, like malaria, HIV/AIDS, Mycobacterium tuberculosis, typhus, and cholera, are endemic in certain regions of the globe. The AIDS epidemic, the SARS epidemic, recurring influenza pandemics, and more recently, the SARS-CoV-2 pandemic, pose significant public health concerns.

A key lesson from the COVID-19 pandemic is twofold. Firstly, the warnings by public health authorities about the inevitability of a pandemic, especially since the start of the 21st century, were well-founded. Despite events like the SARS-CoV-1 epidemic of 2003 or the H1N1 influenza pandemic of 2009 somewhat reducing public and political perception of these risks, COVID-19 has demonstrated its reality. Secondly, such events are likely to recur due to increased interactions between humans and animal reservoirs of pathogens, influenced by expanding human settlements and climate change. For instance, rising mean temperatures have expanded the range of ticks carrying Crimean-Congo Hemorrhagic Fever, with the first case in continental Europe detected in 2011 [2]. Additionally, unprecedented levels of human movement further facilitate pathogen spread.

Mathematical models for the dynamics of infectious diseases have been studied since 1760 [28]. In recent decades, mathematical and statistical models of infectious disease propagation have played different and complementary roles in aiding public health and policy planning during epidemics and pandemics. These models help predict the number of cases and illustrate the effect of public health policies on the spread of emerging infectious diseases. Earlier mathematical models were based on the classical Kermack and McKendrick model [130] while more recent models have been used to analyze the impact of stay-at-home measures or confinement on outbreak trajectories

[26, 210], calculate the basic reproduction number, and integrate individual movement patterns.

Mathematical models make predictions based on the incorporation of knowledge about the disease transmission process. In contrast, statistical models make predictions based on available data. Unfortunately, these two approaches are somewhat disconnected in the field of epidemiology. While mathematical models typically assume a probability that a contact between an individual susceptible to a disease and one infectious with it results in a new infection, the nature of this probability is left open. Statistical models, on the other hand, assume that data have specific properties but omit details about the underlying processes that generate the data.

Nowhere is the disconnect between disciplines and approaches more important than during international-level outbreaks of new or re-emerging diseases. Consider, for instance, the case of SARS, which spread worldwide in 2002-2003. The overall number of cases was not very high, just over 8,000. Mathematical models for this type of spread existed at the time, although they have greatly improved since. To "parameterize" these models, i.e., to set their operational parameters, much more information than what was available was needed. The scarcity of data also posed challenges for statistical models, but statistics has tools that can handle this data paucity. In both cases, however, methodological problems arise when an epidemic event of international significance occurs, especially concerning newly emerging or re-emerging diseases.

Each epidemic spread event represents a single realization of a complex stochastic process involving both local (infection) and global (travel of infected individuals) events. Data are only partially observed since typically only individuals presenting acute symptoms enter the health system and contribute to the data. To accurately predict the "trajectory" of an epidemic, we need to incorporate data as it is produced and reported, as well as expert knowledge about the disease transmission process.

There are many questions public health authorities can ask themselves during an epidemic [42]:

- How many individuals will be infected, looking at both hidden infections and known infections? How many of these individuals will require treatment?
- How many hospital beds are required at any particular time?
- How long will the epidemic last?
- What non-pharmaceutical interventions could be taken to decrease the spread of the disease?
- How effective are the current public health measures, and when should they be adjusted or lifted?
- What groups of people should be prioritized for a vaccination campaigns and how could this campaign be implemented?

- How can international collaboration and data sharing be enhanced to improve epidemic response?
- How should we prepare for potential secondary outbreaks or waves of infection?

The thesis will introduce deterministic models, with a particular emphasis on SICR (Susceptible-Infectious-Cases-Recovered/Removed) models.

Despite the abundance of data related to the COVID-19 pandemic, this data is not consistent across jurisdictions. When COVID-19 was first identified in Canada, all jurisdictions implemented PCR tests only for individuals with symptoms who had either traveled internationally in the 14 days prior or had been in contact with COVID-19 positive patients. This approach meant that cases were under-detected (resulting in partially observed data), as only individuals who were tested entered the health system and thus contributed to the available data.

This thesis exemplifies the interdisciplinary approach that defines Individual Interdisciplinary Studies by bridging mathematics, public health, computer science, and policy analysis to address complex epidemic challenges. The original research proposal inherently recognized that epidemic modeling requires integration across multiple domainsâĂŤcombining mathematical rigor with epidemiological understanding, statistical analysis with expert knowledge, and theoretical frameworks with practical public health applications.

The emergence of COVID-19 transformed this thesis from a theoretical interdisciplinary study into a real-time demonstration of how different fields must work together during public health emergencies. Through Dr. Arino's involvement with the Public Health Agency of Canada, I became directly engaged in pandemic response efforts, requiring rapid integration of mathematical modelling with policy needs and public health decision-making. One crucial aspect of this work involved the integration of expert knowledge from public health authorities, made possible through an internship I completed as part of the NSERC CREATE Visual and Automatic Disease Analytics (VADA) program. The internship took place in Dr. Michael Li's lab at the University of Alberta and included collaboration with Dr. Marie Betsy Varughese, who was affiliated with Alberta Health Services at that time (she is now with the Institute of Health Economics). This collaboration expanded the interdisciplinary scope to include provincial health system perspectives, demonstrating how interdisciplinary research must adapt dynamically to serve urgent societal needs.

The chapters of this thesis reflect this interdisciplinary breadth, with each contribution spanning multiple fields. The epidemiological modelling work combines mathematical analysis with public health policy evaluation, the variant dynamics research integrates theoretical predictions with real-world case studies, and the temporal network analysis bridges pure mathematical concepts with software development. Perhaps most significantly, the use of AI and natural language processing

techniques for literature review demonstrates how methodological innovation can emerge from synthesizing approaches across disciplines. Rather than simply applying multiple disciplinary perspectives to a single problem, this work shows how interdisciplinary thinking creates entirely new approaches that no single field could generate alone, particularly when responding to rapidly evolving public health crises.

As such the research portion of the thesis is divided as follows:

- Using AI for assisted literature review (Chapter 3).
- Assessing the effects of non-pharmaceutical interventions on an epidemic wave (Chapter 4).
- Understanding the dynamics of variants of concern during an epidemic wave to better prepare for the next wave (Chapter 5).
- Utilising social network analysis to observe changes in a network (Chapter 6).

1.3 COVID-19 modelling literature review

The literature review in a thesis is usually in the Introduction chapter, however, as we have used AI to perform this literature review we decided that it is more fitting to be in the Chapter where we explain how the literature review was performed. As such please see 3.3 for the review.

2

Preliminaries

Overview of the chapter This chapter introduces fundamental concepts in mathematical epidemiology, network theory, Bayesian inference and Natural Language Processing (NLP), which will be used in the remainder of the manuscript.

Due to the interdisciplinary nature of this work, the thesis begins by providing a catalogue of definitions and basic tools to establish a common foundation across the diverse fields involve

The chapter begins by introducing ordinary differential equations (ODEs), which are widely used in modelling infectious diseases. The ODE section covers stability analysis and key epidemiological terms like epidemics and pandemics. An emphasise is placed on the Susceptible-Infected-Recovered (SIR) model because of its role in predicting disease dynamics. Social Network Analysis (SNA) is introduced and focuses on metrics like degree centrality and community detection algorithms. Bayesian inference and Markov Chain Monte Carlo (MCMC) methods are discussed for parameter estimation. In natural language processing (NLP), advancements in transformer models such as LLaMA are showcased for their impact on tasks like text summarization.

2.1 Ordinary differential equations

The source for this section is [198]

A first-order **ordinary differential equation (ODE)** takes the following form

$$\frac{d}{dt}x(t) = g(t, x(t)) \tag{2.1.1}$$

or, for short,

$$x' = g(t, x), \tag{2.1.2}$$

where $t \in \mathbb{R}$ is the **independent variable**, $x(t)$ is a **dependent variable** and $g : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ is a **vector field**. Contrary to algebraic equations, the unknown $x(t)$ is thus here a function (of t).

If g does not depend explicitly on time, then (2.1.2) is **autonomous** and can be written as

$$x' = g(x). \tag{2.1.3}$$

Our focus in this section is on autonomous ODEs, as nonautonomous ODEs, where $g(t, x)$ explicitly depends on t, introduce additional complexities compared to autonomous systems.

The general solution to (2.1.3) is

$$x(t) = \int_{t_0}^{t} g(u)\, du. \tag{2.1.4}$$

In the study of ODE, initial value problems (IVPs) and the well-posedness of these problems are fundamental concepts.

Definition 2.1 (Initial value problem)**.** *An **initial value problem** consists of an ODE $x'(t) = g(x)$, along with an initial condition $x(t_0) = x_0$, i.e., a system of the form*

$$x' = g(x) \tag{2.1.5a}$$

$$x(t_0) = x_0. \tag{2.1.5b}$$

The concept of flow is essential to describe the evolution of solutions over time.

Definition 2.2 (Flow)**.** *Consider an initial value problem of the form* (2.1.5)*. The **flow** of IVP* (2.1.5) *is $\psi(x_0)$, representing the solution of the ODE over time given an initial condition, provided that solutions to the differential equation exist and are unique.*

Definition 2.3 (Well-posededness)**.** *Let $x'(t) = g(x)$ be an ODE or a system of ODEs. Then we say that the ODE(s) is **well-posed** if there exists a unique solution and for ODEs describ-*

ing populations, the solutions remain bounded and are non-negative for all non-negative initial conditions.

The existence and uniqueness of solutions to IVPs are guaranteed by the Cauchy-Lipschitz theorem under certain conditions.

Theorem 2.4 (Cauchy-Lipschitz)**.** *Let $x'(t) = g(x)$ be a differential equation with $x \in \mathbb{R}^n$, such that g is a differentiable function whose derivative is continuous, i.e., $g \in C^1$. Then there exists a unique solution such that $x(t_0) = x_0$, where $t_0 \in \mathbb{R}$ and $x_0 \in \mathbb{R}^n$, defined on the largest interval I containing t_0 on which $g \in C^1$.*

Equilibrium points of ODEs are critical in understanding the behaviour of dynamical systems.

Definition 2.5 (Equilibrium point)**.** *Let $x'(t) = g(x)$ be an ODE such that $x \in \mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R}^n$. We say that x^* is an **equilibrium** solution if $g(x^*) = 0$.*

The stability of equilibrium points is a significant aspect of the qualitative analysis of ODEs. Generally, three types of stability are considered, reflecting increasingly strong conditions on the behavior of solutions near an equilibrium. The first, Lyapunov stability, means that if a solution starts close to the equilibrium point, it remains close as time progresses; however, the solution does not necessarily move closer to the equilibrium. The second, asymptotic stability, strengthens this by requiring that solutions not only remain close but also tend to the equilibrium as time progresses. The strongest form, global asymptotic stability, requires that all solutions in the considered domain, regardless of their initial condition, eventually converge to the equilibrium point, indicating the equilibrium's dominance over the system's long-term behaviour.

Definition 2.6 (Locally stable equilibrium point)**.** *Assume $t \in \mathbb{R}$, and let $\psi(t)$ be the flow of (2.1.2). We say that an equilibrium solution x^* is **locally stable** if, for all $\epsilon > 0$, there exists $\delta = \delta(\epsilon) > 0$ such that for all $x \in \mathcal{N}_\delta(x^*)$ and $t \geqslant 0$, there holds*

$$\psi(t, x) \in \mathcal{N}_\epsilon(x^*),$$

where $\mathcal{N}_\delta(x^*)$ is a δ-neighborhood of x^*. The equilibrium point is unstable if this does not hold.

Definition 2.7 (Locally asymptotically stable equilibrium point)**.** *An equilibrium solution x^* is **locally asymptotically stable** if it is locally stable, and there exists a $\delta > 0$ such that for all $x \in \mathcal{N}_\delta(x^*)$, the trajectories $\psi(t, x)$ approach x^* as $t \to \infty$.*

Formally, for every $\epsilon > 0$, there exists $\delta = \delta(\epsilon) > 0$ such that for all $x \in \mathcal{N}_\delta(x^*)$ and $t \geqslant 0$,

$$\lim_{t \to \infty} \psi(t, x) = x^*.$$

Definition 2.8 (Hyperbolic fixed point)**.** *A **hyperbolic equilibrium point** for* (2.1.2) *is a point at which the eigenvalues of the Jacobian matrix evaluated at x^*, $Df|_{x^*}$, all have nonzero real part.*

The sign of the real parts of the eigenvalues of the Jacobian matrix determines the stability of a hyperbolic fixed point. If all eigenvalues have negative real parts, the equilibrium is locally asymptotically stable. If at least one eigenvalue has a positive real part, the equilibrium is unstable.

Definition 2.9 (Globally asymptotically stable equilibrium point)**.** *Let x^* be an equilibrium of* (2.1.2)*. Then the equilibrium point x^* is **globally asymptotically stable** if it is locally asymptotically stable for all initial conditions $x_0 \in \mathbb{R}^n$.*

LaSalle's invariance principle is a fundamental theorem in dynamical systems theory that establishes the asymptotic stability of solutions within a region defined by a Lyapunov function. It provides a powerful tool for analyzing the long-term behaviour of nonlinear systems, particularly in systems where explicit Lyapunov functions may be challenging to construct. LaSalle's invariance principle states that for an autonomous differential equation of the form $x' = f(x)$, where $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable if $V(x)$ is a continuously differentiable function such that:

1. $V(x) > 0$ for all $x \neq 0$,
2. $V'(x) \leqslant 0$ for all x,
3. $\{x \in \mathbb{R}^n | V'(x) = 0\}$ is contained in a region where $V(x)$ is non-increasing,

then, every solution $x(t)$ of the differential equation tends asymptotically to the largest invariant set in $\{x|V(x) = 0\}$ as $t \to \infty$.

2.2 Medical epidemiological terms

The source for this section is [23].

Epidemiological phenomena can be categorized into the following three main groups: epidemic, pandemic and endemic.

Epidemics are characterized by a sudden and often dramatic surge in the number of disease cases within a specific geographic region or population group, exceeding the expected baseline. Notable examples include the 2013-2016 Western African Ebola virus epidemic. Public health interventions are often rapidly implemented to contain epidemics and mitigate their public health consequences.

Pandemics are epidemics that spread over a wide geographical area, often affecting multiple countries or continents. The ongoing COVID-19 pandemic exemplifies the far-reaching effects of infectious diseases. Pandemics pose unique challenges for public health authorities due to their scale and complexity, necessitating coordinated international efforts to control transmission and minimize morbidity and mortality.

Endemics refer to the persistent presence of a disease or infectious agent within a given geographic area or population group. Malaria, for instance, is endemic in many tropical regions, necessitating sustained prevention and control measures. Endemic diseases may exhibit fluctuations in intensity over time but tend to be relatively stable within the affected population. Public health strategies for endemic diseases focus on long-term surveillance, prevention, and management to minimize disease burden and improve overall population health outcomes.

Epidemiological measurement are fundamental tools used to quantify the occurrence and distribution of diseases within populations. A few medical epidemiological terms are:

Incidence refers to the number of new cases of a disease occurring within a population during a specified period. It provides crucial insights into the rate at which individuals are being affected by the disease. Mathematically, incidence can be calculated as the number of new cases divided by the population at risk during a defined time period. For example, if a community experiences 100 new cases of influenza over a one-month period, and the population at risk is 10,000 individuals, the incidence rate would be 100 cases per 10,000 population per month. Incidence is often expressed per 100,000 people, especially at the regional or country level. The **population at risk** refers to the group of individuals susceptible to developing the disease within a defined time frame.

Prevalence represents the number or proportion of individuals in a population who have the disease at a given point in time or over a specified period. It provides information about the overall disease burden within the population. Prevalence can be calculated as the number of existing cases divided by the total population at risk. For instance, if a survey conducted in a community identifies 500 individuals with diabetes out of a total population of 5,000, the prevalence of diabetes in that community would be 10%.

Other significant epidemiological measures besides incidence and prevalence include: mortality rate, case fatality rate (or rate), attack rate, standardised mortality ratio, years of potential life lost.

2.3 Mathematical epidemiology

Mathematical epidemiology employs mathematical models and statistical methods to understand the spread and control of infectious diseases within populations. By using differential equations, probability theory, and computational modelling, researchers can simulate how diseases transmit through communities, predict outbreak patterns, and evaluate the effectiveness of intervention strategies like vaccination campaigns or quarantine measures. In Sections 2.4 and 2.5, we take a more detailed look at specific epidemiological models, but start here by general considerations about mathematical models used in epidemiology.

2.3.1 Compartmental models

Infectious diseases are often modelled using compartmental models, where the population is divided into distinct compartments based on disease status (e.g., susceptible, infectious, recovered). These models employ differential equations to describe the rates of transfer between compartments, with time t as the independent variable.

Compartmental models employ differential equations to mathematically describe the progression of infectious diseases within a population. A fundamental tool of epidemiological modelling, the **Susceptible-Infectious-Recovered (SIR)** model is particularly applicable to diseases that confer lasting immunity upon recovery. While the SIR model is most useful for understanding long-term trends, it can also provide valuable insights into disease dynamics over finite time horizons, informing public health interventions during the course of an outbreak. This model divides the population into three compartments: susceptible individuals who can contract the disease, infectious individuals capable of transmission, and recovered individuals immune to reinfection. By tracking the movement of individuals between these compartments over time, the SIR model offers valuable insights into disease dynamics and informs public health interventions [44].

2.3.2 Disease transmission mechanisms

Vertical Transmission is the process where an infected individual transmits the disease to their offspring during pregnancy, childbirth, or breastfeeding (in the case of mammals) or during reproduction (e.g., through eggs in mosquitoes).

Horizontal Transmission has the infection spread directly or indirectly from one infected individual to a susceptible individual. This can occur through various means such as:

- Direct contact: Physical contact with an infected person, such as touching, kissing, or sexual contact.

- Indirect contact: Contact with contaminated objects or surfaces touched by an infected person.
- Airborne transmission: Inhalation of respiratory droplets expelled by an infected individual when coughing, sneezing, or talking.
- Vector-borne transmission: Transmission through the bite of an infected insect or arthropod, such as mosquitoes or ticks.

The **force of infection** is a critical concept in epidemiology. It represents the rate at which susceptible individuals become infected and is influenced by several factors:

- Prevalence of infectious in the population, $I(t)/N(t)$, where $I(t)$ is the number of the infectious individual at time t and $N(t)$ is the total population at time t,
- Contact rate c this factor captures the frequency of contact between susceptible and infectious individuals within the population,
- Transmission probability per contact β.

The force of infection is not constant and is expected to change with the total population size and structure of interactions within the population. Densely populated areas with frequent close contact often experience a higher force of infection compared to sparsely populated regions.

2.3.3 Equilibrium states in epidemiological models

In most epidemiological models, we commonly focus on two key equilibrium states: the **disease-free equilibrium (DFE)** and the **endemic equilibrium (EE)**. The DFE is the state where the disease is absent, meaning that all infected compartments are at zero, and the total population consists only of susceptible individuals and, potentially, immune individuals. In contrast, the EE is the state where the infection persists within the population, resulting in a positive number of infectious individuals at equilibrium. It is important to note, however, that not all models exhibit both the DFE or EE states, as is observed for instance with the Kermack-McKendrick model in Section 2.4.

2.4 The Kermack-McKendrick SIR model

The first documented result in mathematical epidemiology dates back to the 18th century when Daniel Bernoulli [43] defended the practice of smallpox vaccination. The foundations of mathematical epidemiology, as we know it today, were established by R.A. Ross, W.H. Hamer, A.G.

McKendrick, and W.O. Kermack between 1900 and 1935 [41]. The McKendrick and Kermack model divides the total population into three compartments. Compartment S represents individuals who are susceptible and at risk of infectious disease, including the entire population at the beginning of an epidemic wave or pandemic. Individuals in compartment I have the potential to spread the disease within the community. Compartment R contains individuals who have recovered from the viral infection and are no longer infectious.

The primary route of disease transmission occurs through contacts between susceptible individuals in S and infectious individuals in I, which is modelled as $\beta I(t)S(t)$. The transmission coefficient β can be influenced by various factors, including the average number of contacts among individuals in the population and the average probability of transmission for each contact. These factors may depend on both the infectivity of individuals in I and the susceptibility of individuals in S during each contact. The value of β is averaged over individual variations in the population. For individuals in compartment I who are recently infected and not yet infectious, their infectivity is considered to be zero. The parameter γ represents the recovery rate, and $1/\gamma$ is the mean infectious period. This model is often applied to diseases that confer natural immunity and significantly influence the immune system's response to reinfection, such as measles and chickenpox, as well as finite-time epidemics. The model equations are (see Figure 2.1 for the model diagram):

$$\frac{dS}{dt} = -\beta SI, \tag{2.4.1a}$$

$$\frac{dI}{dt} = \beta SI - \gamma I, \tag{2.4.1b}$$

$$\frac{dR}{dt} = \gamma I, \tag{2.4.1c}$$

with nonnegative initial conditions $S(0), I(0), R(0) \geqslant 0$, $S(0) + I(0) + R(0) = N(0) > 0$.

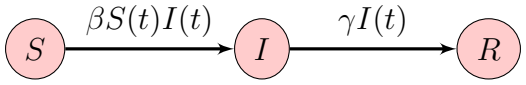


Figure 2.1: Transfer diagram for the SIR model without demography.

System (2.4.1) is deterministic; the behaviour of the model is completely determined by its history and by the rules that govern the development of the model.

The Kermack-McKendrick SIR model has had numerous extensions to incorporate vaccination, quarantine, isolation, antiviral treatment, periodicity (or seasonality), or detected cases over the decades. In Chapter 4 we present extensions of the SIR model.

2.4.1 Well-posedness

In ordinary differential equations usually well-posedness means a system has a unique solution. In our setting, mathematical population dynamics, well-posedness is assumed to mean that a system has a a unique solution, that the solution of the system is non-negative for all non-negative initial conditions, and that the system is bounded. In this section we will show that the Kermack-McKendrick SIR model is well-posede [41, 86].

Let $x(t) = (S(t), I(t), R(t))^T \in \mathbb{R}^3$. Then the system can be written as $\frac{dx}{dt} = f(x)$, where the components of f are

$$\begin{aligned} f_1 &= -\beta SI, \\ f_2 &= \beta SI - \gamma I, \\ f_3 &= \gamma I. \end{aligned}$$

We have that f_i $(i = 1, 2, 3)$ are continuous functions on $\mathbb{R}^3$. Furthermore, $\frac{\partial f_i}{\partial S}$, $\frac{\partial f_i}{\partial I}$, $\frac{\partial f_i}{\partial R}$ exist, which gives us that f_i are differentiable and continuous functions. Thus, $f_i \in C^1(\mathbb{R}^3)$, implying that there exists a unique solution to the system for any initial condition $(S(0), I(0), R(0))$ for all t.

To show that the solutions of (2.4.1) are non-negative, assume that the initial conditions are non-negative, i.e., $S(0) \geqslant 0$, $I(0) \geqslant 0$, $R(0) \geqslant 0$. Setting $S = 0$ in (2.4.1) gives $S' = 0$. Because the initial condition is non-negative and because $S' = 0$ at $S = 0$, we have that S cannot cross into the negative region. Thus, S is non-negative for all $t \geqslant 0$.

Setting $I = 0$ we get $I' = 0$. Because our initial condition is non-negative and because $I' = 0$ at $I = 0$, we have that I cannot cross into the negative region. Thus, I is non-negative for all $t \geqslant 0$.

Lastly, setting $R = 0$ we get $R' = \gamma I$. Since I is non-negative, $R' \geqslant 0$, and hence $R(t)$ is an increasing function, implying that $R(t)$ is non-negative for all $t \geqslant 0$.

To determine the evolution of the total population, we sum the equations in (2.4.1) and get

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

Integrating $N(t)$ with respect to t we get that $N(t) = N$, where N is a constant. Thus, the total population is constant, and for any t we have that

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

System (2.4.1) is well-posed since the solutions exist and are unique, remain non-negative for

non-negative initial conditions for all $t \geqslant 0$, and the total population size is constant.

2.4.2 Disease free equilibrium (DFE)

We need to calculate the equilibrium points [41]. The equilibrium points are found when we set the equations in (2.4.1) equal to zero, giving:

$$0 = -\beta SI, \tag{2.4.3a}$$

$$0 = \beta SI - \gamma I, \tag{2.4.3b}$$

$$0 = \gamma I. \tag{2.4.3c}$$

From $0 = \gamma I$, we conclude that $I^* = 0$. Setting $I^* = 0$ in $0 = -\beta SI$, we find that S^* is arbitrary, with $0 \leqslant S^* \leqslant S(0)$. This restriction is because S is monotonically decreasing; thus, $S(0)$ is the maximum.

From the above, we have that the SIR model without demography has a continuum of equilibria, which is called the disease-free equilibrium:

$$DFE = (S^*, 0, N - S^*) \quad (0 \leqslant S^* \leqslant S(0),\ R^* = N - S^*).$$

Theorem 2.10 ([189])**.** *The disease-free equilibrium of* (2.4.1) *is locally stable but not locally asymptotically stable.*

Understanding the stability of the DFE is crucial for public health interventions. A locally stable DFE suggests that the disease can be eradicated through measures that keep the number of infected individuals are low. However, the DFE is not locally asymptotically stable, meaning the system may return to the DFE after small perturbations, but it does not necessarily stay there over time. Consequently, even a small reintroduction of infected individuals can cause the system to deviate from the DFE.

This result implies that while the disease can be controlled to remain at zero infections under small perturbations, larger disturbances might lead to its re-emergence. Therefore, understanding the dynamics around the DFE helps in designing effective control strategies and anticipating potential resurgences of infections.

2.4.3 The basic reproduction number

The basic reproductive number, $\mathcal{R}_0$, serves as a critical metric in understanding the potential spread of infectious diseases within a population [41]. Defined as the expected number of secondary infections resulting from the introduction of a single infectious individual into an entirely

susceptible population, $\mathcal{R}_0$ aids in forecasting the trajectory of an infectious outbreak.

If $\mathcal{R}_0 < 1$, each infectious individual infects, on average, less than one person, indicating that the epidemic is quite likely to go extinct. Conversely, if $\mathcal{R}_0 > 1$, each infectious individual infects, on average, more than one person, suggesting that an epidemic is quite likely to occur.

To determine $\mathcal{R}_0$, we use the next-generation matrix method [21]. For this, we make a distinction between new infections and the other changes in the population. To use the method of [21], consider an infectious disease transmission model and let $\mathcal{I} \in \mathbb{R}^n$ be the $n \times 1$ matrix of infected compartments, $\mathcal{S} \in \mathbb{R}^m$ be the $m \times 1$ matrix of susceptible compartments, and $\mathcal{R} \in \mathbb{R}^k$ be the $k \times 1$ matrix of recovered or immune compartments. Consider D, a $m \times m$ diagonal matrix where the entries $\sigma_i > 0$ are the relative susceptibilities of the corresponding susceptible class, and assume that $\sigma_1 = 1$. Let Π be an $n \times m$ matrix where $\Pi(i, j)$ represents the fraction of the jth susceptible compartment that goes into the ith infective compartment on becoming infected. Let b be a $1 \times n$ matrix of relative horizontal transmissions. The general incidence function, $\beta(\mathcal{I}, \mathcal{S}, \mathcal{R}) = \beta$, depends on the population in the infective population compartment and the total population size. Let $\mathcal{V}$ be an $n \times n$ matrix that describes the transmissions between the infected compartments, including removals. Let $\mathcal{W}$ be a $k \times n$ matrix with entries $\mathcal{W}(i, j)$ representing the rate at which members of the jth disease compartment go into the ith removed compartment upon recovery.

Hence, the model has the following form:

$$\mathcal{I}' = \Pi D \mathcal{S} \beta b \mathcal{I} - \mathcal{V}\mathcal{I} \tag{2.4.4a}$$

$$\mathcal{S}' = -D\mathcal{S}\beta b\mathcal{I} \tag{2.4.4b}$$

$$\mathcal{R}' = \mathcal{W}\mathcal{I}, \tag{2.4.4c}$$

wherethe intial conditions are non-negative, i.e., $\mathcal{I}_0, \mathcal{S}_0, \mathcal{R}_0 \geqslant 0$ such that at least one component of $\mathcal{I}_0$ is positive.

The point $(0, \mathcal{S}^*, \mathcal{R}^*)$ can be a continuum of disease free equilibria. Let $\mathcal{F} = \Pi D \mathcal{S}^* \beta b$ be an $n \times n$ matrix. Then the reproductive number is

$$\mathcal{R}_0 = \rho(\mathcal{F}\mathcal{V}^{-1}) = \beta b \mathcal{V}^{-1} \Pi D \mathcal{S}^*,$$

where ρ is the spectral radius of the matrix $\mathcal{F}\mathcal{V}^{-1}$.

Using this method we can find the reproductive number of (2.4.1). This system has one infected component, one susceptible component, and one immune component. The DFE takes the form $(0,\ S^*,\ N - S^*)$. We compute $\mathcal{R}_0$ as follows:

$$\mathcal{R}_0 = \beta b_{1\times 1} V^{-1}_{1\times 1} \Pi_{1\times 1} D_{1\times 1} S^* = \frac{\beta}{\gamma} S^*.$$

2.5 SIR with demography

This section explores SIR model with demography, which is very similar to the model from Section 2.4 except that now demographical processes are taken in account. We will see that this system has a very different behaviour then the previously shown one. We assume that the disease is not transmitted vertically from parent to offspring; all newborns are considered susceptible. The constant birth rate is denoted by b, and all compartments (Susceptible, Infected, Recovered) experience natural death at rate d.

An extended version of this SIR model incorporating additional complexities is employed in Chapter 4 and Chapter 5.

We have that β denotes the transmission parameter, γ the per capita rate of recovery. The model is:

$$\frac{dS}{dt} = b - \beta SI - dS \tag{2.5.1a}$$
$$\frac{dI}{dt} = \beta SI - \gamma I - dI \tag{2.5.1b}$$
$$\frac{dR}{dt} = \gamma I - dR, \tag{2.5.1c}$$

with initial conditions $S(0)$, $I(0)$, $R(0) > 0$ and we assume that $S(0) + I(0) + R(0) > 0$ to avoid the trivial case.

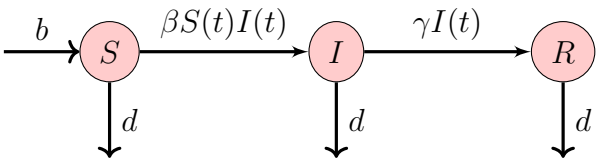


Figure 2.2: Transfer diagram for the SIR model with demography.

2.5.1 Well-posness of the model

This section establishes the well-posedness of the model, ensuring the existence, uniqueness, non-negativity, and boundedness of solutions.

Existence and uniqueness of the solution: Let $x(t) = (S(t), I(t), R(t))^T \in \mathcal{R}^3$. The system can be written as $\frac{dx}{dt} = f(x)$ where the components of f are

$$f_1 = b - \beta SI,$$

$$f_2 = \beta SI - \gamma I - dI,$$
$$f_3 = \gamma I - dR.$$

We have that f_i ($i = 1, 2, 3$) are continuous functions on $\mathbb{R}^3$. Furthermore, $\frac{\partial f_i}{\partial S}$, $\frac{\partial f_i}{\partial I}$, $\frac{\partial f_i}{\partial R}$ exists, which gives us that f_i are differentiable and continuous functions. Thus $f_i \in C^1(\mathbb{R}^3)$ which implies that there exists a unique solution to the system for any initial condition ($S(0)$, $I(0)$, $R(0)$).

Nonnegativity: Given initial conditions $S(0) \geqslant 0$, $I(0) \geqslant 0$, $R(0) \geqslant 0$ to (2.5.1), the solutions of (2.5.1) $S(t)$, $I(t)$, $R(t)$ are all nonnegative for all $t \geqslant 0$.

To show this assume that the initial conditions are nonnegative, i.e., $S(0) \geqslant 0$, $I(0) \geqslant 0$, $R(0) \geqslant 0$. Setting $S = 0$ in system (2.5.1) gives $S' = b > 0$. Because our initial condition is nonnegative and because $S' = 0$ at $S = 0$ we have that S cannot cross into the negative region. Thus S is nonnegative for all $t \geqslant 0$.

Setting $I = 0$, we get $I' = 0$. Because the initial condition is nonnegative and $I' = 0$ at $I = 0$, we have that I cannot cross into the negative region. Thus S is nonnegative for all $t \geqslant 0$.

Setting $R = 0$, we get $R' = \gamma I$. Since I is nonnegative, it follows that $R' \geqslant 0$ and hence $R(t)$ is an increasing function which implies that $R(t)$ is nonnegative for all $t \geqslant 0$.

Boundedness: To get the evolution of the total population, we sum the equations in 2.5.1 and get

$$\frac{dN}{dt} = \frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt} = b - dN(t).$$

Integrating $N(t)$ with respect to t we get that $N(t) = \frac{b}{d} + ke^{-dt}$, where k is a constant. We know that e^{-dt} is a decreasing function and, also a decaying function, since $e^{-dt} \to 0$ as $t \to \infty$. The maximum of e^{-dt} is 1. Thus

$$N(t) = \frac{b}{d} + ke^{-dt} \leqslant \frac{b}{d} + k.$$

System (2.5.1) is well-posed since the solutions exist and are unique, remain nonnegative for nonnegative initial conditions for all $t \geqslant 0$ and the total population size is bounded.

2.5.2 Disease free equilibrium (DFE)

We need to calculate the equilibrium points, which are found when we set the equations in (2.5.1) equal to zero, giving:

$$\begin{aligned} 0 &= b - \beta SI - dS, \\ 0 &= \beta SI - \gamma I - dI, \\ 0 &= \gamma I - dR. \end{aligned} \tag{2.5.2}$$

Let (S^*, I^*, R^*) be the equilibrium point. The system 2.5.1 is at a DFE if it is at an equilibrium and $S^* = 0$. Thus the DFE is (b/d, 0, 0). For the existence of S^*, we require $d > 0$.

2.5.3 Endemic equilibrium

The endemic equilibrium occurs when the system reaches a steady state while maintaining a nonzero infected population $\bar{I} \neq 0$. To find this equilibrium, we set the time derivatives in (2.5.1) to zero.

First, from the equation for I, we have:

$$\beta SI - \gamma I - dI = 0 \implies I(\beta S - \gamma - d) = 0.$$

Since we are considering the endemic case where $\bar{I} \neq 0$, this implies:

$$\bar{S} = \frac{\gamma + d}{\beta}$$

Next, substituting $\bar{S}$ into the equation for S, we solve for $\bar{I}$:

$$b - \beta SI - dS = 0 \implies \bar{I} = \frac{1}{\gamma + d}\left(b - \frac{d}{\beta}(\gamma + d)\right).$$

Similarly, solving for $\bar{R}$, we obtain:

$$\bar{R} = \frac{\gamma}{d(\gamma + d)}\left(b - \frac{d}{\beta}(\gamma + d)\right)$$

Thus the endemic equilibrium is

$$(\bar{S},\, \bar{I},\, \bar{R}) = \left(\frac{\gamma + d}{\beta}, \frac{1}{\gamma + d}\left(b - \frac{d}{\beta}(\gamma + d)\right), \frac{\gamma}{d(\gamma + d)}\left(b - \frac{d}{\beta}(\gamma + d)\right)\right).$$

2.5.4 The basic reproduction number

To find the reproduction number we use the same method presented in 2.4.3. System (2.5.2) has one infected component, one susceptible component, and one immune component. The method outlined in [21] re-orders the compartments with the infected compartment first, followed by the susceptible compartment and the immune compartments. Thus the DFE takes the form

$$\left(\frac{b}{d},\, 0\,, 0\right).$$

We compute $\mathcal{R}_0$ as follows:

$$\mathcal{R}_0 = \beta b_{1\times1} V_{1\times1}^{-1} \Pi_{1\times1} D_{1\times1} S^*,$$

where

$$\begin{aligned}
&\beta \text{ a constant} \\
&b_{1\times1} = 1 \\
&V_{1\times1} = (d+\gamma) \\
&\Pi_{1\times1} = 1 \\
&D_{1\times1} = 1.
\end{aligned}$$

Thus

$$\mathcal{R}_0 = \beta \times 1 \times \frac{1}{d+\gamma} \times 1 \times \frac{b}{d} = \frac{\beta}{(d+\gamma)} S^*.$$

For $\mathcal{R}_0$ to exists we require $\gamma > 0$ because γ represents the recovery rate of infected individuals and ensures a finite infectious period and a biologically meaningful expression for $\mathcal{R}_0$.

If $\mathcal{R}_0 < 1$, then the DFE is locally asymptotically stable and if $\mathcal{R}_0 > 1$, then the endemic equilibrium is locally asymptotically stable.This conclusion follows from the method in [21].

While it can be shown that when $\mathcal{R}_0 < 1$ all solutions converge to the disease-free equilibrium, implying global asymptotic stability, this result is not used in this thesis and is therefore omitted. Similarly, in many epidemiological models, all solutions tend to an endemic equilibrium when $\mathcal{R}_0 > 1$, indicating global asymptotic stability of the endemic equilibrium. However, since this is not relevant to our analysis, we do not provide further details here.

2.6 Network theory

The source for this section is [65].

Social network analysis (SNA) is an analytical tool used to map and measure social relationships. This multidisciplinary field incorporates elements of social science, mathematics, statistics, and computer science. In this context, a network refers to a collection of interconnected entities, called **nodes** or **vertices**, with relationships or interactions, called **edges**, between them. These relationships can represent various phenomena, such as social connections between individuals, protein interactions in biological systems, or transportation routes between cities. Networks can be represented by graphs, and the term is often used to describe complex systems with a network structure, emphasizing the real-world applicability of the concept [143].

This framework and these concepts will be used extensively in Chapter 6

Networks can be categorized based on whether they are oriented (directed) or not, and whether they are weighted or not. These characteristics play a crucial role in defining the structure and behaviour of the network:

1. **Orientation**:

   - In a **directed network** (or **digraph**, for directed graph), the relationships or connections between nodes have a specific direction. This means that existence of an edge from node A to node B does not imply existence of one from B to A. Directed networks are used to model asymmetric relationships and dependencies in various contexts, such as social networks with follower-followee relationships or information flow in a network.
   - In an **undirected network**, the relationships between nodes are symmetric, meaning if there is a connection between node A and node B, there is a connection between B and A as well. Undirected networks are commonly used to represent mutual relationships or symmetric interactions, like friendship in a social network or physical connections in a transportation system

2. **Weight**:

   - In a **weighted network**, each edge is assigned a weight or value that quantifies the strength, significance, or distance associated with the relationship between nodes. These weights can represent various attributes, such as the cost of travel between cities, the intensity of communication between individuals, or the similarity between items in a recommendation system. Weighted networks provide a more nuanced representation of relationships.
   - In an **unweighted network**, all edges are treated as having equal importance, with no assigned weights. This simplifies the network structure, and connections are typically represented as binary, indicating only the presence or absence of a relationship.

2.6.1 Network level properties

Several fundamental metrics provide valuable insights into the structure and connectivity of networks. The **number of nodes** is the count of nodes or points within the graph, while the **number of edges** is the quantity of connections or links linking these vertices. The **density** of a graph is the ratio of the actual number of edges to the maximum possible number of edges. The **mean degree** is the average node degree, computed as the total sum of degrees divided by the number of nodes. The **girth** is the length of the shortest cycle in the graph, where a *cycle* is a closed path

that starts and ends at the same node without repeating any edges or intermediate nodes. The **degree** of a node refers to the number of edges connected to it; in directed graphs, this includes both in-degree and out-degree. The **diameter** is the longest shortest path between any two nodes. The **global efficiency** measures how efficiently information or resources can traverse the entire network. The **average path length** is the mean shortest path length between all node pairs. In directed graphs, **reciprocity** measures the proportion of directed edges that are reciprocated. Beyond these vertex and network-level properties, it is also important to consider **diameter**, which is the maximum value of eccentricity, where eccentricity is defined as the greatest distance between a given node and any other node in the network. These metrics collectively provide a comprehensive understanding of network structure and behaviour.

Connectedness describes the extent to which nodes in a network can reach one another. A network is **connected** if, for every pair of nodes, there exists a path linking them. If no such path exists for at least one pair, the network is **disconnected**. In directed graphs, connectedness is classified into two types. A network is *weakly connected* if replacing all directed edges with undirected ones results in a connected graph. It is *strongly connected* if every node has a directed path to every other node. A **strongly connected component** is a maximal subnetwork where each node can reach every other node via directed paths. If a network consists of multiple strongly connected components, it is disconnected.

The **connectedness measure** quantifies the degree of disconnection by assessing the proportion of node pairs that are not mutually reachable relative to the maximum possible number of such pairs [106]. It is defined as:

$$C_K = \frac{D}{n(n-1)/2},$$

where n is the number of nodes, D is the number of pairs of nodes that are not mutually reachable and $n(n-1)/2$ is the maximum number of nodes unable to reach another point in the graph. Connectedness scores range from 0 for a fully disconnected network to 1 for a completely connected one.

2.6.2 Centralities

In a network, nodes can offer valuable insights into the network structure and dynamics. It is important to note that influence and importance lack precise definitions; to operationalize these concepts in a network context, we require a mathematical framework. **Centrality** serves as an intuitive metric to describe the significance or influence of a vertex in the connected structure of a graph [140].

In graph theory, centrality defines the position of a vertex within a network. The question of "which vertex is more important in a network" is addressed through a function on the vertices of a

graph. The function's values are expected to generate a ranking that identifies the most important nodes. Various centrality measures capture different aspects of a network. Degree centrality considers the number of direct connections a node has. Closeness centrality is based on the inverse of the average shortest path distance from a node to all other nodes. Betweenness centrality quantifies how often a node appears on shortest paths between other nodes; and eigenvector centrality measures a node's influence based on the importance of its neighbours, computed as the components of the leading eigenvector of the network's adjacency matrix [24].

The most common centrality measures include degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality [177].

Degree Centrality is defined as the number of edges connected to a vertex, serving as a measure of immediate connection in a network. It is based on the concept of node degree from graph theory, where a node's degree reflects its number of direct links. When the network is directed, we distinguish between in-degree (edges directed toward the vertex) and out-degree (edges directed away from the vertex). Degree centrality extends this concept by interpreting a node's connectivity as an indicator of its influence or importance within the network.

Closeness centrality of a vertex in a connected graph is defined as the inverse of the sum of the distances from the vertex to all other vertices. In other words, the normalized closeness of a vertex is the average length of the shortest path between the vertex and all other vertices. Closeness captures how efficiently the entire graph can be traversed from a given vertex, emphasizing the notion that a vertex is central if it is close to many other vertices.

Betweenness centrality is computed based on the shortest paths in a connected graph, considering the number of shortest paths that pass through a vertex while minimizing the number of edges or the sum of edge weights along the path. A high betweenness centrality indicates that a vertex has significant control over the network and plays a crucial role in maintaining overall connectivity. If such a vertex is removed from the network, the risk of overall disconnection increases.

Vertices with high betweenness centrality can be considered as super-connectors. These super-connectors are essential for the social integration of new entrants into the network, but their removal could pose a higher risk of connective disruption. Identifying super-connectors involves finding vertices with high betweenness centrality.

2.6.3 Communities

A network exhibits a **community structure** when its vertices can be grouped into densely connected subsets with comparatively fewer connections between these subsets. In practice, vertices may belong to multiple communities.

Community detection is crucial for simplifying network analysis, allowing communities to be treated as meta-vertices with distinct properties. Identifying these structures is especially relevant in applications such as disease spread modelling [141].

Detecting communities is challenging due to unknown community numbers and varying sizes. Common approaches fall into two categories: (a) **Agglomerative Methods**, which start with isolated vertices and iteratively add edges, and (b) **Divisive Methods**, which begin with a complete graph and remove edges to reveal communities [158].

Louvain Algorithm The Louvain algorithm [37] optimizes modularity, a measure comparing actual vs. expected edges within communities. It iteratively assigns vertices to communities, merging them into meta-vertices until modularity reaches a local maximum. A limitation is that it may form weakly connected communities, and the final partitioning can depend on the order of vertex updates [184].

Leiden Algorithm To address Louvain's weaknesses, the Leiden algorithm [183] improves connectivity by refining partitions after each iteration. It consists of three steps: (1) moving vertices to maximize modularity, (2) refining partitions by splitting weakly connected groups, and (3) aggregating the network based on refined communities. Leiden guarantees better-connected communities but increases computational complexity.

Infomap Algorithm Infomap [164] models community detection as a data compression problem, minimizing the description length of a random walk across the network. It assigns vertices to communities based on the frequency of visits in a simulated walk. This method effectively captures hierarchical structures but can be sensitive to parameter tuning.

Leading Eigenvector Method This method splits the graph recursively using the leading eigenvector of the modularity matrix [142]. It is useful for detecting well-separated clusters but cannot split tightly connected groups.

Clique-Based Methods A **clique** is a subset of vertices where each pair is connected. Maximal cliques cannot be expanded further, while maximum cliques contain the most vertices among all

cliques. Clique-based community detection methods define communities as unions of overlapping cliques, often using the Bron-Kerbosch algorithm or k-clique percolation [68].

2.7 Bayesian inference

The concepts introduced in this section are applied in Chapter 4 and Chapter 5 for parameter estimation.

Mathematical models rely on parameter estimation to achieve the best fit for observed data. Two primary approaches exist: Bayesian inference and frequentist methods, with the latter using optimization techniques.

Bayesian inference uses Bayes' Theorem to estimate parameters, incorporating both observed and unobserved data within a joint probability framework that includes prior distributions [188]. In contrast, frequentist inference relies solely on observed data. A key challenge in Bayesian inference is the lack of a universal standard for choosing prior distributions, which can introduce subjectivity and variability in results across researchers [83]. However, this subjectivity also enables **sensitivity analysis**, where different plausible priors are tested to examine how sensitive the model's conclusions are to those choices. This helps assess the robustness of the findings and is especially useful when prior knowledge is uncertain or limited.

Bayesian inference follows three main steps: (1) determining the prior distribution through data collection or literature review, (2) evaluating the likelihood function based on observed data, and (3) applying Bayes' Theorem to derive the posterior distribution.

Formally, given observed data $\mathbf{y}$ and parameters $\theta = (\theta_1, \ldots, \theta_n)$, the likelihood of the observed data given the parameters is denoted as $p(\theta \mid \mathbf{y})$. Applying Bayes' Theorem yields the posterior distribution, $\pi(\theta, |, \mathbf{y})$, expressed as:

$$p(\theta \,|\, \mathbf{y}) = \frac{p(\mathbf{y} \,|\, \theta)\pi(\theta)}{p(\mathbf{y})} \propto L(\theta \,|\, \mathbf{y})p(\theta) = \pi(\theta \,|\, \mathbf{y}), \tag{2.7.1}$$

where $p(\theta \,|\, \mathbf{y})$ is the posterior distribution, $L(\theta \,|\, \mathbf{y})$ is the likelihood function, $\pi(\theta \,|\, \mathbf{y})$ is the unnormalized posterior distribution, and $p(\mathbf{y})$ is the prior distribution.

A key advantage of Bayesian inference is its adaptability to new data, unlike frequentist methods, which require predefined experimental designs. However, frequentist inference provides a single optimal estimate without quantifying parameter uncertainty, which Bayesian analysis inherently captures [83].

2.7.1 Sampling Techniques - Markov chain Monte Carlo

Once the model and likelihood functions have been established, the next task involves fitting the observed data to the model to derive parameter estimates. The primary aim is to gauge the complete posterior distribution, often summarized through the posterior mean, median, and credible intervals. However, it is important to note that the exact calculation of the posterior distribution is typically unattainable due to its intricate and multi-dimensional nature.

A method employed for inferring the posterior distribution is **Markov chain Monte Carlo (MCMC)** [76].

MCMC provides a computational approach for posterior inference through simulation. The method combines two key concepts: Markov chains, where the probability of the next state depends only on the current state

$$P(X_t|X_{t-1}, \ldots, X_1) = P(X_t|X_{t-1}),$$

and Monte Carlo integration for estimating complex integrals through sampling.

MCMC constructs a Markov chain that converges to the target posterior distribution. Once the chain reaches its stationary distribution, samples can be drawn to estimate posterior statistics. While MCMC remains the most widely used algorithm in Bayesian inference, traditional methods like Metropolis-Hastings face significant limitations, particularly regarding convergence assessment and the requirement for extensive sampling to ensure reliable approximation of the target distribution. [51, 188]

Traditional MCMC algorithms encounter substantial difficulties when dealing with nonidentifiability. Nonidentifiability of model parameters refers to the situation when infinitely many values of parameters can produce the best model fit of the data, while different choices of best-fit parameter values may lead to significantly different model predictions.This creates highly correlated, complex posterior surfaces that are challenging to sample efficiently. Nonidentifiability can be addressed through model simplification, advanced calibration algorithms, or by incorporating additional data sources [60].

The affine invariant ensemble MCMC algorithm, developed by Goodman and Weare [81], addresses key limitations of traditional MCMC methods, particularly for high-dimensional, multimodal, or strongly correlated distributions. Rather than employing a single chain, this approach utilizes multiple "walkers" (separate Markov chains) that explore the parameter space cooperatively. Each walker proposes new positions based on both its current state and the collective positions of other walkers in the ensemble. This coordinated approach captures global distributional structure more effectively than traditional random-walk proposals. The algorithm maintains the standard Metropolis-Hastings acceptance criterion while leveraging ensemble information to generate more informed proposals.

Assessing performance of an MCMC

To evaluate the performance of the MCMC, trace plots are instrumental. These two-dimensional plots show the parameter values (on the y-axis) at each iteration (on the x-axis). Trace plots provide valuable insights into the mixing process of the Markov chain, which refers to the ability of the chain to explore the entire parameter space. If the chain fails to mix well, it can get "stuck" in certain regions, and as a result, it will not explore the parameter space effectively.

Poor mixing is indicated by trace plots where the chain shows little movement across the parameter space or oscillates between a few distinct values. In such cases, the MCMC may require more iterations to explore the space adequately. Figures 2.3a and 2.3b illustrate two scenarios of mixing: the first figure shows a trace plot with good mixing, where the parameter values fluctuate in an even manner, suggesting that the chain is exploring the parameter space efficiently; the second figure shows poor mixing, where the chain gets stuck in two distinct regions, demonstrating inefficient exploration.

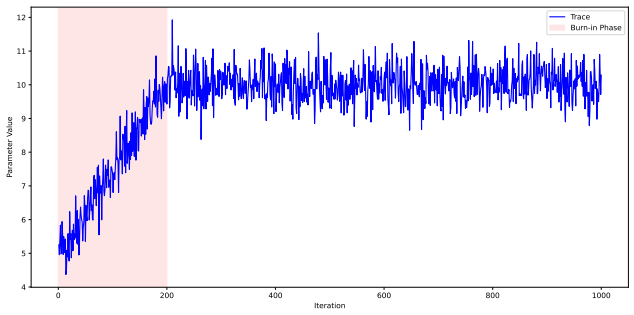
Trace plots also serve to determine whether the Markov chain has reached stationarity, whether the distribution of parameter values has stabilized. Before this point, the chain is still "burning in", meaning that the initial iterations do not represent the target distribution and must be discarded. This period is known as the burn-in phase, and the number of iterations required for burn-in depends on how quickly the chain reaches stationarity.

To quantify convergence, the $\hat{R}$-statistic (also known as the Gelman-Rubin statistic) is commonly used. The $\hat{R}$-statistics compares the variance between chains (how different the chains are from each other) to the variance within each chain (how consistent the parameter estimates are within a single chain). This statistic compares the variance within each chain to the variance between chains. When $\hat{R}$ approaches 1, it indicates that the chains have converged to the same distribution. Values significantly greater than 1 suggest that the chains have not yet converged. Figures 2.4a and 2.4b show how the $\hat{R}$-statistic behaves under different mixing conditions. The first figure, Figure 2.4a shows that $\hat{R}$ starts at 1.5 during the burn-in phase and gradually decreases to 1, indicating that the chains have converged to the stationary distribution. The second figure, Figure 2.4b, however, illustrates poor mixing: the $\hat{R}$ value remains high and fluctuates, indicating that the chains have not converged, and the sampling process is inefficient.

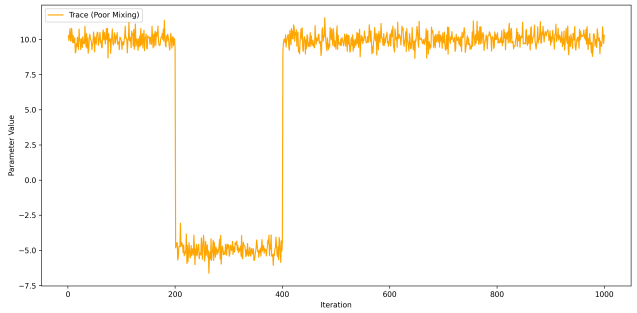
2.8 Natural language processing and large language models

The following concepts are applied in Chapter 3.

Natural Language Processing (NLP) is a field of artificial intelligence (AI) that focuses on

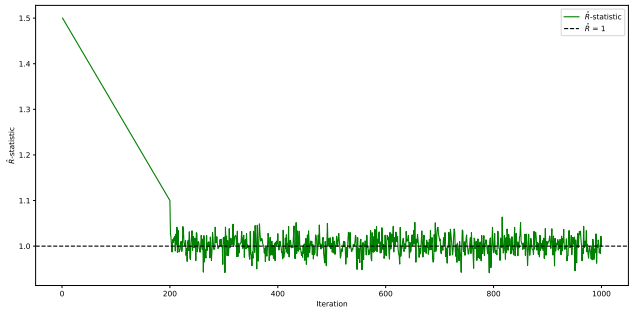


(a) Good mixing of parameter values over 1000 iterations, with burn-in phase (shaded red) discarded.

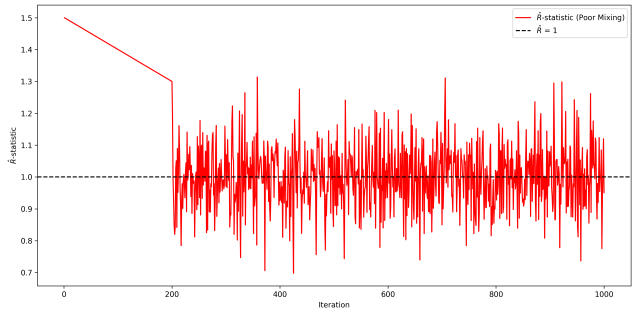


(b) Poor mixing of parameter values, showing the chain stuck in two regions. The burn-in phase is not shown here to emphasize the persistent poor mixing throughout the chain

Figure 2.3: Comparison of MCMC trace plots illustrating (a) good mixing and (b) poor mixing of parameter values over 1000 iterations. The burn-in phase is shown only in the good mixing plot to highlight stable convergence behaviour.



(a) Convergence of $\hat{R}$ starting at 1.5 and approaching 1, indicating good mixing. Burn-in phase is included and visible here.



(b) $\hat{R}$ values fluctuate above 1, indicating poor mixing and convergence. Burn-in phase is not shown here, which can affect interpretation.

Figure 2.4: Comparison of $\hat{R}$-statistic plots illustrating (a) good mixing and convergence with visible burn-in phase, and (b) poor mixing with absence of burn-in phase visualization. The burn-in phase is typically discarded to improve chain convergence assessment, but it is not shown in the poor mixing plot, which may obscure early convergence issues.

enabling computers to understand, interpret, and generate human language in a way that is both meaningful and contextually appropriate [98]. Significant advancements in NLP have been driven by improved computational power, linguistic theories, and machine learning techniques [55]. A key breakthrough in recent years is the development of large language models (LLMs).

The roots of NLP trace back to the work of Alan Turing in the 1950s and 1960s [98], who laid the groundwork for the field of machine intelligence with his seminal contributions "Computing Machinery and Intelligence" [186]. Early efforts were focused on rule-based systems that utilized predefined grammatical rules and dictionaries to process text [98].

With the emergence of transformer models like BERT [64], GPT [159], and T5 [160], NLP has seen breakthroughs in attention mechanisms, enabling the models to better capture long-range dependencies and achieve state-of-the-art performance across a variety of tasks.

NLP encompasses a diverse array of tasks, ranging from fundamental syntactic and semantic analysis to more intricate language understanding and generation capabilities. Here are some examples:

- Text Understanding and Information Extraction: Algorithms for tasks such as named entity recognition, sentiment analysis, and part-of-speech tagging [124].
- Text summarisation (see below)
- Machine Translation: Development of systems for automatic translation of text between languages, leveraging statistical and neural machine translation approaches [32].
- Speech Recognition: Utilizing algorithms to convert spoken language into written text, a pivotal technology in voice-activated systems and virtual assistants [98].

Text summarization, a critical NLP task, has evolved from statistical methods to advanced machine learning techniques. It is categorized into extractive summarization, which selects key sentences from a text, and abstractive summarization, which generates new sentences that capture the essence of the original document [165].

The rise of LLMs has transformed NLP by enabling models to learn intricate patterns from massive datasets, leading to breakthroughs in tasks such as language translation, question answering, and text generation [64]. These models excel at capturing long-range dependencies and understanding context, surpassing earlier models' limitations [64].

2.8.1 LLaMA

Meta AI's **LLaMA (Large Language Model Meta AI)**, introduced in February 2023 [182], has established itself in the LLMs field. It prioritizes both achieving state-of-the-art performance and promoting model efficiency.

At the core of LLaMA's training methodology is the process of aggregating vast datasets comprising publicly available text and code, totalling trillions of tokens (words) [182].

While praised for its commitment to open science principles, LLaMA's accessibility is subject to restrictions. While the code for running queries with the model is open, allowing researchers to interact with LLaMA. However, Meta AI controls access to the trained model weights, which hold the key information LLaMA has learned.

LLaMA's suite encompasses four models varying in size, ranging from 7 billion to 65 billion parameters. This diversity empowers researchers to select a model commensurate with their computational resources and task demands. Notably, the 13B parameter model has demonstrated superior performance compared to the considerably larger GPT-3 (175B parameters) across various NLP benchmarks, indicative of LLaMA's efficiency-driven approach yielding competitive outcomes even with modest model sizes [182].

2.8.2 Text classification

Text classification is a fundamental NLP task that enables machines to categorize text documents into predefined labels based on their content.

For example, in sentiment analysis, text classification helps determine whether a customer review expresses positive or negative sentiment [154]. It is also widely used in spam filtering, where classifiers analyze known spam characteristics to filter unwanted emails [17]. Additionally, text classification aids in topic labeling, automatically organizing documents into relevant topics to improve information retrieval [193].

Text classification can be approached through three main techniques:

1. **Supervised Learning**: This approach uses labeled data to train models like Naive Bayes, Support Vector Machines (SVM), and k-nearest Neighbors (KNN), which map text features to categories. It provides high accuracy with sufficient labeled data but can be costly and time-consuming to create the labels [123].

2. **Rule-based Classification**: This method uses manually crafted rules to classify text based on predefined features or keywords. It is flexible and fast for specific tasks but requires significant manual effort to create and maintain the rules, with accuracy depending on the rules' complexity.

3. **Unsupervised Learning**: In this approach, algorithms like Latent Dirichlet Allocation (LDA) identify patterns in unlabeled data, grouping similar texts together. While less accurate than supervised methods, it can uncover hidden topics and themes in large datasets, offering valuable insights without requiring labeled data [173].

3

Assisted literature review

Overview of the chapter: This chapter presents a novel approach to conducting a comprehensive literature review on COVID-19 epidemiological modelling. In academic research, literature reviews serve as a critical foundation for contextualizing a research problem, identifying gaps in existing knowledge, and developing a coherent theoretical framework. However, the unprecedented volume of publications on COVID-19-related topics poses significant challenges to traditional synthesis methods, which are often time-consuming, labour-intensive, and prone to selection bias. To overcome these barriers, we adopted a computational approach using natural language processing (NLP) techniques and automated data retrieval tools. Specifically, we utilized the Semantic Scholar API within a Python environment to systematically extract over 10,000 peer-reviewed journal articles relevant to metapopulation modelling, Bayesian inference, case infection dynamics, and variants of concern in the context of SARS-CoV-2. This large corpus necessitated the use of summarisation strategies to effectively distil, cluster, and interpret trends within the literature.

Ultimately, this chapter demonstrates how computational tools can enhance the efficiency, transparency, and scope of academic literature reviews, particularly in fast-moving fields like pandemic modelling, by enabling researchers to analyse thousands of documents.

3.1 Introduction

In the realm of academic research, a comprehensive literature review stands as a pivotal component, offering a critical analysis and synthesis of existing studies and literature on a specific subject. This process not only contextualizes the chosen topic but also identifies gaps, ultimately establishing

the theoretical framework for the ensuing study. Synthesizing and summarizing findings from a multitude of sources are integral aspects of this review, allowing for the creation of connections between diverse studies and the construction of a cohesive narrative.

For this assisted literature review, we utilized the Semantic Scholar API in Python to systematically gather relevant research articles. Our query was structured to capture key works in metapopulation modelling, Bayesian inference, modelling, case infection dynamics and variance of concern related to COVID-19 and SARS-CoV-2. Our query yielded an extensive collection of over 10,000 journal articles aligned with our search criteria. Faced with the task of synthesizing and summarizing such a voluminous corpus of over 10,000 articles, the question emerged: "How can we effectively distil this wealth of information?". Since this thesis has already taken a somewhat different direction than initially envisioned, for the literature review, we will be taking a novel route using data science methods. Recognizing the significant time and effort required to navigate through copious amounts of long-form information, we emphasize the utility of summarisation techniques.

To visualize the publication trends over time, we generated a time series plot representing the monthly frequency of publications from 1970 onwards. The time series plot in Figure 3.1 illustrates the growing volume of research in this field, highlighting a surge in mathematical modelling studies in recent decades. More specifically, we observed a sharp increase in publications from 2019 onward, corresponding to the COVID-19 pandemic.

Figure 3.2 presents the monthly frequency of COVID-19-related mathematical modelling studies. This dramatic rise reflects the intensified research efforts to understand, predict, and mitigate the spread of SARS-CoV-2. Notably, while the publication rate appears to decline in recent months, this trend may be influenced by data availability and indexing lags.

3.2 Flowchart

This subsection details the systematic process employed to conduct the literature review for this study. Refer to Section 2.8 for the methodology and an overview of Natural Language Processing (NLP). The flowchart below (Figure 3.4) summarizes the key steps involved.

1. Define Aims of Review: The initial step involved outlining the specific objectives for the review. The identified aims were to critically assess existing research on COVID-19 epidemiological modelling, to inform the development of novel modelling methodologies or interventions in the context of infectious diseases.

2. Identify Keywords: Following the outlined aims, a meticulous process was undertaken to identify relevant keywords and synonyms associated with the research topic. This involved

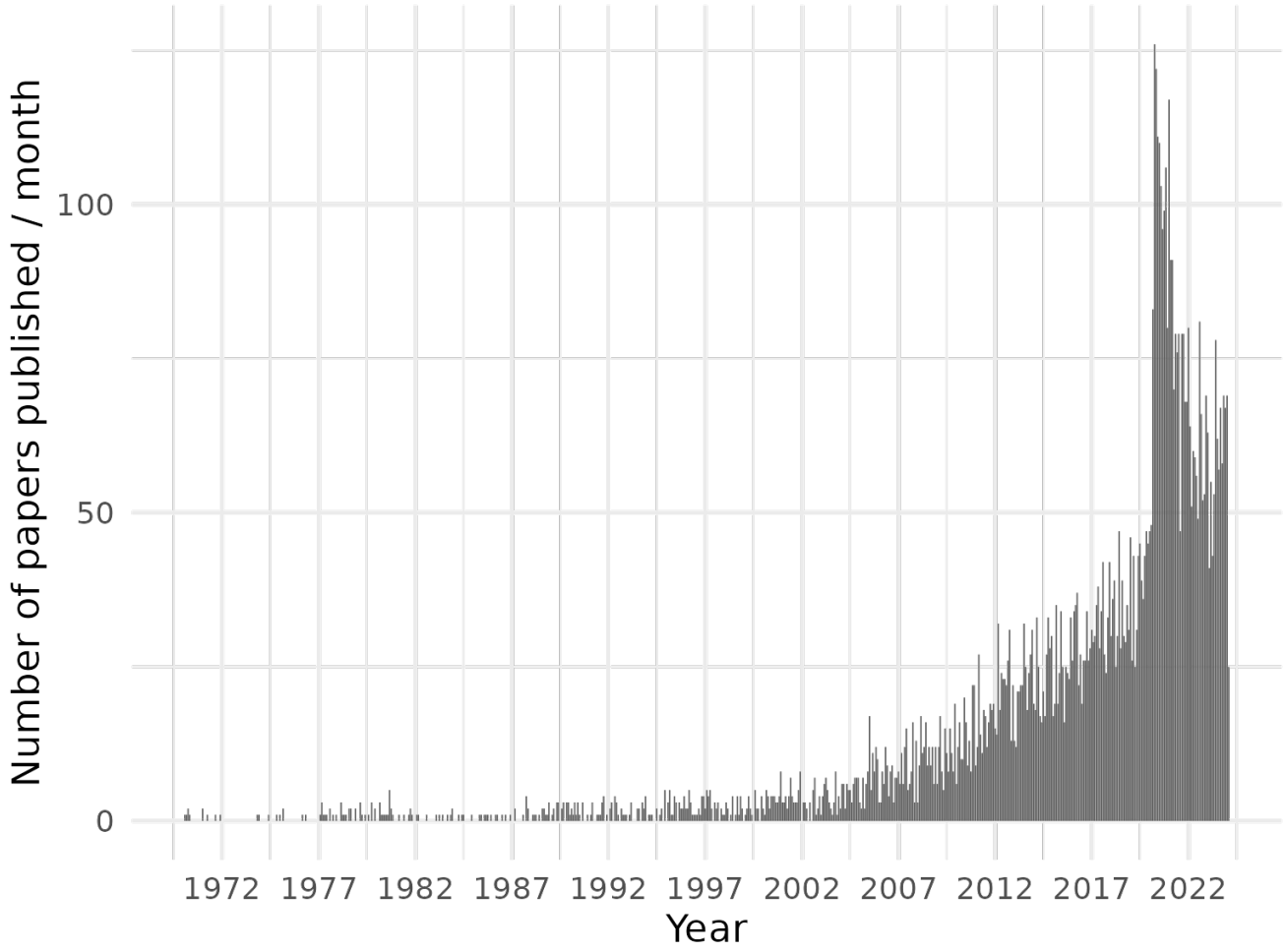


Figure 3.1: Number of articles published related to mathematical epidemiology since 1970.

Figure 3.2: Number of articles that have been published related to mathematical modelling of COVID-19 per month since January 2020.

consulting pertinent academic databases, and repositories, to compile an exhaustive list of keywords and employing Boolean operators (e.g., AND, OR, NOT) to refine the search strategy. This ensured a comprehensive yet targeted search strategy. For our review we have used the following:

```
(covid 19 metapopulation) | (sars cov 2 metapopulation) |
(covid 19 modeling) | (covid 19 modelling) |
(covid 19 bayesian inference) | (sars cov 2 modeling) |
(sars cov 2 modelling) | (sars cov 2 bayesian inference) |
(covid 19 case infection) | (sars cov 2 case infection) |
(covid 19 mathematical modeling) |
(sars cov 2 mathematical modeling) |
(sars cov 2 variance of concern) |
(covid 19 variance of concern)
```

3. Search for Articles: The Semantic Scholar API emerged as the platform of choice for searching scholarly articles based on the identified keywords. This selection was underpinned by several factors, including the platform's comprehensive coverage of scientific literature span-

ning diverse disciplines and features such as easy access to full-text articles and citation information, facilitating efficient retrieval of scholarly resources.

4. Classify Articles (PDFs): Text Rule-Based Classification (SIR-Type, Logistic Regression, etc.). Articles were classified based on model types using a text rule-based approach. Unaccessible or non-machine-readable PDFs or broken URLs were marked for potential revisitation.

5. Classify Articles (Abstracts): The same approach was applied to abstracts to ensure comprehensive categorization of articles based on model types. Based on the abstracts analyzed using our text classification system, the distribution of identified models is as follows:

   - Model not identified: 7245
   - SIR-Type/deterministic: 2014
   - Stochastic: 459
   - ARIMA: 174
   - Logistic Regression: 47
   - Random Forest: 41
   - Support Vector Machine: 12
   - Agent Based: 5

6. Final Classification: Classifications obtained from both PDFs and abstracts were amalgamated for each article, with precedence given to PDF classifications when available. Articles lacking classification from either source were categorized as "Model not identified".

   Following the two-stage classification process, the final distribution shown in Figure 3.3 reveals SIR-type/deterministic models represent the most prevalent approach among classified articles, accounting for approximately 41.2% of the identified models, followed by stochastic models at 5.1%. Machine learning approaches, while present, constitute a smaller proportion of the literature, with logistic regression, random forest, and support vector machine models representing less than 1% each of the total classified articles.

7. Manual Review: For articles classified as "Unclassified", manual review was initially considered, involving a researcher's assessment to determine relevance and potentially assign model type classification, particularly when relying solely on abstracts due to unavailable or unaccessible or non-machine-readable PDFs, which could be attributed to various factors such as changed URLs, articles behind paywalls requiring university library access, or technical issues encountered during retrieval using Python scripts. However, given the substantial number of

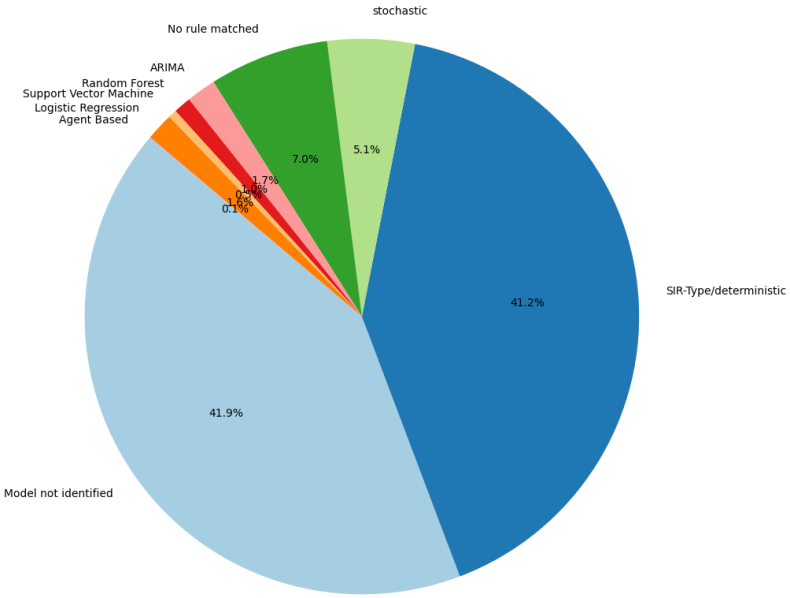


Figure 3.3: Pie chart showing the distribution of model types identified in the abstracts. The chart illustrates the prevalence of various modelling approaches, including deterministic models like SIR-type, stochastic models, and machine learning methods such as Logistic Regression, Random Forest, and Support Vector Machine.

relevant articles successfully identified and classified through the automated screening process, manual review of unclassified articles was deemed unnecessary for the scope of this research. The volume of papers captured through the classification methodology was considered sufficient.

8. Select Articles of Interest: Based on final classifications, a subset of articles was manually selected from each category. Random selection was applied within categories, with proportionally more articles selected from the SIR-type/deterministic model category given its greater relevance to the thesis objectives. This approach balanced representative coverage across model types with focused depth in the most pertinent methodological area.

   Limitations and Bias Acknowledgment: While systematic procedures were implemented to minimize bias throughout the literature review process, it is important to acknowledge that complete objectivity in literature selection and interpretation remains challenging. Potential sources of bias include the inherent focus on SIR-type models aligned with thesis objectives, and possible systematic biases in the automated classification algorithms employed.

9. Summarize Abstracts: Utilizing advanced Natural Language Processing (NLP) techniques, text summarisation was performed on selected articles' abstracts to distil key information for efficient review and analysis.

10. Improve Summary Quality and AI Enhancement: Following the Selection of Articles of Interest, artificial intelligence was re-employed to perform three distinct analytical tasks on the selected subset of articles. First, AI was utilized to enhance the quality and conciseness of abstract summaries, improving their readability and coherence while reducing length without compromising essential information. Second, the selected articles were systematically categorized into thematic clusters using AI-driven content analysis to identify and group articles by research focus and methodological similarities. Finally, AI was employed to extract and synthesize key findings from each selected article, generating comprehensive lists of principal conclusions and insights to facilitate systematic comparison and analysis across the literature subset.

We utilized a rule-based classification to automatically categorize the types of models used in our collection of journal articles. By employing regular expressions, we were able to match specific keywords and phrases within the articles, enabling us to classify them into distinct categories, including:

1. Logistic Regression: Articles mentioning "Logistic Regression" or variations with different spacings;
2. Random Forest: Articles mentioning "Random Forest" or variations with different spacings;
3. Support Vector Machine: Articles mentioning "Support Vector Machine" or variations with different spacings;
4. SIR-Type/Deterministic: Articles mentioning "SIR", "SIRD", "SEIR", "SLIR", or "deterministic model";
5. ARIMA: Articles mentioning "ARIMA";
6. Stochastic: Articles mentioning "stochastic";
7. Agent Based: Articles mentioning "Agent Based" or variations with different spacings.

While our utilization of rule-based classification proved effective for categorizing the types of models discussed in our collection of journal articles, it's important to acknowledge several limitations inherent to this approach [110]:

- Dependency on Predefined Rules: Rule-based classification relies heavily on explicitly defined rules, which must be carefully crafted and updated to accurately capture the nuances of the data. This dependency introduces the risk of overlooking important patterns or failing to adapt to changes in the dataset.

- Limited Adaptability: Rule-based systems may struggle to accommodate variations or new patterns not explicitly covered by the predefined rules. This lack of adaptability can result in misclassifications or incomplete categorization of articles, especially when dealing with complex or evolving domains.

- Challenges in Handling Ambiguity: Natural language is inherently ambiguous, and text data often contains synonyms, variations, or misspellings. Rule-based systems may face difficulty in handling such ambiguity, leading to inaccuracies in classification.

3.3 Literature Review

3.3.1 Mechanism for generating this literature review

The literature review, including major findings, below was generated with the assistance of large language models (LLMs), using techniques described in Chapter 3. While efforts were made to ensure accuracy and clarity, it is important to note that the interpretations and formulations presented here are ultimately shaped by the capabilities and limitations of AI technology at the time of writing.

The following sections organize the literature into thematic categories. While we have aimed to present the papers in a coherent and logical manner, it is important to note that some studies could fit into multiple themes. As such, despite our efforts to categorize them accurately, the division remains somewhat *ad hoc* and may not fully capture the complexity or multifaceted contributions of each study. The thematic structure presented here is intended to serve as a helpful framework for understanding the literature but should not be viewed as a definitive or exhaustive categorization.

We have made efforts to minimize bias in the generation of this review. However, it is necessary to acknowledge that the model's design could introduce certain biases or gaps in the selection and interpretation of sources. The papers included were selected based on their relevance to the topic; however, due to the automated nature of the process, some studies may not have been considered, or others may have been prioritized differently than they would have been by a human researcher.

3.3.2 Impact of non-pharmaceutical interventions (NPIs)

Since the onset of the outbreak in Wuhan, numerous modelling groups worldwide have delved into investigating how the implementation of Non-Pharmaceutical Interventions (NPIs) contributes to the reduction in confirmed cases. Nonetheless, predicting the severity of COVID-19, or any pandemic for that matter, and quantifying the impact of NPIs remain formidable challenges for both public health officials and mathematical modellers. Various models have been devised and

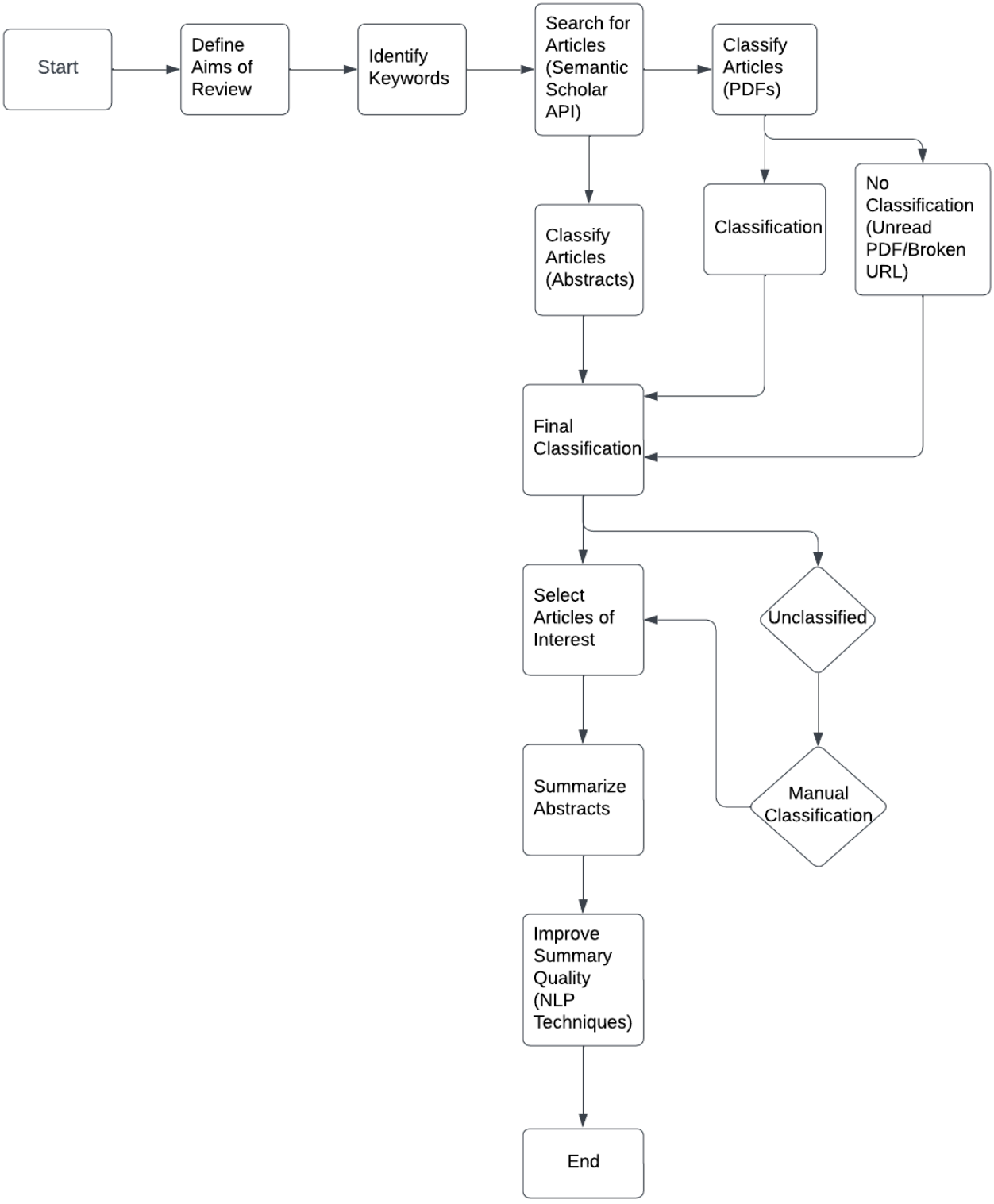


Figure 3.4: The Paper Pipeline: a systematic approach to finding, classifying, and summarizing research papers on COVID-19 epidemiological modelling. This process culminates in the comprehensive literature review presented in the next section.

applied to explore the effects of social distancing, testing, and isolation on the dynamics of COVID-19.

For example, Tuite et al. [185] conducted a study investigating the impact of fixed-duration interventions and dynamic interventions on the number of COVID-19 cases in Ontario, Canada. In another study by Mizumoto et al. [133], a mathematical model and time series incidence data were utilized to model the outbreak aboard the Diamond Princess Cruise Ship, revealing that quarantine measures implemented by the Japanese government led to a decrease in the effective reproductive number. Additionally, Hellewell et al. [84] employed a stochastic model to quantify the impact of contact tracing and isolation on the outbreak.

The role of mobility restrictions and social distancing in controlling COVID-19 was emphasized in several studies, for example, Badr et al. [30] used daily mobility data to derive a social distancing metric and assessed its impact on new infection rates across the 25 most affected US counties. Their findings underscored the crucial role of reduced mobility in decreasing COVID-19 case growth rates.

Ko et al. [103] model the impact of NPIs and vaccination during the Delta variant outbreak in South Korea. Their study finds that younger age groups have higher transmission rates and emphasizes the importance of timely and intense NPIs, vaccination speed, and screening to prevent new epidemic waves. Similarly, Kraay et al. [105] examine the timing and conditions for safely relaxing non-pharmaceutical interventions (NPIs) in the U.S. as vaccination efforts increase. They find that vaccination can reduce deaths and health system burdens, but the speed of vaccine rollout is crucial for determining when NPIs can be safely relaxed. Their analysis supports a two-dose vaccination strategy, with a recommended initial delay of at least three months before easing restrictions.

Alternative strategies to traditional quarantine were explored by Foncea et al. [70], who proposed an alternative to quarantine by proposing periodic testing for COVID-19 contacts as a mitigation strategy. Through an analysis of data from over 150,000 individuals in South America, they demonstrate that periodic testing reduces transmission risk by 84%, nearly matching the effectiveness of quarantine while mitigating the significant social and economic impacts associated with it. Further advancing this field, Kucharski et al. [107] also evaluated testing, isolation, and contact tracing strategies in the UK, concluding that combined isolation and tracing were more effective than mass testing alone.

Adapting models to specific sociocultural contexts is vital for optimizing intervention strategies. Ardila et al. (2020) [20] highlights the importance of mathematical models for managing COVID-19, with emphasis on adapting models to sociocultural contexts to improve public health measures. Additionally, interventions can have broader benefits, as Ting Xu et al. (2021) examines the impact of COVID-19 control measures on air quality in Taiyuan, China. It uses Gray Relational Analysis (GRA) and an improved seagull optimization algorithm combined with Support Vector Regression

(SVR) to predict Air Quality Index (AQI) with high accuracy [206].

Major Findings:

- Early and rigorous control measures significantly mitigate spread.
- Combined isolation and contact tracing more effective than mass testing alone.
- Social distancing metrics strongly correlate with decreased case growth.
- Quarantine alternatives like periodic testing can be similarly effective while reducing socioeconomic impact.

3.3.3 Stochastic and Bayesian approaches

Stochastic and Bayesian modeling techniques have emerged as powerful tools for understanding the dynamics of COVID-19. These approaches address the inherent uncertainty and variability in epidemiological data, offering robust frameworks for parameter estimation, trend analysis, and decision-making.

Bayesian methods have proven effective in addressing data limitations and providing improved estimates in various contexts. Furstova et al. [74] propose a Bayesian approach to seroprevalence studies, allowing for more accurate estimates of condition prevalence based on binary tests. By leveraging Bayes' Theorem, the authors address potential limitations and counterintuitive outcomes in test results, providing practical examples of their method.

Kumar et al. [109] analyze COVID-19 trends using a spline-based time series model within a Bayesian framework. Their research, published in the Japanese Journal of Statistics and Data Science, demonstrates that incorporating seasonal components improves the model's accuracy over traditional time series approaches, offering better insights for predicting COVID-19 trends.

Bayesian approaches have also been applied to patient outcomes and viral mutations. Yanuar, Deva, and Maiyastri [209] focused on modeling the length of hospital stay for COVID-19 patients in West Sumatra using quantile regression and Bayesian quantile approaches. Their findings indicated that the Bayesian approach outperformed the standard quantile regression method, offering a better fit with narrower confidence intervals. Key factors influencing hospital stay duration included age, diagnosis status, and discharge status. Zhao et al. [211] used a Bayesian phylogenetic approach to link the D614G substitution in the SARS-CoV-2 spike protein with increased transmissibility, providing insights into viral mutations' impact. Dehning et al. (2020)[62] paper discusses Bayesian inference in SIR models to estimate the reproduction number of SARS-CoV-2, addressing data limitations and cross-validating results with alternative data sources. The article "Parametric Modeling Approach to Covid-19 Pandemic Data" by Badmus et al. (2021) Stochastic models and advanced parametric approaches have also played a vital role in analyzing COVID-19

data. [29] introduces the Extended Rayleigh Lomax distribution for modeling skewed survival data. Derived using the beta logit function, the authors present key statistical properties like probability density functions, cumulative density functions, and moment-generating functions. Through maximum likelihood estimation, they compare their proposed distribution against other models, using criteria such as Anderson-Darling, Cramer-Von Mises, Kolmogorov-Smirnov, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Consistent Akaike Information Criterion (CAIC). Their findings reveal that this distribution outperforms other models when applied to COVID-19 death cases in Nigeria.

A. Ibrahim et al. (2020) study evaluates the impact of population density on COVID-19 spread using MATLAB and the stringency index model. It finds that population density does not significantly impact control measures in the initial month but suggests stricter regulations in denser populations [92]. Abdullah A. Al-Shammari et al. [10] (2020) discusses the challenges of COVID-19 control in developing countries, emphasizing shortages of testing supplies and competition. It highlights the role of mathematical modeling in forecasting and guiding public health policies.

Major Findings:

- Bayesian methods improve accuracy in prevalence estimation.
- Incorporating seasonal components enhances prediction accuracy.
- Bayesian approaches provide narrower confidence intervals.
- Better handling of uncertainty in epidemiological parameters.

3.3.4 Compartmental and advanced models

Several studies have employed compartmental and advanced models to understand and predict the dynamics of COVID-19. These models incorporate key factors such as undetected cases, reinfection, vaccination, and socioeconomic impacts, making them essential tools for guiding public health decisions.

Kamara et al. [99] use a modified SLIAR model to predict vaccine coverage needed to eradicate COVID-19 in Sierra Leone. Their study concludes that 58% of the population must be vaccinated, with facemasks playing a limited role in controlling the virus. Similarly, Ngonghala et al. [145] developed a modified SIR model incorporating features relevant to COVID-19 transmission dynamics and control, such as the quarantine of suspected cases and the isolation/hospitalization of confirmed COVID-19 cases. Ivorra et al. [94] presented the θ-SEIHRD model, which includes undetected infections and varying clinical conditions, demonstrating good alignment between model predictions and reported data in China. Kucharski et al. [108] used a stochastic SLIR model to

study transmission in Wuhan, noting the potential for underestimation in reported cases and the risks of international spread.

Sugiyanto et al. [178] explore the stability of mathematical models for interactions between target cells and COVID-19 infected cells. Their analysis reveals that parameters like the effective surface area of the network and initial target cell counts significantly affect equilibrium stability, offering insights for developing prevention and control strategies. Micheletti et al. [131] review various mathematical models used to understand the spread of COVID-19. They emphasize the need to consider social and economic contexts and adapt models as new data becomes available, demonstrating the potential of these models in informing public health policies. In a separate study, Saraiva et al. [169] used a mathematical modeling approach to project the evolution of COVID-19 in Campo Grande, Brazil. The piecewise growth model they developed was able to estimate the pandemic's peak and predict the number of patients requiring intensive care. Weekly technical reports were provided to the Health Department of Campo Grande to aid in decision-making regarding disease prevention.

Nishimoto and Inoue [146] develop a model explaining the repetitive waves of the COVID-19 pandemic using an activator-inhibitor system, advocating for proactive PCR testing after peak stages to mitigate recurrences. Lotfi et al. [119] propose integrating mathematics and medicine to better understand and treat COVID-19, focusing on the interaction between SARS-CoV-2 and immune cells to improve model predictions and treatment strategies. Wu, Zhang, and Cai [204] propose efficient algorithms for sparse topic modeling under the pLSI model, demonstrating rate-optimality and establishing valid confidence intervals for word-topic and topic-document matrices, supported by an analysis of the COVID-19 Open Research Dataset. Li and Ma [115] model COVID-19 transmission, revealing that while single asymptomatic contact creates a control threshold, multiple contacts cause complex bifurcations. Their work includes a detailed bifurcation and phase diagram, emphasizing the model's biological relevance. Dehning et al. (2020): This paper discusses Bayesian inference in SIR models to estimate the reproduction number of SARS-CoV-2, addressing data limitations and cross-validating results with alternative data sources [62].

Dynamic population models have also played a pivotal role in capturing the evolving risks of infection. Cooper et al. [56] introduced a novel SIR model that treats the susceptible population as a dynamic variable, enabling adjustments over time to account for new infections. By applying their model to data from January to June 2020, they predicted COVID-19 dynamics through September 2020, emphasizing that early and rigorous control measures significantly mitigate the spread. Wang et al. (2022) used a SEIRE model to study COVID-19 with reinfection. They found that when the basic reproduction number $\mathcal{R}_0 > 1$, the system reaches a stable endemic equilibrium, but eradication requires $\mathcal{R}_0 < \mathcal{R}_c$. A new concept,"robustness", was introduced to assess the difficulty of disease elimination [194].

Vaccination has been a key focus in many studies. Ghostine et al. (2021) proposed an extended SEIR model with vaccination to forecast COVID-19 in Saudi Arabia. Using ensemble Kalman filtering, the model predicted trends up to two weeks and explored vaccination's effects on controlling the pandemic's spread [79].

Watson et al. [197] analyzed the global impact of vaccination using an age-structured SEIRS model, estimating significant reductions in COVID-19 deaths due to vaccination, while highlighting challenges in low-income countries. Watson et al. [196] applied an age-structured SEIR model to estimate underreported COVID-19 deaths in Damascus, Syria, while Ghosh and Ghosh [78] introduced a refined SEIR model, demonstrating the efficacy of confinement, testing, and vaccination in reducing transmission. Mbabazi et al. (2020) applied an SEIR model to analyze COVID-19 interventions in Uganda. The study found that strict early interventions, such as lockdowns, were effective in reducing transmission rates in resource-limited settings [129].

Anastassopoulou et al. [15] employed a SIDR model to analyze epidemiological data from Hubei, China, highlighting the importance of accurate data for effective outbreak prediction and control, while Hu et al. [91] developed a time-dependent compartmental model, providing insights into asymptomatic and undetected cases, which informed control strategies and vaccination programs in Wuhan. Ryan H. Wilkinson (2021) paper discusses the SIR-compartment model and its effectiveness in approximating models with heterogeneous contacts. Wilkinson demonstrates that the SIR model can still provide accurate approximations and interprets parameters from regression analysis within the context of heterogeneous dynamics [199]. L. Russo et al. (2020) uses compartmental modeling to estimate the start of the COVID-19 outbreak in Lombardy, Italy, and predicts its fade-out by late May or early June based on current measures [166]. Vaidya et al. [187] investigate within-host dynamics of SARS-CoV-2 in ferrets using mathematical models and experimental data. Their study estimates key viral dynamic parameters and performs a global sensitivity analysis to identify factors impacting viral infection characteristics. The results suggest that ferrets are a suitable animal model for studying SARS-CoV-2 dynamics, providing valuable insights for the pre-clinical development of antiviral agents. Oud et al. [151] introduced a fractional order model to account for quarantine, isolation, and environmental viral load, offering new insights into pandemic management.

Advances in fractional-order modeling have further refined the understanding of disease dynamics. Verma et al. (2023) analyzed a fractional-order model for COVID-19, showing it offered better predictive capabilities than traditional integer-order models, enhancing understanding of disease dynamics [192]. Wasim Ahmad et al. (2021) paper presents a fractional order mathematical model using Caputo derivatives to describe the spread of COVID-19. It incorporates eight different classes and employs fractional Taylor's method for approximation and simulation over a 50-day period [8]. Croccolo (2020) developed a percolation-type model to study COVID-19 in-

fections on random graphs. The model provided insights into outbreak likelihood and informed public health strategies for containment [57].

Yadav et al. (2022) modeled global COVID-19 dissemination post-Omicron using multipronged approaches, showing the advantages of fractional models for understanding the virus's spread and forecasting its future impact more accurately [207].

R. M. Jena et al. (2023) presents a time-fractional order mathematical model for COVID-19, incorporating vaccination using non-singular kernel functions. It provides a framework for understanding transmission dynamics with vaccination effects [95]. Ghaffari and Saadati [77] present a $*$-fuzzy measure model for COVID-19 that incorporates data uncertainty, offering more accurate predictions of the disease's spread, identifying high-risk areas, and evaluating intervention effectiveness. This model represents a significant improvement over traditional approaches, with potential applications in controlling and preventing COVID-19. The application of control theory to understand the variation in SARS-CoV-2 infections and virulence is discussed by Sarma et al. [170]. The authors propose that the extreme variation in infections is due to two key mechanisms: a sparsely expressed host receptor and potent suppression of interferon. Their model unifies previously unexplained features of the pandemic and predicts future viruses that may cause pandemics, while also identifying potential interventions to mitigate such threats.

Wang, Washington, and Weber [195] use complex systems analysis, including Fourier and wavelet techniques, to study COVID-19 case progression globally and regionally. Their methods reveal valuable insights into population responses, aiding public health decision-making.

Al-Tuwairqi and Al-Harbi (2022) proposed a time-delayed COVID-19 model with vaccination, highlighting the importance of vaccination timing and scale in controlling the epidemic spread through simulations [11]. Basnarkov (2020) developed an epidemic spreading model for COVID-19, incorporating delayed infection onset and asymptomatic cases. The model, analyzed in different versions, highlighted relationships between epidemic thresholds and susceptible populations, especially in weak epidemics. Eigenvector centrality was identified as an indicator of infection risk [34].

I. A. Lakman et al. (2020) reviews mathematical tools for predicting COVID-19's progression in the Russian Federation. It presents ARIMA, SIRD, and Holt-Winters models, noting their high accuracy in predicting morbidity and mortality [112]. Min Lu and H. Ishwaran (2021) develops a competing risk compartmental model to analyze COVID-19 dynamics in the U.S., including cure and death rates and the effect of vaccination. It identifies wave patterns and the impact of vaccination on suppressing further waves [120]. Balayla et al. (2020) analyzed the relationship between positive predictive value (PPV) of screening tests and disease prevalence, identifying a prevalence threshold critical for interpreting COVID-19 test results and optimizing public health interventions [33].

Data-driven approaches and uncertainty handling have also been critical. Saeed et al. (2022): The study proposes a complex fuzzy hypersoft set (CFHS) framework for diagnosing COVID-19, addressing uncertainty in medical data. The CFHS mapping links symptoms to medicines and predicts recovery time based on patient records [167]. Major Findings:

- Modified compartmental models better capture COVID-19 dynamics.
- Inclusion of undetected cases improves prediction accuracy.
- Dynamic population variables enhance model realism.
- Reinfection modeling crucial for endemic equilibrium understanding.

3.3.5 Model Optimization and Real-Time Forecasting

Fox et al. [72] explored optimizing model ensembles for outbreak forecasting, offering valuable insights into enhancing accuracy in collaborative forecasting efforts. This work underscores the importance of leveraging diverse model outputs to improve predictive reliability. Similarly, P'eni et al. [156] employed a model predictive control approach to address the evolving dynamics of pandemics, aligning model outcomes with governmental responses to effectively manage disease spread. In another study, Long [118] developed a real-time analytical tracker for COVID-19, incorporating data visualization techniques, ARIMA forecasting, and logistic regression models to derive actionable insights from pandemic data. Adding to these advancements, Zeyi Liu et al. [116] utilized ensemble learning classifiers to estimate parameters for two-level individual-based models, effectively capturing the spatiotemporal spread of COVID-19 in Wuhan and other cities. Their findings highlight the effectiveness of ensemble classifiers in modeling complex virus transmission dynamics.

- Ensemble approaches improve forecast accuracy
- Predictive control helps manage evolving dynamics
- Real-time tracking essential for response optimization
- Machine learning enhances parameter estimation

3.3.6 Emerging variants and risk factors

Yang et al. [208] assessed the trade-off between transmissibility and virulence in SARS-CoV-2, proposing that more virulent strains, while more fatal, may be less transmissibleâĂŤa finding with implications for public health policies and vaccine strategies. Zhao et al. [211] employed

a Bayesian phylogenetic approach to link the D614G substitution in the SARS-CoV-2 spike protein to increased transmissibility, offering valuable insights into the effects of viral mutations. In Ontario, Canada, Betti et al. [36] developed a mathematical model to estimate the potential dominance of mutant COVID-19 variants over wild-type strains, stressing the continued need for non-pharmaceutical interventions (NPIs) alongside vaccination efforts to prevent future outbreaks. Similarly, Cohen et al. [54] constructed a mechanistic model to analyze SARS-CoV-2 immune memory, variant emergence, and vaccine efficacy, offering critical insights into the timing of booster doses and the interplay between natural and vaccine-derived immunity. Giordano et al. [80] examined the effects of vaccination and NPIs on curbing the spread of COVID-19 variants in Italy, demonstrating the necessity of maintaining both strategies to control the pandemic amid emerging infectious variants. Ciupeanu et al.[53] investigated the dynamics of variants of concern (VOCs) during the COVID-19 pandemic, focusing on factors influencing their dominance and coexistence. Their findings emphasize that a variant's transmissibility advantage and the initial number of infections play key roles in outbreak size. Early public health interventions targeting new VOCs were shown to effectively limit epidemic scales, offering valuable guidance for managing future outbreaks as COVID-19 transitions to an endemic phase.

- More virulent strains might be less transmissible
- Mutations can significantly impact transmissibility
- Variants pose significant threats to vaccination efforts
- Immune memory crucial for variant protection

3.3.7 Regional studies and data-driven approaches

Mandal et al. [122] combined serology with case detection to manage restrictions in India, while Williams [200] linked COVID-19 infection risk to antibody concentration and affinity. Moriconi [136] extended Mario Wuethrich's model for COVID-19 by incorporating exponentially decreasing intensity, applying it to data from Italy and China to focus on confirmed cases and adaptable effects like fatalities and recoveries.

Griette, Demongeot, and Magal [82] introduced a two-step method to analyze COVID-19 data in France, first using a phenomenological model to capture epidemic and endemic periods, and then applying a mathematical model to estimate key parameters. Atangana and Araz [25] modeled the third waves of COVID-19 in Turkey, Spain, and Czechia using piecewise differential operators, effectively capturing pandemic dynamics.

Aldila et al. [12] assessed early detection and vaccination strategies in Jakarta, Indonesia, finding that while these strategies reduced cases and hospitalizations, they had limited impact on

mortality, suggesting a need for a more comprehensive approach. Perera et al. [157] analyzed COVID-19 data and socio-economic factors in Sri Lanka, clustering districts to inform preventive measures. Pathak et al. [155] used an exponentiated exponential model to estimate the basic reproduction number (R0) for COVID-19 in Kerala, India.

Usoro and A.E. [7] analyzed COVID-19 patterns in Nigeria, showing a three-wave trajectory and advocating for consistent control measures. Song et al. [175] forecasted COVID-19 trends in India, predicting a 43% attack rate by mid-2021, while Chatterjee et al. [50] emphasized the importance of early lockdowns in India. Herrera [85] studied the influence of local mobility patterns and socioeconomic conditions on COVID-19 spread in Santiago, Chile.

Vasconcelos et al. [190] analyzed the COVID-19 epidemic across Brazil, offering insights for region-specific public health strategies. Brugnano et al. [46] developed a model for forecasting COVID-19 cases in Italy, including undiagnosed infections. JosephI et al. [96] analyzed COVID-19 trends in Nigeria, identifying patterns to guide policy. Miyamoto [132] used a logistic model to study COVID-19 waves in Japan, finding that the first wave fit well but the second wave deviated in the decline phase.

Karaulov et al. [100] presented a model to assess COVID-19 incidence across Russian regions. Bosetti et al. [39] modeled mass testing in France during an epidemic rebound, showing its effectiveness in reducing infections. Mandal et al. [121] evaluated vaccination strategies in India, suggesting targeted vaccination for key workers and individuals with comorbidities to reduce symptomatic cases and mortality.

Sarma et al. [170] applied control theory to explain variations in SARS-CoV-2 infections and virulence, predicting future viruses causing similar outbreaks. Mohamed et al. [135] proposed a more efficient method for estimating daily recovery cases in Egypt. Aphale et al. [19] advocated for a lockdown in Pune, India, while Aristov et al. [22] predicted recovery trajectories in different countries.

Obasi and Nwaka [147] proposed a model for COVID-19 transmission dynamics in Nigeria, emphasizing the role of hygiene and early detection. Ayoub et al. [27] explored the effectiveness of prioritizing vaccination based on antibody status in Qatar. Ng et al. [144] examined a pooling method for RT-PCR tests in Hong Kong, significantly reducing testing volumes in low-prevalence scenarios.

Mukandavire et al. [137] estimated the basic reproductive number in South Africa, concluding that a highly effective vaccine would have been essential to control the outbreak. Soubeyrand et al. [176] forecasted COVID-19 mortality trends in different countries, finding that second-line European countries experienced milder mortality patterns.

- Regional variations significantly impact outbreak patterns.

- Socioeconomic conditions influence contagion patterns.
- Underreporting significant in certain regions.
- Local mobility patterns crucial for spread.

3.3.8 Decision-making and statistical models

Almagrabi et al. [13] introduced a Q-linear Diophantine fuzzy emergency decision support system tailored for COVID-19. By combining fuzzy logic with linear programming, the system facilitates complex decision-making in emergencies. A case study on patient triage highlighted its effectiveness in addressing uncertain, time-sensitive decisions, showcasing its potential to improve emergency management during pandemics.

Khedhiri [102] compared various statistical models for COVID-19 death counts, highlighting the advantages of zero-inflated models in addressing excess zero counts. Scrucca [172] proposed COVINDEX, based on a GAM beta regression model, which offers valuable insights for improving public health policies. Fonseca et al. [71] evaluated several growth functions to model COVID-19 data, concluding that the Gompertz function was the most effective in predicting epidemic trajectories.

- Zero-inflated models better account for death counts.
- Gompertz function most efficient for epidemic trajectories.
- Fuzzy logic enhances emergency decision-making.
- Complex statistical approaches improve prediction. accuracy

3.3.9 Innovative statistical and parametric models

Almetwally et al. [14] proposed the modified Kies inverted Topp-Leone (MKITL) distribution for predicting COVID-19 mortality rates in the UK and Canada. This model, which combines the features of the inverted Topp-Leone and modified Kies distributions, was shown to outperform traditional distributions in fitting mortality data. Similarly, Xin, Zhou, and Mekiso (2022) introduced the "new generalized-X" (NG-X) family of distributions, which includes the NG-Weibull model. With its heavy-tailed characteristics, the NG-Weibull model was found to be superior in fitting COVID-19 data compared to three other models [205].

In addition to these approaches, Gallardo et al. [75] developed two parametric quantile regression models for double-bounded response data, utilizing the power Johnson SB distribution. These models were applied to COVID-19 mortality data across several countries, demonstrating

their effectiveness in capturing relationships between various variables. On a similar note, Lahcene [111] explored probability distributions for modeling COVID-19 data, focusing on suspected, recovered, and deceased patients. His study highlighted the crucial role of herd immunity in stabilizing infection rates and daily deaths, further contributing to the understanding of epidemic dynamics.

Meanwhile, Escamilla et al. [67] constructed probabilistic models to forecast COVID-19 cases and deaths in Mexico. Using exponential regression and negative binomial regression, their initial predictions were later revised based on updated data from the federal health sector, demonstrating the adaptability of their models. Expanding on forecasting, M.K. and WiliÅĎski [134] applied a Gaussian mixture model to predict future COVID-19 waves in Poland. Their findings indicated that current and past waves provide valuable information for predicting future outbreaks, thus offering insights for ongoing pandemic management.

In a global context, Pan [153] conducted a quantitative analysis using machine learning to predict future COVID-19 case numbers in eight representative countries. The study underscored the ongoing severity of outbreaks in countries such as the United States, Spain, and Brazil, providing crucial data for future responses. Furthermore, Bosse et al. (2023) [40] demonstrated that log transformation of counts enhances the evaluation of predictive models in epidemiological contexts. This transformation improved model rankings and performance evaluations, particularly when using metrics like CRPS and WIS in COVID-19 forecasts.

In the realm of patient outcomes, Cui et al. [59] developed a hierarchical Gaussian process model to predict COVID-19 patient outcomes. Their model, which shows promise for real-time applications, highlights the potential of advanced machine learning techniques in pandemic forecasting.

On the methodological front, Murakami and Matsui [138] improved over-dispersed Poisson regression with a log-Gaussian approximation to address identifiability issues, providing an accurate analysis of COVID-19 data in Japan. Further contributing to statistical modeling, Al-Ani (2021) applied three nonlinear growth modelsâĂŤGompertz, Richards, and WeibullâĂŤto analyze COVID-19 cases in Iraq. Among them, the Weibull model provided the best fit, predicting 114,907 cases by mid-August 2020, with an epidemic peak in early July [9]. Lobato et al. [117] and Hong et al. [90] introduced models designed to handle uncertainties in reported data and temporal variations in transmission rates. Their work enhanced parameter estimation and pandemic forecasting accuracy, addressing key challenges in epidemic modeling.

In the area of risk management, Surowiec and Warowny [179] applied the Value at Risk method to estimate COVID-19 death rates. This method proved effective in managing pandemic resources and offering a clearer picture of the risk landscape.

Finally, Chang et al. (2020) [49] used the CovidSIMVL model to analyze COVID-19 transmission chains and network structures. Their work questioned the statistical validity of superspreading

individuals, relying on contact tracing data to assess epidemic progression and transmission dynamics.

- New distributions better capture COVID-19 data skewness.
- Parametric approaches improve mortality prediction.
- Heavy-tailed distributions essential for outbreak modelling.
- Novel regression approaches enhance relationship understanding.

4

Quantifying the effects of public health interventions in Alberta during the first wave of COVID-19

Overview of the chapter In this chapter, we investigate the effects of the non-pharmaceutical interventions (NPIs) implemented by Alberta Health during the first wave of COVID-19. These interventions, social distancing measures and testing policies, aimed to reduce transmission and mitigate the epidemic's burden on the healthcare system. Understanding how these interventions influenced the course of the epidemic is crucial for informing future public health strategies, particularly in the face of ongoing or emerging infectious disease threats. Rather than solely focusing on fitting the epidemic curve of COVID-19 cases in Alberta, this study emphasizes the use of parameter estimates derived from data fitting to explore a variety of hypothetical scenarios. This approach shifts the focus from retrospective description to prospective exploration, allowing us to evaluate how different combinations or intensities of NPIs could alter epidemic outcomes. By doing so, the study provides valuable insights into the relative effectiveness of these interventions, supporting evidence-based decision making.

We start by calibrating our epidemiological model to observed data from the first wave. This allows us to capture key transmission dynamics using time-dependent parameters that reflect how the environment and public behaviour evolved during the pandemic. These calibrated parameters serve as a foundation for our scenario simulations where parameters are held constant to systematically assess intervention impacts. This distinction between initial parameter estimation and

subsequent scenario analysis allows for a rigorous yet interpretable modelling framework.

This is joint work, in preparation to being submitted to a journal, with Dr. Michael Li's group from the University of Alberta (Dr. Weston C Roda, Donglin Han) and Dr. Marie Betsy Varughese (previously Alberta Health Services) from Institute for Health Economics.

4.1 Model Structure

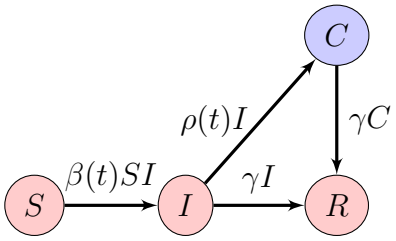


Figure 4.1: Flow diagram for the SICR model, an extension of the SIR model which records the number of cases.

This chapter employs a SICR compartmental model with time-dependent parameters and without demography to depict the transmission dynamics of COVID-19 in Alberta. The SICR model extends the classical SIR framework by explicitly tracking confirmed cases, which aligns with the available public health surveillance data.

The compartment S represents all individuals susceptible to COVID-19, effectively initially encompassing the entire population of Alberta. Compartment I consists of the so-called **hidden infections**, including individuals who are asymptomatically infected and those who have symptoms but have not yet been tested. Individuals in compartment I can spread the disease in the community. Compartment C includes individuals who have tested positive for COVID-19 through a PCR test and by public health restrictions, are under strict isolation orders for 10-14 days. Lastly, the compartment R consists of individuals who have recovered from COVID-19 and have acquired immunity, thus being protected from reinfection.

It is crucial to differentiate between a *confirmed case* and an *infection* in this context. An infected individual is someone exposed to and infected by the SARS-CoV-2 virus, who in turn will have a symptomatic or asymptomatic infection. A *confirmed case* refers specifically to an infected individual confirmed positive through testing and reported to public health authorities. Identified cases are either hospitalized or self-quarantined, effectively being removed from transmission. Since SARS-CoV-2 is a novel virus, we assume the entire population is initially susceptible.

Figure 4.1 provides a visual representation of the model structure. Model (4.1.1) consists in a system of differential equations. Table 4.1 details parameters used.

Parameter	Description
$\beta(t)$	Time-dependent transmission rate
b	Initial transmission rate $\beta(0)$
q	Fraction of reduction in transmission due to lock-down
s	Fraction of increase in transmission due to open-up
time_{1b}	Time point parameter in the definition of $\beta(t)$
time_{2b}	Second time point parameter in the definition of $\beta(t)$
$\rho(t)$	Time-dependent case-infection ratio
f	Scaling factor for $\rho(t)$
time_1	Time point parameter in the definition of $\tilde{\rho}(t)$
time_{1b}	Second time point for $\tilde{\rho}(t)$
time_3	Third time point for $\tilde{\rho}(t)$
time_4	Fourth time point for $\tilde{\rho}(t)$
time_5	Fifth time point for $\tilde{\rho}(t)$
height_val	Level parameter in the definition of $\tilde{\rho}(t)$
const_val	Offset constant in $\tilde{\rho}(t)$
start_val	Initial value in $\tilde{\rho}(t)$
γ	Recovery rate
p	$1/p$ is the variance of the case data noise
p_2	$1/p_2$ is the variance of the $\tilde{\rho}$ data noise

Table 4.1: Biological meaning of the SIRC model parameters.

Based on the transfer diagram in Figure 4.1, the following system of differential equations is derived:

$$\begin{aligned} S' &= -\beta(t)IS \\ I' &= \beta(t)IS - \rho(t)I - \gamma I \\ C' &= \rho(t)I - \gamma C \\ R' &= \gamma I + \gamma C \end{aligned} \tag{4.1.1}$$

Recovery from COVID-19 can occur through two main pathways. Firstly, individuals, particularly those who are asymptomatic, may recover from the infection without undergoing testing. This process is represented in the model by γI, where $1/\gamma$ is the average infectious period. Secondly, individuals who test positive and are subsequently isolated or self-isolate in compartment C can recover during this isolation period, whether at home or in hospitals. Recovery from compartment C occurs at the rate γC. We assume that infected individuals and cases recover at the same rate. Recovered individuals are then included in compartment R.

To streamline the model, we assume negligible natural birth and death rates by setting them to zero, as we are modelling a single wave of COVID-19 transmission over a relatively short time frame. We further assume that individuals who recover from infection acquire full immunity against reinfection during the period under consideration.

The principal route of transmission of an infectious disease is contact of susceptible individuals in S with infectious individuals in I. This process is modelled by $\beta(t)I(t)S(t)$. The transmission coefficient $\beta(t)$ can be influenced by many factors, including the average number of contacts among individuals in the populations and the average probability of transmission for each contact, which may depend on both the infectivity of individuals in I and the susceptibility of individuals in S during each contact. The value of $\beta(t)$ is averaged over individual variations in the population. For individuals in compartment I who are recently infected and not yet infectious, infectivity is considered to be zero.

Social distancing restrictions and relaxations will cause changes in the transmission coefficient $\beta(t)$.

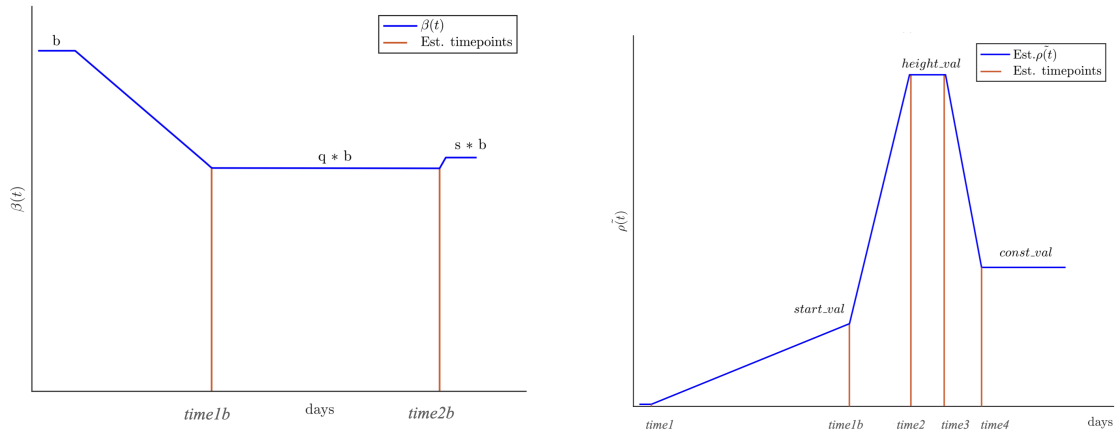
The time-dependence in $\beta(t)$ is informed by the timing and changes in non-pharmaceutical interventions taken by public health authorities and is defined as a function of time t (see Figure 4.2a).

The transmission rate $\beta(t)$ remained constant from March 9 to March 14, 2020. Over the subsequent three days, Alberta implemented a range of additional public health measures, including the prohibition of attendance at venues such as public recreation centres, casinos, bingo halls, bars, nightclubs, fitness centres, arenas, museums, and indoor children's play centres. During this period, the transmission rate significantly decreased and remained low.

Beginning on May 14, 2020 (which is time time_{1b} in in the the Figure 4.2), some of these public health orders were gradually lifted, permitting the reopening of daycare centres, out-of-school care facilities, day camps, post-secondary institutions, places of worship, hair salons, restaurants, and other businesses, all operating at 50% capacity. Consequently, in our model, the transmission rate increased slightly towards the end of the first wave.

The transfer term $\rho(t)I(t)$ from compartment I to C is the daily number of positive tests or *daily reported cases*, which is part of the public health data we will use for model calibration, and parameter $\rho(t)$ is the case-infection ratio: the ratio between daily new case reports and the number of people living with the infectious disease infection in the community on that day. It is an indicator of the efficiency of public health surveillance, and its values answer the question: for each infectious disease case identified, how many *hidden infections* are there in the community on a particular day?

During the initial wave of the pandemic, public testing policies required individuals experiencing symptoms to schedule a PCR test, either by contacting the 811 Healthline or by completing an online self-assessment form. The time-dependent parameter $\rho(t)$ incorporates changes in public health-seeking behaviour and individual testing behaviour, as individuals need to engage with



(a) $\beta(t)$. (b) $\tilde{\rho}(t)$.

Figure 4.2: (a) Piecewise linear functions used to describe changes in (a) transmission $\beta(t)$ according to social-distancing policy adjustments and (b) daily rate $\tilde{\rho}(t)$ of COVID-19 tests and health-seeking behaviour in the population, as informed by public health data. Time points on the horizontal axis correspond to the dates of policy changes and are allowed to vary around those dates during model calibration.

public health authorities to arrange a test. As such we write the parameter $\rho(t)$ as

$$\rho(t) = f\,\tilde{\rho}(t), \tag{4.1.2}$$

where $\tilde{\rho}(t)$ is defined as a piece-wise linear function of time t since testing capacity changes will influence the case-infection ratio $\rho(t)$ (see Figure 4.2b). The scaling factor f is the average probability of infected individuals to seek a COVID-19 test.

The $\tilde{\rho}$-data is the rate of positive tests among all individuals who sought a COVID-19 PCR test and is defined as

$$\tilde{\rho} = \frac{\text{Newly reported positive cases at time } t}{\text{Total number of individuals seeking testing at time } t}, \tag{4.1.3}$$

where the total number of individuals seeking testing at time t is the total number of Healthline calls and completed online self-assessment forms at time t.

4.2 Computational analysis

4.2.1 Final size of the epidemic, attack rate and $\mathcal{R}_0$

To gain analytical insight into the long-term behavior of our system, we consider a simplified version of the model with constant parameters. This allows us to derive an expression for the *final size* of the epidemic, the total number of individuals infected over the course of the outbreak, as well as the *attack rate*, which quantifies its overall impact on the population.

Analytically solving the full system is challenging due to the presence of time-varying parameters and additional compartments (C for confirmed cases and R for recoveries). However, since the equations for C and R do not feed back into the dynamics of S and I, we may omit them without loss of generality for this analysis. The resulting two-equation system still captures the essential infection dynamics needed to study the total number of infections.

We further assume that the parameters β, ρ, and γ are constant. This simplification enables a tractable derivation of the final size relation. While these parameters may vary in real-world settings, due to interventions, behavioural responses, or changes in detection policies, treating them as constant yields a first-order approximation of average epidemic behaviour and clarifies the influence of key parameters, such as the basic reproduction number $\mathcal{R}_0$.

We note that in a more realistic model with time-dependent parameters, the final size could differ from the one predicted here. Nonetheless, the simplified model provides a valuable baseline for understanding how transmission and removal rates shape epidemic outcomes.

Hence we will use the following system and the classical approach to compute the final size:

$$S' = -\beta SI \tag{4.2.1a}$$

$$I' = \beta SI - \rho I - \gamma I. \tag{4.2.1b}$$

We know that $S' < 0$ for all t, hence S is monotonically decreasing and

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

Summing both sides of (4.2.1), we get

$$S'(t) + I'(t) = -(\rho + \gamma)I(t).$$

Integrating from 0 to ∞ with respect to t gives

$$\int_0^\infty \frac{d}{dt}\left(S(t) + I(t)\right) dt = -\int_0^\infty (\rho + \gamma)I(t)dt.$$

Now the left-hand side gives

$$\begin{aligned}\int_0^\infty \frac{d}{dt}\left(S(t)+I(t)\right)dt &= S_\infty + I_\infty - S_0 - I_0 \\ &= S_\infty - S_0 - I_0, \quad \text{as } I_\infty = 0.\end{aligned}$$

The right-hand side gives

$$-\int_0^\infty (\rho+\gamma)I(t)dt = -(\rho+\gamma)\hat{I}.$$

Thus equating the two sides gives

$$S_\infty - S_0 - I_0 = -(\rho+\gamma)\hat{I}. \tag{4.2.2}$$

Now consider $S' = -\beta SI$ and divide by S:

$$\frac{S'(t)}{S(t)} = -\beta I.$$

Integrating from 0 to ∞ with respect to t gives

$$\ln(S_\infty) - \ln(S_0) = -\beta\hat{I}. \tag{4.2.3}$$

Using (4.2.2) and (4.2.3), expressing them in terms of $-\hat{I}$ and equating gives

$$\begin{aligned}\frac{\ln(S_\infty)-\ln(S_0)}{\beta} &= \frac{S_\infty - S_0 - I_0}{(\rho+\gamma)} \\ \implies \ln(S_\infty) - \ln(S_0) &= \frac{\beta}{\rho+\gamma}\left(S_\infty - S_0 - I_0\right).\end{aligned}$$

The final size equation of the system is thus given by

$$(\ln(S_0) - \ln(S_\infty))S(0) = (S_\infty - S(0))\mathcal{R}_0 + I_0\mathcal{R}_0, \tag{4.2.4}$$

where $\mathcal{R}_0 = \dfrac{\beta}{\rho+\gamma}S_0$.

The **attack rate** is given by the fraction of the population that becomes infected during the outbreak:

$$\frac{S_0 - S_\infty}{S_0 + I_0}. \tag{4.2.5}$$

This equation provides an expression for the proportion of the population that is susceptible after the outbreak, and the attack rate quantifies the overall impact of the epidemic in terms of the

fraction of the population that became infected.

4.2.2 Impact of time dependence on the model

In the non-time-dependent model, we assume that β, ρ, and γ are constant over time. This assumption allows us to derive the final size relation, which shows how the total number of susceptibles S_∞ and initial conditions are related through the basic reproduction number $\mathcal{R}_0$.

However, in reality, the parameters β and ρ can vary over time due to changes in behaviour, interventions, or other factors. When these parameters become time-dependent, the analysis becomes significantly more complex:

- Changing Dynamics: The rates of infection and recovery are no longer constant, which means the differential equations need to account for these changes, complicating the integration process.
- Inaccurate Predictions: The final size relation derived under the assumption of constant parameters might no longer hold, leading to inaccurate predictions of the epidemic's progression and final size.
- Complex Modelling: Incorporating time-dependent parameters often requires more sophisticated modelling techniques, such as numerical simulations, which can handle varying rates but at the cost of analytical simplicity.

4.2.3 Fitting of the model for baseline

To model the transmission dynamics of COVID-19 in Alberta, we employed a simplified version of the model described in Section 4.1. This simplification was necessary due to non-identifiability issues: the daily reported case counts can only inform the product $\rho p_t q I(t)$, and a key factor is the duration individuals remain in the infected compartment I, which is determined by the outflow rates from I.

Due to the lack of reliable recovery data during the first wave in Alberta, we do not explicitly include the recovered compartment $R(t)$ in our model fitting or simulations. Instead, we adopt a simplified SIC framework, which captures the dynamics of susceptible, infected, and confirmed case populations. This approach allows us to calibrate the model using only the available daily reported case data, while implicitly accounting for recovery processes without explicitly modelling them.

The SIC model is considerably simpler than most compartmental models used in the infectious disease literature, such as SLIR or SLIAR-type models. (We use the notation SLIR or SLIAR

instead of SEIR or SEIAR; see [47] for a discussion of why the use of the letter E can be problematic.) More complex models typically include additional compartments and transitions, leading to a greater number of parameters that must be estimated from data. Since the public health data used for model calibration remains the same, introducing more parameters increases the risk of non-identifiability, which in turn raises the uncertainty of model predictions and parameter estimates [163].

In particular, the parameters β and ρ in model (4.1.1) are non-identifiable when only daily reported case data is available, as such data can only inform the product $\rho I(t)$, the combination of two unknowns ρ and $I(t)$.

Accordingly, our fitting process used the SIC model (see Figure 4.3 for the flow diagram), governed by the following equations:

$$\begin{aligned} S' &= -\beta(t)IS \\ I' &= \beta(t)IS - \rho(t)I - \gamma I \\ C' &= \rho(t)I. \end{aligned} \tag{4.2.6}$$

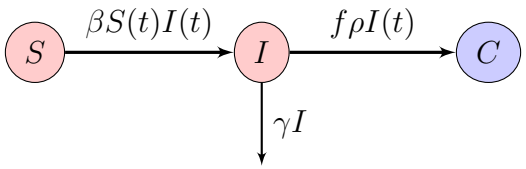


Figure 4.3: Simplified SIC model transfer diagram depicting transitions between compartments S (Susceptible), I (Infected), and C (Cases). In this model, recovered individuals are not explicitly considered.

Timeline of implementations of public-health measures are used to define the time-dependent transmission rate $\beta(t)$. Newly reported positive COVID-19 cases and total number of Healthline calls and COVID-19 online self-assessment forms were used to produce the data for $\tilde{\rho}(t)$ according to the definition in (4.1.2). For parameters in Table 4.2, their values were estimated by fitting model outcomes to daily reported new positive COVID-19 cases data and COVID-19 testing data. Our model calibration used a Bayesian inference-based Markov chain Monte Carlo algorithm (see Chapter 2.7)

The model calibration yielded a time-dependent case-infection ratio $\rho(t)$, which reflects changes in detection and reporting of COVID-19 cases during the first wave in Alberta (Figure 4.5). The shape of $\rho(t)$ was predefined and informed by observed trends in Healthline calls and self-assessment data, the calibration process allowed for flexibility in its timing and magnitude. As a result, the fitted $\rho(t)$ provides a good approximation of observed detection patterns, partly because it was constructed to align closely with the data. Despite this constraint, the model captures key

Parameter	Description	Best-fit value	95% credible Interval	Prior
$\beta(t)$	Time-dependent transmission rate	time-varying	time-varying	time-varying
γ, γ	Recovery rates	0.1429	(0.1260, 0.1920)	$(\frac{1}{11}, \frac{1}{5})$
$\rho(t)$	Time-dependent case-infection ratio	time-varying	time-varying	time-varying
f	Scaling factor for ρ	1.1013	(0.8146, 1.4857)	U(0,1.5)
$beta$	Initial transmission rate $\beta(0)$	6.195e-08	(5.63e-08, 7.84e-08)	U(1e-9, 1e-7)
q	Fraction of reduction in transmission due to lock-down	0.6548	(0.5751, 0.8163)	U(0.01, 1)
s	Fraction of increase in transmission due to open-up	0.6864	(0.6054, 0.9666)	U(0.01, 1)
p	$1/p$ is the variance of the case data noise	8.7933	(6.5232, 14.5960)	U(1, 79)
p_2	$1/p_2$ is the variance of the $\tilde{\rho}$ data noise	0.0042	0.0031, 0.0062)	U(0.0001, 0.03)
$time_1$	Parameter in the definition of $\tilde{\rho}(t)$	3.2	(1.1168, 4.6156)	U(1, 10)
$time_{1b}$	Same as above	36.9	(35.1107, 36.9854)	U(10, 37)
$time_2$	Same as above	46.2	(44.1507, 49.0937)	U(39, 55)
$time_3$	Same as above	52.0	(52.0024, 54.4615)	U(52, 60)
$time_4$	Same as above	57.2	(57.0013, 58.6041)	U(57, 72)
$height\_val$	Same as above	0.0895	(0.0795, 0.0989)	U(0.001, 0.1)
$const\_val$	Same as above	0.0375	(0.0317, 0.0431)	U(0.001, 0.1)
$start\_val$	Same as above	0.0223	(0.0185, 0.0268)	U(0.01, 0.04)
$time1beta$	Parameter in the definition of $\beta(t)$	29.2070	(29.0017, 31.7245)	U(29, 32)
$time2beta$	Same as above	66.5656	(65.1864, 67.9948)	U(65, 68)
R_0	Basic reproduction number	1.8861	(1.5663, 2.0029)	-

Table 4.2: Estimated model parameters derived from confirmed case data, including best-fit values, 95% credible intervals, and prior distributions used for Bayesian inference

shifts in testing and reporting behaviour over time, offering a realistic representation of how case ascertainment evolved throughout the outbreak.

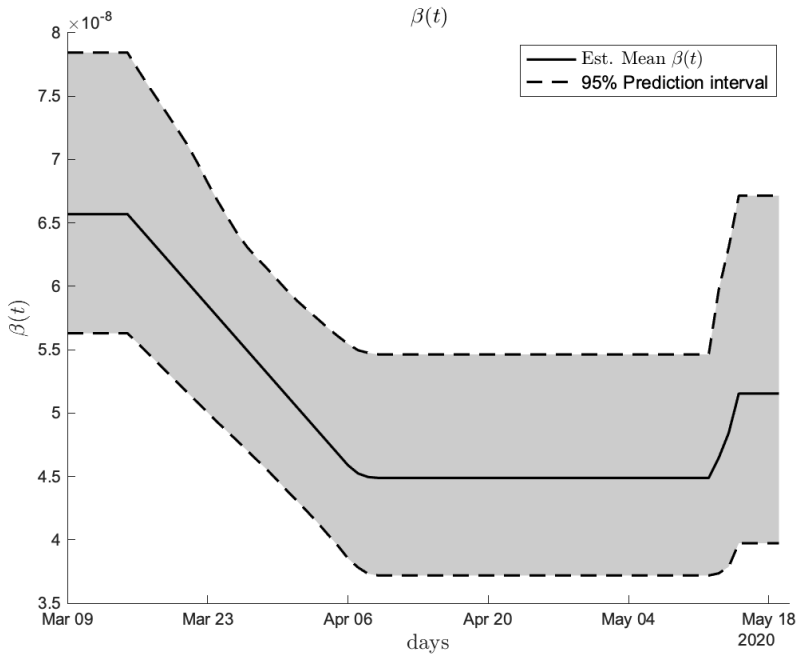


Figure 4.4: Estimated time-dependent $\beta(t)$. The plot shows the calibrated evolution of $\beta(t)$. Shaded regions indicate the 95% credible intervals (95% CrI) of the model estimations.

The model estimations successfully reproduced the time course of the first wave of the COVID-19 pandemic in Alberta during Spring 2020. Figure 4.6a shows the model fitting to daily reported new COVID-19 cases, where the solid line represents the model's mean prediction, and the shaded regions indicate the 95% credible intervals (CI).

Figure 4.6 highlights a key finding from our model estimations: the peak times for hidden infections $I(t)$ and for identified cases $C(t)$ do not coincide. The peak for hidden infections occurs approximately one week earlier than the peak for identified cases. This discrepancy contrasts with many model predictions in the literature where identified cases and total infections peak simultaneously. Several factors might contribute to the later peak in identified cases: symptoms typically appear about a week after infection; enhanced testing during the period when hidden infections peak and begin to decline could lead to a continued rise in identified cases; and health-seeking behaviour, driven by the number of identified cases (as hidden infections are not known), may further skew the correlation between the two peak times. Our analysis reveals a substantial number of unreported cases, highlighting the challenges in capturing the true scale of the pandemic. Figure 4.7 shows the proportion of hidden infections relative to the total number of infections over time. This measure highlights the role of hidden infections in the overall case count at each time

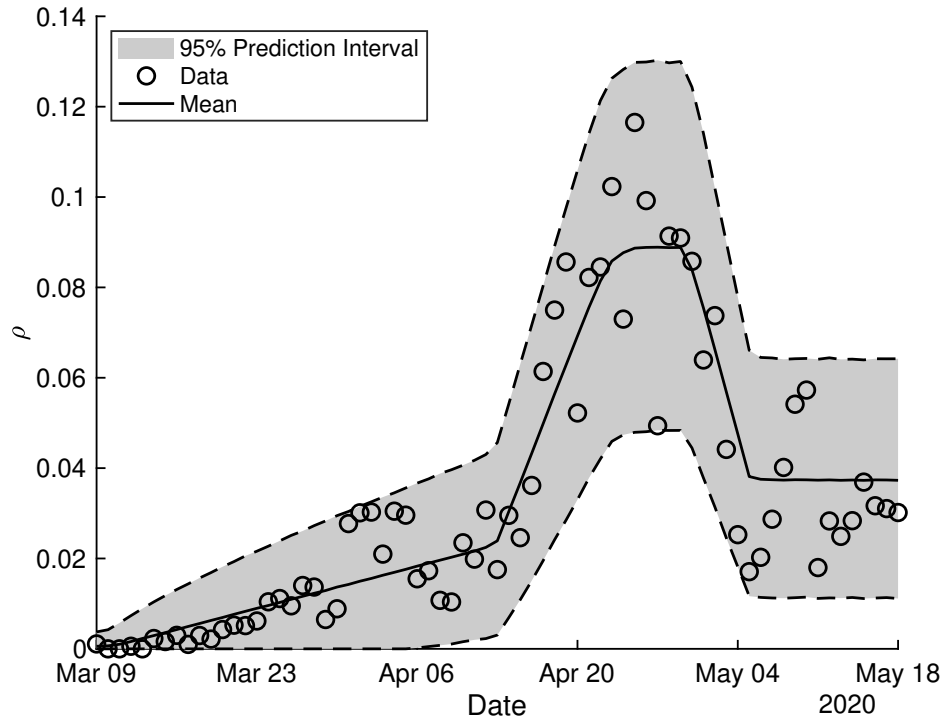
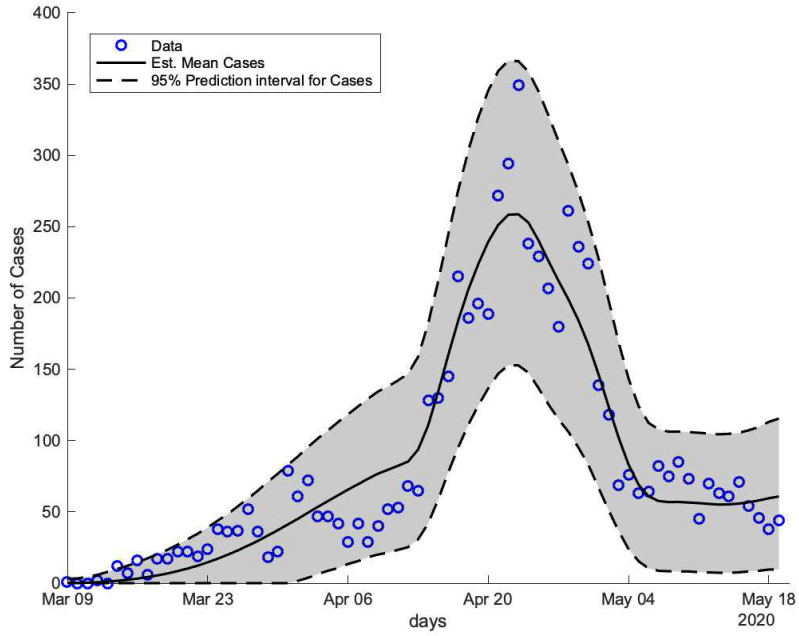
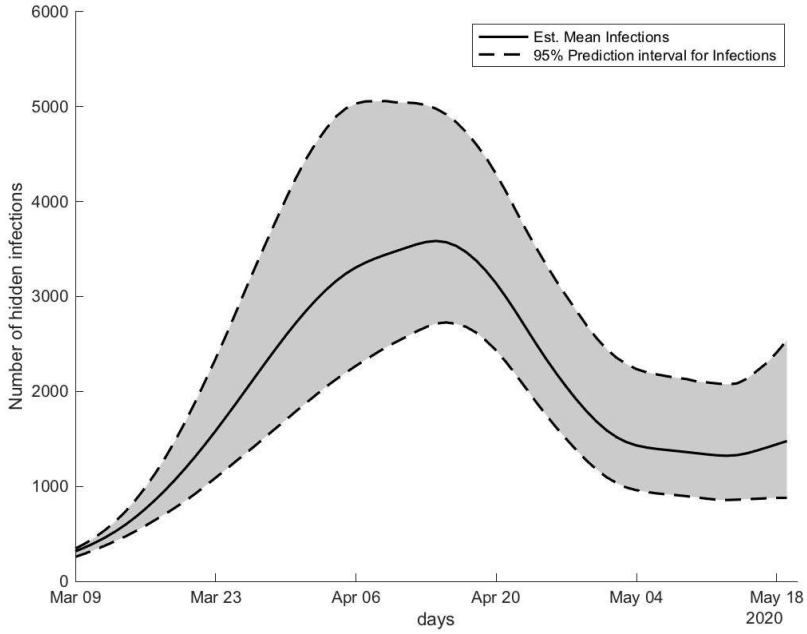


Figure 4.5: Estimated time-dependent case-infection ratio $\rho(t)$. The plot shows the calibrated evolution of $\rho(t)$, defined as the ratio of newly reported positive cases to the total number of Healthline calls and completed online self-assessment forms at time t. This ratio reflects changes in the detection and reporting of COVID-19 cases during the first wave in Alberta in Spring 2020.



(a) Model fitting to daily reported new COVID-19 cases in Alberta, this is compartment C in our model.



(b) Estimated number of hidden infections $I(t)$.

Figure 4.6: Baseline model calibration and fitting results for the first wave of the COVID-19 pandemic in Alberta. The top panel illustrates the model fitting to daily reported new COVID-19 cases in Alberta. The solid line represents the model's mean prediction, while the shaded regions indicate the 95% credible intervals (CI). The bottom panel represents the model's estimation of the number of COVID-19 cases that were not detected by public health surveillance during the first wave.

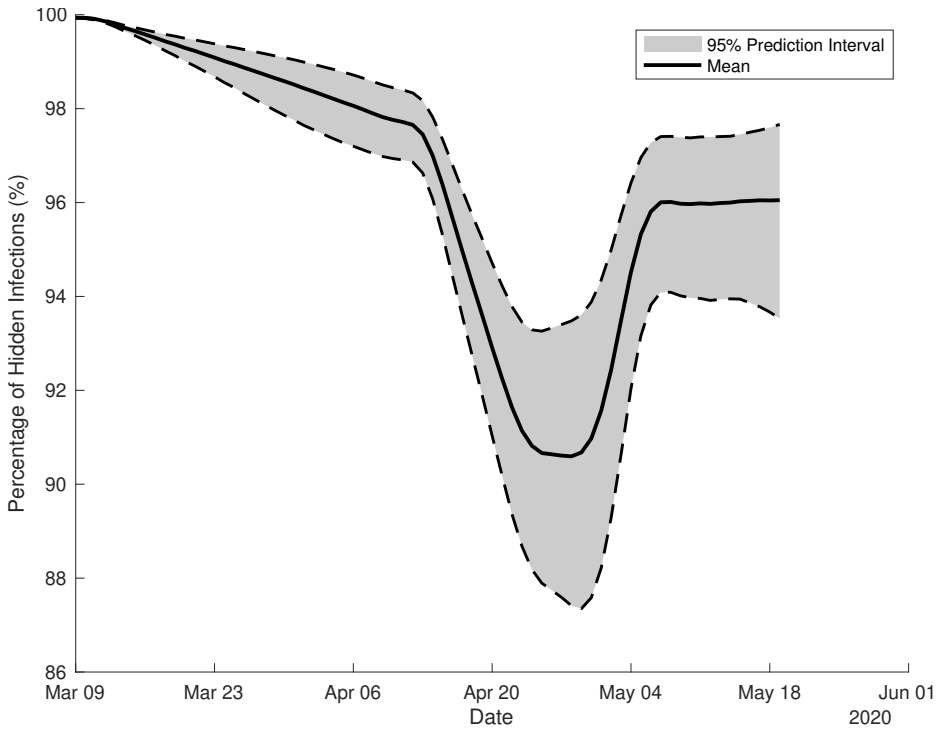


Figure 4.7: Proportion of hidden infections (i.e., infections that are unreported or undetected) as a percentage of the total infections, which includes both reported and hidden cases. This proportion is calculated as $\frac{I(t)}{I(t)+C(t)} \times 100$

point, providing a clearer picture of the relationship between reported and unreported cases.

Implications of baseline model findings

These results have important implications for both public health surveillance and epidemic modelling. First, the discrepancy between the peaks of hidden infections and reported cases emphasizes the inherent delay in using case data to monitor real-time transmission. This suggests that policies relying solely on observed cases may lag behind actual transmission dynamics, potentially delaying timely interventions. Second, our estimation of a substantial proportion of undetected infections, particularly during the peak of the outbreak, highlights the critical role of asymptomatic or mildly symptomatic spread in shaping the epidemic trajectory. This underscores the need for broader testing strategies and real-time proxies, such as syndromic surveillance or waste-water analysis, to better track hidden transmission. Finally, the calibrated time-dependent $\rho(t)$ can inform future models that aim to correct for under-ascertainment, enabling more accurate forecasting and resource planning during emerging waves of infection.

4.2.4 Effects of social distancing, testing and isolation

This section presents the core results of my original research on the impact of varying social distancing and testing strategies on the spread of COVID-19 in Alberta. While time-dependent parameters were initially used to capture the dynamic nature of transmission during model fitting, we transitioned to using constant parameters for the purposes of scenario analysis. This methodological shift streamlined the analytical process, allowing for easier interpretation of results and enabling direct comparisons across different intervention strategies.

The total population was set to 4,371,323, with a predefined time horizon of 72 days, which is how long the first COVID-19 wave lasted in Alberta. Key parameters where set to the best-fit values obtained from the fitting, the recovery rate (γ) was set to 0.1429 and the scaling factor (f) was set to 1.1013. The target basic reproduction number ($\mathcal{R}_0$) was 1.8861. Our script uses `linspace` to define ranges for β and ρ, generating 10,000 evenly spaced values within specified intervals for each parameter. We then employed an iterative process to identify valid parameter combinations where $\mathcal{R}_0$ was within 0.01 of the target, storing only these valid combinations for further analysis.

To simulate a variety of social distancing and testing strategies, we applied multipliers to the constant values of β and ρ. Specifically, we explored increases of 10%, 20%, 30%, and 40% as well as decreases of 10% and 20% from the baseline value (i.e., multiplier of 1.0). This systematic perturbation allowed us to simulate realistic behavioural and policy changes. For instance, reducing β reflects enhanced social distancing, while increasing ρ represents improvements in testing and

isolation strategies.

To establish a baseline for comparison, we set β and ρ multipliers to 1, representing the status quo of COVID-19 transmission dynamics in Alberta. This baseline scenario served as a reference point for evaluating the relative effects of different intervention strategies. By comparing alternative scenarios to the baseline, we could quantify and understand the impact of varying social distancing measures and testing strategies on key epidemiological metrics.

In quantifying the effects of varying β and ρ multipliers, we analysed a range of epidemiological metrics, including cumulative infections, peak infections, cumulative cases, and peak cases. By comparing scenario outcomes to the baseline, we could assess the relative impact of different interventions on disease spread. Percentage changes relative to the baseline scenario were calculated to quantify the magnitude of the effects of different intervention strategies, providing valuable insights for policymakers and public health officials.

Finally, to infer the potential effects of different social distancing and testing strategies on disease transmission dynamics, we applied the percentage changes obtained from the analysis to the outputs of the time-dependent fitting model. This allowed us to predict the potential outcomes of alternative intervention strategies and gain insights into the effectiveness of different public health measures in controlling the spread of COVID-19 in Alberta.

The scenarios presented in the following tables were selected to illustrate the spectrum of observed outcomes. The scenario labels (e.g., Scenario 8, Scenario 15, etc.) correspond to specific combinations of β and ρ multipliers in the full grid of simulations, which were indexed sequentially for tracking purposes. Each scenario represents a unique simulated intervention strategy based on these parameter multipliers.

The results presented in Tables 4.3 and 4.4 highlight the impact of social distancing measures on the dynamics of disease transmission. Increasing social distancing, indicated by a decrease in β, led to substantial reductions in key epidemiological metrics, including cumulative infections, peak infections, cumulative cases, and peak cases (see Table 4.3). Conversely, relaxing social distancing measures, denoted by an increase in β, resulted in notable increases in these metrics (see Table 4.3). While the impact of varying testing strategies (adjusting ρ) was observable, it was comparatively smaller than the effects of social distancing measures (Tables 4.3 and 4.4). Figure 4.8 provides a visual representation of these effects across the range of scenarios tested.

To further understand the impact of testing strategies (ρ) on disease transmission dynamics while keeping social distancing measures (β) at baseline levels, we explored additional scenarios through simulations. The scenarios involved variations in testing effectiveness while maintaining constant social distancing measures. The results presented in Tables 4.5 and 4.6 reveal insights into how changes in testing strategies influence key epidemiological metrics.

Figure 4.9 visualises the effects of varying testing across different scenarios.

Scenario	β Mult.	ρ Mult.	Infections Cumulative (%)	Infections Peak (%)	Cases Cumulative (%)	Cases Peak (%)
Scenario 8	1.1	1	23.56	24.78	16.74	24.70
Scenario 15	1.2	1	36.36	48.41	25.82	48.16
Scenario 29	1.4	1	48.76	92.00	35.40	90.84
Scenario 36	0.9	1	-38.80	-25.54	-32.40	-25.53
Scenario 43	0.8	1	-80.15	-61.22	-76.39	-57.35

Table 4.3: Percentage change in epidemiological metrics (cumulative infections, peak of the number of infections, cumulative cases and peak of the number of cases) relative to baseline scenario for various β while ρ is kept at baseline.

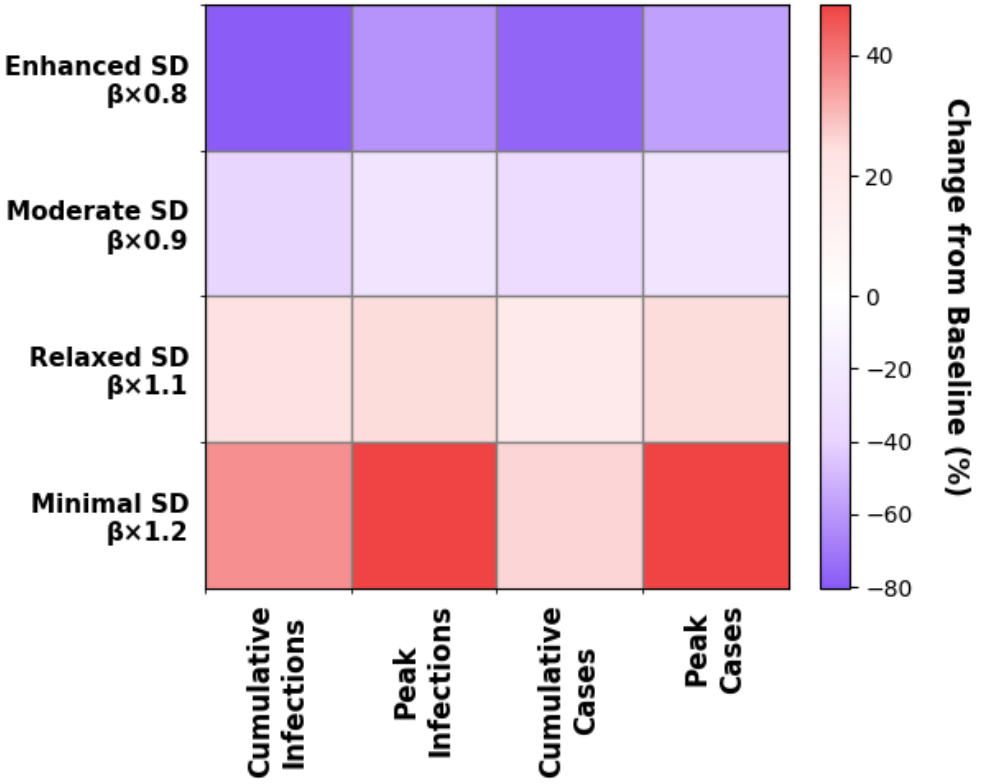


Figure 4.8: Heatmap showing the impact of social distancing measures (varying β multiplier) on key epidemiological outcomes. Darker blue indicates greater reductions relative to baseline; red indicates increases. Enhanced social distancing ($\beta \times 0.8$) reduces all metrics substantially, whilst minimal social distancing ($\beta \times 1.2$) increases cases and infections substantially.

Scenario	β Mult.	ρ Mult.	Infections Cumulative	Infections Peak	Cases Cumulative	Cases Peak
Baseline	1	1	146,850	3,585.40	6,028	258.91
Scenario 8	1.1	1	181,445.91	4,473.80	7,037.30	322.85
Scenario 15	1.2	1	200,249.24	5,320.96	7,584.39	383.59
Scenario 29	1.4	1	218,457.26	6,883.89	8,161.96	494.11
Scenario 36	0.9	1	89,874.02	2,669.75	4,074.65	192.81
Scenario 43	0.8	1	29,148.56	1,390.49	1,423.10	110.43

Table 4.4: Absolute values of cumulative and peak infections and cases for different scenarios compared to the baseline. The table illustrates how changes in β affect the overall and peak disease metrics

Scenario	β Mult.	ρ Mult.	Infections Cumulative (%)	Infections Peak (%)	Cases Cumulative (%)	Cases Peak (%)
Scenario 2	1	0.9	6.06	4.45	-3.85	-3.96
Scenario 3	1	0.8	12.25	9.12	-8.92	-9.03
Scenario 4	1	0.7	18.60	14.19	-15.24	-15.27
Scenario 5	1	0.6	25.15	19.63	-22.86	-22.72
Scenario 6	1	1.1	-5.96	-4.09	2.63	3.11
Scenario 7	1	1.2	-11.84	-7.99	4.03	5.25

Table 4.5: Percentage change in epidemiological metrics (cumulative infections, peak of the number of infections, cumulative cases and peak of the number of cases) relative to baseline scenario for various ρ while β is kept at baseline.

Scenario	β Mult.	ρ Mult.	Infections Cumulative	Infections Peak	Cases Cumulative	Cases Peak
Baseline	1	1	146,850	3,585.40	6,028	258.91
Scenario 2	1	0.9	155,749.45	3,744.78	5,796.04	248.66
Scenario 3	1	0.8	164,841.24	3,912.48	5,490.19	235.52
Scenario 4	1	0.7	174,169.41	4,094.09	5,109.04	219.37
Scenario 5	1	0.6	183,781.58	4,289.28	4,650.27	200.07
Scenario 6	1	1.1	138,100.67	3,438.87	6,186.52	266.95
Scenario 7	1	1.2	129,460.82	3,299.06	6,271.15	272.51

Table 4.6: Absolute values of cumulative and peak infections and cases for different scenarios compared to the baseline. The table illustrates how changes in ρ affect the overall and peak disease metrics.

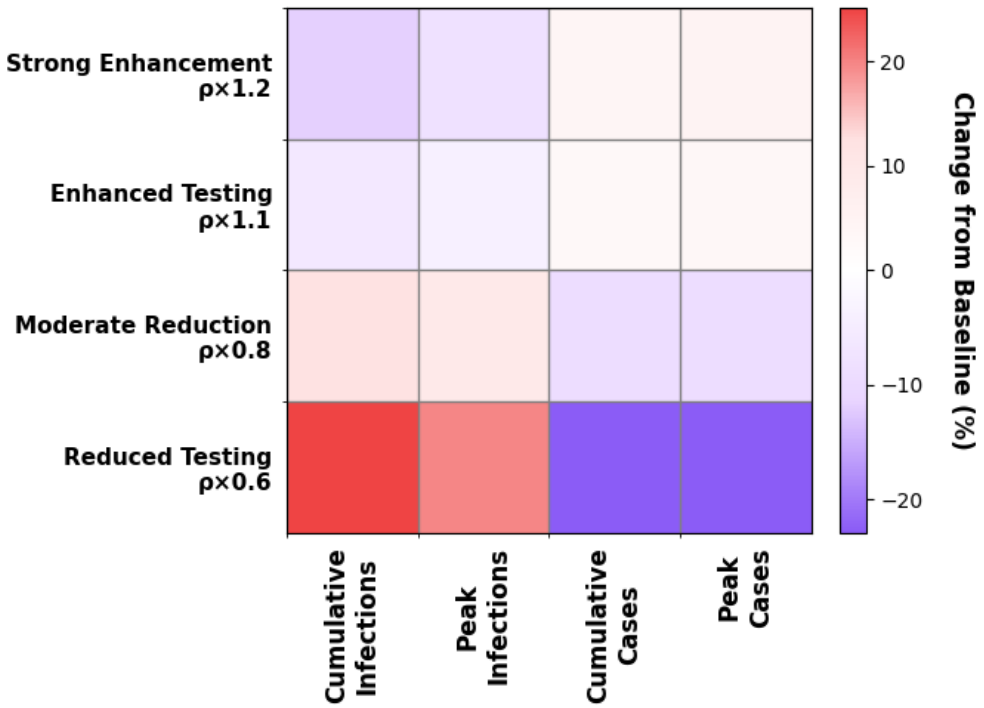


Figure 4.9: Heatmap showing the impact of testing and isolation effectiveness (varying ρ multiplier) on key epidemiological outcomes. Blue indicates reductions; red indicates increases relative to baseline. Enhanced testing ($\rho \times 1.2$) provides modest but consistent reductions across all metrics, whilst reduced testing ($\rho \times 0.6$) increases both infections and cases.

To comprehensively understand the dynamics of COVID-19 transmission and the effectiveness of public health interventions, we examined various scenarios where both social distancing measures (β) and testing strategies (ρ) were simultaneously varied. These scenarios were selected to provide nuanced insight into how alterations in both β and ρ values collectively influence critical epidemiological metrics. For instance, in Scenario 10, characterized by a β multiplier of 1.1 and a ρ multiplier of 0.8, there was a notable 34.20% increase in cumulative infections and a corresponding 34.46% surge in peak infections. Similarly, Scenario 17, marked by a β multiplier of 1.2 and a ρ multiplier of 0.8, demonstrated a substantial 46.24% rise in cumulative infections and a remarkable 58.70% escalation in peak infections. Conversely, Scenario 33, which involved an increased β value of 1.4 and a decreased ρ value of 0.6, exhibited a striking 68.14% increase in cumulative infections and a remarkable 114.96% surge in peak infections. Interestingly, despite the significant escalation in infections, there was a slight decrease of 5.19% in cumulative cases, coupled with a 33.74% decline in peak cases. These diverse scenarios underscore the intricate interplay between social distancing measures and testing effectiveness in shaping the trajectory of COVID-19 outbreaks, emphasizing the importance of adaptive and evidence-based public health interventions.

Scenario	β Mult.	ρ Mult.	Infections Cumulative (%)	Infections Peak (%)	Cases Cumulative (%)	Cases Peak (%)
Scenario 10	1.1	0.8	34.20	34.46	3.12	11.27
Scenario 17	1.2	0.8	46.24	58.70	9.78	30.41
Scenario 24	1.3	0.8	53.33	81.56	13.96	48.38
Scenario 33	1.4	0.6	68.14	114.96	33.74	37.33
Scenario 35	1.4	1.2	40.35	82.20	50.88	111.68

Table 4.7: Percentage change in cumulative and peak infections and cases relative to the baseline scenario for different combinations of β and ρ. The table demonstrates the combined efficiency on key epidemiological metrics.

Scenario	β Mult.	ρ Mult.	Infections Cumulative	Infections Peak	Cases Cumulative	Cases Peak
Baseline	1	1	146,850	3,585.40	6,028	258.91
Scenario 10	1.1	0.8	197073.357	4820.995338	6216.312269	288.0805003
Scenario 17	1.2	0.8	214756.5318	5690.203101	6617.561077	337.6503211
Scenario 24	1.3	0.8	225165.6296	6509.519667	6869.436816	384.1694328
Scenario 33	1.4	0.6	246914.7492	7707.15916	5715.317588	346.2596368
Scenario 35	1.4	1.2	206097.0982	6532.621254	9095.080682	548.0557102

Table 4.8: Absolute values of cumulative and peak infections and cases for various combinations of social distancing β and testing strategies ρ compared to the baseline scenario. This table illustrates the significant effects of altering both social distancing measures and testing strategies on the overall and peak disease metrics.

Figure 4.10 visualises how these interventions interact. These diverse scenarios underscore the intricate interplay between social distancing and testing effectiveness in shaping outbreak trajectories.

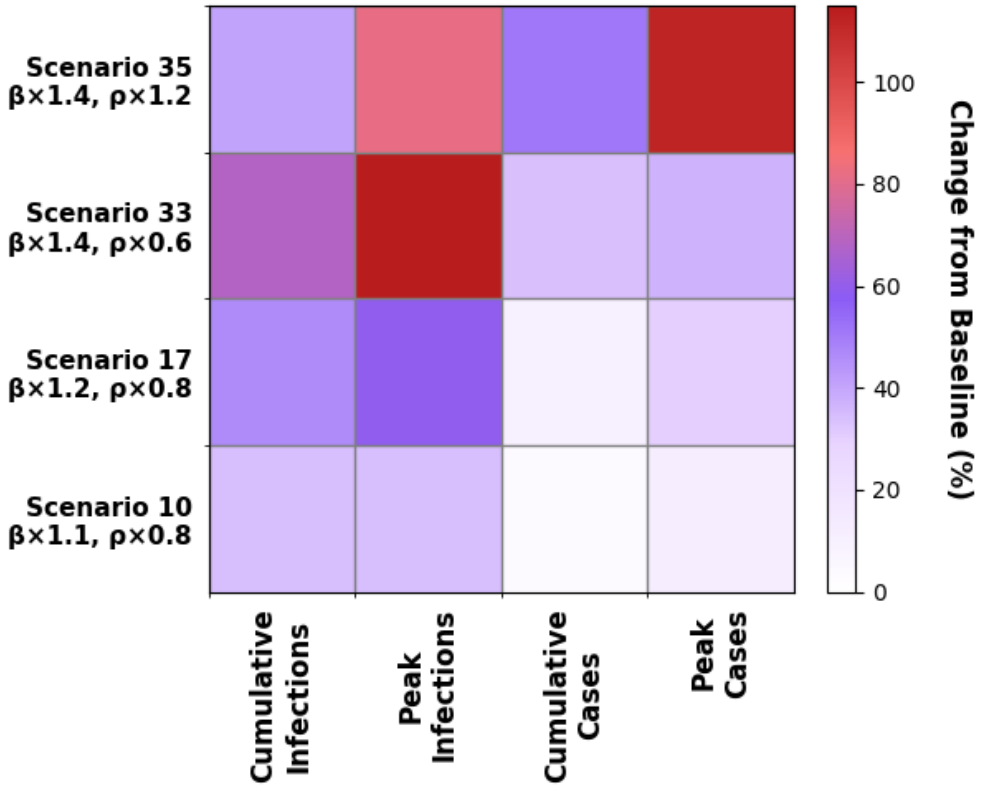


Figure 4.10: Heatmap showing combined effects of social distancing (β multiplier) and testing (ρ multiplier) on key epidemiological outcomes. Notable scenarios: Scenario 10 ($\beta \times 1.1$, $\rho \times 0.8$) shows moderate increases across infections; Scenario 33 ($\beta \times 1.4$, $\rho \times 0.6$) shows dramatic increases in infections but paradoxical case reductions due to detection changes; Scenario 35 ($\beta \times 1.4$, $\rho \times 1.2$) shows enhanced detection partially offsetting increased transmission.

4.2.5 Partial rank correlation coefficients

Sensitivity analysis is essential for validating the reliability and robustness of a model and identifying key parameters that influence its output. Traditional methods, such as local sensitivity analysis, one-at-a-time analysis, global sensitivity analysis, meta-model methods, and Monte Carlo simulations [18], are effective for time-independent parameters but are not suitable for our SIC model's time-dependent parameters $\beta(t)$ and $\rho(t)$.

To address this, we analysed the sensitivity of $\beta(t)$ and $\rho(t)$ by varying their overall profiles over time, which indirectly assesses their impact on model outputs. This approach also helps evaluate the sensitivity related to the timing and extent of public health interventions. Additionally, we

performed sensitivity analysis on other time-independent parameters, such as the initial infected population I_0 and the recovery rate γ.

For the sensitivity analysis, we employed Partial Rank Correlation Coefficients (PRCC) for global sensitivity analysis [126].

Figure 4.11 presents the results of the sensitivity analysis using Partial Rank Correlation Coefficients (PRCC), a method that quantifies the strength and direction of the monotonic relationship between model parameters and outputs while accounting for the influence of other parameters. The analysis focuses on four key model outputs: the proportion of the population infected, the case-infection ratio, total infections, and total cases. Parameters are grouped into three categories: red for the time-dependent transmission coefficients $\beta(t)$, blue for the time-dependent case-infection ratio $\rho(t)$, and orange for the initial number of hidden infections I_0 and the mean recovery rate γ.

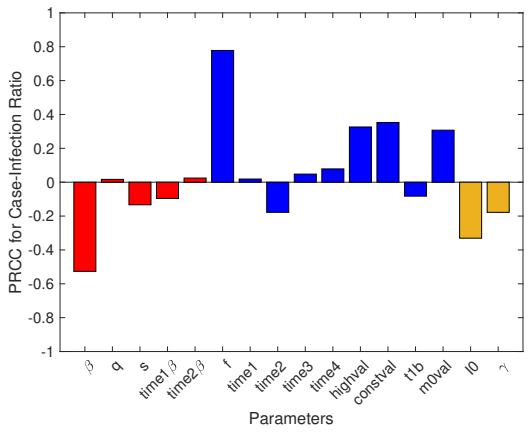
The PRCC analysis for the case-infection ratio (Figure 4.11a) reveals several parameters exhibiting significant sensitivity. Notably, the parameter β (initial transmission) shows negative sensitivity. This suggests that as the initial transmission rate β increases, the case-infection ratio tends to decrease. Conversely, parameter f shows positive sensitivity, indicating that increases in these parameters lead to a decrease in the case-infection ratio.

In examining the proportion of infected individuals (Figure 4.11b), the output shows strong positive sensitivity to the transmission rate parameter, β. This indicates that changes in β critically influence the proportion of the population that becomes infected. Conversely, the output is negatively sensitive to parameter f, which represents the average probability of infected individuals seeking a COVID-19 test. This suggests that as f increases, the proportion of infected individuals tends to decrease.

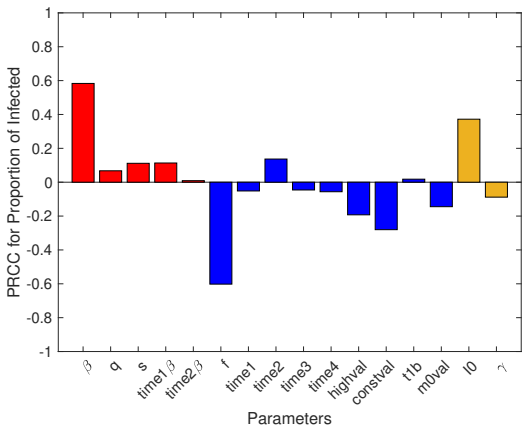
Similarly, in the analysis of total infections (Figure 4.11c), the number of infections is highly sensitive to β, reinforcing the key role of the transmission rate in driving infection dynamics.

For the sensitivity of reported cases (Figure 4.11d), the output again exhibits positive sensitivity to β, consistent with earlier findings. This pattern underscores the critical role of transmission in elevating both infections and reported case counts. The output also shows negative sensitivity to f, indicating that higher testing rates may reduce observed cases. Additionally, the recovery rate parameter, γ, has a negative sensitivity, implying that increased recovery reduces the number of cases. This highlights the potential impact of interventions that improve recovery or reduce infectious periods in managing the epidemic.

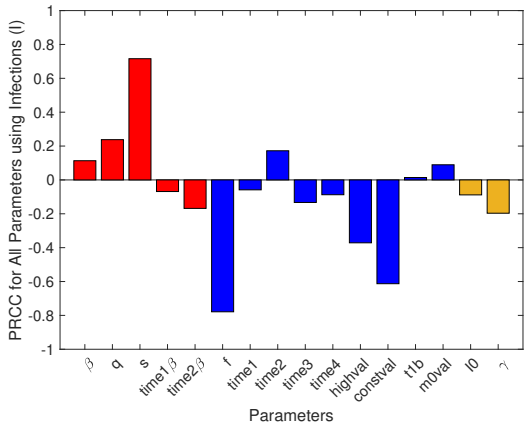
Overall, the consistent positive sensitivity of multiple outputs to β underscores its central role in driving disease spread, emphasizing the importance of targeting transmission rate reduction in control efforts.



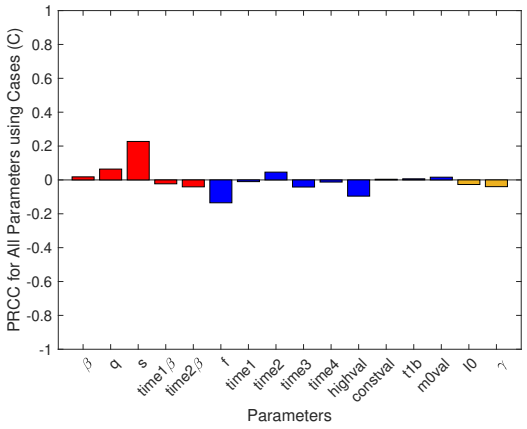
(a) Case-infection ratio.



(b) Proportion of infected.



(c) Infected.



(d) Cases.

Figure 4.11: Sensitivity analysis results using Partial Rank Correlation Coefficients (PRCC) for four model outputs. The red parameters are for time-dependent transmission coefficients $\beta(t)$, blue parameters are for time-dependent case-infection ratio $\rho(t)$, and orange parameters are for the initial number of hidden infections I_0 and the mean recovery rate γ.

4.2.6 Extension of the PRCC method

To provide a more comprehensive assessment of parameter influence across multiple outcomes, we extended the traditional Partial Rank Correlation Coefficient (PRCC) method by utilizing the absolute values of PRCCs and summing them across all outcomes. While PRCC traditionally assesses both the strength and direction of correlations between input parameters and model outputs, focusing on the absolute values enables us to capture the overall strength of each parameter's impact, regardless of whether the effect is positive or negative. This modification aligns conceptually with global sensitivity analysis techniques, such as variance-based methods like Sobol analysis, where the absolute magnitude of sensitivity indices is used to quantify the total influence of parameters on model behaviour [168]. By concentrating on magnitude, this approach provides a clearer measure of parameter contributions, particularly when the interest lies in the overall impact rather than its directional nature. It also facilitates the identification of parameters with the most significant overall impact, highlighting the sensitivity of the model to variations in input parameters.

Furthermore, we introduced a normalization step to calculate the percentage contribution of each parameter to each outcome. This was achieved by dividing the absolute PRCC values by the total sum of absolute PRCCs for the given outcome. Normalization is a common practice in global sensitivity analysis, where relative contributions of each parameter to the overall model variance are computed, as in Sobol indices [168]. Although not part of the standard PRCC framework, this adaptation allows for a logical comparison of parameter importance across multiple outcomes, facilitating the identification of parameters with the most significant influence.

While the underlying PRCC method is well-established, the specific adaptation I propose in this section, combining absolute PRCC values with a normalization step to produce relative percentage contributions, is, to my knowledge, not commonly reported in the literature. This approach provides a practical extension that enhances interpretability and comparison of parameter sensitivities across multiple model outputs without the computational overhead of fully variance-based methods. My contribution lies in demonstrating the utility of this extended PRCC method in the context of a complex epidemiological model, providing an intuitive and computationally efficient means of quantifying and comparing parameter influence across diverse outcomes.

This customization of PRCC offers both computational efficiency and clarity in interpreting parameter influence, making it a practical choice for large-scale models with multiple outputs. Compared to Sobol analysis, which provides greater insight into parameter interactions but often requires significantly more computational resources, the extended PRCC method offers a more accessible yet still robust sensitivity analysis. The Morris method, by contrast, delivers a more qualitative assessment but lacks the quantitative depth offered by both PRCC and Sobol indices. By incorporating absolute values and normalization into PRCC, we achieve a balance between computational feasibility and the ability to quantify and compare parameter impacts across various

outcomes.

However, it is important to recognize the limitations of this extended PRCC method. PRCC is most effective when the relationships between parameters and outcomes are monotonic; it may underestimate the importance of parameters in scenarios with non-linear or highly non-monotonic relationships, potentially leading to biased conclusions about parameter sensitivity [125]. Additionally, PRCC does not explicitly account for parameter interactions. When input parameters are strongly correlated, interpreting PRCC values becomes more complex, as the impact of one parameter may be confounded by its correlation with another. While variance-based methods such as Sobol analysis are better suited to handle these interactions, careful interpretation is needed when using PRCC in such contexts.

Despite these considerations, the use of absolute values and normalization in PRCC enables us to focus on the magnitude of influence, providing a clearer and more intuitive measure of parameter importance. These extensions enhance the traditional PRCC method, allowing for a rigorous sensitivity analysis while maintaining computational efficiency, making it an appropriate tool for complex models.

The analysis presented in Figure 4.12 was conducted to evaluate the relative influence of various model parameters on key outcomes, including the proportion of infected individuals, the case-infection ratio, the total number of infections, and the total number of reported cases. Using PRCC analysis, we quantified the sensitivity of each outcome to changes in the input parameters, enabling a rigorous evaluation of the factors that significantly impact the model's behaviour.

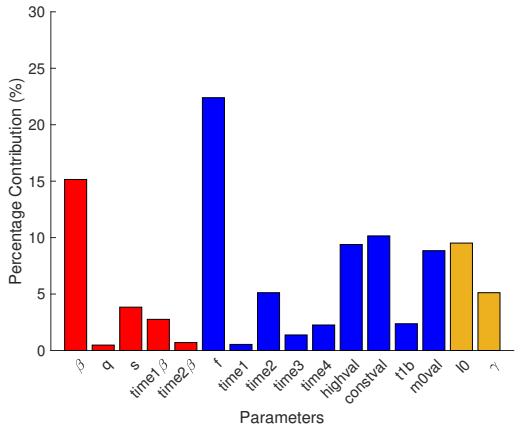
To achieve this, the absolute values of the PRCCs for each parameter were computed for the four key outcomes: proportion of infected individuals, case-infection ratio, total infections, and total cases. The absolute sum of these PRCC values across all outcomes was then calculated to determine the total sensitivity of the model to each parameter, providing a basis for understanding which parameters had the greatest overall impact on the model.

Finally, the percentage contribution of each parameter to each outcome was determined by dividing the absolute PRCC values by the total sum of absolute PRCCs for that outcome and then multiplying by 100. This normalization step enabled a clearer comparison of parameter impacts within each specific outcome, offering a more intuitive interpretation of which parameters were the most significant contributors to the model dynamics.

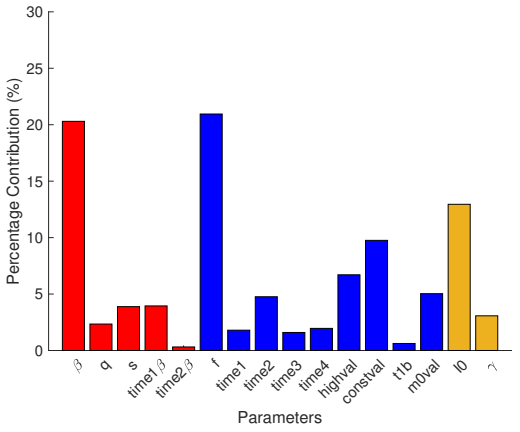
4.2.7 Extended Fourier amplitude sensitivity test

To further explore the impact of input parameters on model outputs, a complementary sensitivity analysis using the extended Fourier Amplitude Sensitivity Test (eFAST) was conducted .

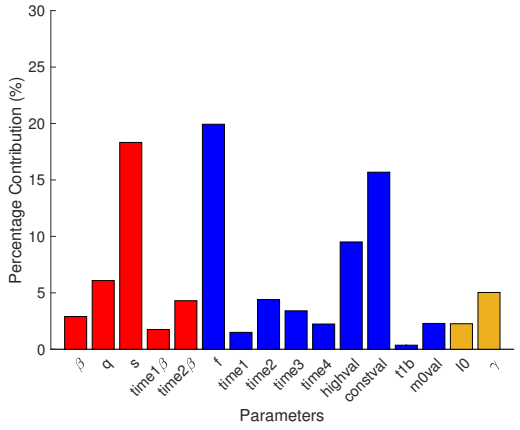
The **Extended Fourier Amplitude Sensitivity Test (eFAST)** [63] is a global sensitivity analysis method that uses variance decomposition to assess the impact of model parameters on



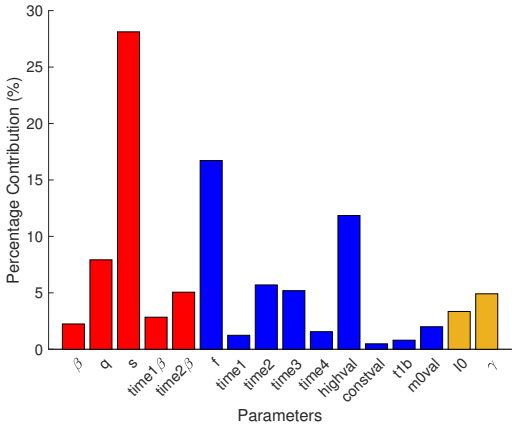
(a) Case infection ratio.



(b) Proportion of infected.



(c) Infected.



(d) Percentage contribution for case.

Figure 4.12: Percentage contributions of parameters to key model outcomes. This analysis extends the traditional Partial Rank Correlation Coefficient (PRCC) method by using absolute PRCC values, enabling a clearer comparison of parameter influence across multiple outcomes. Each panel represents the normalized percentage contribution of model parameters to (a) the case-infection ratio, (b) the proportion of infected individuals, (c) the total number of infections, and (d) the total number of reported cases. The red parameters are for time-dependent transmission coefficients $\beta(t)$, blue parameters are for time-dependent case-infection ratio $\rho(t)$, and orange parameters are for the initial number of hidden infections I_0 and the mean recovery rate γ.

a specific outcome. It analyzes model outputs by varying parameters at distinct frequencies and applies Fourier analysis to determine the contribution of each parameter's frequency component to the total variance.

FAST uses the total-order sensitivity index to rank the importance of parameters, capturing both direct and interaction effects. While the sum of these indices equals 1 for linear models, it can exceed 1 for nonlinear models, making eFAST especially useful for models with non-monotonic responses where other methods, like partial rank correlation coefficients (PRCC), may be less effective.

The methodology involves applying sinusoidal functions (search curves) for each parameter, with phase-shifted resampling to increase efficiency. Fourier analysis is then used to calculate the sensitivity index, followed by statistical tests like ANOVA and Dunnett's test to assess whether the indices for each parameter significantly differ from those of a dummy parameter.

eFAST does not require posterior samples like PRCC, making it well-suited for deterministic models or situations where posterior distributions are unavailable or unnecessary.

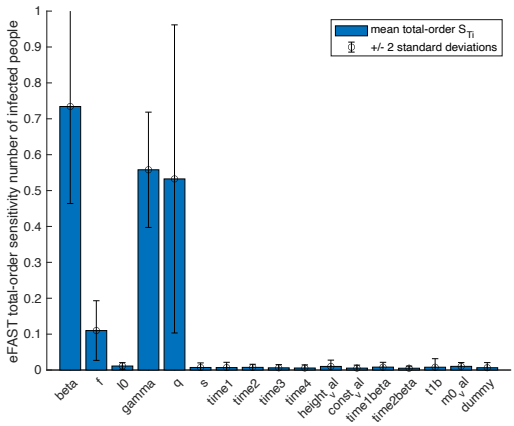
The analysis entailed generating 40 search curves with 105 samples per curve and utilizing up to four Fourier coefficients to minimize frequency interferences. Frequencies for each parameter were selected based on the number of parameters and sample size, with the model being evaluated multiple times per parameter combination to ensure robust results. Frequencies and parameter distributions were adjusted iteratively, with the model being rerun until the outputs remained consistent across multiple evaluations. Sensitivity indices for both first-order and total-order effects were calculated, focusing on the number of infected individuals, cumulative cases, and the proportion of infected individuals.

Table 4.9 presents the results of the global sensitivity analysis conducted using eFAST. The values in the table are First-Order Sensitivity Indices (S_i) for each variable across three key outcomes: hidden infections, total cases, and the proportion of infected individuals. These indices, often referred to as first-order effects, measure the direct contribution of each input variable to the variance in the output, ignoring any interactions with other parameters. In other words, S_i quantifies the extent to which a variable, by itself, influences the model's output.

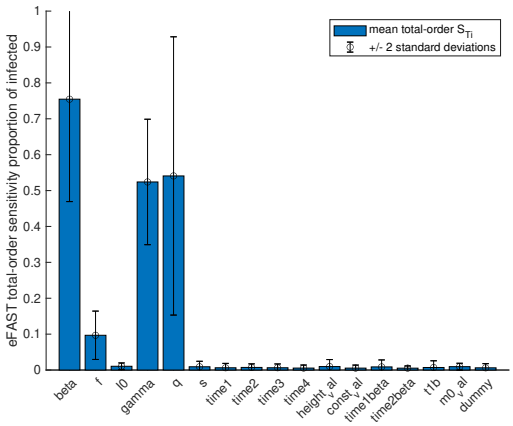
Higher S_i values indicate that a given variable has a greater direct impact on the corresponding outcome. In this analysis, beta has the largest S_i across all three outputs, suggesting that the initial transmission rate is the most influential parameter driving hidden infections, total cases, and the proportion of infected individuals. Other parameters, such as γ (recovery rate) and q (fraction of reduction in transmission due to lock-down), also exhibit moderate sensitivity.

This analysis highlights the importance of focusing on high-S_i parameters when calibrating or optimizing model performance, as these parameters are the primary drivers of the system's behaviour. Lower-S_i parameters, while still part of the model, have relatively minor direct effects on the outcomes.

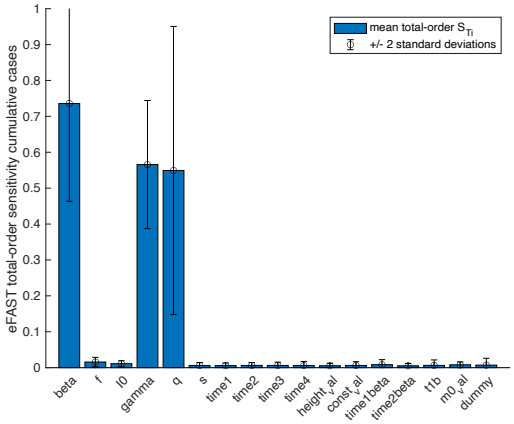
The results of the total-order sensitivity analysis are depicted in Figure 4.13. Figures 4.13a, 4.13b, 4.13c shows that the parameter beta has the highest S_{T_i} suggesting it is the most influential parameter in determining the number of infected people.Parameters gamma and q also exhibit significant sensitivity, indicating their importance in the model.



(a) Number of infected people.



(b) Proportion of infected.



(c) Cumulative cases

Figure 4.13: eFAST sensitivity analysis showing total-order sensitivity S_{T_i} for different metrics. The blue bars indicate the mean total-order sensitivity, while the error bars represent ± 2 standard deviations. (a) Sensitivity analysis for the number of infected people. (b) Sensitivity analysis for the proportion of infected. (c) Sensitivity analysis for cumulative cases.

Variable	Hidden Infection	Cases	Proportion of infected
β	0.447	0.460	0.463
f	0.044	0.004	0.037
I_0	0.003	0.003	0.003
γ	0.255	0.240	0.233
q	0.207	0.241	0.221
s	0.000	0.000	0.001
time_1	0.000	0.000	0.000
time_2	0.001	0.000	0.001
time_3	0.000	0.000	0.000
time_4	0.000	0.000	0.000
height_val	0.002	0.000	0.002
const_val	0.000	0.000	0.000
time1beta	0.001	0.001	0.001
time2beta	0.000	0.000	0.000
t_{1b}	0.000	0.000	0.000
m0_val	0.003	0.001	0.002
dummy	0.000	0.000	0.000

Table 4.9: Median First-Order Sensitivity Indices (S_i) for each variable from the eFAST Global Sensitivity Analysis. The table presents the sensitivity indices for three key outputs: Hidden Infections, Total Cases, and the Proportion of Infected Individuals. The indices were computed using the eFAST method, with median values reported across all sample points.

4.3 Discussion

In this study, we employed time-dependent parameters initially to accurately model the evolving dynamics of COVID-19 transmission in Alberta. This approach allowed us to account for temporal variations in factors such as social distancing behaviours, testing capacity, and public health interventions, providing a realistic representation of the pandemic's progression during the initial wave.

As our focus transitioned toward scenario analysis to evaluate different intervention strategies, we recognized the need for a more streamlined and interpretable modelling framework. To this end, we shifted from time-dependent to constant parameter values. This methodological adjustment enabled more straightforward comparisons between intervention scenarios and helped isolate the specific effects of changes in public health measures on epidemiological outcomes.

By adopting constant parameters in our scenario analysis, we enhanced the interpretability of our results, enabling policymakers to grasp the potential implications of different intervention scenarios more readily. Moreover, this approach provided a practical means of assessing the robustness of our findings across different contexts and time-frames, ensuring consistency and facilitating replication in future studies or different geographical settings.

To assess the influence of model parameters, we conducted a series of sensitivity analyses. Initially, we used PRCC to examine how individual parameters influenced four key outcomes: the proportion of infected individuals, the case-infection ratio, total infections, and total reported cases. PRCC enabled us to quantify the monotonic relationships between inputs and outputs while accounting for interdependencies among parameters.

We then extended the PRCC method by using the absolute values of PRCCs and normalizing them to calculate the percentage contribution of each parameter to each outcome. This modification allowed us to compare the magnitude of parameter influence across multiple outputs and highlighted the dominant role of specific variables, particularly the initial transmission rate β.

Finally, we performed a global sensitivity analysis using the Extended Fourier Amplitude Sensitivity Test (eFAST). The eFAST analysis reaffirmed the central importance of β, with high first-order and total-order sensitivity indices across all outcomes. Parameters such as γ (recovery rate) and q (representing transmission reduction due to lockdown) also showed moderate influence, particularly in the total-order indices that capture interaction effects.

A comparison of the three sensitivity approaches reveals important insights. Naive scenario comparisons (Section 4.2.4) offered intuitive but less rigorous insights, helping highlight the impact of interventions on outcomes through manually adjusted parameters. PRCC (Section 4.2.5) provided a computationally efficient method to assess monotonic sensitivities. The extended PRCC (Section 4.2.6) improved interpretability by summarizing absolute influence across multiple outcomes, offering a useful approximation to full variance-based methods with low computational cost. Lastly, eFAST (Section 4.2.7) served as a rigorous variance-based analysis capable of detecting both nonlinearities and parameter interactions, confirming the robustness of key findings from PRCC and extended PRCC.

Overall, these analyses consistently identified β as the most influential parameter across all methods, underscoring its central role in epidemic propagation. This reinforces the importance of intervention strategies aimed at reducing transmission rates, such as enhanced social distancing, timely lockdowns, or vaccination campaigns. Parameters like γ and q also emerged as meaningful contributors, suggesting a multifaceted approach to epidemic control that combines treatment, testing, and behavioural interventions.

Looking forward, future research could refine these models further by incorporating additional parameters such as vaccination rates and public compliance levels, thereby offering a more comprehensive understanding of pandemic dynamics and strengthening the basis for evidence-based policy recommendations.

5

Mathematical modelling of the dynamics of COVID-19 variants of concern: asymptotic and finite-time perspectives

Overview of the chapter This chapter reproduces my paper published Mathematical modeling of the dynamics of COVID-19 variants of concern: Asymptotic and finite-time perspectives A-S Ciupeanu et. al in Infectious Disease Modelling, No 4, p. 581-596, 2022 [53] with only minor stylistic adjustments. Figures were resized for improved clarity, but their content remains unchanged.

The COVID-19 pandemic has seen multiple waves, in part due to the implementation and relaxation of social distancing measures by the public health authorities around the world, and also caused by the emergence of new variants of concern (VOCs) of the SARS-Cov-2 virus. As the COVID-19 pandemic is expected to transition into an endemic state, how to manage outbreaks caused by newly emerging VOCs has become one of the primary public health issues. Using mathematical modelling tools, we investigated the dynamics of VOCs, both in a general theoretical framework and based on observations from public health data of past COVID-19 waves, with the objective of understanding key factors that determine the dominance and coexistence of VOCs. Our results show that the transmissibility advantage of a new VOC is the main factor for it to

become dominant. Additionally, our modelling study indicates that the initial number of people infected with the new VOC plays an important role in determining the size of the epidemic. Our results also support the evidence that public health measures targeting the newly emerging VOC taken in the early phase of its spread can limit the size of the epidemic caused by the new VOC [202, 203].

5.1 Introduction

The COVID-19 pandemic, caused by the infection of the SARS-Cov-II virus, has become one of the most severe and deadly pandemics in recent history. By May 2022, more than two years after its first known outbreak in December 2019, the WHO reported over 6.28 million COVID-19 deaths and over half a billion confirmed COVID-19 cases [201], while the total number of people infected with COVID-19 is believed to be much greater. It may take many years from now to fully ascertain the health burden and socioeconomic impact of the pandemic.

The COVID-19 pandemic has included multiple waves. These waves are mainly due to the implementation and relaxation of non-pharmaceutical interventions and the emergence of new variants of concern (VOCs). These VOCs have been observed worldwide. A COVID-19 variant contains one or more mutations in its viral genome. Certain COVID-19 variants have higher transmissibility and severity in populations than other COVID-19 variants. Emerging VOCs are those variants that are considered to have a distinguishable and significant health impact. Global travel and the timing of non-pharmaceutical interventions have made it difficult to determine when the importation of new VOCs may enter a given geographic region. The Alpha, Beta, and Gamma variants were first detected between October and November 2020 in the United Kingdom, South Africa, and Brazil, respectively [139]. These three variants then spread to other countries through global travel, and countries around the world experienced dissimilar transmission dynamics of these variants [52].

In Canada, these VOCs were first detected in December 2020 (Alpha), January 2021 (Beta), and March 2021 (Gamma). Each of these variants contributed differently to the transmission dynamics observed in the third wave. More recently, the emergence and spread of the Omicron variant (BA.1 and BA.2) was a major driver of the fifth and sixth waves in Canada. An in-depth understanding of the COVID-19 VOCs transmission dynamics during the previous waves can provide valuable new insights on how to effectively prevent and control future waves of VOCs.

Mathematical modelling has been widely used as a research and policy tool. During the COVID-19 pandemic, academic researchers and government agencies worldwide have used mathematical models to help understand the spatial and temporal dynamics of the COVID-19 disease. These models have incorporated dynamical features motivated by disease dynamics, emerging variants,

vaccine dynamics, and public health policies [26, 48, 104, 113, 202, 203, 210]. The public health policies included the following strategies: travel restrictions, social distancing, isolation, testing for cases, and contact tracing. COVID-19 models have provided valuable insights and evidence that helped inform policies in a continuously changing pandemic.

A common feature of VOCs of COVID-19 has been their increased transmissibility, and they are expected to have a higher basic reproduction number. By the standard theory of the multi-strain competition [16], under general assumptions of strain competition, the emerging VOC with the highest basic reproduction number will be able to invade a population and replace the wild-type or existing variants according to the competitive exclusion principle in ecology [16, 45, 69, 114]. Several mechanisms for co-existence of strains of the pathogen in a population have been established in the literature, including super-infection and co-infection, mutation of one strain to another, cross immunity among strains, and population age heterogeneity in which different strains preferentially infect different age groups (see [128] for reviews). Time-periodic infection rates caused by seasonality and environmental influences have also been shown to lead to strain co-existence [127]. These mechanisms for co-existence are generally not applicable to COVID-19 epidemics, and we assume the competitive exclusion principle holds for COVID-19 VOCs. The mathematical theories of multi-strain competition and competitive exclusion are based on the asymptotic behaviours of solutions to mathematical models when time is infinitely large. For finite-time horizon real-world epidemics such as epidemic waves of COVID-19, variant and strain dominance and coexistence often are not as clearly defined as in the theory of competition. As was shown in [113, 202, 203], public health interventions can play a key role in mitigating, and possibly preventing, an emerging VOC from becoming dominant in the population.

The main objective of our study was to understand the dynamics of variant competition using a two-variant mathematical model and to interpret these behaviours in the context of emerging COVID-19 VOCs. The overall analysis will include both asymptotic behaviours based on the theory of mathematical epidemiology, and finite-time dynamics during a single epidemic wave using public health COVID-19 data from the provinces of Alberta and British Columbia, Canada. Our investigation focused on the following questions:

i) How to interpret the dominance and coexistence of variants within the finite-time horizon of a COVID-19 wave, in comparison to the infinite-time horizon of the asymptotic limits?

ii) What are the different characteristics of variant dominance or coexistence in finite-time and infinite-time horizons?

iii) How to distinguish between dominance and coexistence of variants during a COVID-19 wave?

iv) What public health measures can be implemented to prevent emerging variants from becoming dominant and/or mitigate the spread and size of the resulting epidemic?

Multiple variant dynamics of COVID-19 have not been widely discussed in the modelling literature, especially in the context of the finite-time dynamics during an epidemic or a single epidemic wave. With both an asymptotic and finite-time perspective, our study enriches the theory of variant/strain competition and it provides actionable insights to public health interventions related to preventing, mitigating, and managing emerging COVID-19 VOCs.

In the next section, we illustrate the derivation of a general two-variant model for infectious diseases that includes COVID-19 as a special case. Section 3 provides a detailed mathematical analysis of the model and Section 4 presents numerical simulation results using public health data from Alberta, Canada.

5.2 Derivation of the model

We developed an SIRS type of compartmental model for the transmission of two viral variants in the population. Historically, COVID-19 in Canada has presented with a clearly identified dominant variant and a newly emerging VOC. This has motivated our consideration of using a two variant rather than a multi-variant model. Furthermore, the mathematical analysis was simplified using a two-variant model. To further reduce the technicality in the mathematical analysis, we considered a model of SIRS type. More complex models such as SEIR and SEIAR types that include latent and asymptomatic compartments have been used for COVID-19 dynamics. In [163], the authors have shown that SIR models perform better than SEIR or more complex compartmental models to represent the public health data of COVID-19.

Our mathematical model has four compartments: number of individuals susceptible to the viral infection at time t, $S(t)$; number of individuals infected with variants 1 and 2 at time t and not detected by the public health surveillance, $I_1(t)$ and $I_2(t)$; and number of individuals recovered (detected or undetected from testing) and remain protected against infection at time t, $R(t)$. Individuals detected from testing transition to R and these individuals do not contribute to further transmission. Undetected individuals transition to R based on an average infectious period and these individuals also do not contribute to further transmission.

For COVID-19, the evidence of an individual having co-infection from two or more variants has been rare, with only a number of case reports of co-infection (see [35, 38, 61, 161, 191]). Considering the evidence of co-infection during the COVID-19 pandemic, we assumed that co-infection from two or more variants is negligible and we have a single recovered compartment, R, in the model. The transmission coefficients for variants 1 and 2 are given by β_1 and β_2, respectively. Each transmission coefficient is a product of two factors: the average contact rate between the susceptible and infected individuals, and the probability of transmission per contact. During COVID-19, social-distancing measures and face masking were aimed at reducing the contact rate

among individuals and the probability of transmission, respectively. The parameters γ_1 and γ_2 are the natural recovery rates from variant 1 infection and variant 2 infection, respectively, for infected people who are not detected by the public health surveillance. Parameters ρ_1 and ρ_2 are case-infection ratios for variants 1 and 2, respectively. A COVID-19 case is an individual diagnosed as COVID-19 positive by a PCR test as is recorded in the public health surveillance system. An undetected infection is an individual infected with COVID-19 but not recorded in the public health surveillance system. The case-infection ratio is defined as the number of new cases divided by the number of people living with undetected COVID-19 infections, or the so-called hidden infections. The ratio $1/\rho_i$, $i = 1, 2$, measures the number of undetected COVID-19 infections by variant i in the community for each newly diagnosed variant i case, and it is a measure of the effectiveness of public health surveillance. Daily positive cases identified through testing is denoted by $\rho_1 I_1(t)$ and $\rho_2 I_2(t)$ for variants 1 and 2, respectively. Once detected through testing, individuals are assumed to not infect others until recovered, which is analogous to isolation requirements and/or reduced social interactions while ill. The rate at which individuals lose immunity is given by the parameter δ. The parameter Λ is the influx of susceptible individuals from birth and migration. When modeling a short duration epidemic, such as a COVID-19 wave, births and baseline deaths are often negligible, and, in this case, the parameter Λ is set to 0. Since the COVID-19 pandemic is transitioning to an endemic state, the assumption of $\Lambda > 0$ is helpful to assess the long-term effects of the infection. Also, with the consideration of the COVID-19 endemic state, we assumed the death rates (baseline or infection-related) are positive.

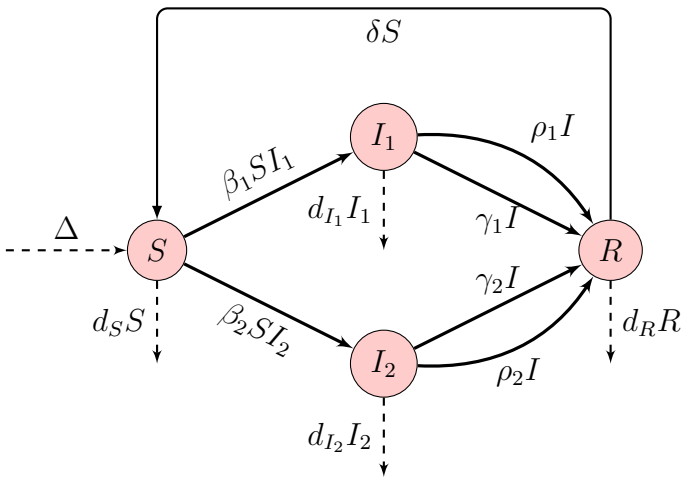


Figure 5.1: Transfer diagram illustrating the SIRS model dynamics with two variants. Nodes represent compartments (S for Susceptible, I_1, I_2 for Infected variants, and R for Recovered).

The model is depicted in the transfer diagram in Figure 5.1. Model parameters are listed in Table 5.1, together with their biological meanings. Based on our assumptions and the transfer diagram in Figure 5.1, the following system of differential equations can be derived for the model:

$$\frac{dS}{dt} = \Lambda - \beta_1 S I_1 - \beta_2 S I_2 - d_S S + \delta R, \tag{5.2.1a}$$
$$\frac{dI_1}{dt} = \beta_1 S I_1 - \gamma_1 I_1 - \rho_1 I_1 - d_{I_1} I_1, \tag{5.2.1b}$$
$$\frac{dI_2}{dt} = \beta_2 S I_2 - \gamma_2 I_2 - \rho_2 I_2 - d_{I_2} I_2, \tag{5.2.1c}$$
$$\frac{dR}{dt} = \gamma_1 I_1 + \gamma_2 I_2 + \rho_1 I_1 + \rho_2 I_2 - d_R R - \delta R. \tag{5.2.1d}$$

Table 5.1: Model Parameters for the SIR Framework with Two Infected Compartments: Biological Interpretations and Descriptions. The model describes the dynamics of two variants of an infectious disease, each with its own set of transmission and recovery rates.

Parameter	Description
β_1	transmission rate of variant 1
β_2	transmission rate of variant 2
γ_1	recovery rate of variant 1
γ_2	recovery rate of variant 2
ρ_i	case-infection ratio for variant i, $i = 1, 2$
δ	rate of immunity loss
d_S	background death rate of susceptible people
d_{I_1}	death rate of infected people (variant 1)
d_{I_2}	death rate of infected people (variant 2)
d_R	death rate of recovered people
S_0	initial susceptible population size
I_{01}	initial number of infections for variant 1
I_{02}	initial number of infections for variant 2
R_0	initial people immune

5.3 Model analysis

By examining the direction of the vector field of system (5.2.1) on coordinate subspaces of $\mathbb{R}^4$, we can verify that the nonnegative orthant $\mathbb{R}^4_+$ is positively invariant under the flow of model (5.2.1), namely, solutions with nonnegative initial conditions will remain nonnegative, and the model is well-posed.

Adding all equations in system (5.2.1) leads to

$$(S + I_1 + I_2 + R)' = \Lambda - d_s S - d_{I_1} I_1 - d_{I_2} I_2 - d_R R \leqslant \Lambda - d(S + I_1 + I_2 + R),$$

where $d = \min\{d_s, d_{I_1}, d_{I_2}, d_R\} > 0$. This implies that $\limsup_{t\to\infty}(S(t)+I_1(t)+I_2(t)+R(t)) \leqslant \Lambda/d$. We study the system (5.2.1) in the following feasible region:

$$\Gamma = \{(S, I_1, I_2, R) \in \mathbb{R}^4_+ \mid S + I_1 + I_2 + R \leqslant \frac{\Lambda}{d}\}, \tag{5.3.1}$$

which is positively invariant and contains the global attractor of model (5.2.1) in $\mathbb{R}^4_+$.

5.3.1 Equilibria and stability analysis

Model (5.2.1) always has the disease-free equilibrium $P_0 = (\frac{\Lambda}{d_S}, 0, 0, 0)$. Since waning immunity is included, when the infection is not present, previously acquired immunity will be lost at the equilibrium P_0 and the entire population will be susceptible. There are two possible single-variant equilibria: $P_1 = (\bar{S}, \bar{I}_1, 0, \bar{R})$ and $P_2 = (\bar{\bar{S}}, 0, \bar{\bar{I}}_2, \bar{\bar{R}})$, where

$$\bar{S} = \frac{\gamma_1 + \rho_1 + d_{I_1}}{\beta_1} \quad \text{and} \quad \bar{\bar{S}} = \frac{\gamma_2 + \rho_2 + d_{I_2}}{\beta_2}. \tag{5.3.2}$$

The single-variant equilibrium P_1 exists in Γ if

$$\mathcal{R}_{01} := \frac{\beta_1}{\gamma_1 + \rho_1 + d_{I_1}} \frac{\Lambda}{d_S} > 1, \tag{5.3.3}$$

it is outside of $\mathbb{R}^4_+$ if $\mathcal{R}_{01} < 1$, and it coincides with P_0 if $\mathcal{R}_{01} = 1$. Similarly, P_2 exists if

$$\mathcal{R}_{02} := \frac{\beta_2}{\gamma_2 + \rho_2 + d_{I_2}} \frac{\Lambda}{d_S} > 1, \tag{5.3.4}$$

it is outside of $\mathbb{R}^4_+$ if $\mathcal{R}_{02} < 1$, and it coincides with P_0 if $\mathcal{R}_{02} = 1$.

We note that, for $i = 1, 2$, threshold parameter $\mathcal{R}_{0i}$ is the basic reproduction number for the variant i when it is the only variant present in the population. Let

$$\mathcal{R}_0 := \max\{\mathcal{R}_{01}, \mathcal{R}_{02}\}. \tag{5.3.5}$$

Then $\mathcal{R}_0$ is the basic reproduction number for the two-variant model (5.2.1), namely, it measures the average number of the secondary infections caused by a single infective with either variants during its entire infectious period.

Can a coexistence equilibrium $P^* = (S^*, I_1^*, I_2^*, R^*)$ exist when $\mathcal{R}_0 > 1$? Assuming $I_1^*, I_2^* > 0$, then S^* needs to satisfy

$$\beta_1 S^* = \gamma_1 + \rho_1 + d_{I_1} \quad \text{and} \quad \beta_2 S^* = \gamma_2 + \rho_2 + d_{I_2}$$

simultaneously. This is only possible if $(\gamma_1 + \rho_1 + d_{I_1})/\beta_1 = (\gamma_2 + \rho_2 + d_{I_2})/\beta_2$, namely, when $\mathcal{R}_{01} = \mathcal{R}_{02}$. Furthermore, in such a case, infinitely choices of $I_1^*, I_2^* > 0$ are possible as solutions of a linear equation

$$\Big[\beta_1 S^* - \frac{\delta}{\delta + d_R}(\gamma_1 + \rho_1)\Big] I_1^* + \Big[\beta_2 S^* - \frac{\delta}{\delta + d_R}(\gamma_2 + \rho_2)\Big] I_2^* = \Lambda.$$

and accordingly, a line segment defined by this equation connects P_1 and P_2 and consists entirely of positive equilibria (see Appendix for a proof).

5.3.2 Stability analysis

Local stability analysis of equilibria can be carried out using the method of linearization and the Routh-Hurwitz criteria. We state the following result that summarizes the existence and stability of the equilibria. Technical proofs are presented in the Appendix. It can be shown that similar results hold for SEIR or more complex models of this type.

Theorem 5.1. *Let $\mathcal{R}_{01}, \mathcal{R}_{02}, \mathcal{R}_0, \bar{S}$, and $\bar{\bar{S}}$ be defined in* (5.3.2) - (5.3.5)*. The following statements hold.*

I If $\mathcal{R}_0 < 1$, then the disease-free equilibrium P_0 is the only equilibrium in the feasible region Γ and it is asymptotically stable.

II If $\mathcal{R}_0 > 1$, then P_0 is unstable. Furthermore,

(a) if $\mathcal{R}_{02} < 1 < \mathcal{R}_{01}$, then the single-variant equilibrium $P_1 = (\bar{S}, \bar{I}_1, 0, \bar{R})$ exists and is asymptotically stable, while P_2 does not exist in Γ.

(b) if $1 < \mathcal{R}_{02} < \mathcal{R}_{01}$, then both single-variant equilibria P_1 and P_2 exists in Γ. P_1 is asymptotically stable while P_2 is unstable in the direction pointing to the interior of Γ.

(c) if $\mathcal{R}_{01} < 1 < \mathcal{R}_{02}$, then the single-variant equilibrium $P_2 = (\bar{\bar{S}}, 0, \bar{\bar{I}}_2, \bar{\bar{R}})$ exists and is asymptotically stable, while P_1 does not exist in Γ.

(d) if $1 < \mathcal{R}_{01} < \mathcal{R}_{02}$, then both single-variant equilibria P_1 and P_2 exists in Γ. P_2 is asymptotically stable while P_1 is unstable in the direction pointing to the interior of Γ.

(e) if $\mathcal{R}_{01} = \mathcal{R}_{02} > 1$, then both P_1 and P_2 exist, and there exists a line segment in Γ that consists entirely of positive equilibria and connects P_1 and P_2. Each positive equilibrium is neutrally stable in the direction of the line segment, and is asymptotically stable in directions transversal to the line segment.

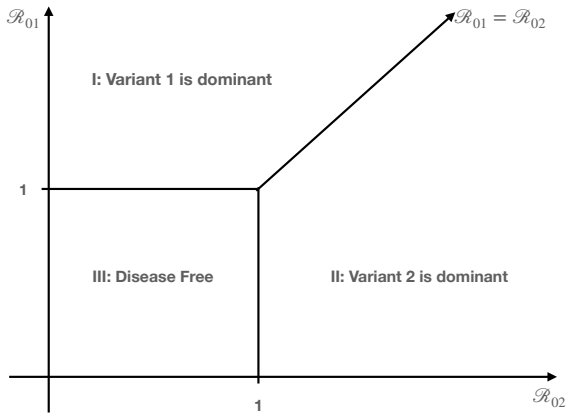


Figure 5.2: A diagram illustrating the results in Theorem 5.1. In region I, where $\mathcal{R}_{01} > \mathcal{R}_{02} > 1$ holds, variant 1 will become dominant and eventually drive the variant 2 to extinction; in region II, where relation $\mathcal{R}_{02} > \mathcal{R}_{01} > 1$ holds, variant 2 is dominant and drives variant 1 to extinction; and in region III, both $\mathcal{R}_{01}$ and $\mathcal{R}_{02}$ are less than 1, and neither variant can establish itself in the population and the disease dies out. On the half line defined by $\mathcal{R}_{01} = \mathcal{R}_{02} > 1$, both variants are able to coexist in the population.

Biologically, results in Theorem 5.1 on the existence and stability of equilibria infer that outcomes of the variants in the population are determined by the variant-specific reproduction numbers $\mathcal{R}_{01}$ and $\mathcal{R}_{02}$ as defined in (5.3.3) and (5.3.4). As illustrated in Figure 5.2, in region I, relation $\mathcal{R}_{01} > \mathcal{R}_{02} > 1$ holds, both variants are able to establish in the population. Variant 1 has the larger basic reproduction number and will become dominant and eventually drive the variant 2 to extinction; in region II, the reverse relation $\mathcal{R}_{02} > \mathcal{R}_{01} > 1$ holds, and variant 2 will be dominant and drive variant 1 to extinction; and in region III, both $\mathcal{R}_{01}$ and $\mathcal{R}_{02}$ are less than 1, and neither variant can establish itself in the population and the disease dies out. This is consistent with $\mathcal{R}_0 = \max\{\mathcal{R}_{01}, \mathcal{R}_{02}\} < 1$.

We pay a special attention to the case in the diagram in Figure 5.2 when $\mathcal{R}_{01} = \mathcal{R}_{02} > 1$. This is the only scenario under which both variants can coexist in the population under our model assumptions. Mathematically, a line of equilibria (case (e) in Theorem 5.1) is non-generic and not all equilibria on the line will survive under small perturbations. Furthermore, the half line given by $\mathcal{R}_{01} = \mathcal{R}_{02} > 1$ has measure 0 in the 2-dimensional parameter region $\{(\mathcal{R}_1, \mathcal{R}_2) \mid \mathcal{R}_{01} > 1, \mathcal{R}_{02} > 1\}$. These facts suggest that coexistence of the two variants in the sense of positive asymptotic limits is unlikely.

In real-world epidemics such as the COVID-19, dominance and coexistence of variants are often discussed within a finite time horizon (e.g. a single epidemic wave or pandemic) rather than an infinitely long time (or within asymptotic limits). In this paper, we use *practical* and *theoretical* dominance and coexistence to distinguish between finite and infinite time analysis and

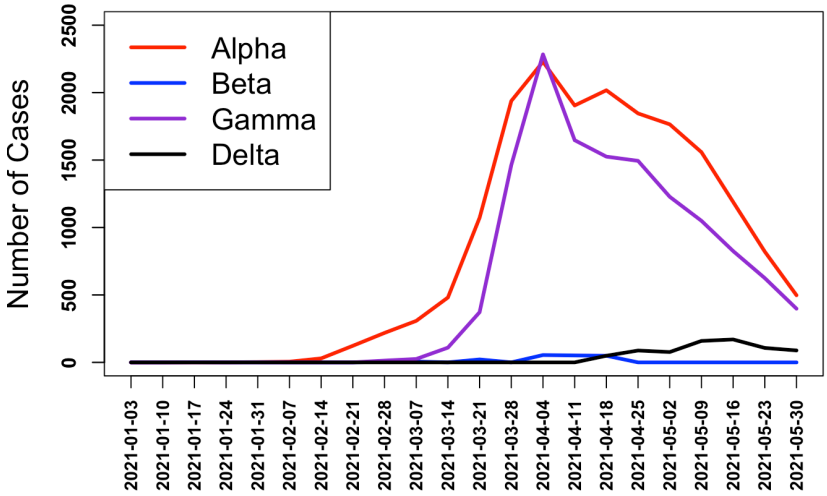
observations.

Public health data during the COVID-19 pandemic provides real-world examples of practical dominance and coexistence between variants and the original strain. Figure 5.3 and Figure 5.4 describe incident cases and percentage contributions of cases by VOCs for COVID-19 in British Columbia and Alberta (Jan 1 to May 30, 2021), respectively. In Figure 5.3, the Alpha variant and Gamma variant appear to demonstrate practical coexistence during the second COVID-19 wave, having comparable levels of variant-specific incident cases and percent contributions. In contrast, the Beta variant was not able to establish itself in the population during the same period, and both Alpha and Gamma variants showed practical dominance over the Beta variant. Based on our model analysis and results in Theorem 5.1, the VOC data from British Columbia in Figure 5.3 suggests that the reproduction numbers of Alpha and Gamma variants are similar, and both are larger than the reproduction number of the Beta variant.

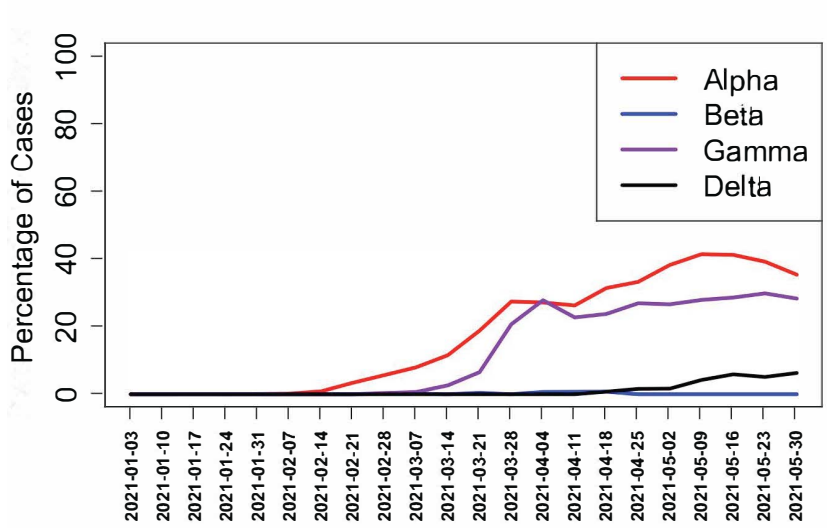
In Figure 5.4, the VOC data for Alberta during the same time period as Figure 5.3 describes a different situation. The Alpha variant showed a practical dominance over both Beta and Gamma variants in terms of incident cases and percentage case contributions. Although the VOC data from British Columbia showed practical coexistence of the Alpha and Gamma variant suggesting similar reproduction numbers, this was not observed in Alberta during the same time period. Alberta experienced practical dominance of the Alpha variant following the original strain. How can the mathematical theory of competition of variants be used to explain the apparently different situations of VOCs shown in Figures 5.3 and 5.4? Can variant-specific public health interventions explain the different situations in the two provinces? We carried out numerical investigations to address these questions in Section 5.4.

5.4 Numerical investigations and implications for endemic states of the COVID-19 pandemic

Numerical simulations were carried out using model (5.2.1) to investigate questions related to the dominance and coexistence of variants raised in the previous two sections. Subsection 5.4.1 illustrates analytical results from numerical simulations of theoretical (long-term) dominance and coexistence as described in Theorem 5.1. Parameter values in these simulations were independent of time and they are described in figure captions. Subsection 5.4.4 focuses on simulations to demonstrate practical (short-term) dominance and coexistence of variants. In these simulations, parameter values are fitted to public health data from Alberta, Canada, and include time-dependency to account for phased implementation and relaxation of public health measures (including testing policy changes).

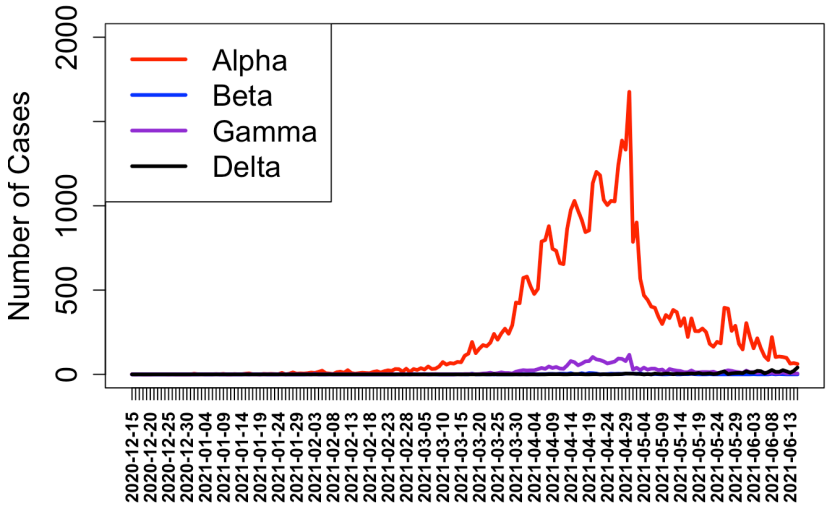


(a) Incident cases of COVID-19 VOCs in British Columbia.

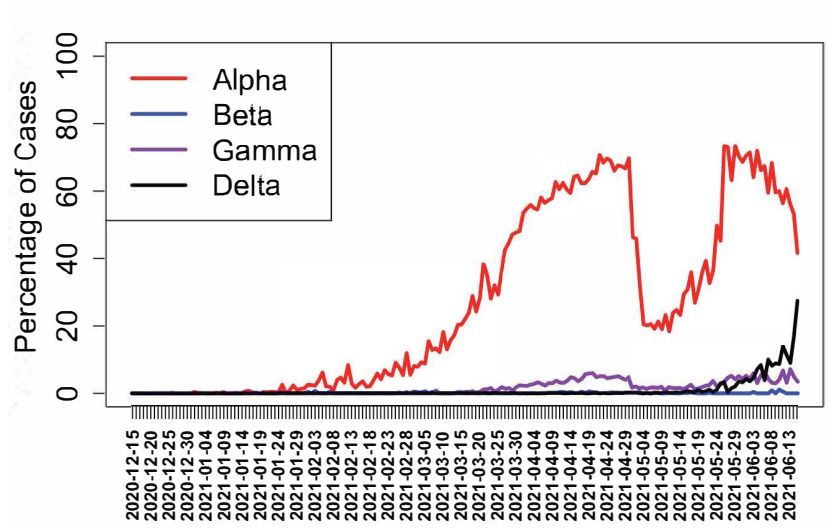


(b) Percentage of cases of COVID-19 VOCs in British Columbia.

Figure 5.3: Weekly reported public health data in the Province of British Columbia, Canada, shows that variants Alpha and Gamma are able to coexist at comparable levels, while variant Beta was not able to establish itself in the population. The data covers the period from January 3 to June 20, 2021. Source of data: http://www.bccdc.ca/health-info/diseases-conditions/covid-19/about-covid-19/variants, accessed on June 25, 2021.



(a) Number of incident cases of COVID-19 VOCs in Alberta.



(b) Percentage of cases of COVID-19 VOCs in Alberta.

Figure 5.4: Daily reports of public health data in the Province of Alberta, Canada, show that the Alpha variant dominated the Gamma variant in both (a) case numbers and (b) case percentages. This is in clear contrast to the variants situation in British Columbia as shown in Figure 5.3. The data covers the period from December 15 2020 to June 15, 2021. Source of data: https://www.alberta.ca/stats/covid-19-alberta-statistics.htm#variants-of-concern, accessed on June 25, 2021. Note: the drop in the top curve in (b) is artificial and was due to the temporary stoppage of typing of variants during that period.

5.4.1 Theoretical dominance and coexistence of variants

Our analytic results in Theorem 5.1 show that the variant with a larger basic reproduction number will become dominant, and will drive the main circulating variant to extinction as time tends to infinity. Furthermore, when the two variants have the same basic reproduction number, they can coexist as time tends to infinity. These long-term behaviours are termed *theoretical* dominance and coexistence dynamics. To illustrate the concepts, numerical simulations of number of the number (right) and percentage (left) of hidden infections by each of the variants are shown in Figures 5.5 and 5.6.

Impact of $\mathcal{R}_0$ on the theoretical dominance of variants

We assume that $\gamma_1 = \gamma_2$ and $\rho_1 = \rho_2$, and then select $\beta_2 = 1.5\beta_1$ so that $\mathcal{R}_{02} = 1.5\mathcal{R}_{01}$. By Theorem 5.1, we expect that variant 2 will dominate variant 1, irrespective of their initial values $I_{10} > 0$ and $I_{20} > 0$, and that $I_2(t) \to \bar{I}_2 > 0$, and $I_1(t) \to 0$ as $t \to \infty$ (see Figure 5.5). From the simulation results, we observe the following:

(1) The number of infection $I_2(t)$ is much larger than $I_1(t)$ during the course of the epidemic, suggesting that variant 2 dominates variant 1 (Figure 5.5 (a)).

(2) The asymptotic limit $\bar{I}_2 > 0$ in Figure 5.5 (a) appears to be very close to 0 in comparison to the peak value of $I_2(t)$ because of the scale, the infected populations $I_1(t)$ and $I_2(t)$ both appear to converge to 0 at the end of the epidemic. This leaves doubt about the dominance of variant 2 based on simulations of $I_1(t)$ and $I_2(t)$.

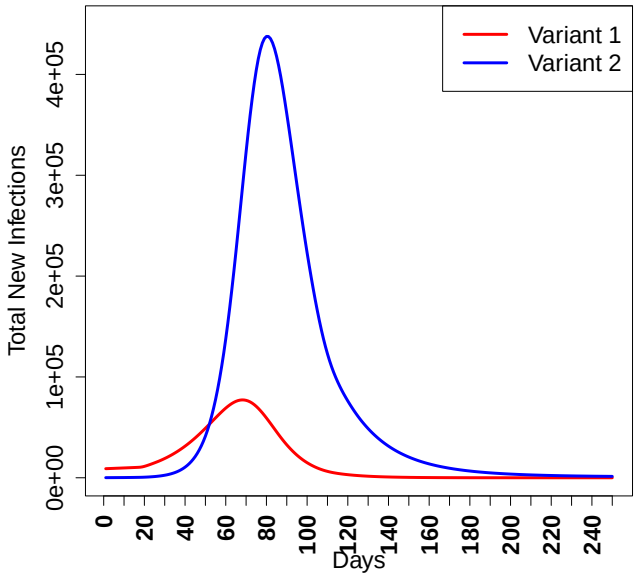
(3) The percentage contributions of $I_1(t)$ and $I_2(t)$ in Figure 5.5 (b) demonstrate a clear dominance of variant 2 over variant 1.

In conclusion, using percentage contributions of $I_i(t)$ instead of their numbers can better identify the dominant variant for analyzing both numerical simulations (Figure 5.5) and public-heath data (see Figures 5.3 and 5.4).

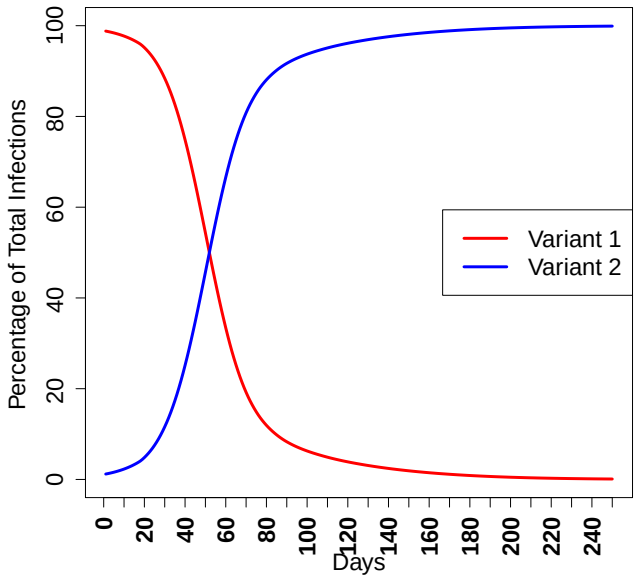
Theoretical coexistence of VOCs

In Figure 5.6, the parameter values included the following assumptions: $\gamma_1 = \gamma_2$, $\rho_1 = \rho_2$, and $\beta_1 = \beta_2$ so that $\mathcal{R}_{01} = \mathcal{R}_{02}$. Based on Theorem 5.1, we expect that both variants will coexist, and $I_1(t)$, $I_2(t)$ both converge to positive limits as $t \to \infty$. Furthermore, the limits are determined by the initial conditions. From our simulation results in Figure 5.6 we observed the following:

(1) In Figure 5.6 (a), $I_1(t)$ is much larger than $I_2(t)$ during the course of the epidemic, and the variant 1 appear to dominate variant 2.



(a) Number of infections of variant 1 ($I_1(t)$) and variant 2 ($I_2(t)$).



(b) Percentage contributions of $I_1(t)$ and $I_2(t)$ to the total infections.

Figure 5.5: Simulations of model (5.2.1) that demonstrate theoretical dominance of variant 1 by variant 2 when $\mathcal{R}_{02} = 1.5\mathcal{R}_{01}$. Parameter values used for simulations are $\beta_2 = 1.5\beta_1$, $\gamma_2 = \gamma_2 = 0.1, \rho_2 = \rho_1$, $I_{01} = 9000$ and $I_{02} = 100$. We note that even when the initial number of infected of variant 2 is smaller than that of variant 1, variant 2 still becomes dominant in the long term because of its basic reproduction number is larger.

(2) In Figure 5.6 (a), variant 1 has a larger initial condition than variant 2, and variant 1 produces a much larger epidemic than variant 2. This illustrates the conclusion (e) of Theorem 5.1 that the limits of number of infected for coexisting variants are dependent on the initial conditions.

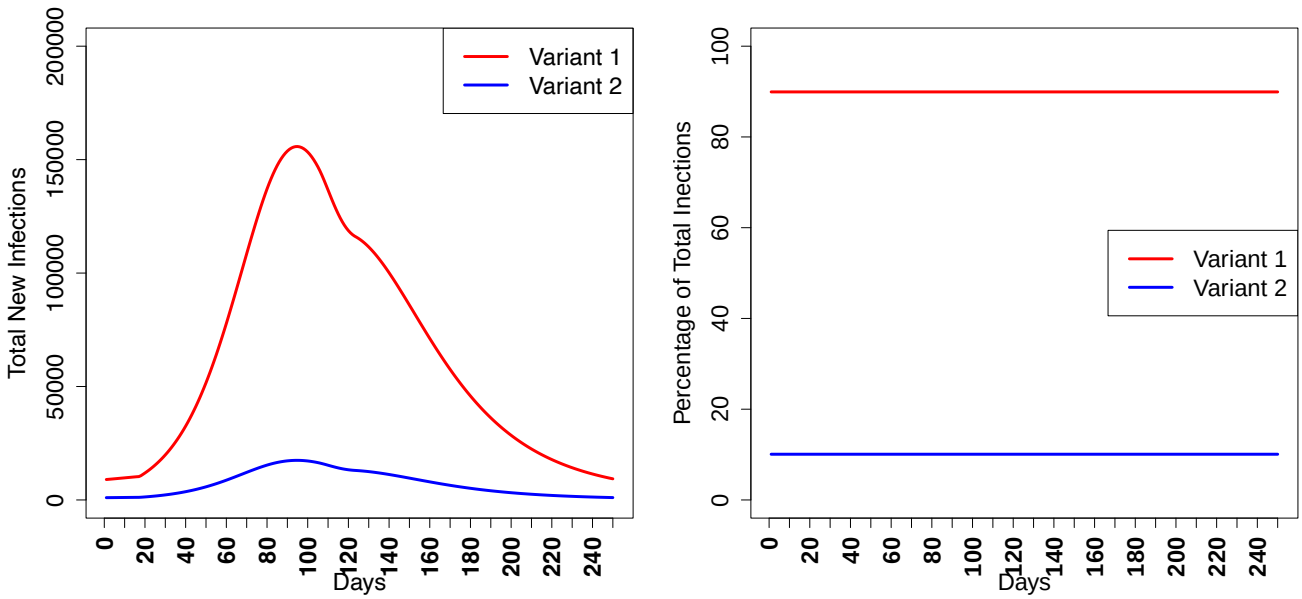
(3) In Figure 5.6 (b), the percentage contributions of I_1 and I_2 illustrate that both variants take a positive percentage throughout the epidemic. In such a case, we say that variant 2 has a low-level coexistence with variant 1.

Overall, two variants coexist when $\mathcal{R}_{01} = \mathcal{R}_{02}$. In this case, the variant with a greater number of initially infected will have a higher epidemic curve and percentage contribution of total infections. The significance of this theoretical result for real-world epidemics is that when there are multiple variants having similar reproductive numbers (e.g. Omicron BA.4 and BA.5) at an early stage within a similar time period, the variant with the largest initial infections will have the greatest percentage contribution of total infections. In situations where VOCs differ in their severity in terms of symptoms and case-fatality, control measures during the early stage of an epidemic or wave can be designed to limit the importation and spread of a more severe variant. If control measures cannot be implemented that target all circulating variants, suppressing the most severe variant would allow for a milder variant to take a higher percentage contribution and help reduce severe outcomes.

5.4.4 Practical dominance and coexistence of VOCs during the COVID-19 pandemic

In this subsection, we will examine the concepts of dominance and coexistence of new variants during real-world epidemics, or epidemic waves that typically only last for only a finite time. The COVID-19 pandemic provided an ideal context for such a study, since each epidemic wave was caused by a new variant that became dominant during the epidemic wave. We will use public health data in the provinces of Alberta and British Columbia, Canada, to inform the model simulations.

The main objective of the following numerical investigations is to provide plausible explanations for differences in transmission dynamics in finite time (a few months) between Alpha and Gamma variants in Alberta and British Columbia, as shown in Figures 5.3 and Figure 5.4. In these simulations, it is important to determine conditions that allow for variant dominance or the coexistence of two variants during an epidemic wave. Since these dynamics are challenging to observe through infections alone (see Figures 5.5 and 5.6), percentage contributions of variants were also provided. This is observed in Alberta's variants data in Figure 5.4, where, the Alpha variant appears to dominate the Gamma variant in the number of cases (Figure 5.4 (a)). However,



(a) Number of infections of variant 1 ($I_1(t)$) and variant 2 ($I_2(t)$).

(b) Percentage contributions of $I_{01}(t)$ and $I_{02}(t)$ to the total infections.

Figure 5.6: Simulations of model (5.2.1) demonstrating the coexistence of variants 1 and 2 when $\mathcal{R}_{01} = \mathcal{R}_{01}$. Parameter values used in the simulations are $\beta_2 = \beta_2$, $\gamma_2 = \gamma_2 = 0.1, \rho_2 = \rho_1$, $I_{01} = 9000$ and $I_{02} = 100$. Simulation results show that the variant 2 with smaller initial condition I_{02} has a smaller limit in both numbers (a) and percentages (b).

it is clear from the percentage contribution by variant (Figure 5.4 (b)) that the Gamma variant showed low-level coexistence with the Alpha variant. This shows the percentage contribution by variant is a better gauge for variant coexistence than incident case numbers.

To realistically model the COVID-19 dynamic, the incorporation of public health interventions such as social distancing, lock-down measures, testing, quarantine, isolation, and contact tracing are important. Some key parameters require time dependent characteristics to reflect the policy changes at different phases of the pandemic. For example, time-dependent transmission coefficients $\beta_i(t)$ will reflect impacts on transmission such as social distancing and lock-down measures. Time-dependent $\rho_i(t)$ will account for the impacts of COVID-19 testing, contact tracing, quarantine, and isolation measures. Accordingly, we will carry out our simulations using the following modified model:

$$\frac{dS}{dt} = -\beta_1(t)SI_1 - \beta_2(t)SI_2 - d_S S + \delta R, \tag{5.4.1a}$$

$$\frac{dI_1}{dt} = \beta_1(t)SI_1 - \gamma_1 I_1 - \rho_1(t)I_1 - d_{I_1} I_1, \tag{5.4.1b}$$

$$\frac{dI_2}{dt} = \beta_2(t)SI_2 - \gamma_2 I_2 - \rho_2(t)I_2 - d_{I_2} I_2, \tag{5.4.1c}$$

$$\frac{dR}{dt} = \gamma_1 I_1 + \gamma_2 I_2 + \rho_1(t)I_1 + \rho_2(t)I_2 - d_R R - \delta R. \tag{5.4.1d}$$

We note that the influx of susceptibles Λ and non-COVID-19 death are set to zero since the epidemic only lasted three months and the impact of birth and background death on the size of the susceptible population were negligible during epidemic.

Determination of $\beta_i(t)$ and $\rho_i(t)$ in the model

A step-wise function used in time-dependent transmission parameters, $\beta_i(t)$, $i = 1, 2$, represented the easing lock-down measures in Alberta between January and June 2021. The baseline transmission value was obtained from the endpoints of prior modeling results based on Alberta public health data in January 2021. One step-wise increase of 30% was introduced 18 days after the simulation start date of January 25, 2021. The simulation end date was June 30, 2021.

The time-dependent case-infection ratio $\rho_i(t)$, $i = 1, 2$, represented the effects of population health-seeking behaviours and behavioural change during the COVID-19 pandemic. They were generated and scaled within the simulations. The minimum and maximum values for $\rho_i(t)$ was obtained from prior fitting results between March and May 2020, informed by the testing and health link call data from Alberta Health Services. During each simulation, a burn-in run would help generate a case detection curve, which would be scaled such that the minimum and maximum

values would be limited to previous fitting results. The minimum value was limited between 2.5×10^{-2} and 3.8×10^{-2}. The maximum value was limited between 6.6×10^{-2} and 7.9×10^{-2}.

The Affine Invariant Ensemble Markov Chain Monte Carlo (MCMC) algorithm [162, 163] was the calibration procedure used to estimate time dependent terms $\rho_1(t)I_1(t)$ and $\rho_2(t)I_2(t)$, baseline transmission rates, the infectious period, and initial conditions (I_1, I_2, S and R) using confirmed cases for variant 1 and 2 on day t, respectively. The Matlab 2020a software was used to run the calibration procedures. Prior distributions of parameters at the start of simulations (January 25, 2021) were informed from epidemiological information and previous calibration results conducted before January 2022.

Dominance of Alpha variant over the Beta variant during the third wave in Alberta and British Columbia

From Figures 5.3 and 5.4, it is apparent that the Alpha variant dominated the Beta variant during the third COVID-19 waves in Alberta and British Columbia, in both case number and case percentage. This suggested that the basic reproduction number of the Alpha variant was sufficiently larger than that of the Beta variant in both provinces.

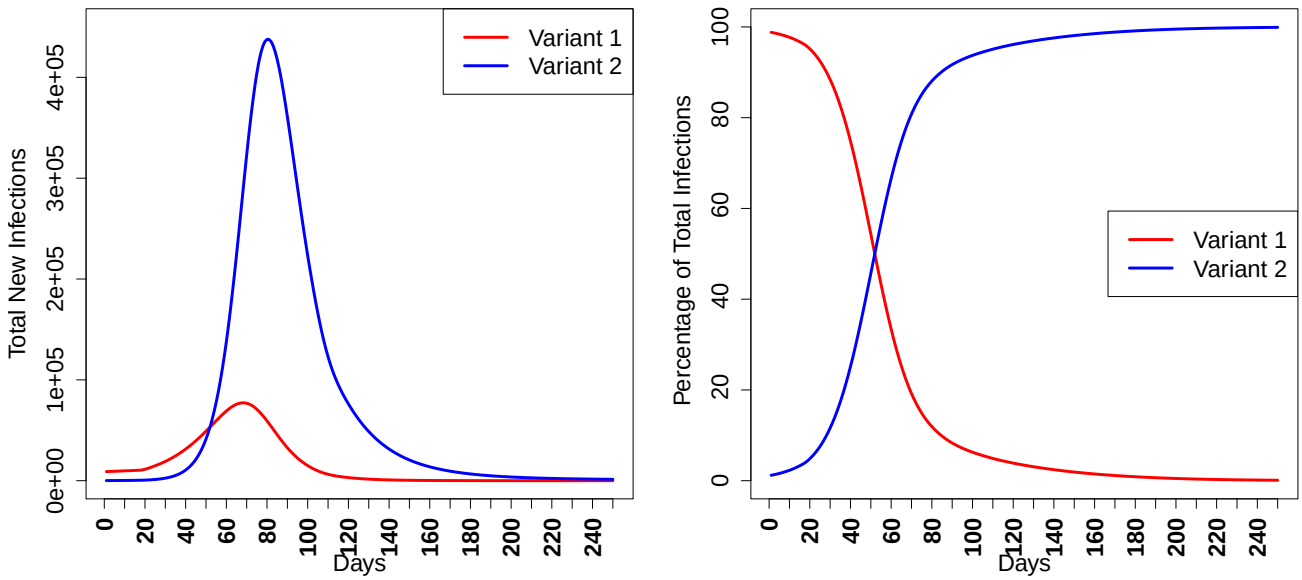
Based on the numerical simulations using model (5.4.1) (see Figure 5.7), results showed that when the basic reproduction number of the Alpha variant (variant 2) is 50% larger than the Beta variant (variant 1), namely, $\mathcal{R}_{02} = 1.5\mathcal{R}_{01}$, the Alpha variant will dominate the Beta variant even when the Alpha variant had a smaller initial number of infected people.

Practical coexistence of the Alpha and Gamma variants during the third wave of COVID-19 epidemic in British Columbia

Figure 5.3 describes VOC trends for COVID-19 from British Columbia during the third wave. Case numbers and percentage contributions by variant type show that the Alpha and Gamma variants coexisted during the third wave and dominated the Beta variant.

From our theoretical results in Theorem 5.1, two variants can coexist long-term if and only if they have the same basic reproduction number. The observed coexistence of the Alpha and Gamma variants during the third COVID-19 wave in British Columbia suggests that both the Alpha and Gamma variants may have similar basic reproduction numbers. We performed numerical simulations on the model (5.4.1) to verify this possibility. From our simulation results in Figure 5.8, we observed the following:

(1) When variant 1 (Alpha) and variant 2 (Gamma) have similar basic reproduction numbers, $\mathcal{R}_{01} = 1.06\mathcal{R}_{02}$, they can coexist, in both the infection numbers and percentage contributions (see Figure 5.8).

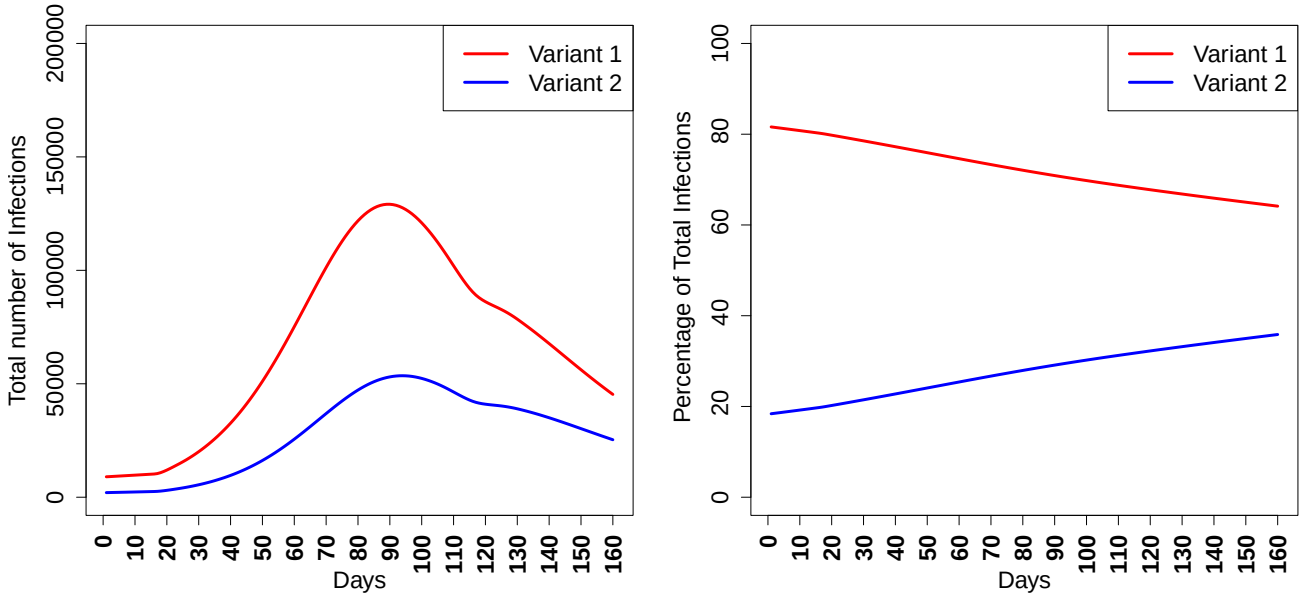


(a) Number of infections of variant 1 ($I_1(t)$) and variant 2 ($I_2(t)$).

(b) Percentage contributions of $I_1(t)$ and $I_2(t)$ to the total infections.

Figure 5.7: Simulations of model (5.4.1) demonstrating the dominance of the Beta variant (variant 1) by the Alpha variant (variant 2) in Alberta as observed in Figure 5.4, with a relation $\mathcal{R}_{02} = 1.5\mathcal{R}_{01}$. Parameter values used for simulations are $\beta_2 = 1.5\beta_2$, $\gamma_2 = \gamma_2 = 0.1, \rho_2 = \rho_1$, $I_{01} = 9000$ and $I_{02} = 100$. Even when the initial number of the infected is much lower for the Alpha variant, its sufficiently larger basic reproduction number allows the Alpha variant to become dominant.

(2) In Figure 5.8, variant 2 had a slightly larger basic reproduction number but a smaller number of initially infected people compared to variant 1. Based on this simulation, variant 1 had greater number of infections and percentage contributions compared to Variant 2. This suggested that when the basic reproduction numbers are comparable between two variants, the variant with the more infected people early on will produce a larger epidemic wave. This agrees with our observations of theoretical coexistence in Sub-section 5.4.1. The use of real data asserts the importance of early public health responses during an epidemic given an emerging variant.



(a) Number of infections of variant 1 ($I_1(t)$) and variant 2 ($I_2(t)$).

(b) Percentage contributions of $I_1(t)$ and $I_2(t)$ to the total infections.

Figure 5.8: Simulation results of model (5.4.1) demonstrating the coexistence of the Alpha variant (variant 1) and Gamma variant (variant 2) when they have similar basic reproduction numbers ($\mathcal{R}_{02} = 1.06\mathcal{R}_{01}$), as can be observed in both (a) case numbers and (b) case percentages. Parameter values used in the simulations are $\beta_2 = 1.06\beta_2$, $\gamma_2 = \gamma_2 = 0.1, \rho_2 = \rho_1$, $I_{01} = 9000$ and $I_{02} = 2000$.

Early public health responses given an emerging variant can help control replacement dynamics in a population: Alberta's experience with COVID-19 variants, Alpha and Gamma

During the third wave, the Gamma and Alpha variant coexisted in British Columbia, while in Alberta, coexistence of the Gamma variant occurred at low-levels with the Alpha variant dominating

the epidemic wave (see Figures 5.3 and 5.4). The practical coexistence of the Alpha and Gamma variants in both provinces suggest that the two variants have similar basic reproduction numbers. In subsection 5.4.1, simulations showed that the variant with more initially infected people can lead to a larger epidemic size (see Figure 5.6). The following simulations aim to address differences in public health responses that may have resulted in variant dynamics observed in both provinces.

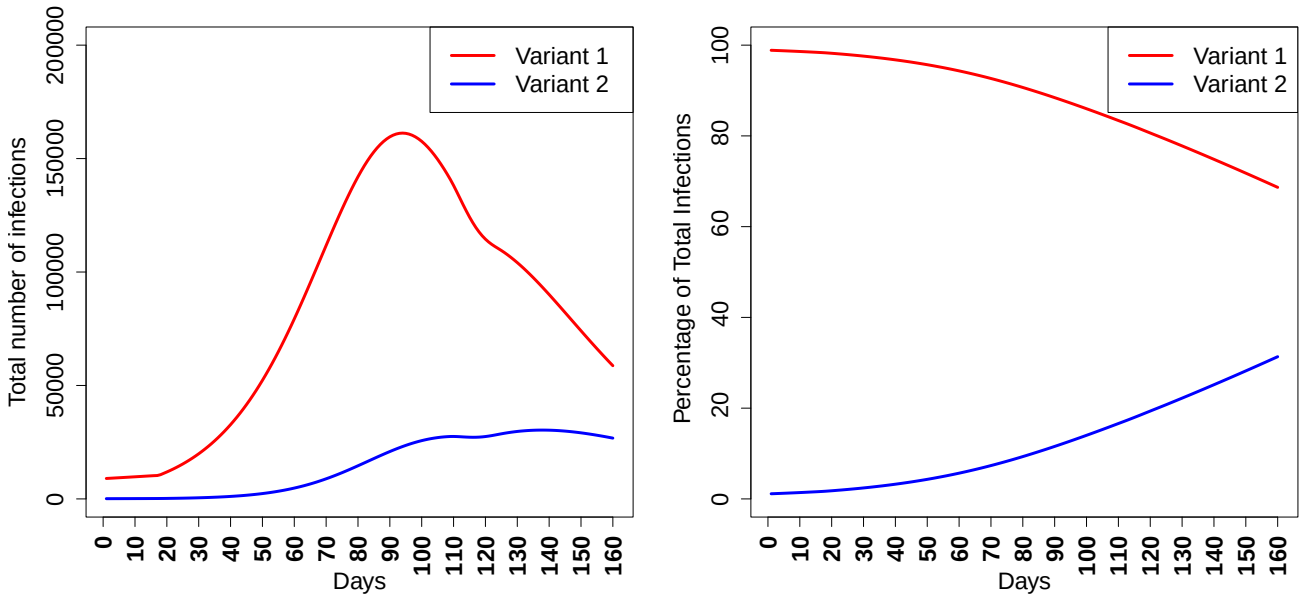
As outbreaks of the Gamma variant emerged in Alberta, enhanced public health responses were implemented to contain its spread through increased testing, contact tracing, and isolation measures ([87],[58]). Can variant-specific public health measures explain the low-level coexistence of the Gamma variant in Alberta? We carried out numerical simulations to investigate the impacts of containment measures (i.e. testing, contact tracing, and isolation) on variant dynamics.

These simulations aimed to investigate the impacts of increasing the case-infection ratio $\rho_i(t)$, which incorporates the effects of case detection through testing, contact tracing, and isolation. Variant 1 (Alpha) and variant 2 (Gamma) had similar reproduction numbers since they were able to coexist in both provinces. To reflect enhanced surveillance of the emerging variant 2, the case-infection ratio $\rho_1(t)$ was increased to two times of $\rho_1(t)$. In Figure 5.9, simulations show low-level coexistence of the variant 1 and variant 2, with the latter was at much lower levels for both infection numbers and percentage contributions. An important result in Figure 5.9 was that while variant 1 and variant 2 had a similar basic reproduction number, the targeted increase in testing, contact tracing, and isolation of the emerging variant was effective to mitigate the total number of infections.

These simulations results highlighted that when an emerging variant has a similar reproduction number as the existing variant, early public health measures that increase case-infection ratios (i.e. testing, contact tracing, and isolation) of the emerging variant can prevent its replacement and/or allow for low level coexistence in the population. In addition, if the emerging variant poses a greater population-level risk of severity, this targeted approach can be effective to reduce the level of severe disease outcomes in the community. This was particularly evident in Alberta, where targeted public health measures that increased case-infection-ratios for the emerging Gamma variant was effective at mitigating its total infections and percent contributions (Figure 5.4).

5.5 Summary and discussions

We investigated disease transmission dynamics of two variants (strains) using deterministic modelling and focused on coexistence of variants and dominance of one variant over the other. Analysis and simulations considered asymptotic limits (termed theoretical coexistence) and finite-time horizon for e.g. within a single epidemic wave (termed practical coexistence). While the simulations focused on COVID-19 specific data, the lessons learned can offer insights for general infectious



(a) Number of infections of variant 1 ($I_1(t)$) and variant 2 ($I_2(t)$).

(b) Percentage contributions of $I_1(t)$ and $I_2(t)$ to the total infections.

Figure 5.9: Simulation results of model (5.4.1) demonstrating that variant-specific public health interventions can prevent the Gamma variant from taking hold in Alberta. We have assumed that the Gamma variant (variant 2) has a slightly higher transmission rate β_2 than the transmission rate β_1 of the Alpha variant (variant 1), $\beta_2 = 1.17\beta_1$. We also assumed that the Gamma variant has a higher $\rho_2(t)$ than the Alpha variant, $\rho_2(t) = 2\rho_1(t)$. The initial number of infected are chosen as $I_{01} = 9000$ and $I_{02} = 100$. The choices of $\rho_i(t)$ and I_{0i} reflect the additional effort in testing, DNA typing, contact tracing, and isolation directed at cases of the Gamma variants implemented in Alberta.

disease epidemics and public health interventions.

The theoretical concept of coexistence of variants and dominance of a variant in mathematical epidemiology is based on the asymptotic limit in an infinite time horizon. The outcomes are determined by the variant-specific basic reproduction numbers, irrespective of the respective number of initially infected individuals. In the case of simple competition, the principle of competitive exclusion prevails: a variant with the largest basic reproduction number dominates and drive the other variants to extinction. The theory focuses on the asymptotic limits without providing much information about finite-time relations among the variants, and it is less useful when applied to actual epidemics such as seasonal influenza and epidemic waves of COVID-19. One useful insight from this theory based on infinite-time horizon is the possibility that variants can coexist when their reproduction numbers are similar and when they emerge concurrently. When applied to real-world epidemics, this can help us to infer that two variants with comparable number of confirmed cases should have similar basic reproduction numbers. In Figure 5.8, we similar reproductive number and different initial conditions for infectious people.

Real-world epidemics typically last for only a short period of time and the number of infected caused by each variant will rise at the start, peak, then decline to low levels. Assessing patterns of coexistence or dominance in simulations was challenging using the number of infections alone. Percentage contributions was a better outcome to assess coexistence and dominance relations. Unlike the asymptotic theory of variant dominance that depends solely on the variant-specific basic reproductive numbers, in finite-time variant dynamics, the number of initially infected individuals can also play an important role to determine which variant may cause an higher level of epidemic or a greater final size.

The dynamics of the Alpha and Gamma variant in Alberta and British Columbia during the third wave was of particular interest in this analysis since coexistence of both variants were observed in both provinces. However, the percent contributions of these variants were shared more equally in British Columbia [88] compared to Alberta (Figure 5.3 and Figure 5.4). Despite magnitude differences, the overall coexistence observed indicated that both variants likely had similar basic reproduction numbers based on theoretical insights from mathematical epidemiology. Despite comparable reproductive numbers for both variants, early and enhanced containment measures (e.g. testing, contact tracing, and/or isolation) targeted on the emerging Gamma variant may have also been more effective in keeping infectious cases at lower levels in Alberta. Given variants with comparable reproductive numbers, there may be some interaction between initial size of infectious people (i.e. initial number and size of outbreaks) (Figure 5.8) and the effectiveness of containment measures (including asymptomatic testing [73]) (Figure 5.9) that could also explain differences in variant dynamics between Alberta and British Columbia.

Overall, the mathematical analysis and simulations of real-world examples such as COVID-

19 provided valuable insights for future events involving variants emerging at similar times with comparable reproductive numbers. Earlier containment measures that effectively target emerging variants that pose a greater risk for severity can impact finite-time variant replacement dynamics by forcing the more severe variant to coexist at a much lower level in a population compared to less severe variants. While differences in percent contributions of the Alpha and Gamma variants between Alberta and British Columbia were described using modeling, there may be other factors not captured in the mathematical model that could play a role such as vaccine coverage and importation through travel. In Canada, as we transition from the pandemic to endemic phase of COVID-19, the expectation of emerging VOCs is a reality. These valuable insights offered from the analysis of past waves can help with the future management of COVID-19.

Appendix

In this section, we provide the technical details for the proof of stability results in Theorem 5.1. The Jacobian matrix of model (5.2.1) at a point $P = (S, I_1, I_2, R)$ is

$$J(P) = \begin{bmatrix} -\beta_1 I_1 - \beta_2 I_2 - d_S & -\beta_1 S & -\beta_2 S & \delta \\ \beta_1 I_1 & \beta_1 S - \gamma_1 - \rho_1 - d_{I_1} & 0 & 0 \\ \beta_2 I_2 & 0 & \beta_2 S - \gamma_2 - \ \rho_2 - d_{I_2} & 0 \\ 0 & \gamma_1 + \rho_1 & \gamma_2 + \rho_2 & -\delta - d_R \end{bmatrix}.$$

We recall that the variant-specific basic reproduction numbers are:

$$\mathcal{R}_{01} = \frac{\beta_1}{\gamma_1 + \rho_1 + d_{I_1}} \frac{\Lambda}{d_S}, \quad \mathcal{R}_{02} = \frac{\beta_2}{\gamma_2 + \rho_2 + d_{I_2}} \frac{\Lambda}{d_S},$$

and the basic reproduction number for model (5.2.1) is $\mathcal{R}_0 = \max\{\mathcal{R}_{01}, \mathcal{R}_{02}\}$.
Stability of P_0. The Jacobian matrix at $P_0 = (S^0, 0, 0, 0)$, $S^0 = \Lambda/d_S$, is

$$J(P_0) = \begin{bmatrix} -d_S & -\beta_1 S^0 & -\beta_2 S^0 & \delta \\ 0 & \beta_1 S^0 - \gamma_1 - \rho_1 - d_{I_1} & 0 & 0 \\ 0 & 0 & \beta_2 S^0 - \gamma_2 - \ \rho_2 - d_{I_2} & 0 \\ 0 & \gamma_1 + \rho_1 & \gamma_2 + \rho_2 & -\delta - d_R \end{bmatrix}.$$

The eigenvalues of $J(P_0)$ are $\lambda_1 = -d_S$, $\lambda_2 = -\delta - d_R$, $\lambda_3 = \beta_1 S^0 - \gamma_1 - \rho_1 - d_{I_1}$, and $\lambda_4 = \beta_2 S^0 - \gamma_2 - \rho_2 - d_{I_2}$. Therefore,

(1) P_0 is asymptotically stable if $\lambda_3 < 0$ and $\lambda_4 < 0$, namely if $\mathcal{R}_{01} < 1$ and $\mathcal{R}_{02} < 1$. This is

equivalent to $\mathcal{R}_0 = \max\{\mathcal{R}_{01}, \mathcal{R}_{02}\} < 1$.

(2) P_0 is unstable if either $\lambda_3 > 0$ or $\lambda_4 > 0$, or equivalently, if $\mathcal{R}_0 > 1$.

Stability of P_1. Suppose that $\mathcal{R}_{01} > 1$. Then $\bar{S} = (\gamma_1 + \rho_1 + d_{I_1})/\beta_1 < \frac{\Lambda}{d_S}$, and $P_1 = (\bar{S}, \bar{I}_1, 0, \bar{R})$ exists in $\mathbb{R}^4_+$. The Jacobian matrix at P_1 is

$$J(P_1) = \begin{bmatrix} -\beta_1\bar{I}_1 - d_S & -\beta_1\bar{S} & -\beta_2\bar{S} & \delta \\ \beta_1\bar{I}_1 & \beta_1\bar{S} - \gamma_1 - \rho_1 - d_{I_1} & 0 & 0 \\ 0 & 0 & \beta_2\bar{S} - \gamma_2 - \ \rho_2 - d_{I_2} & 0 \\ 0 & \gamma_1 + \rho_1 & \gamma_2 + \rho_2 & -\delta - d_R \end{bmatrix}.$$

One of the eigenvalues is $\mu_4 = \beta_2\bar{S} - \gamma_2 - \rho_2 - d_{I_2}$, corresponding to an eigenvector that is transversal to the SI_1R-subspace ($I_2 = 0$) of $\mathbb{R}^4$, which is the invariant subspace of model when only the variant 1 is present.

The remaining three eigenvalues, μ_1, μ_2, μ_3, are eigenvalues of the 3×3 sub-matrix of $J(P_1)$

$$\begin{aligned} M &= \begin{bmatrix} -\beta_1\bar{I}_1 - d_S & -\beta_1\bar{S} & \delta \\ \beta_1\bar{I}_1 & \beta_1\bar{S} - \gamma_1 - \rho_1 - d_{I_1} & 0 \\ 0 & \gamma_1 + \rho_1 & -\delta - d_R \end{bmatrix} \\ &= \begin{bmatrix} -\beta_1\bar{I}_1 - d_S & -\beta_1\bar{S} & \delta \\ \beta_1\bar{I}_1 & 0 & 0 \\ 0 & \gamma_1 + \rho_1 & -\delta - d_R \end{bmatrix}, \end{aligned}$$

since $\beta_1\bar{S} - \gamma_1 - \rho_1 - d_{I_1} = 0$. We will apply Routh-Hurwitz criteria to show that all eigenvalues of M have negative real parts.

First, $\mathrm{tr}(M) = -\beta_1\bar{I}_1 - d_S - \delta - d_R < 0$. Next,

$$\begin{aligned} \det(M) &= \beta_1\bar{I}_1(\gamma_1 + \rho_1)\delta - \beta_1\bar{I}_1\beta_1\bar{S}(d_R + \delta) \\ &= -\beta_1\bar{I}_1[d_R(\gamma_1 + \rho_1 + d_{I_1}) + \delta d_{I_1}] < 0, \end{aligned}$$

and thus the first two Routh-Hurwitz conditions hold. The sum of all 2×2 principal minors of M

$$a_2 = \beta_1\bar{I}_1\beta_1\bar{S} + (\beta_1\bar{I}_1 + d_S)(\delta + d_R) > 0,$$

and, using $\beta_1\bar{S} = \gamma_1 + \rho_1 + d_{I_1}$, we have

$$\begin{aligned}\mathrm{tr}(M)a_2 &= -(\beta_1\bar{I}_1 + d_S + \delta + d_R)[\beta_1\bar{I}_1(\gamma_1 + \rho_1 + d_{I_1}) + (\beta_1\bar{I}_1 + d_S)(\delta + d_R)] \\ &< -\beta_1\bar{I}_1(\gamma_1 + \rho_1 + d_{I_1})(\delta + d_R) < -\beta_1\bar{I}_1[(\gamma_1 + \rho_1 + d_{I_1})d_R + \delta d_{I_1}] \\ &= \det(M).\end{aligned}$$

We have verified all three Rough-Hurwitz conditions for M, and hence the eigenvalues μ_1, μ_2, and μ_3 of $J(P_1)$ have negative real parts.

Based on the preceding discussion, the stability of P_1 is determined by the sign of $\mu_4 = \beta_2\bar{S} - \gamma_2 - \rho_2 - d_{I_2}$. It can be verified that $\mu_4 < 0$ if and only if $\mathcal{R}_{02} < \mathcal{R}_{01}$, and thus P_1 is asymptotically stable if $\mathcal{R}_{02} < \mathcal{R}_{01}$, and unstable if $\mathcal{R}_{02} > \mathcal{R}_{01}$.

The stability of P_2 when $\mathcal{R}_{02} > 1$ can be analyzed similarly.

Stability of positive equilibrium P^* when $\mathcal{R}_{01} = \mathcal{R}_{02}$. Under the assumption that

$$\mathcal{R}_{01} = \frac{\beta_1}{\gamma_1 + \rho_1 + d_{I_1}} = \mathcal{R}_{02} = \frac{\beta_2}{\gamma_2 + \rho_2 + d_{I_2}},$$

a positive equilibrium $P^* = (S^*, I_1^*, I_2^*, R^*)$ exists, where S^*, I_1^*, I_2^*, and R^* satisfy equations

$$\begin{aligned}&S^* = \bar{S} = \frac{\gamma_1 + \rho_1 + d_{I_1}}{\beta_1} = \frac{\gamma_2 + \rho_2 + d_{I_2}}{\beta_2}, \\ &(\gamma_1 + \rho_1 + d_{I_1})I_1^* + (\gamma_2 + \rho_2 + d_{I_2})I_2^* - \delta R^* = \Lambda - d_S S^*, \\ &(\gamma_1 + \rho_1)I_1^* + (\gamma_2 + \rho_2)I_2^* - (d_R + \delta)R^* = 0.\end{aligned} \tag{5.5.1}$$

There are infinitely many solutions to this linear system, and they all lie on the 3d hyperplane $S = \bar{S}$ in $\mathbb{R}^4$. On the 3-dimensional hyperplane $S = \bar{S}$, the solutions of linear system (5.5.1) lie on the line of intersection of the two 2-dimensional planes defined by the last two equations of system (5.5.1), whose normal vectors (on the hyperplane $S = \bar{S}$) are

$$N_1 = (\gamma_1 + \rho_1 + d_{I_1}, \gamma_2 + \rho_2 + d_{I_2}, -\delta), \quad N_2 = (\gamma_1 + \rho_1, \gamma_2 + \rho_2, -(d_R + \delta)).$$

Therefore, the directional vector of the line of equilibria (in the 3d hyperplane $S = \bar{S}$) is the cross product

$$\begin{aligned}v = N_1 \times N_2 = \Big(&-(\gamma_2 + \rho_2)d_R - (d_R + \delta)d_{I_2}, (\gamma_1 + \rho_1)d_R + (d_R + \delta)d_{I_1}, \\ &(\gamma_2 + \rho_2)d_{I_1} - (\gamma_1 + \rho_1)d_{I_2}\Big).\end{aligned}$$

In $\mathbb{R}^4$, vector v is given as

$$v = \Big(0, -(\gamma_2+\rho_2)d_R - (d_R+\delta)d_{I_2}, (\gamma_1+\rho_1)d_R + (d_R+\delta)d_{I_1}, \\ (\gamma_2+\rho_2)d_{I_1} - (\gamma_1+\rho_1)d_{I_2}\Big).$$

The Jacobian matrix of any positive equilibrium $P^* = (S^*, I_1^*, I_2^*, R^*)$ on the line is

$$J(P^*) = \begin{bmatrix} -\beta_1 I^* - \beta_2 I^* - d_S & -\beta_1 S^* & -\beta_2 S^* & \delta \\ \beta_1 I_1^* & 0 & 0 & 0 \\ \beta_2 I_2^* & 0 & 0 & 0 \\ 0 & \gamma_1+\rho_1 & \gamma_2+\rho_2 & -d_R-\delta \end{bmatrix}.$$

Here, we have used the relations $\beta_1 S^* = \gamma_1 + \rho_1 + d_{I_1}$ and $\beta_2 S^* = \gamma_2 + \rho_2 + d_{I_2}$. Straightforward calculations show that

$$J(P^*)v = 0,$$

which implies that the directional vector v of the line of equilibria is an eigenvector of $J(P^*)$ with eigenvalue 0, at each equilibrium on the line. This shows that each positive equilibrium P^* is neutrally stable in the direction v of the line of equilibria.

Next, we show that the remaining eigenvalues of $J(P^*)$ all have negative real parts, for all positive equilibria P^*. The characteristic polynomial of $J(P^*)$ is

$$\begin{aligned} &|\lambda I - J(P^*)| \\ &= \begin{vmatrix} \lambda + \beta_1 I_1^* + \beta_2 I_2^* + d_S & \beta_1 S^* & \beta_2 S^* & -\delta \\ -\beta_1 I_1^* & \lambda & 0 & 0 \\ -\beta_2 I_2^* & 0 & \lambda & 0 \\ 0 & -\gamma_1-\rho_1 & -\gamma_2-\rho_2 & \lambda + d_R + \delta \end{vmatrix} = \lambda P(\lambda), \end{aligned}$$

where $P(\lambda)$ is the following cubic polynomial

$$\begin{aligned} P(\lambda) = {} & \lambda^3 + \lambda^2(\beta_1 I_1^* + \beta_2 I_2^* + d_S + d_R + \delta) \\ & + \lambda\big[(d_R+\delta)(\beta_1 I_1^* + \beta_2 I_2^* + d_S) + \beta_1 I_1^* \beta_1 S^* + \beta_2 I_2^* \beta_2 S^*\big] \\ & - \delta\big[\beta_1 I_1^*(\gamma_1+\rho_1) + \beta_2 I_2^*(\gamma_2+\rho_2)\big] + (d_R+\delta)(\beta_1 I_1^* \beta_1 S^* + \beta_2 I_2^* \beta_2 S^*). \end{aligned}$$

The remaining three eigenvalues of $J(P^*)$ are roots of the polynomial $P(\lambda)$. We use the Routh-Hurwitz conditions for cubic polynomials $P(\lambda) = \lambda^3 + a_1\lambda^2 + a_2\lambda + a_3$, namely, $a_1 > 0, a_3 > 0$, and $a_1 a_2 > a_3$, to show that all roots of $P(\lambda)$ have negative real parts.

It is clear that $a_1 = \beta_1 I_1^* + \beta_2 I_2^* + d_S + d_R + \delta > 0$. Also,

$$\begin{aligned}
a_3 &= -\delta\big[\beta_1 I_1^*(\gamma_1 + \rho_1) + \beta_2 I_2^*(\gamma_2 + \rho_2)\big] + (d_R + \delta)(\beta_1 I_1^* \beta_1 S^* + \beta_2 I_2^* \beta_2 S^*) \\
&= -\delta\big[\beta_1 I_1^*(\gamma_1 + \rho_1) + \beta_2 I_2^*(\gamma_2 + \rho_2)\big] \\
&\quad + (d_R + \delta)\big[\beta_1 I_1^*(\gamma_1 + \rho_1 + d_{I_1}) + \beta_2 I_2^*(\gamma_2 + \rho_2 + d_{I_2})\big] \\
&\geqslant -\delta\big[\beta_1 I_1^*(\gamma_1 + \rho_1) + \beta_2 I_2^*(\gamma_2 + \rho_2)\big] \\
&\quad + \delta\big[\beta_1 I_1^*(\gamma_1 + \rho_1 + d_{I_1}) + \beta_2 I_2^*(\gamma_2 + \rho_2 + d_{I_2})\big] > 0.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
a_1 a_2 - a_3 &= (\beta_1 I_1^* + \beta_2 I_2^* + d_S + d_R + \delta)(d_R + \delta)(\beta_1 I_1^* + \beta_2 I_2^* + d_S) \\
&\quad + (\beta_1 I_1^* + \beta_2 I_2^* + d_S + d_R + \delta)(\beta_1 I_1^* \beta_1 S^* + \beta_2 I_2^* \beta_2 S^*) \\
&\quad + \delta\big[\beta_1 I_1^*(\gamma_1 + \rho_1) + \beta_2 I_2^*(\gamma_2 + \rho_2)\big] - (d_R + \delta)(\beta_1 I_1^* \beta_1 S^* + \beta_2 I_2^* \beta_2 S^*) \\
&= (\beta_1 I_1^* + \beta_2 I_2^* + d_S + d_R + \delta)(d_R + \delta)(\beta_1 I_1^* + \beta_2 I_2^* + d_S) \\
&\quad + (\beta_1 I_1^* + \beta_2 I_2^* + d_S)(\beta_1 I_1^* \beta_1 S^* + \beta_2 I_2^* \beta_2 S^*) \\
&\quad + \delta\Big[\beta_1 I_1^*(\gamma_1 + \rho_1) + \beta_2 I_2^*(\gamma_2 + \rho_2)\Big] > 0.
\end{aligned}$$

The Routh-Hurwitz conditions are verified, and $J(P^*)$ has an eigenvalue 0 in the direction of the line of equilibria, and the remaining three eigenvalues have negative real parts. Accordingly, P^* is asymptotically stable in the directions transversal to the line of equilibria. This completes the proof of Theorem 5.1.

6

Quantifying Change in a Network

Overview of the chapter: This chapter outlines how the research developed from initial work in data processing and social network analysis into the creation of a Python package for analysing dynamic networks. The project was motivated by the role of human mobility in the spread of infectious diseases, where transportation networks can accelerate transmission and connect distant regions.

The motivation for this work comes from the recognition that human mobility plays a crucial role in the transmission of infectious diseases. Transport networks do not just connect places, they can accelerate outbreaks by linking distant regions and enabling pathogens to travel faster and farther than they would otherwise. This is particularly evident in air travel, where an infected passenger has the potential to expose others not only during the flight but also at airports and connecting hubs. In contrast, a person biking across a city poses little transmission risk en route. That difference makes the structure and behaviour of transportation networks highly relevant to epidemiological modelling.

However, modelling these networks accurately is not straightforward. There are two main obstacles: first, data availability. While commercial providers offer rich flight data, these are often expensive and come with strict licensing restrictions. Second, there's the issue of time. Most available data are static snapshots, which fail to capture how travel patterns shift, something we saw dramatically during the COVID-19 pandemic when flight routes were suspended, travel bans were imposed, and overall traffic plummeted in a matter of weeks.

To address these challenges, I developed a Python package that makes it easier to analyse

networks that change over time. Two case studies demonstrate its functionality: one based on the Helsinki city bike share system, and the other using air traffic data from the ADS-B dataset. These two examples demonstrate the functionality of this package.

The theories utilized in this chapter are detailed in Chapter 2.6.

6.1 Introduction

Understanding how infectious diseases spread across regions requires a realistic representation of human mobility. Travel plays a central role in the propagation of epidemics: it not only moves infected individuals from one location to another, but in some cases facilitates further transmission along the way. The structure and dynamics of transportation networks, especially air travel, can significantly amplify the speed and scale of an outbreak, turning local epidemics into global crises.

For example, an individual travelling by bicycle may carry an infection from one point to another, but poses little risk of transmission en route. In contrast, an infected person on a commercial flight can expose fellow passengers and airport workers during boarding, the flight itself, and while in transit at hub airports. Thus, air travel acts not just as a conduit but as a multiplier of disease spread, linking distant communities in ways that can accelerate the pace of transmission. Capturing these dynamics in models is essential for making accurate predictions and designing effective interventions.

Incorporating travel into epidemiological modelling introduces two main challenges. The first is data availability. Several commercial providers such as IATA, OAG, FlightAware, and Flightradar24 offer detailed information about flight schedules and passenger volumes. However, these datasets are expensive and licensed under restrictive conditions that limit their use and reproducibility in academic research. Alternatively, open-access datasets like those provided by the OpenSky Network [171] are freely available but require significant preprocessing, and the data they offer may be incomplete or only released monthly. In Section 6.4, we describe the dataset used in our case study and the steps taken to make it suitable for analysis.

The second challenge is the temporal nature of mobility. Most datasets, including those from IATA, are compiled retrospectively, often based on data from the previous year. However, during the COVID-19 pandemic, the structure of the global air transportation network changed rapidly, flights were cancelled, travel bans were enacted, and volumes dropped dramatically. Any meaningful model must account for these sudden shifts, which are not captured in standard, static datasets.

Our aim here is to derive tools to consider the evolution of networks with time. The research in this chapter, like the thesis program itself, evolved from simply considering how to process data and use social network analysis to analyse a dataset used in a model, to creating a Python package

to make analysing temporal (dynamic) networks more accessible.

We used Python for our analysis, leveraging a variety of libraries to handle data manipulation, visualization, and network analysis. The key libraries used include:

1. folium (v0.16.0) for interactive mapping and geospatial data visualization,
2. pandas (v2.2.2) for data manipulation and analysis,
3. igraph (v0.11.5) for complex network analysis,
4. matplotlib (v3.8.4) for creating static, animated, and interactive visualizations,
5. seaborn (v0.13.2) for statistical data visualization,
6. numpy (v1.26.4) for numerical computing,
7. networkx (v3.2.1) for creating, analyzing, and visualizing complex networks,
8. plotly (v5.22.0) for creating interactive visualizations.

6.2 Description of the functions

We review here the different functions used to study the networks and their evolution. It is important to note that the functions presented in this section reflect their state at the time of writing the case studies. The Python scripts containing these functions has continued to evolve, aligning with the purpose of making the analysis of temporal (dynamic) networks more accessible.

6.2.1 `network_properties`

Understanding the structure and dynamics of networks is pivotal in various fields, from social network analysis to biological systems and network infrastructure. The `network_properties` function plays a role in this context by systematically analysing a collection of networks, extracting many properties of these networks.

Purpose of the Function The primary purpose of the `network_properties` function is to compute and optionally visualise a wide range of metrics for a set of networks. These metrics provide detailed insights into the structural characteristics and potential functional behaviours of the networks. By analysing these properties, researchers can detect patterns, compare different networks, and track changes over time.

Input and Setup The function accepts a list of networks as input. Optionally, it allows specifying a filename to save numerical results and a directory path for saving visualisations. It ensures the specified directory exists or creates it if necessary, facilitating organised storage of results.

Network Properties Calculation The function computes a comprehensive set of properties for each graph. These properties include:

- Number of Nodes: Indicates the size of the network.
- Number of Edges: Reflects the connectivity level of the network.
- Density: Measures how densely connected the network is.
- Degree Distribution: Provides statistics about the node degrees (connections).
- Strongly Connected Components: Represents subgraphs where every node is reachable from every other node within the same subgraph.
- Girth: The length of the shortest cycle in the graph, important for understanding cyclic structures.
- Diameter and Average Path Length: Indicate the longest shortest path and the average shortest path between nodes, respectively, providing insights into network navigability.
- Mean Degree: Average number of connections per node, giving a sense of overall connectivity.
- Reciprocity: Measures the likelihood of mutual connections, relevant in directed networks.
- Transitivity (average clustering coefficient): Also known as the clustering coefficient, it measures the tendency of nodes to form clusters or triangles.
- Additionally, we check if the network is bipartite, connected, a directed acyclic graph (DAG), directed, named, simple (without loops and multiple edges), weighted, or if it has multiple edges.

The function processes a collection of graphs, which can be stored in multiple formats. One such format is a pickle file, a Python-specific serialization method that efficiently preserves the state of complex objects like graphs, enabling quick storage and retrieval. Alternatively, the graphs can be stored in a simple list structure, with each graph represented as an individual element.

Within the function, each graph in the list is systematically iterated over, and key network properties âĂŤ such as centralities, clustering coefficients, or other graph metrics âĂŤ are computed.

These computed properties are then appended to their respective lists, allowing for organized accumulation of data.

Once the iteration is complete, the compiled data is structured into a pandas DataFrame. This organization into a DataFrame ensures that the network properties are readily accessible, facilitating both granular exploration and broad statistical evaluations.

Data Organisation and Storage If a filename is provided, the numerical results are saved to a CSV file, ensuring the data can be easily shared and referenced. Flattening the degree vectors by calculating their mean values simplifies the DataFrame, making it more manageable.

Visualisation Visual representation of network properties is integral to understanding and communicating the findings. The function generates plots for each numerical property (excluding vectors like degree distributions), saving these visualisations as PDF files in the specified directory. These plots help identify trends, outliers, and comparative differences across the networks.

The function includes an argument, `visualisation`, which is set to `TRUE` by default, that determines whether plots are generated. If visualizations are not required, setting the `visualisation` argument to `FALSE` will skip the plot generation, focusing solely on the numerical analysis of the network properties.

Significance The `network_properties` function is a useful tool for network analysis, offering several significant benefits:

1. Comprehensive Analysis: Calculating a wide range of properties, provides a holistic view of the network's structure.

2. Automation: Automating the computation and visualisation process saves time and reduces the potential for human error.

3. Comparative Insights: The ability to analyse multiple networks simultaneously allows for comparative studies, essential in understanding how different networks relate or how a single network evolves.

4. Scalability: Suitable for large datasets, it can handle numerous graphs and extensive properties, making it applicable for big data analytics in network science.

5. Versatility: Applicable across various domains, from social network analysis to biological networks, providing valuable insights in each field.

The `network_properties` function can be an essential tool for researchers studying network dynamics. The `network_properties` function systematically calculates and optionally visualizes a wide array of network metrics, aiding in the analysis of network structures and their changes over time. The function includes an argument, `visualisation`, which is set to `TRUE` by default. This determines whether plots are generated, providing flexibility for users to choose whether or not to create visual representations based on their specific needs. This function enhances the ability to detect patterns, compare different networks, and gain insights into the underlying principles governing complex systems.

6.2.2 `calculate_centralities`

Centrality measures provide insights into the importance and influence of individual nodes within a network and as such, are extremely important when studying a network. The `calculate_centralities` function is designed to systematically compute various centrality metrics for a collection of networks.

Purpose of the Function The primary purpose of `calculate_centralities` is to compute a comprehensive set of centrality measures for each node in a list of network. Centrality measures such as degree centrality, closeness centrality, betweenness centrality, and others provide critical insights into the roles and importance of nodes within the networks. These measures may help researchers identify key nodes, understand network robustness, and explore the network's structural and functional dynamics.

Input and Setup The function accepts a list of networks as its input. It initialises an empty list to store the centrality measures for each graph. Additionally, an optional filename parameter allows specifying where to save the computed centralities.

Handling Non-Simple Graphs When dealing with non-simple graphs, which may include multiple edges between the same nodes and loops (edges that connect a node to itself), the function ensures proper handling to maintain the integrity of centrality calculations. This involves combining multiple edges between the same pair of nodes by summing their weights and removing self-loops. This simplification ensures that the calculated centrality measures accurately reflect the network's structure without redundant or self-referential connections.

Centrality Measures Calculation For each graph in the input list, the function calculates the following centrality measures:

- Degree Centrality: Measures the number of direct connections a node has, indicating its immediate influence.
- Closeness Centrality: Reflects how close a node is to all other nodes in the network, indicating its efficiency in spreading information.
- Betweenness Centrality: Quantifies the number of times a node acts as a bridge along the shortest path between two other nodes, indicating its role in information flow.
- Eigenvector Centrality: Measures a node's influence based on the connectivity of its neighbours, indicating its importance in the overall network structure.
- PageRank: Evaluates the significance of a node based on the quality and quantity of its links, famously used by Google for ranking web pages and was named after Larry Page, one of the co-founders of Google.
- Harmonic Centrality: Considers the closeness of a node to all other nodes, offering a variant of closeness centrality that is less sensitive to disconnected components.
- Eccentricity: The greatest distance from a node to any other node in the network, providing a measure of the node's reach.
- Clustering Coefficient: Indicates the degree to which nodes tend to cluster together, reflecting the local density of connections.
- HITS Authority and Hub Scores: Measures from the Hyperlink-Induced Topic Search algorithm, where authority scores indicate the value of a node as a source of information and hub scores reflect its value as a connector to information sources.

Each graph's centrality measures, along with node labels, are stored in a dictionary and appended to the centralities list.

Data Organisation and Storage The collected centrality data is normalised into a flat structure for easier manipulation and analysis. This involves creating a list of dictionaries where each dictionary contains the centrality measures for a single node across all graphs. The flattened data is then converted into a pandas DataFrame.

Significance The `calculate_centralities` function computes a wide range of centrality measures. This detailed node analysis provides a comprehensive view of each node's role and importance within the network.

6.2.3 `communities_measures`

Community detection in networks is a fundamental task that helps in understanding the modular structure and functional units within complex networks. The `communities_measures` function is designed to apply multiple community detection algorithms across a series of networks, analyse the resulting community structures, and provide comprehensive visual and numerical insights.

Purpose of the Function The primary purpose of the `communities_measures` function is to compute and optionally visualise community structures within a set of networks using various community detection algorithms. By analysing these community structures, researchers can uncover patterns of modularity, detect significant subgroups within the network, and study how these structures evolve over time or under different conditions.

Input and Setup The function accepts a list of networks and an optional directory path for saving visualisations. It ensures the specified directory exists or creates it if necessary.

Handling Non-Simple and Undirected Graphs In the context of community detection, it is essential to address the complexities presented by non-simple graphs, which may include multiple edges between nodes and loops. The incorporates a process of simplification to manage these complexities. This process involves aggregating the weights of multiple edges between the same pair of nodes and eliminating loops. Moreover, certain community detection algorithms (specifically Walktrap, Fast Greedy, Label Propagation, and Spinglass) necessitate the use of undirected graphs. Consequently, the function converts directed graphs to undirected graphs when employing these algorithms. This conversion is essential to facilitate the appropriate analysis of the network's community structure, thereby ensuring that the detection outcomes are both meaningful and reflective of the network's true dynamics.

Community Detection Algorithms The function employs several well-known community detection algorithms, including:

- Leiden: Algorithm optimized for finding well-defined communities.
- Louvain: Popular method for detecting communities by optimizing modularity.
- Walktrap: Uses random walks to detect communities.
- Fast Greedy: Greedy optimization method for community detection.
- Label Propagation: Fast method based on label propagation.
- Spinglass: Method based on statistical mechanics.

Properties Calculation For each graph in the input list, the function constructs an undirected graph and applies each community detection algorithm. It then collects and organises the results into a pandas DataFrame, capturing the community assignment for each node within each graph.

Data Organization and Storage The calculated community assignments are saved to CSV files, one for each algorithm. Additionally, the function computes and saves statistics such as the number and maximum size of communities for each graph.

Visualization Visual representation of community measures is crucial for understanding and communicating the results. The function generates time series plots for the number and maximum size of communities, saving these visualisations as PNG files in the specified directory.

Significance By applying multiple community detection algorithms, it provides a robust analysis of the network's structure.

6.2.4 `plot_community_evolution`

Understanding how communities within a network evolve over time or across different conditions is crucial for many areas of research. The `plot_community_evolution` function leverages Plotly to create interactive visualisations that animate the evolution of community structures detected by various algorithms.

Purpose of the Function The `plot_community_evolution` function is designed to visualise the dynamic changes in community structures within a series of networks. Animating these changes provides a clear and engaging way to observe how communities form, merge, split, and evolve over time, offering valuable insights into the temporal dynamics of networked systems.

Input and Setup The function takes a list of networks and a community detection algorithm name. It initializes variables to store results and determines the corresponding algorithm function.

Handling Non-Simple and Undirected Graphs The function simplifies networks with multiple edges or loops by combining edge weights. Additionally, it converts directed networks to undirected ones for community detection algorithms requiring this format.

Community Detection Algorithms : The function supports various algorithms such as Edge Betweenness, Walktrap, Fast Greedy, Label Propagation, Spinglass, Leiden, and Louvain, using `igraph` functions.

Computing Communities For each graph, the function computes communities using the specified algorithm.

Visualizing the Evolution It utilises Plotly to create an interactive animation showing community evolution over graphs. Nodes are coloured based on community membership.

User Controls : Playback controls allow users to play, pause, and restart the animation.

Significance The function offers:

- Interactive Visualization: Engaging visualizations aid in understanding community dynamics.
- Temporal Insights: Observing community evolution over time facilitates dynamic network analysis.
- Algorithm Comparison: Comparison of community detection algorithms.

The function initialises results storage, computes communities, and creates Plotly animations. It utilises random node positions if not provided and generates node traces and edge traces for each frame of the animation. Playback controls are implemented using Plotly's features.

6.2.5 `vertex_properties`

Analysing vertex properties within networks provides insights into the structural and functional roles of individual nodes. The `vertex_properties` function is designed to calculate and visualise a comprehensive set of vertex-centric properties for a specific node across multiple networks.

Purpose of the Function The `vertex_properties` function aims to compute and visualise a wide range of vertex-centric properties for a particular node across a series of networks. By tracking these properties, researchers can gain insights into how the node's role and influence evolve over time or under varying conditions.

Input and Setup The function accepts a list of networks, the label of the node of interest, a filename for saving the results, and an optional directory path for saving visualisations. It ensures the specified directory exists or creates it if necessary.

Handling Non-Simple and Undirected Graphs The function simplifies networks with multiple edges or loops by combining edge weights.

Properties Calculation For each graph in the input list, the function verifies the presence of the node of interest. If the node is present, it calculates the following properties:

- Diversity: Measures the node's role in connecting diverse parts of the network.
- Authority Score: Evaluates the node's value as a source of information.
- Hub Score: Reflects the node's value as a connector within the network.
- Betweenness: Quantifies the node's role as a bridge in the shortest paths between other nodes.
- Closeness: Indicates the node's efficiency in spreading information within the network.
- Constraint: Measures the extent to which a node is constrained by its neighbours.
- Coreness: Represents the node's position within the network's core structure.
- Eccentricity: The greatest distance from the node to any other node in the network.
- Eigenvector Centrality: Indicates the node's influence based on its connections and the importance of its neighbours.
- Harmonic Centrality: Considers the node's closeness to all other nodes, adjusted for disconnected components.
- PageRank: Evaluates the node's significance based on the quality and quantity of its links.
- Strength: Measures the sum of the weights of the edges connected to the node.
- Transitivity (Local Clustering Coefficient): Measures the tendency of the node to form clusters with its neighbours.

The function compiles these properties into a dictionary, which is then converted into a pandas DataFrame.

Data Organization and Storage The DataFrame is saved to a CSV file using the specified filename, ensuring the data can be easily accessed for further analysis.

Visualization Visual representation of vertex properties is crucial for understanding their temporal or contextual changes. The function generates plots for each property over time (i.e., across different graphs) and saves these visualisations as pdf files in the specified directory. Additionally, it creates a combined plot displaying all properties on a single graph to facilitate comparative analysis.

Significance By targeting a specific node, the function provides insights into its role and importance within the network.

6.3 Case Study: Helsinki City Bikes

The data used in this section can be found on kaggle [5].

In the realm of infectious disease modelling, human mobility plays a crucial role in understanding disease transmission dynamics.

Transportation networks, consisting of nodes (locations) and edges (routes), represent pathways along which individuals move within urban environments. Traditional networks include physical tracks like roads and rails, alongside less tangible routes such as air and sea corridors. Unlike fixed routes typical of scheduled public transit services, bike-sharing and car-sharing systems operate on-demand, allowing for dynamic, spatially flexible mobility patterns. The distinctiveness of bike-sharing networks stems from their self-organizing nature; network edges are defined by user interactions rather than predetermined routes. This dynamic configuration continually evolves due to the habitual movements of users, resulting in usage patterns that progressively shape the network's structure over time.

Helsinki City Bikes are shared bicycles available to the public in the Helsinki and Espoo metropolitan areas. Since its inception in 2016, Helsinki City Bikes has grown from a pilot project with just 46 stations in Helsinki to a robust network serving both the Helsinki and Espoo metropolitan areas. By 2019, the network had expanded significantly, adding approximately one hundred stations annually between 2017 and 2019. Despite a modest increase in 2020, with only 7 additional stations, the network comprised a total of 3,510 bikes distributed across 350 stations by the end of that year.

To use the city bikes, citizens can purchase access for a day, a week, or the entire cycling season, which lasts roughly from April to November. All passes include an unlimited number of 30-minute bike rides. For an extra fee of 1 € per hour, users can extend their rides. Bikes are picked up and returned to stations located throughout Helsinki and Espoo.

6.3.1 Description of the data

The dataset describes trips with rental bicycles in Helsinki, capturing various aspects of each trip, including spatial and temporal details, trip metrics, and environmental conditions. Table 6.1 provides an overview of the variables included in the dataset. The data spans multiple seasons, including the following periods:

- 2016: May to October,

- 2017: May to October,
- 2028: April to October,
- 2019: April to October,
- 2020: March to October.

Column Name	Data Type	Description
departure	datetime64[ns]	Timestamp of departure.
return	datetime64[ns]	Timestamp of return.
departure_id	object	Unique identifier for the departure.
departure_name	object	Name of the departure location (if available).
return_id	object	Unique identifier for the return.
return_name	object	Name of the return location (if available).
distance (m)	float64	Distance traveled in meters.
duration (sec.)	float64	Duration of the trip in seconds.
avg_speed (km/h)	float64	Average speed of the trip in kilometers per hour.
departure_latitude	float64	Latitude coordinate of the departure location.
departure_longitude	float64	Longitude coordinate of the departure location.
return_latitude	float64	Latitude coordinate of the return location.
return_longitude	float64	Longitude coordinate of the return location.
Air temperature (degC)	float64	Air temperature in degrees Celsius (if available).

Table 6.1: Variables in the Helsinki City Bike-sharing dataset, encompassing timestamps, identifiers, location coordinates, and trip metrics essential for analysing urban mobility patterns within the Helsinki and Espoo metropolitan areas.

Analyzing the data reveals that the initial dataset of Helsinki city bike trips contains some anomalies, such as negative and excessive large distances. Table 6.2 provides descriptive statistics, including the count, mean, standard deviation, minimum, 25% percentile, median, 75% percentile and maximum for both distance and duration before and after filtering out all negative distances.

Although the Helsinki city bike-sharing system has expanded significantly since its inception, the characteristics of individual bike trips have shown consistent patterns over the years. Analysis of trips from the past five years reveals that the average ride duration is approximately 10 minutes (Figure 6.1a) and an average distance of approximately 2167 meters (Figure 6.1b). However, it is important to note that the distribution of trip durations is right-skewed, with the majority of trips lasting between 4 to 8 minutes and covering an average distance of around 1700 meters. This skewed distribution can be attributed to the system's pricing structure. Users can purchase access for a day, a week, or the entire cycling season, with each pass including an unlimited number of 30-minute bike rides. To avoid extra charges for extended rides, many users aim to

Statistic	Initial		Filtered	
	Distance (m)	Duration (sec.)	Distance (m)	Duration (sec.)
Count	12,157,458	12,157,458	11,769,717	11,769,717
Mean	2,295.28	959.78	2,373.44	976.48
Std. Dev.	24,520.67	7,346.53	24,696.88	7,426.38
Min	-4,292,467	0	1	2
25%	1,000	344	1,057	360
Median	1,739	586	1,794	600
75%	2,869	971	2,924	983
Max	3,681,399	5,401,659	3,681,399	5,401,659

Table 6.2: Descriptive statistics of Helsinki city bike-sharing network trips before and after filtering. The statistics include count, mean, standard deviation (Std. Dev.), minimum (Min), 25th percentile (25%), median (50%), 75th percentile (75%), and maximum (Max) for distance in meters and duration in seconds.

complete their trips within the 30-minute window. This incentive likely influences the observed trip characteristics, resulting in shorter, more frequent rides.

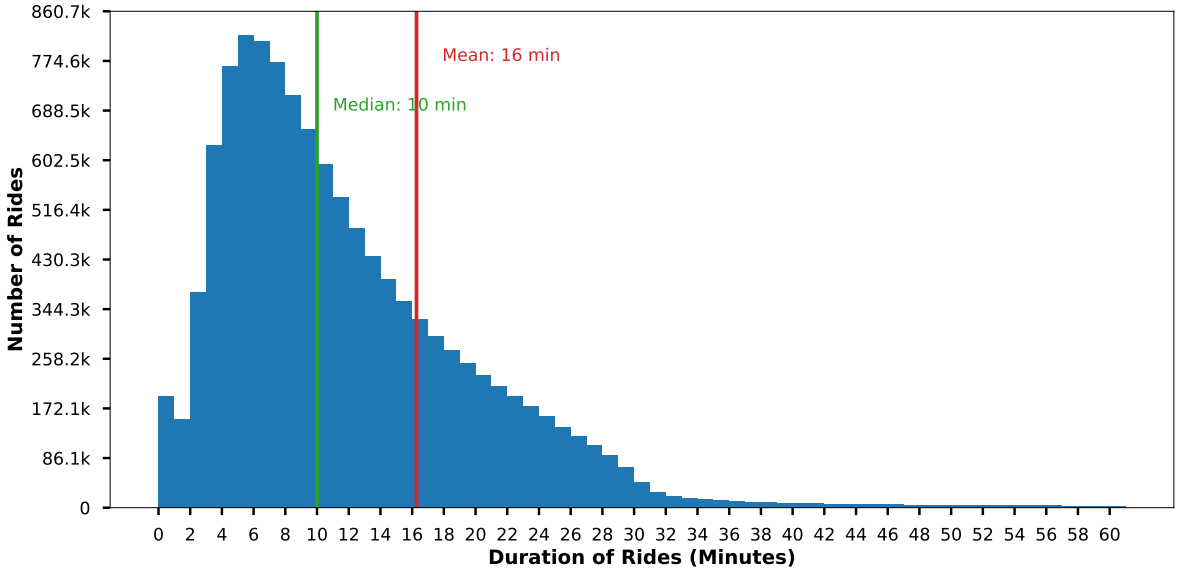
Figure 6.2a depicting daily bike trips since the system's launch illustrates the significant impact of network expansion on citizen usage. Notably, 2020 marked the first year of a decline in bike usage, which can be attributed to population movement restrictions during the COVID-19 pandemic.

6.3.2 Data wrangling

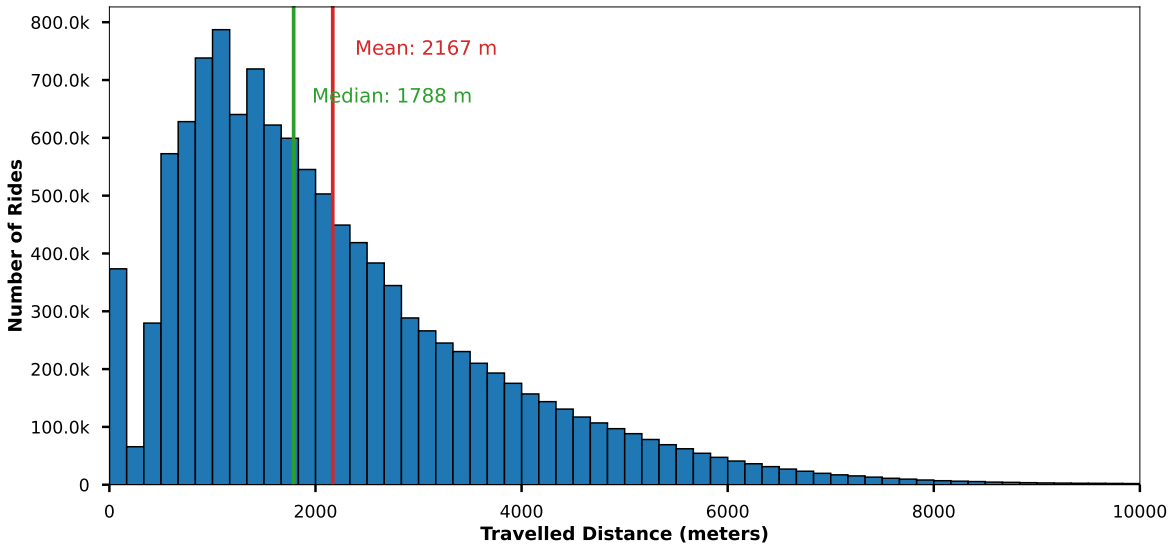
The initial phase involved loading the dataset with careful attention given to ensuring the "departure" and "return" columns were appropriately parsed as datetime objects to facilitate temporal analysis. Subsequently, filtering out rows where the recorded distance travelled was non-positive, thereby restricting the dataset to valid ride records only.

Furthermore, we aggregated the dataset into discrete monthly intervals, grouping rides that occurred within the same month. It is important to note that the data covers continuous time, reflecting each borrowed bike over the following periods:

- 2016: May to October,
- 2017: May to October,
- 2018: April to October,
- 2019: April to October,
- 2020: March to October

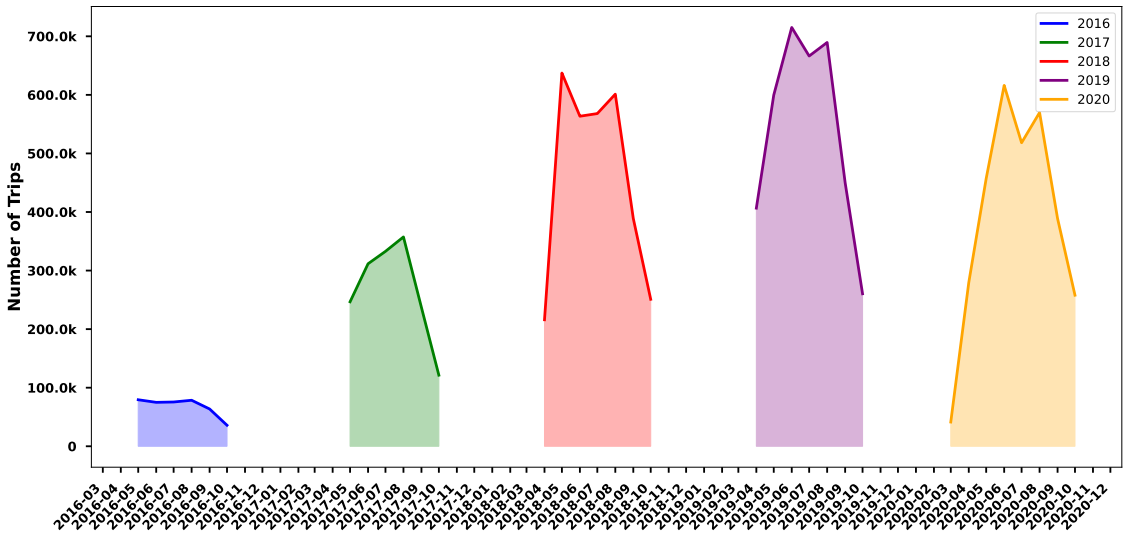


(a) Distribution of trip durations in minutes.

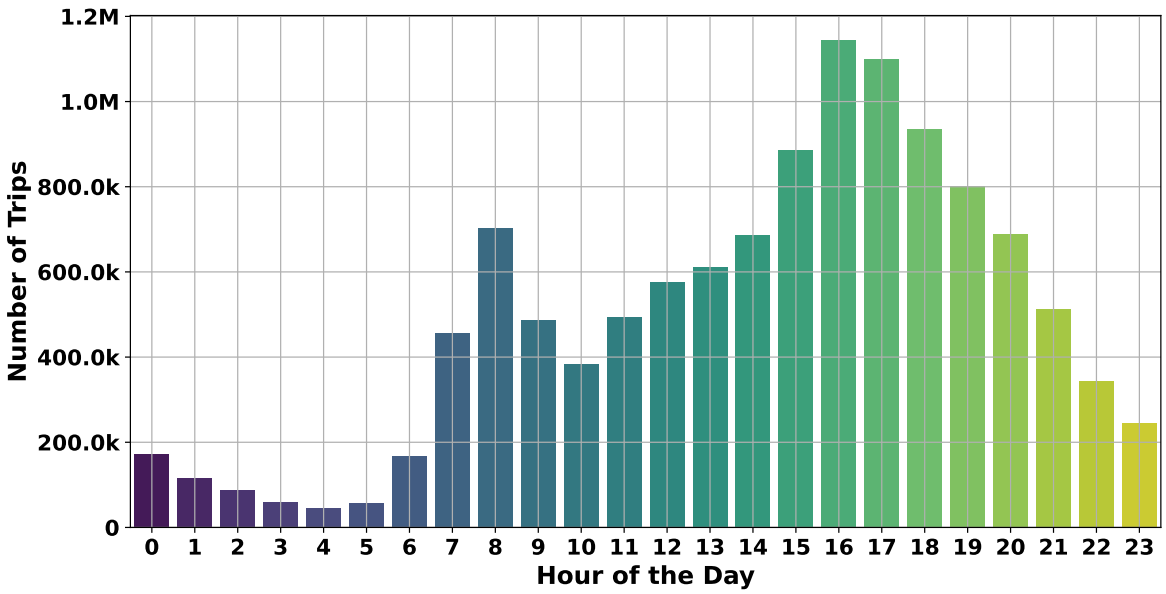


(b) Distribution of travelled distances in meters.

Figure 6.1: Exploratory analysis of bike trips in the Helsinki City Bike-sharing system. The top panel illustrates the distribution of trip durations, highlighting that most trips last between 4 to 8 minutes, with an average duration of approximately 10 minutes. The bottom panel depicts the distribution of trip distances, showing an average distance of approximately 2167 meters. The skewed distribution of trip durations can be attributed to the pricing structure of the bike-sharing system, which incentives users to complete their trips within 30 minutes to avoid additional charges.



(a) Daily number of trips.



(b) Hourly distribution of trips.

Figure 6.2: Exploratory analysis of bike trips in Helsinki. The top panel shows the total number of trips per day, providing an overview of daily usage patterns. The bottom panel displays the distribution of trips across different hours of the day, highlighting peak usage times and periods of lower activity.

6.3.3 Evolution of the network

Figure 6.3a illustrates the number of trips in the Helsinki city bikes-sharing network over a period ranging from May 2016 to December 2020. Figure 6.3b illustrates the number of stations in the Helsinki city bikes-sharing networks over a period ranging from May 2016 to December 2020.

The network was constructed using trip data from Helsinki city bikes-sharing data. Each CSV file, representing monthly data, was processed to create a directed graph where nodes represent bike stations and edges represent trips between stations. Node positions were derived from station coordinates, ensuring geographical accuracy. Edges were added between nodes with weights corresponding to the total number of trips recorded between each station pair over the month. This approach ensured that the resulting network captured the flow of bike trips across Helsinki, facilitating analysis of station connectivity and traffic patterns over time.

In Figure 6.4a, we observe the temporal changes in the mean degree of the networks. This metric can be interpreted as the average number of connections that a bike station has with other stations over time. It provides insight into how the connectivity of bike stations evolves, indicating changes in the overall network structure and potentially reflecting shifts in usage patterns or network expansions.

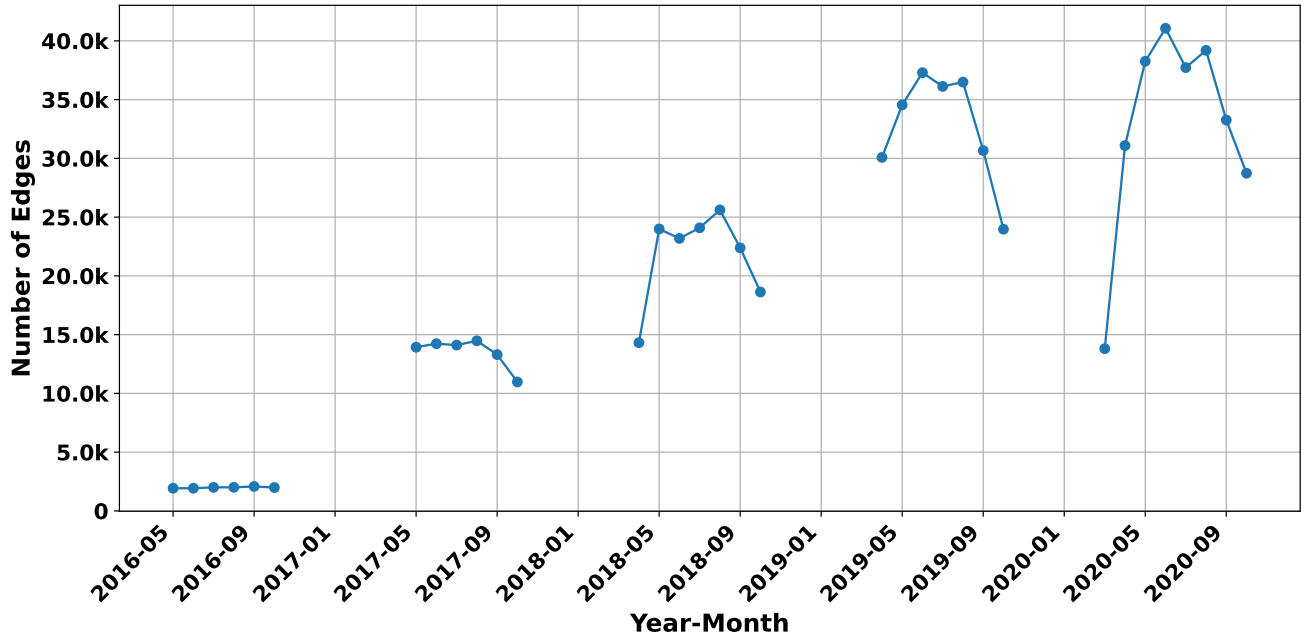
Figure 6.4b shows the average clustering coefficient, starting near 1 and decreasing over time, indicating a trend in the bike city network. Initially, the network featured highly clustered nodes, suggesting numerous closely connected groups. However, as more bike stations were added over time, the network structure evolved, with connections becoming more dispersed.

A notable drop in the average clustering coefficient occurred during the COVID-19 pandemic, reflecting significant changes in network dynamics possibly due to altered usage patterns or infrastructure adjustments. This period likely saw reduced connectivity between bike stations, contributing to the observed decrease in clustering.

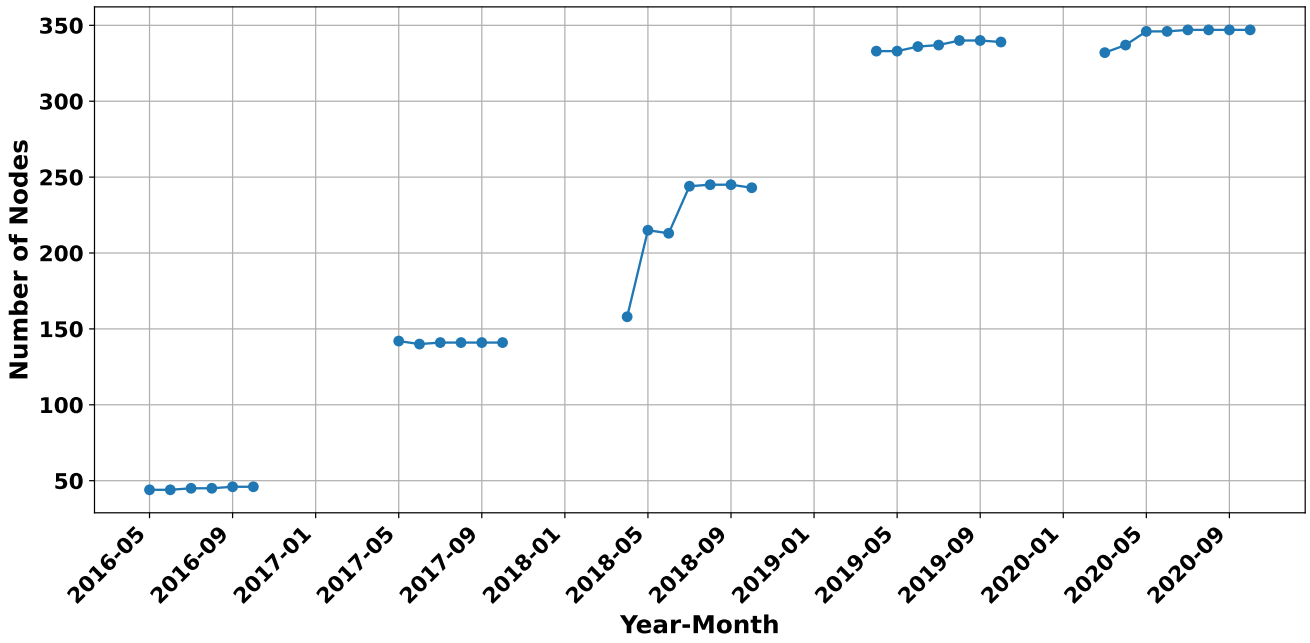
These findings suggest that the bike city network underwent substantial transformations over time, influenced by both network expansion and external events such as the COVID-19 pandemic, impacting the network's overall structure and connectivity patterns.

In Figure 6.5, we present the evolution of the number of communities and the size of the largest community in the Helsinki bike share network as detected by three popular community detection algorithms: Louvain, Leiden, and Walktrap. The plots reveal both the temporal variation in the number of communities and the corresponding fluctuations in the size of the largest community over the analysed period.

Notably, while the size of the maximum community remains relatively similar across all three methods, the total number of detected communities varies dramatically. This divergence highlights a critical point in community detection: different algorithms can yield vastly different partitions of the same network because they are designed to prioritize different structural properties.

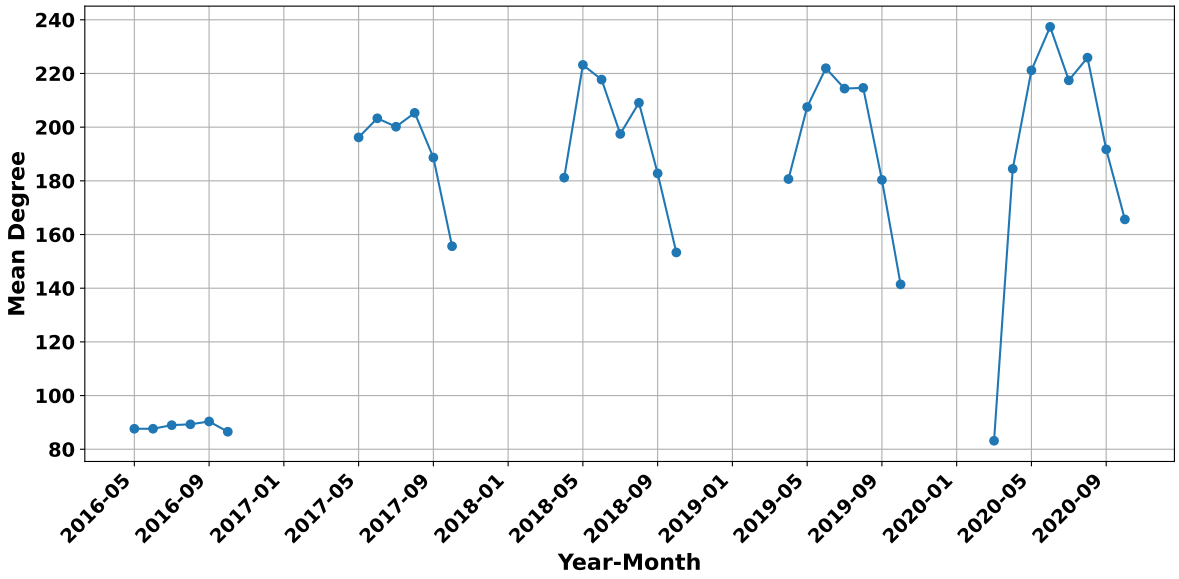


(a) Evolution of the number of edges.

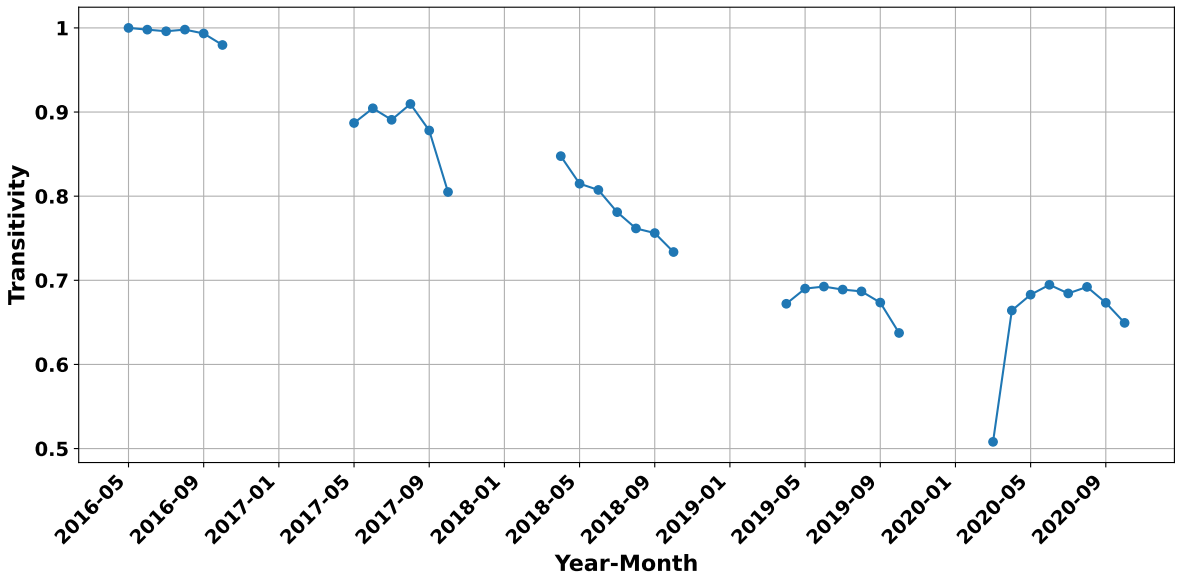


(b) Evolution of the number of nodes.

Figure 6.3: Temporal analysis of network structures in Helsinki. The top panel illustrates the changes in the number of edges, indicating the connectivity between nodes, while the bottom panel shows the changes in the number of nodes, representing the entities within the networks.



(a) Mean total degree over time.



(b) Transitivity (average clustering coefficient) over time.

Figure 6.4: Temporal analysis of network metrics in Helsinki. The top panel shows the mean total degree, indicating the average number of connections per node over time. The bottom panel displays the transitivity, revealing the extent of clustering among nodes in the network over time.

For example, Louvain and Leiden optimize modularity, which tends to favor partitions that group together nodes with dense internal connections and sparse connections to other groups. These methods are more likely to produce a smaller number of larger communities, sometimes merging together distinct but loosely connected regions. On the other hand, Walktrap uses short random walks to detect local cohesion, which can be sensitive to even subtle variations in the local connectivity. As a result, it tends to identify a much larger number of smaller, tightly-knit communities, especially in networks with significant local structure.

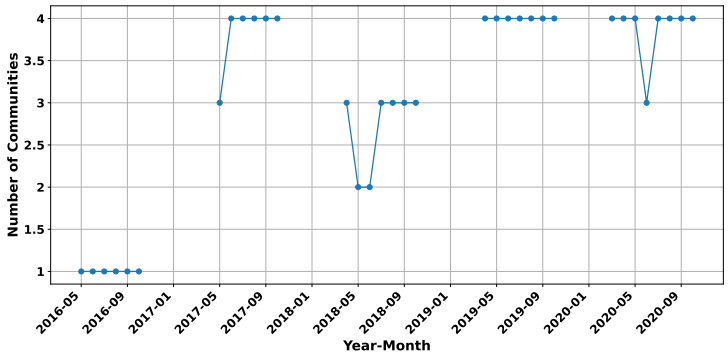
The fact that the maximum community size remains stable while the total number varies suggests that all three algorithms consistently identify a dominant hub or core in the network, likely reflecting central bike stations with high traffic. However, beyond this core, the interpretation of smaller, peripheral structures diverges. This makes clear that the number of communities is not an absolute value, but a function of algorithmic assumptions and resolution limits.

These differences underscore the importance of algorithm selection and the need to consider multiple approaches when interpreting community structures. In practice, this variation can be viewed not as a limitation, but as an opportunity to uncover multi-scale structure in the network: broad patterns from Louvain or Leiden, and localized dynamics from Walktrap.

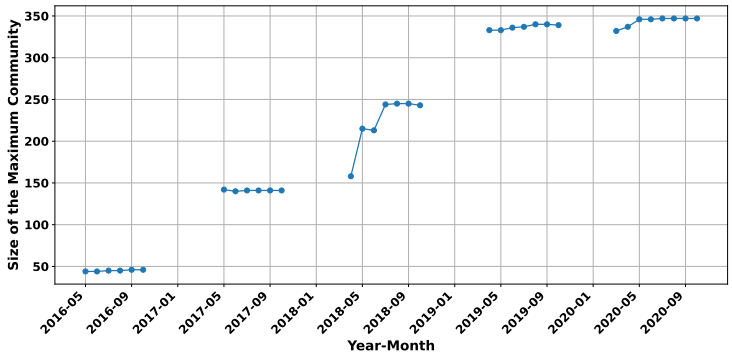
Figures 6.6 provide a comprehensive view of the monthly dynamics of edge formations and dissolutions within the Helsinki city bike-sharing network. 6.6a illustrates the number of edges that were dissolved each month, while 6.6b depicts the number of edges formed within the same period.

These visualizations are crucial in understanding the temporal evolution of the network. The dissolution and formation of edges reflect changes in connectivity and interaction patterns among nodes. Figure 6.6a shows a trend of edge dissolution, with noticeable peaks indicating months where a significant number of connections were terminated. Conversely, Figure 6.6b highlights the periods of network growth, with peaks representing months where many new connections were established. These peaks may correlate with events or periods of increased activity, such as the start of the biking season, public initiatives to promote cycling, or improvements in bike-sharing infrastructure.

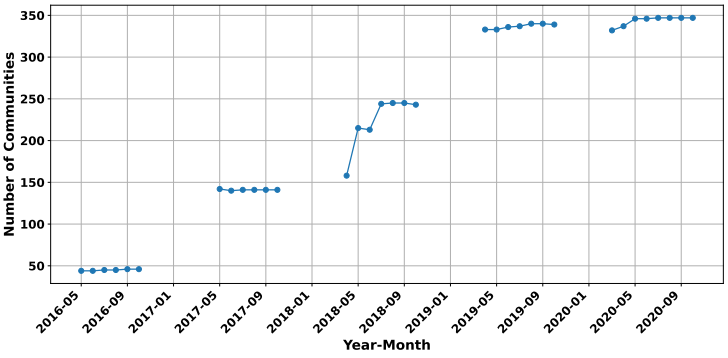
Betweenness centrality is a measure of the extent to which a node lies on the shortest paths between other nodes in a network. A higher betweenness centrality indicates that a location acts as a crucial bridge or connector within the network, facilitating the flow of bike trips between other locations. Figure 6.7a illustrates the maximum betweenness centrality value in the Helsinki city bike-sharing network over 34 months. Initially, the maximum betweenness centrality value is relatively small and almost constant across the first six networks, but the location of the node with the maximum value changes. This suggests that multiple nodes are sharing the role of key connectors early on. From May 2018 onwards, a single location consistently holds the maximum



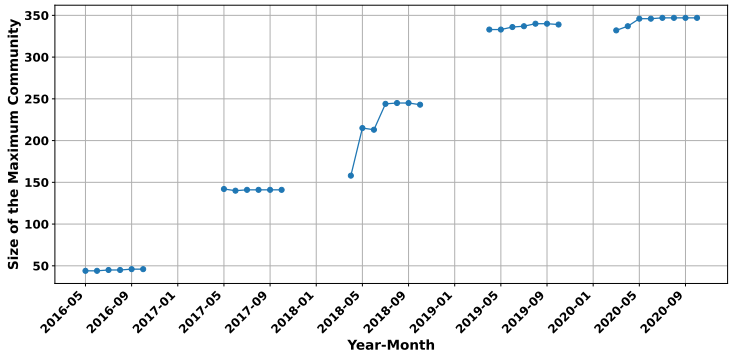
(a) Number of communities (Louvain).



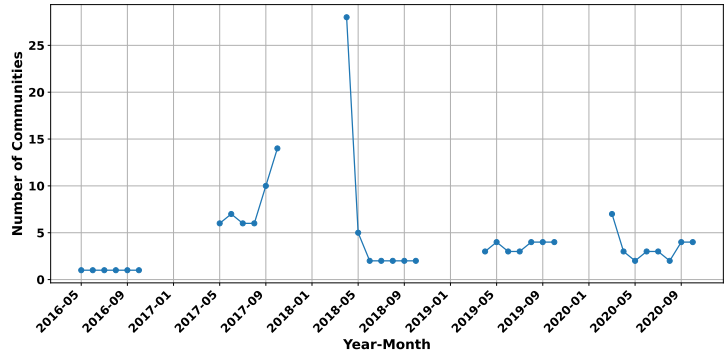
(b) Size of the maximum community (Louvain).



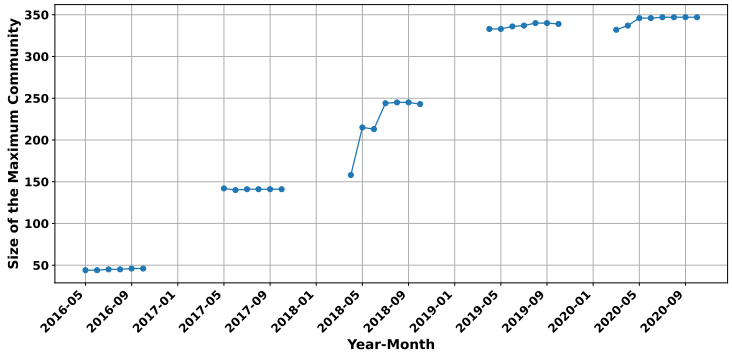
(c) Number of communities (Leiden).



(d) Size of the maximum community (Leiden).

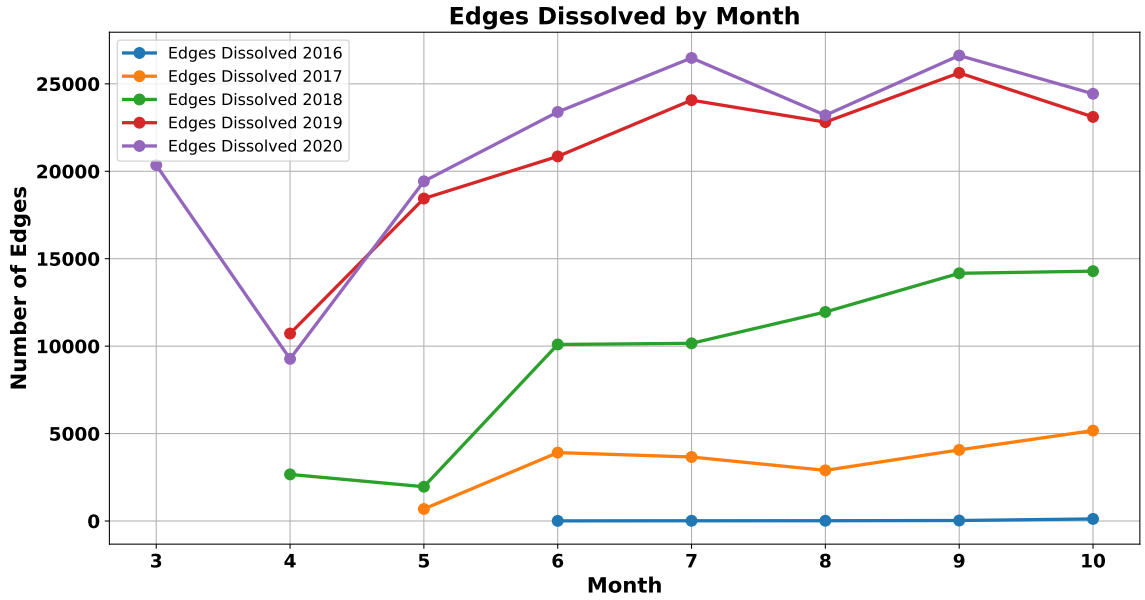


(e) Number of communities (Walktrap).

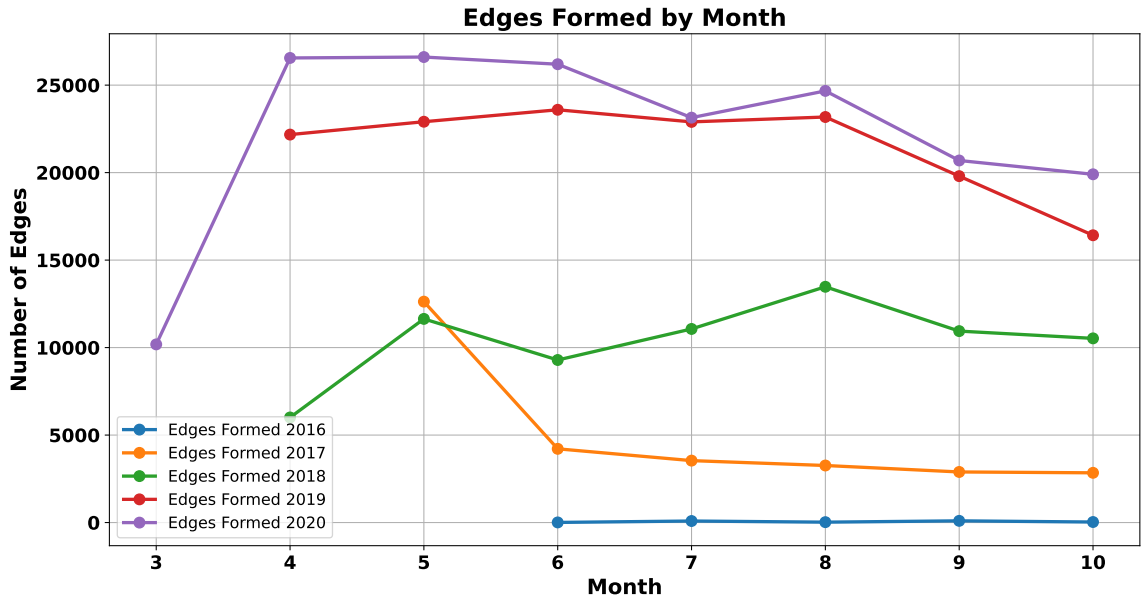


(f) Size of the maximum community (Walktrap).

Figure 6.5: Evolution of the number of communities and size of the maximum community in the Helsinki bike share data obtained using different community detection algorithms (Louvain, Leiden and Walktrap).

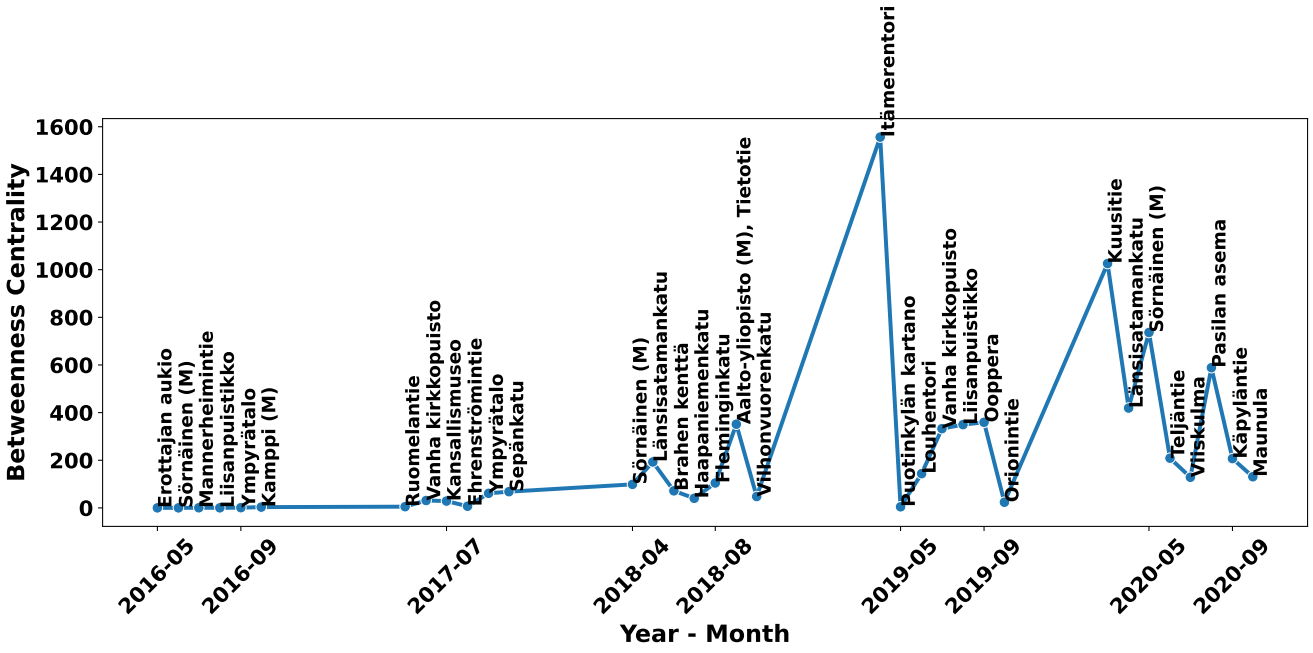


(a) Edges Dissolved in the Network.

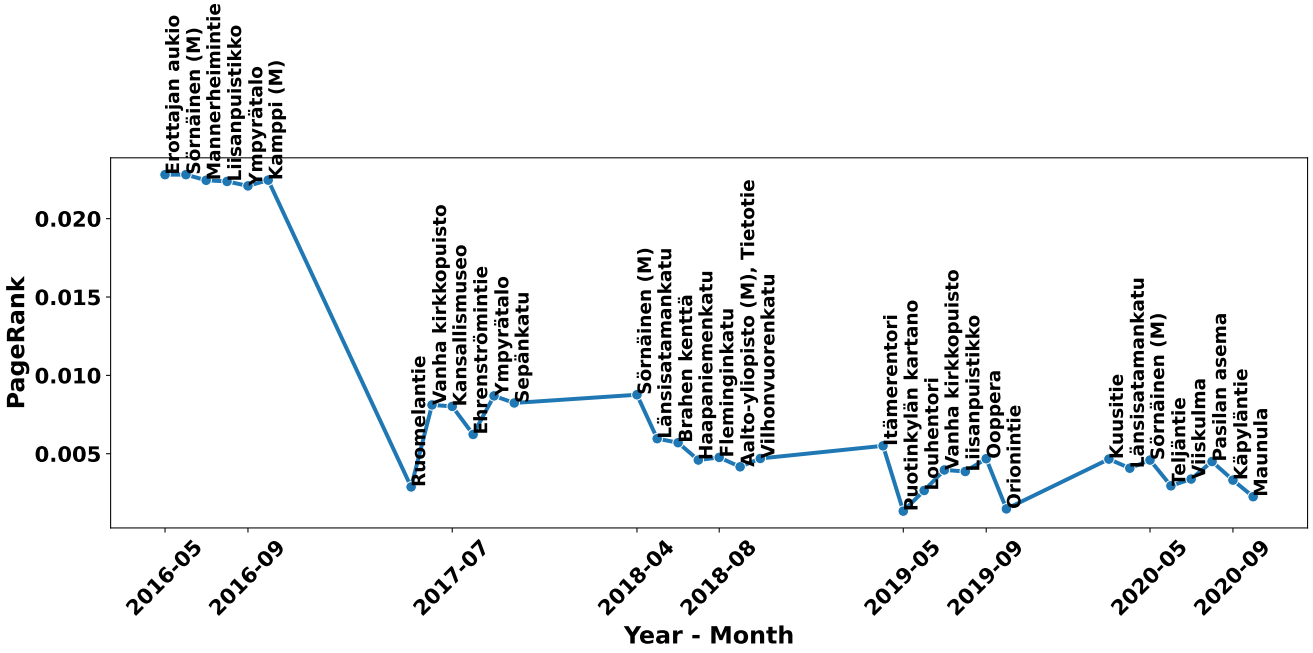


(b) Edges Formed in the Network.

Figure 6.6: Monthly dynamics of the Helsinki network. (a) The number of edges dissolution each month from 2016 to 2020. (b) The number of edge formations each month over the same period. These figures illustrate the fluctuating nature of connections within the network, emphasizing periods of high connectivity changes.



(a) Maximum betweenness for Helsinki city bike-sharing network.



(b) Maximum PageRank for Helsinki city bike-sharing network.

Figure 6.7: Monthly maximum centrality measures for Helsinki city bike-share network. Subfigure (a) shows the maximum betweenness centrality across graphs, while subfigure (b) displays the maximum PageRank centrality. Station names above each data point indicate the station corresponding to the maximum centrality values in each graph, representing monthly snapshots of network properties from the dataset.

betweenness centrality. By tracking the maximum betweenness centrality and the corresponding nodes over time, we can identify key connector locations and understand how their importance evolves. This information is valuable for optimizing network design, ensuring robust connectivity, and identifying potential bottlenecks or critical points of failure.

PageRank centrality is a measure of the importance of nodes within a network, originally developed by Google to rank web pages in search results. The PageRank algorithm assigns a centrality value to each node based on the number and quality of links to it, with higher values indicating more important or influential locations.

Figure 6.7b illustrates the maximum PageRank centrality value in the Helsinki city bike-sharing network. The maximum PageRank value varies over time, reflecting changes in the relative importance of different locations within the network. These fluctuations may be influenced by various factors, such as shifts in user behaviour or changes in network structure. By tracking the maximum PageRank centrality and the corresponding nodes over time, we can identify key locations and understand how their importance evolves. This information can be valuable for network planning and optimization, helping to identify critical hubs and potential areas for improvement.

6.4 Case Study: Air transportation network 2019-2022

6.4.1 Motivation

Understanding how people move across international borders is essential for assessing the risk of emerging or re-emerging infectious diseases spreading between countries. Although the analysis presented here may seem tangential, it directly supports an external line of research not included in this thesis. This work aims to estimate the risk of disease importation during the early stages of an outbreak and was first outlined in our June 2020 report to the Public Health Agency of Canada [97].

To support such assessments, we require dynamic models of cross-border mobility. Access to near-real-time transportation data and robust analytical frameworks enables us to evaluate how travel restrictions impact disease propagation and identify high-risk points of entry. These needs motivate the use of air travel data and temporal network analysis tools, such as the software developed and demonstrated in this thesis.

The external research relies on a deterministic metapopulation model that treats each geographic location as a node and passenger flows as edges. Rather than using potentially unreliable incidence curves, the model focuses on the timing of first reported cases to predict the spatial and temporal spread of a pathogen. The framework extends the classic SLIAR model (Susceptible Latent Infectious Asymptomatic Recovered) by incorporating both symptomatic and asymptomatic

cases, as well as detection probabilities and region-specific progression dynamics.

A central component of the model is the movement matrix, which encodes how individuals travel between regions over time. Due to uncertainties in disease parameters,particularly during the early stages of an outbreak, the model adopts a stochastic optimization approach. This allows it to generate multiple plausible epidemic trajectories and evaluate relative risks of introduction, rather than relying on a single deterministic output. This approach improves robustness and accounts for parameter non-identifiability, providing more realistic risk estimates.

This case study, like the Helsinki bike sharing case study, demonstrates the capabilities of our Python package for temporal network analysis.

6.4.2 Description of the data

The flight data is sourced from Automatic Dependent SurveillanceâĂŞBroadcast (ADS-B) data, which is automatically transmitted by aircraft and includes information about their position and identification. This data can be received by simple ground-based receivers, leading to the development of a community, including many members of the general public, who receive and share data through platforms such as the OpenSky Network [171].

In addition to the more subtle limitations we address later, there are structural limitations to the data worth mentioning at this point. ADS-B, as a system, is progressively mandated by law in various national and transnational jurisdictions. Consequently, data coverage is robust in jurisdictions that mandate or will soon mandate aircraft to be equipped, such as Canada [101] and the USA [6], as well as Europe [66], but is considerably patchier in other locations. For instance, if an aircraft only operates between countries where ADS-B equipment is not mandatory, it is likely not equipped with the required technology and therefore is absent from the database.

The datasets we utilize were derived from OpenSky Network data by the authors of [149]. This data, available on Zenodo, was updated monthly and covers flight information from January 2019 to March 2022. It is important to note that the authors distribute two versions of the data: one under a Creative Commons license with some anonymized fields, and another covered by the OpenSky Network license, free to use for research purposes but subject to other limitations. We use the latter version in this study.

For this case study, we concentrated on flight data covering 2019, 2020 and 2021 corresponding to a full year prior to the COVID-19 pandemic, the transition from regular travel patterns to pandemic-level travel patterns and beginning of recovery of travel patterns, respectively.

For each month, the data consists in a csv file; see Table 6.4 for a sample. The file has the columns in Table 6.3.

The dataset has some limitations. The origin and destination are computed using ADS-B (Automatic Dependent SurveillanceâĂŞBroadcast) trajectories (see [31]) during approach and take

Variable	Meaning
callsign	Flight identifier
number*	Commercial number of the flight
icao24	Transponder unique identification number
registration*	Aircraft tail number
typecode*	Aircraft type
origin*	ICAO code for the origin airport
destination*	ICAO code for the destination airport
firstseen	UTC timestamp of the first message received by OSN
lastseen	UTC timestamp of the last message received by OSN
day	UTC day of the last message received by OSN
latitude_1	First detected position of the aircraft
longitude_1	First detected position of the aircraft
altitude_1	First detected position of the aircraft
latitude_2	Last detected position of the aircraft
longitude_2	Last detected position of the aircraft
altitude_2	Last detected position of the aircraft

Table 6.3: Variables in the data and their meaning. Starred variables, e.g., origin*, can be empty. OSN: OpenSky Network.

callsign	icao24	registration	typecode	origin	destination
HVN19	888152			YMML	LFPG
CES219	780b7e	B-5936	A332	YSSY	EDDF
TGW700	76bcca	9V-OFJ	B788		RJBB
CSN609	781364				KLAX
SVA840	710411			WMKK	WMKK
LAN600	e8027b	CC-BBG	B788	SKBO	KLAX
HVN55	8880f8	VN-A868	B789	YSSY	EGLL
AAR551	71bf94	HL7794	A333		LTBA
CPA343	789202	B-LRU	A359	YMML	EGKK
AAL126P	a999d2	N718AN	B77W	KLAX	KDFW
LAN706	e80450	CC-BGJ	B789	KJFK	LEMD
CCA985	780cb8	B-2487	B748		KSFO

Table 6.4: Sample rows in the dataset. Flight number (usually a very small variation on the callsign), location information (latitude, longitude and altitude) as well as date and time are omitted.

off. This leads to issues if a smaller airport is on the flight path of and close to a larger airport. For instance, many flights landing in CYYT (St. John's International Airport, Newfoundland) fly over CCV4 (Bell Island Airport), 15 kilometres west of the start of the main east-west runway in CYYT and are thus wrongly attributed to that airport in the dataset. Origin and destination are also empty when no airport can be found. Furthermore, no crosschecking with external sources of data has been conducted. The aircraft information comes from the OpenSky dataset, and the fields typecode and registration are empty when the aircraft is not present in the OpenSky dataset. Since not every flight has an aircraft type, we can only obtain lower and upper bounds for volume.

6.4.3 Data cleaning

The data, as provided by the authors of [149], is already cleaned to a large extent; therefore, the cleaning steps are quite limited.

1. Select rows in which both the origin and destination are non-empty. Cross-linking with external sources might be possible in some instances, given, for instance, the tail number or the callsign and flight number, but this is an entirely different project and was not undertaken here.

2. Exclude rows in which the origin and destination airports are identical. These often correspond to leisure personal flights, mostly in the USA. Such flights have no consequence for the global spread of infectious diseases and also have no impact on the overall dynamics of the network since they are not transport flights.

To illustrate the effect of these initial cleaning steps, let us consider the data from January 2019. The data initially had 2,660,901 rows. Of these, 1,341,646 rows were excluded in the cleaning steps because they had either an unknown origin or destination or had the same origin and destination. The data for the remaining months follows a similar pattern; see Table 6.5.

6.4.4 Data wrangling

Once the data has been cleaned, we submit it to processing steps.

1. We add country, continent and country region information for each flight for both the origin and destination, using the data in [152].

2. Using an aircraft capacity dataset [93], we add information about the flight capacities. This provides upper bounds for the number of passengers on each flight. For flights that do not have an entry for the aircraft type, we assign a volume of 2 passengers, the reasoning being that many aircrafts in the database are small personal planes in the USA.

Table 6.5: Data Cleaning Process for the ADS-B Air Transportation Network: Monthly Breakdown of Row Counts, Including Initial Data, Rows with Missing Values (NA), Loops, and Rows with Missing Aircraft Information.

Month	Initial Number of Rows	Number of Rows with NA and loops	Number of NA rows with aircraft
2019-06	2,660,901	1,341,646	251,250
2019-07	2,898,415	1,472,035	287,734
2019-08	2,990,061	1,441,288	287,881
2019-09	2,721,743	1,263,726	178,654
2019-10	2,946,779	1,348,042	271,106
2019-11	2,721,743	1,263,726	254,324
2019-12	2,946,779	1,348,042	271,106
2020-01	2,734,791	1,253,919	70,467
2020-02	2,648,835	1,133,469	249,670
2020-03	2,152,157	921,075	203,858
2020-04	842,905	349,828	77,161
2020-05	1,088,267	458,980	109,234
2020-06	1,444,224	622,932	156,686
2020-07	1,905,528	820,004	82,111
2020-08	2,042,040	872,631	81,664
2020-09	1,930,868	819,864	73,794
2020-10	1,985,145	851,915	223,703
2020-11	1,930,868	819,864	73,794
2020-12	1,985,145	851,915	223,703
2021-01	1,783,384	805,621	221,904
2021-02	1,617,845	715,874	203,027
2021-03	2,079,436	901,919	252,124
2021-04	2,227,362	958,959	269,123
2021-05	2,278,298	939,572	258,830
2021-06	2,540,487	1,044,863	295,776
2021-07	2,840,201	1,173,235	349,230
2021-08	2,794,400	1,169,944	363,986
2021-09	2,523,676	1,043,522	328,727
2021-10	2,726,252	1,135,317	364,723
2021-11	2,523,676	1,043,522	328,727
2021-12	2,726,252	1,135,317	364,723

3. Using the preprocessed data, we consider a subset of airports including airports in Canada, the United States of America and Europe[1]. Note that for USA territories outside of the territorial USA, only Puerto Rico is used; for European countries, only territories close to the continent are included.

6.4.5 Global network evolution

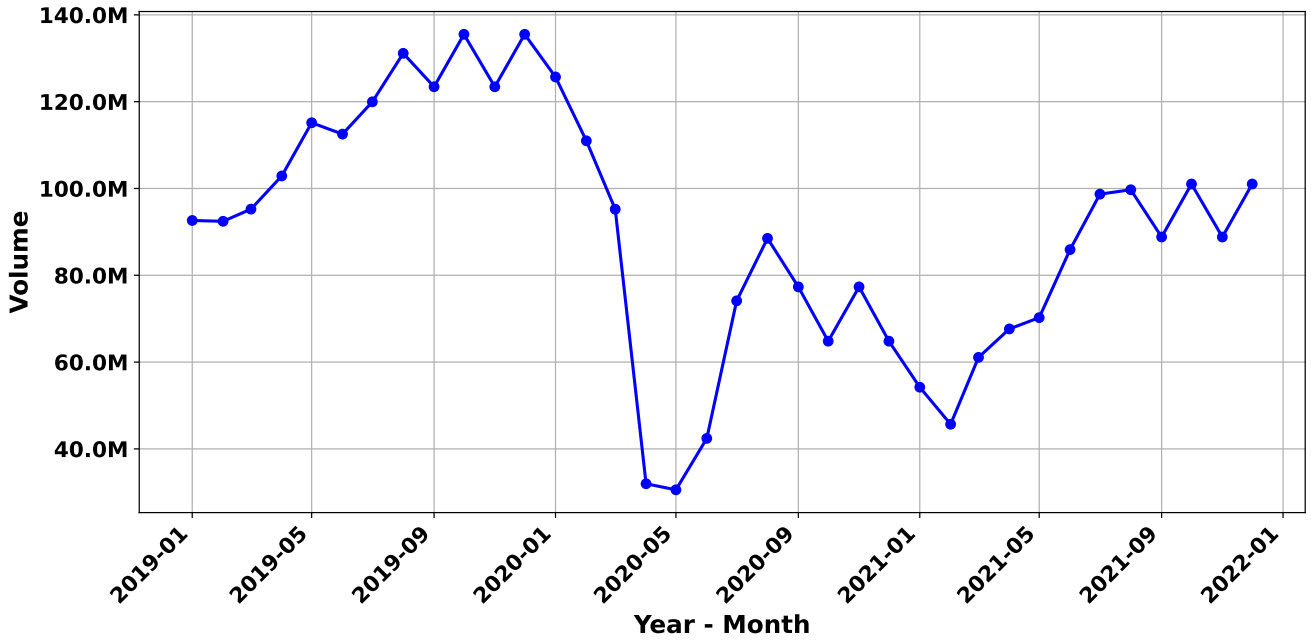


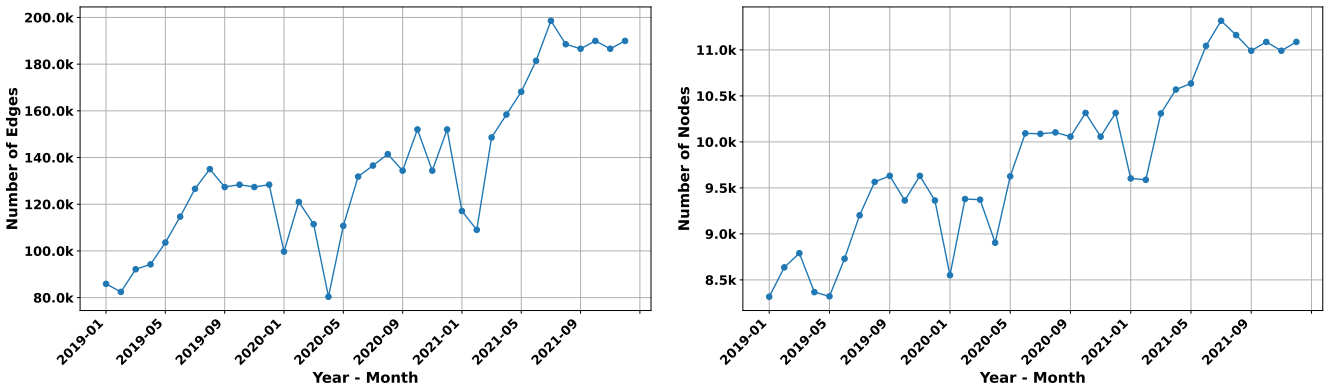
Figure 6.8: Total number of travellers in the ADS-B air transportation network during 2019-2021.

The aviation industry has experienced a significant decline, as depicted in Figure 6.8, which shows a dramatic decrease in the number of passengers in March 2020. This decline is attributed not only to reduced transportation needs due to the pandemic but also to the non-pharmaceutical interventions implemented by countries to curtail travel.

Additionally, Figure 6.9a shows a sharp decrease in the number of active airports in March 2020.

Note that the networks are not strongly connected, Figure 6.10 shows the number of strongly connected components for each month.

[1]Albania, Andorra, Armenia, Austria, Belarus, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Liechtenstein, Lithuania, Luxembourg, Malta, Moldova, Monaco, Montenegro, Netherlands, North Macedonia, Norway, Poland, Portugal, Romania, Russia, San Marino, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, TÃijrkiye, Ukraine, United Kingdom



(a) Evolution of the number of edges.

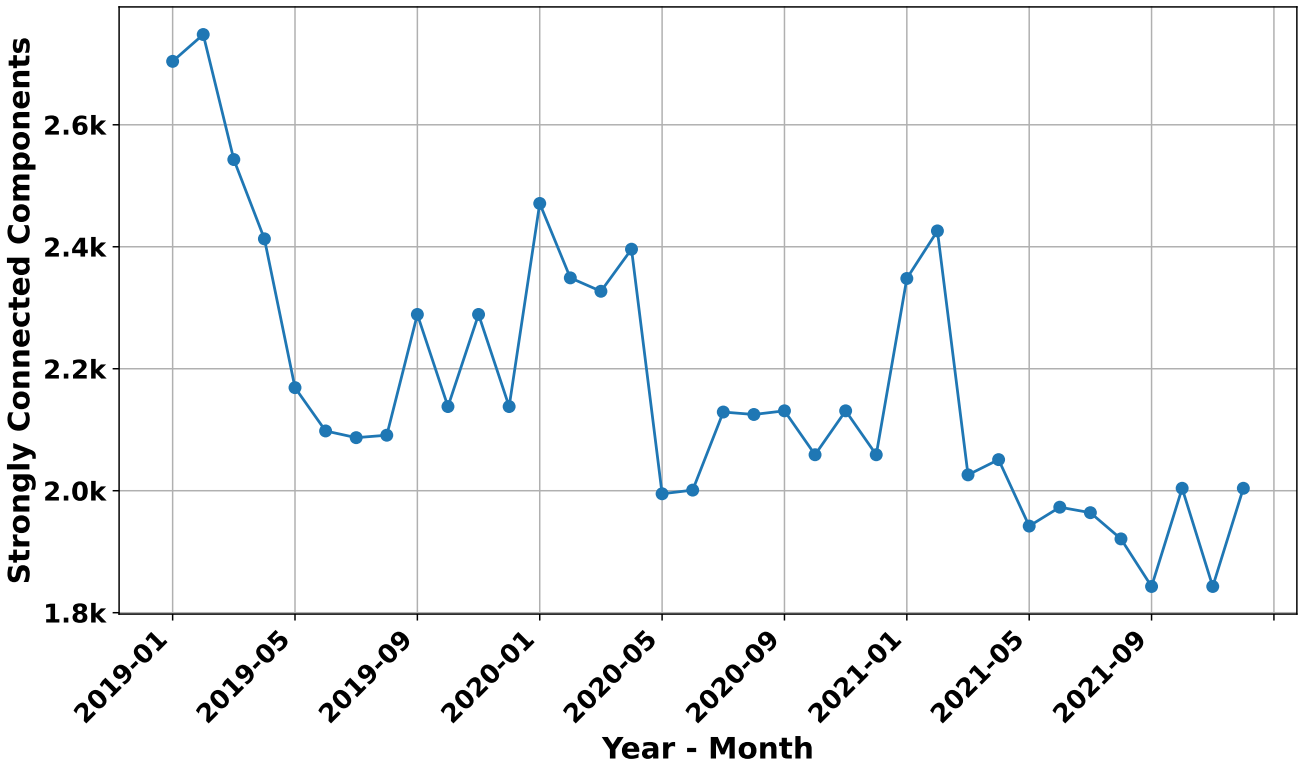
(b) Evolution of the number of nodes.

Figure 6.9: Temporal analysis of network structures in the ADS-B air transportation network. The top panel illustrates the changes in the number of edges, indicating the connectivity between nodes, while the bottom panel shows the changes in the number of nodes, representing the entities within the networks.

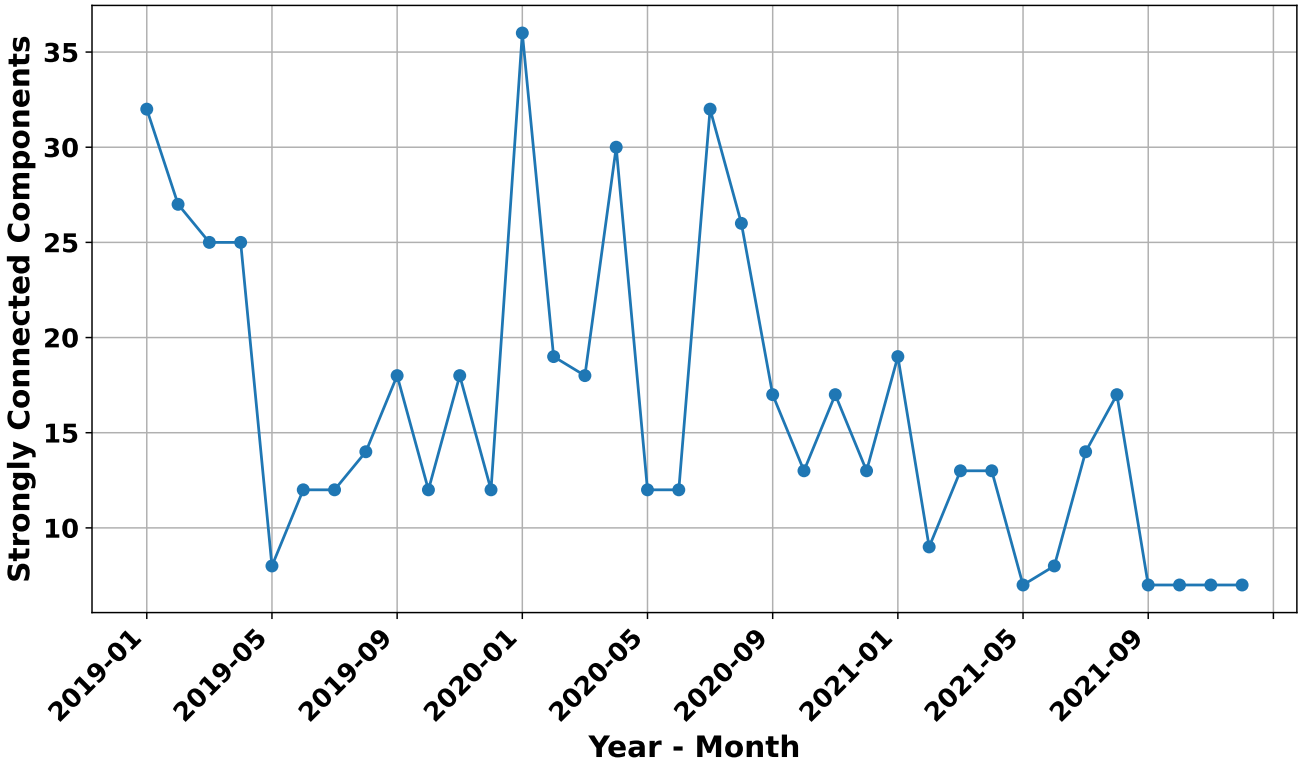
In Figure 6.11, we observe that based on the Louvain algorithm, the Leiden algorithm and the Infomap algorithm, the number of communities changes over time; the size of the maximum community also changes. From an epidemiological perspective, communities represent semi-isolated subnetworks within which disease can spread efficiently, but between which transmission may be more limited. Understanding community structure can inform targeted interventions: travel restrictions between communities may be more effective than restrictions within communities. The temporal changes in community structure observed in both case studies (Figures 6.5 and 6.11) suggest periods when the network became more or less fragmented, with direct implications for containment strategies.

In Figure 6.12a, we observe that Chicago O'Hare International Airport (KORD) consistently exhibits the highest betweenness centrality among United States airports from 2019 to 2021. This reflects its role as a major national and international hub. However, exceptions to this trend occur in specific months, for example, in July 2019 and July 2021, Wittman Regional Airport (KOSH) in Wisconsin displays the highest centrality, and in May 2021, Teterboro Airport (KTEB) in New Jersey takes the lead.

These anomalies illustrate an important point: airports with relatively small passenger volumes can still exhibit high betweenness centrality if they serve as critical connectors within specific regional transportation networks. For instance, Thomson Airport in northern Manitoba (not shown in this figure) has high centrality because it acts as a key hub for travel to and from remote northern communities. Similarly, KOSH's centrality spikes in July due to its role as the host of EAA AirVenture Oshkosh, the world's largest annual aviation gathering. Such examples highlight how centrality is not solely a function of size or traffic volume, but also of an airport's structural

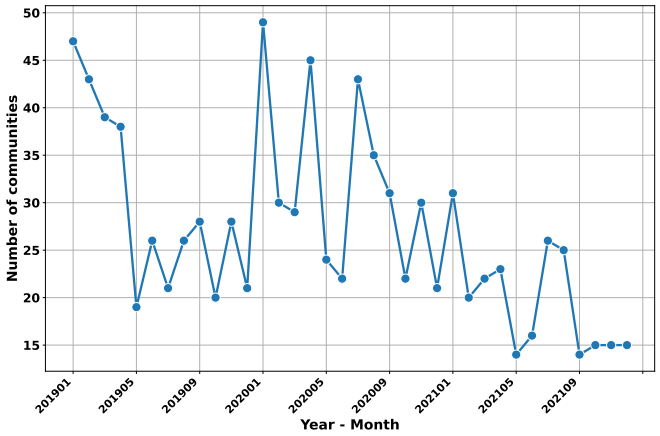


(a) Directed networks.

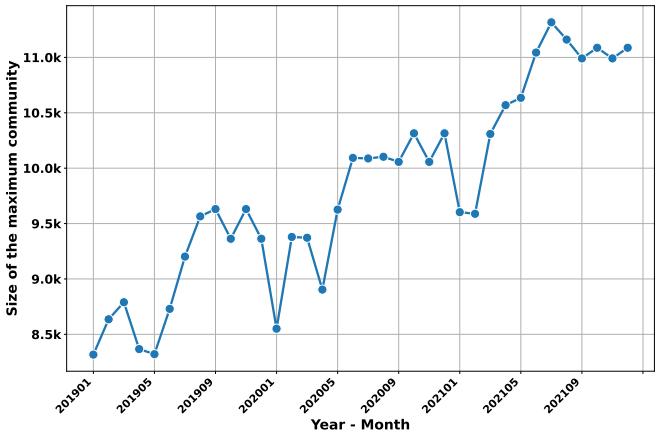


(b) Undirected networks.

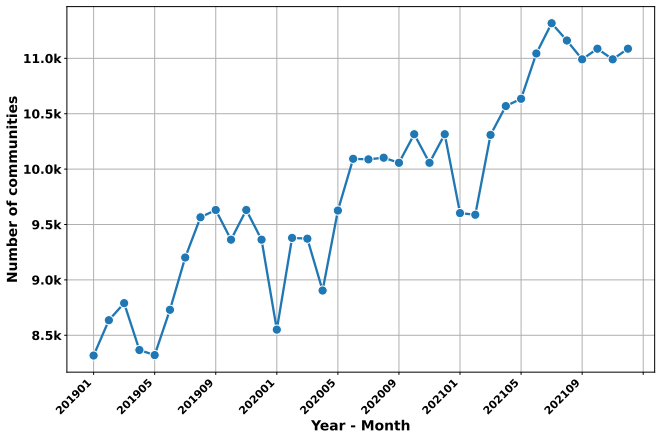
Figure 6.10: Monthly count of strongly connected components in directed and undirected ADS-B air transportation networks.



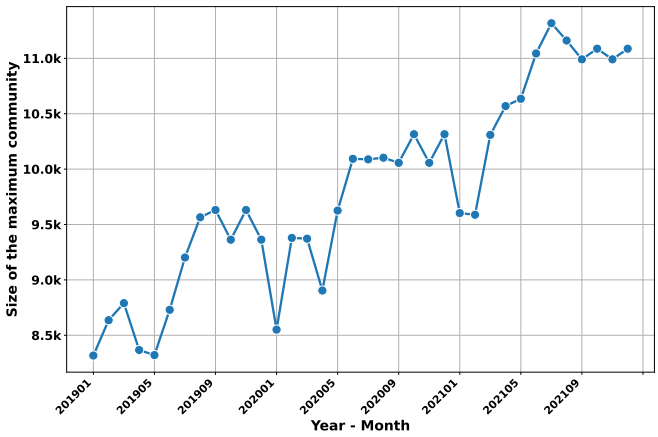
(a) Number of communities (Louvain).



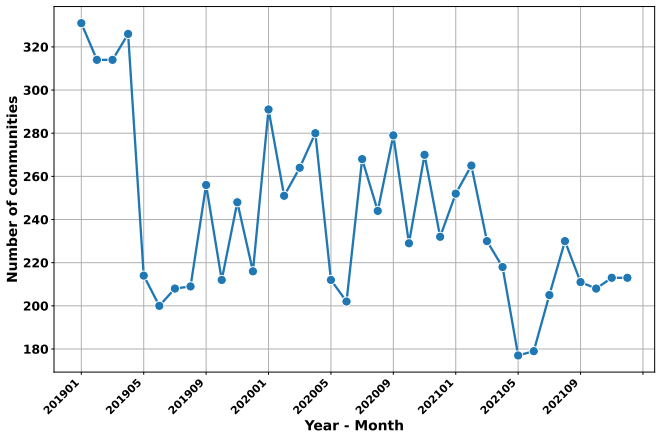
(b) Size of the maximum community (Louvain).



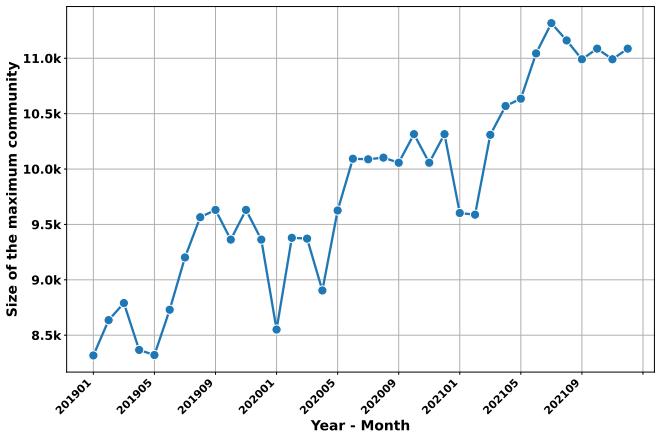
(c) Number of communities (Leiden).



(d) Size of the maximum community (Leiden).



(e) Number of communities (Infomap).



(f) Size of the maximum community (Infomap).

Figure 6.11: Evolution of the number of communities and size of the maximum community in the ADS-B air transportation network data obtained using different community detection algorithms (Louvain, Leiden and Infomap).

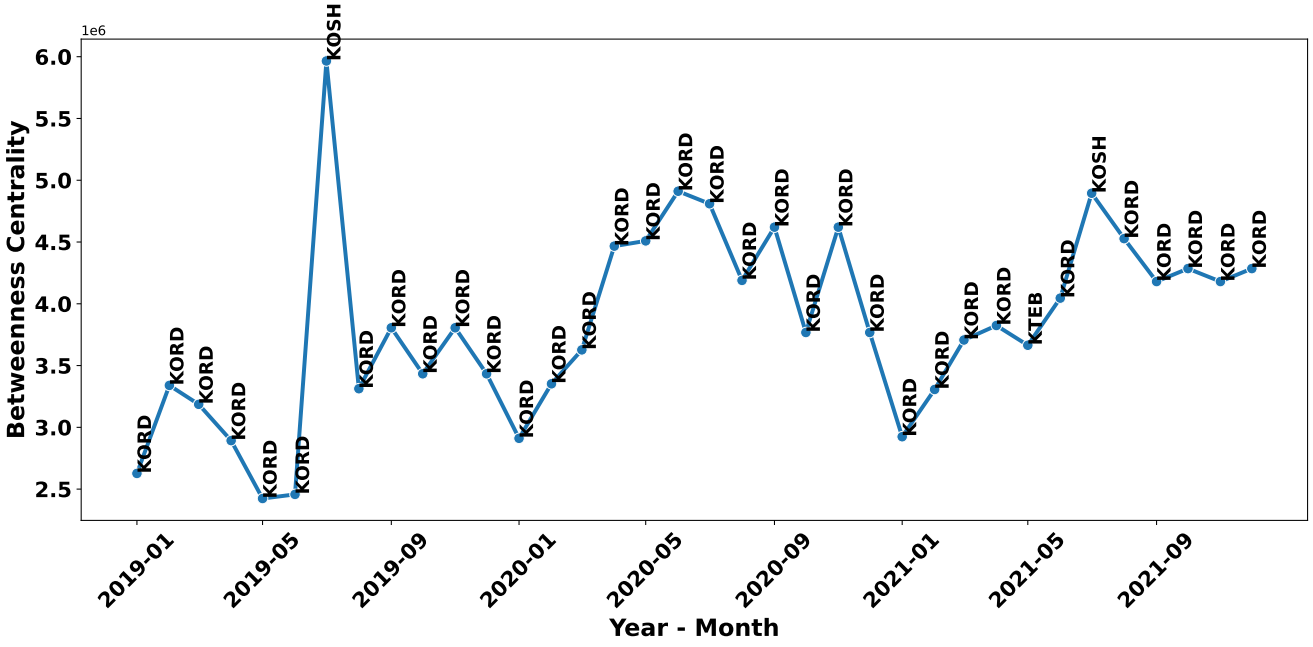
importance within the broader network.

Note that centrality metric in this section are based on the global airline flight networks, and they represent the potential for flow through the networks and although airports with high betweeness centrality are important transit points for passengers, in reality this may not be the case.

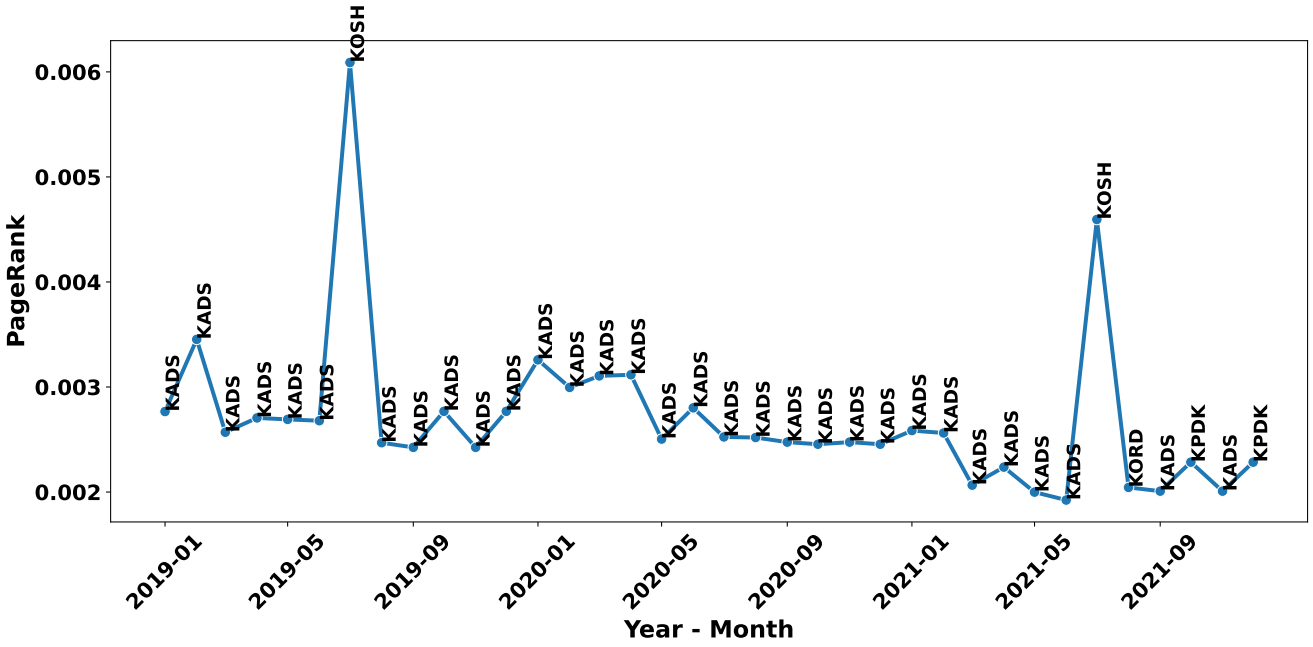
Public health interpretation: Betweenness centrality measures the extent to which a node (e.g., an airport or city) lies on the shortest paths between other nodes. In epidemiological terms, locations with high betweenness centrality act as critical bridges for disease transmission. An infected individual traveling through a high-betweenness node has a higher probability of seeding infections in multiple downstream locations. For example, Figure 6.12a shows that Chicago O'Hare International Airport (KORD) consistently exhibited the highest betweenness centrality in our air transportation network, suggesting its critical role as a potential disease dissemination hub. Changes in betweenness centrality over time (such as the dramatic shifts observed during the COVID-19 pandemic) directly translate to changes in importation risk profiles.

Figure 6.12b illustrates the maximum PageRank centrality value in the ADS-B air transportation network. The PageRank values fluctuate over time, indicating shifts in the relative importance of various locations within the network. This information is crucial for network planning and optimization, as it helps identify critical hubs and potential areas for improvement. Addison Airport (KADS) in Texas consistently demonstrates the highest PageRank centrality, with the exceptions of July 2019 and July 2021 when Wittman Regional Airport (KOSH) in Wisconsin holds the highest centrality, August 2021 when Chicago O'Hare International Airport (KORD) exhibits the highest centrality, and October and December 2021 when Portland International Airport (KPDX) in Oregon shows the highest centrality.

Public health interpretation: PageRank centrality, originally developed for ranking web pages, measures a node's importance based on both the number and quality of its connections. In disease transmission networks, high PageRank indicates locations that serve as epidemiological amplifiers, nodes that are not only well-connected but connected to other well-connected locations, creating pathways for rapid outbreak propagation. During pandemic emergence, PageRank captures what passenger volume alone cannot: the structural role a location plays in shaping transmission chains. A moderately busy airport connected to major international hubs can seeded outbreaks across multiple secondary hubs simultaneously, effectively shortening the time to global dispersal. This cascading effect is particularly critical for diseases with significant asymptomatic transmission, where infected travelers may transit through hub airports undetected. These network bottlenecks become strategic priorities for surveillance and intervention, as disrupting transmission at high-PageRank nodes can disproportionately reduce outbreak spread compared to lower-ranked locations with similar traffic volume.



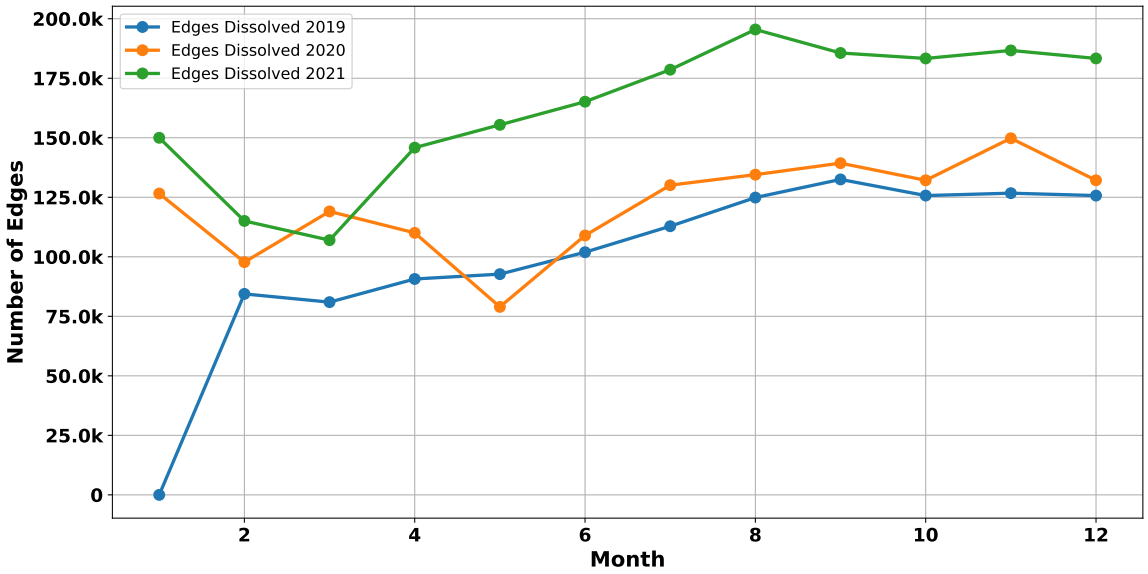
(a) Maximum monthly betweenness centrality.



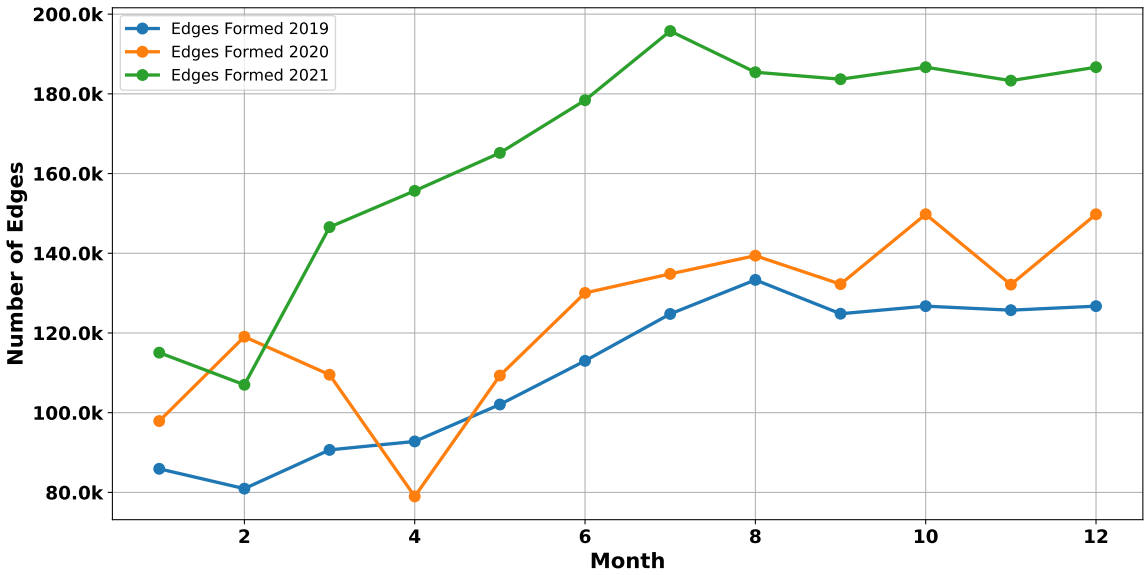
(b) Maximum monthly PageRank centrality.

Figure 6.12: Monthly maximum centrality measures for air transportation networks. Subfigure (a) shows the maximum betweenness centrality across graphs, while subfigure (b) displays the maximum PageRank centrality. Airport names above each data point indicate the airports corresponding to the maximum centrality values in each graph, representing monthly snapshots of network properties from the dataset.

Figure 6.13b shows the the edge formation for 2019, 2020 and 2021, from this we can see that April 2020 has a significant drop of edge formations compared to April 2019 and April 2021. The function used evaluates network objects at multiple time points, providing counts of edge onsets to describe momentary rates of change in the network. Figure 6.13a shows the edge disolution for our networks. Similarly, we observe a notable decrease in edge dissolutions in May 2020 compared to May 2019 and May 2021. The function evaluates network objects similarly, returning counts of edge dissolutions. The ability to track edge formation and dissolution over time (Figures 6.6 and 6.13) captures the creation of new transmission pathways and the elimination of existing ones. During the COVID-19 pandemic, we observed dramatic edge dissolution in April 2020, representing route cancellations and travel restrictions. From an epidemiological standpoint, this quantifies the impact of network-based interventions on transmission potential.



(a) Edges Dissolved in the Network.



(b) Edges Formed in the Network.

Figure 6.13: Monthly dynamics of the ADS-B air transportation network. (a) The number of edges dissolution each month from 2019 to 2021. (b) The number of edge formations each month over the same period. These figures illustrate the fluctuating nature of connections within the network, emphasizing periods of high connectivity changes.

7

Conclusion

This thesis commenced with a primary goal of addressing methodological challenges posed by the scarcity of epidemic data, initially focusing on historical outbreaks like SARS and Ebola. The unexpected arrival of the COVID-19 pandemic required a quick change in research direction, leading to an in-depth exploration of the complex dynamics of infectious diseases. This shift emphasized the development of adaptive modelling strategies and the integration of real-time data constraints into our analytical frameworks.

Reflecting its interdisciplinary nature, this thesis draws upon mathematical epidemiology, network science, data science, and artificial intelligence techniques. For example, AI-driven natural language processing was leveraged to systematically review the rapidly expanding COVID-19 literature, while novel Python tools were developed to analyse temporal network dynamics relevant to disease spread and intervention impact.

In Chapter 4, our initial approach involved utilizing time-dependent parameters to accurately model the evolving dynamics of COVID-19 transmission in Alberta. As my research progressed toward scenario analyses evaluating various intervention strategies, we recognized the need for a more streamlined and interpretable framework. Transitioning from time-dependent to constant parameters facilitated clearer comparisons across scenarios and improved the clarity of our findings.

To simulate a range of social distancing and testing strategies, we introduced multipliers applied to the constant baseline values of the transmission rate, β and the detection rate, ρ. Specifically, we considered increases of 10%, 20%, 30%, and 40% as well as reductions of 10% and 20% relative to baseline (i.e., multiplier of 1.0). This systematic perturbation allowed us to represent plausible behavioural and policy changes.

A baseline scenario with $\beta = 1$ and $\rho = 1$ was used to reflect the status quo in Alberta and served as the reference point for evaluating the relative impact of each intervention scenario. The effects of varying these multipliers were assessed across key epidemiological outcomes, including cumulative and peak infections and cases.

To quantify intervention impacts, we calculated percentage changes in each outcome relative to the baseline. These results were then applied back to the time-dependent fitted model to estimate what the trajectory might have looked like under alternative policy choices. This allowed us to translate scenario comparisons into actionable insights crucial for informing evidence-based public health decisions

Looking forward, future versions of the model could incorporate additional epidemiologically relevant factors, most notably, vaccination. A natural way to include vaccination within the compartmental framework is by introducing a new compartment, V, representing vaccinated individuals. Two additional parameters would be required: one for the vaccination rate, and one to accounting for breakthrough infections. This latter parameter allows us to model reduced susceptibility in the vaccinated population. Incorporating vaccination in this way would influence several key model parameters, including the effective reproduction number and outbreak thresholds. It would also enable scenario-based analysis of vaccine rollout strategies, waning immunity, and the role of breakthrough infections, all of which are crucial for understanding ongoing and future public health risks. While models of this kind have appeared in the literature, our framework could contribute by embedding these features within a scenario-driven, decision-support context specifically calibrated for regional public health planning.

This approach promises a more comprehensive understanding of pandemic dynamics, strengthening the foundation for evidence-based policy recommendations.

Chapter 5 explores how two infectious disease variants interact using mathematical models. The goal is to understand when both variants coexist or when one dominates.

Traditional methods rely on an infinite timeframe to predict dominance based on a variant's basic reproduction number, $\mathcal{R}_0$. However, this doesn't reflect real-world outbreaks with limited durations.

This chapter introduces "practical coexistence", considering dynamics within a single wave. It shows that the number of initially infected individuals, along with $\mathcal{R}_0$, plays a role in dominance or coexistence. It also proposes using percentage contribution, rather than just total infections, to assess coexistence in shorter timeframes.

The chapter analyzes the Alpha and Gamma variants during the third COVID-19 wave in Alberta and British Columbia. While both provinces exhibited coexistence, the percentage contribution of each variant differed. This suggests that despite potentially similar $\mathcal{R}_0$ values, Alberta's early containment measures might have limited the spread of the Gamma variant, leading to a lower contribution compared to British Columbia. This highlights the potential interaction between initial outbreak size, containment strategies, and variant dynamics within a finite timeframe.

These findings offer valuable insights for understanding future outbreaks with multiple variants emerging simultaneously. Early intervention targeting high-risk variants can influence their spread

and promote coexistence with less severe variants. The limitations of the model, such as not fully capturing factors like vaccine coverage, are acknowledged. However, the chapter's insights can inform future COVID-19 management strategies as Canada transitions to an endemic phase, where new variants are expected.

Chapter 6 introduced a novel Python package for temporal network analysis. This package, built upon the principles of graph theory, empowers researchers to analyse and visualise dynamic network properties, encompassing comprehensive metric computation, centrality analysis, community detection, and visualisation of evolving network structures.

The application of this methodological approach to two distinct networks âĂŞ the air transportation network spanning North America and Europe, and the Helsinki city bike network. The analysis revealed temporal variations in network structure, connectivity, and community dynamics. Algorithms like Louvain, Leiden, and Infomap employed within the package provided a deeper understanding of network resilience and functionality.

This chapter makes significant contributions on two fronts. Firstly, it demonstrates the effectiveness of accessible data sources and novel methodological tools in studying transportation networks. Secondly, by revealing the adaptive nature of these networks and the importance of key nodes for resilience, the research offers valuable insights for informing critical decisions in urban planning, public health strategies, and crisis management during disease outbreaks. This improved understanding of transportation networks as complex adaptive systems has the potential to significantly enhance preparedness and response efforts in the face of future public health challenges.

Lessons Learned and Future Directions .

- Interdisciplinary Collaboration related to my work: The COVID-19 pandemic underscored the indispensable role of interdisciplinary collaboration among researchers, public health authorities, and policymakers. These collaborative efforts were instrumental in integrating diverse datasets, expertise, and perspectives into comprehensive modelling frameworks.

- Adaptability: Models must exhibit flexibility to accommodate evolving data and contextual dynamics, reflecting the dynamic nature of disease outbreaks and the varying efficacy of control measures. This adaptability proved crucial for timely interventions and decision-making.

- Public Health Preparedness: Effective pandemic preparedness necessitates robust plans encompassing both pharmaceutical and non-pharmaceutical interventions.

Therefore, planning for the next pandemic should remain a top priority for governments and

public health authorities, and the roles of both non-pharmaceutical and pharmaceutical interventions should be carefully evaluated and incorporated into preparedness plans.

In conclusion, this thesis journey has provided insights into the complex dynamics of infectious diseases while identifying ongoing challenges and emerging opportunities in pandemic response strategies. By documenting these experiences and lessons learned, this work aims to significantly contribute to global efforts in pandemic preparedness and response, ensuring that health systems worldwide are better equipped to navigate the uncertainties posed by future infectious disease threats.

Index

References

[1] Coronavirus (COVID-19) SARS-CoV-2. https://ipac-canada.org/coronavirus-resources#:~:text=Pandemic%20Coronavirus%20(COVID%2D19),11%20million%20in%20central%20China. Accessed on July 28, 2023.

[2] Crimean congo hemorrhagic fever virus spreading in europe due to climate change. https://www.forbes.com/sites/brucelee/2023/07/08/crimean-congo-hemorrhagic-fever-cchf-virus-spreading-in-europe-due-to-climate-change/?sh=6d21289c6a69. Accessed on July 25, 2023.

[3] Swaziland a culture that encourages hiv/aids. https://www.refworld.org/docid/49e6ef2dc.html. Accessed on July 28, 2023.

[4] Updated preparedness and response framework for influenza pandemics. https://www.cdc.gov/flu/pandemic-resources/pdf/mmwr-rr6306.pdf. Accessed on July 25, 2023.

[5] Helsinki city bikes dataset. https://www.kaggle.com/datasets/geometrein/helsinki-city-bikes?resource=download, Accessed: July 8 2024. Kaggle Dataset.

[6] ADMINCOPA . U.s. ads-b mandate. https://copanational.org/u-s-ads-b-mandate/, 2019. Accessed: 2022-07-18.

[7] A.E. U and E.E. J. Modelling COVID-19 pandemic in nigeria using multivariate autoregressive distributed lag-moving average models. *African Journal of Mathematics and Statistics Studies*, 2021.

[8] Ahmad W, et al. Fractional order mathematical modeling of novel corona virus (covidâĂŘ19). *Mathematical Methods in the Applied Sciences*, 2021.

[9] Al-Ani B. Statistical modeling of the novel COVID-19 epidemic in iraq. 2021.

[10] Al-Shammari A. A, et al. COVID-19 transmission and forecasting in kuwait: A mathematical modeling study. *SSRN Electronic Journal*, 2020.

[11] Al-Tuwairqi S. M and Al-Harbi S. K. A time-delayed model for the spread of COVID-19 with vaccination. *Scientific Reports*, 12, 2022.

[12] Aldila D, et al. Impact of early detection and vaccination strategy in COVID-19 eradication program in jakarta, indonesia. *BMC Research Notes*, 14, 2021.

[13] Almagrabi A, et al. A new approach to q-linear diophantine fuzzy emergency decision support system for covid19. *Journal of Ambient Intelligence and Humanized Computing*, 13:1687 – 1713, 2021.

[14] Almetwally E, et al. A new inverted topp-leone distribution: Applications to the COVID-19 mortality rate in two different countries. *Axioms*, 10:25, 2021.

[15] Anastassopoulou C, et al. Data-based analysis, modelling and forecasting of the COVID-19 outbreak. *PLoS One*, 15(3):e0230405, Mar 2020.

[16] Andreasen V, et al. The dynamics of cocirculating influenza strains conferring partial cross-immunity. *J Math Biol*, 35:825–842, 1997.

[17] Androutsopoulos I, et al. Honey pot: A collaborative e-mail spam detection system. In *ACM Transactions on Information Systems (TOIS)*, volume 22, pages 144–178, 2004.

[18] Anstett-Collin F, et al. Sensitivity analysis of complex models: Coping with dynamic and static inputs. *Reliability Engineering & System Safety*, 134:268–275, 2015.

[19] Aphale P, et al. Doubling time and its interpretation for COVID-19 cases: A comparative study in pimpri chinchwad municipal corporation, pune, maharashtra, india. *Journal of Pharmaceutical Research International*, 2021.

[20] Ardila E. K. G, et al. Mathematical model and COVID-19. *Colombia MÃdica : CM*, 51, 2020.

[21] Arino J, et al. A final size relation for epidemic models. *Mathematical Biosciences and Engineering*, 4(2):159–175, 2007.

[22] Aristov V, et al. Application of the kinetic type model for study of a spatial spread of COVID-19. *Computer Research and Modeling*, 2021.

[23] Aschengrau A and Seage G. R. *Essentials of Epidemiology in Public Health*. Jones & Bartlett Learning, fourth edition, 2018.

[24] Ashtiani M, et al. CINNA: an R/CRAN package to decipher Central Informative Nodes in Network Analysis. *Bioinformatics*, 35(8):1436–1437, 09 2018.

[25] Atangana A and ÄřÄ§ret Araz S. Modeling third waves of COVID-19 spread with piecewise differential and integral operators: Turkey, spain and czechia. *Results in Physics*, 29:104694, 2021.

[26] Auger P and Moussaoui A. On the threshold of release of confinement in an epidemic seir model taking into account the protective effect of mask. *Bulletin of Mathematical Biology*, 83(4):25, 2021.

[27] Ayoub H, et al. Epidemiological impact of prioritising SARS-CoV-2 vaccination by antibody status: mathematical modelling analyses. *BMJ Innovations*, 7:327 – 336, 2021.

[28] Bacaër N. *Daniel Bernoulli, d'Alembert and the inoculation of smallpox (1760)*, pages 21–30. Springer London, London, 2011.

[29] Badmus N, et al. Parametric modeling approach to Covid-19 pandemic data. *Open Journal of Statistics*, 2021.

[30] Badr H. S, et al. Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study. *The Lancet Infectious Diseases*, 20(11):1247–1254, Nov 2020.

[31] Baek K.-Y and Bang H.-C. Ads-b based trajectory prediction and conflict detection for air traffic management. *International Journal of Aeronautical and Space Sciences*, 13(3):377–385, 09 2012.

[32] Bahdanau D, et al. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[33] Balayla J, et al. Prevalence threshold and temporal interpretation of screening tests: The example of the SARS-CoV-2 (COVID-19) pandemic. *medRxiv*, 2020.

[34] Basnarkov L. Epidemic spreading model of COVID-19. *arXiv: Physics and Society*, 2020.

[35] BBC News . Covid: Woman aged 90 died with double variant infection. `https://www.bbc.com/news/health-57761343`, 2022. Accessed: 2022-06-10.

[36] Betti M, et al. Could the new COVID-19 mutant strain undermine vaccination efforts? a mathematical modelling approach for estimating the spread of the uk mutant strain using ontario, canada, as a case study. In *medRxiv*, 2021.

[37] Blondel V. D, et al. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, oct 2008.

[38] Bolze A, et al. Evidence for SARS-CoV-2 Delta and Omicron co-infections and recombination. *medRxiv*, 2022.

[39] Bosetti P, et al. Impact of mass testing during an epidemic rebound of SARS-CoV-2: a modelling study using the example of france. *Eurosurveillance*, 26, 2021.

[40] Bosse N, et al. Transformation of forecasts for evaluating predictive performance in an epidemiological context. 2023.

[41] Brauer F. *Mathematical Models in Population Biology and Epidemiology*. Texts in Applied Mathematics, 40. Springer New York, New York, NY, 2nd ed. edition, 2012.

[42] Brauer F. *Mathematical Models in Population Biology and Epidemiology*. Texts in Applied Mathematics, 40. Springer New York, New York, NY, 2nd ed. 2012. edition, 2012.

[43] Brauer F. Mathematical epidemiology: Past, present, and future. *Infectious Disease Modelling*, 2(2):113–127, Feb 2017.

[44] Brauer F, et al., editors. *Mathematical Epidemiology*. Springer Berlin Heidelberg, 2008.

[45] Bremermann H. J and Thieme H. A competitive exclusion principle for pathogen virulence. *Journal of Mathematical Biology*, 27:179–190, 1989.

[46] Brugnano L, et al. The hidden side of COVID-19 spread in italy. *arXiv: Populations and Evolution*, 2020.

[47] Burke D. S. Origins of the problematic e in SEIR epidemic models. *Infectious Disease Modelling*, 9(3):673–679, 2024.

[48] Callaway E. The coronavirus is mutating - does it matter? *Nature*, 585:174–177, 2020.

[49] Chang E, et al. Covidsimvl – transmission trees, superspreaders and contact tracing in agent based models of COVID-19. 2020.

[50] Chatterjee S, et al. Studying the progress of COVID-19 outbreak in india using sird model. *Indian J Phys Proc Indian Assoc Cultiv Sci (2004)*, 95(9):1941–1957, 2021.

[51] Chen M.-H. *Monte Carlo methods in Bayesian computation*. Springer series in statistics. Springer, New York, 2000.

[52] Chen Z, et al. A global analysis of replacement of genetic variants of SARS-CoV-2 in association with containment capacity and changes in disease severity. *Clinical Microbiology and Infection*, 27(5):750–757, 2021.

[53] Ciupeanu A.-S, et al. Mathematical modeling of the dynamics of COVID-19 variants of concern: Asymptotic and finite-time perspectives. *Infectious Disease Modelling*, 7(4):581–596, 2022.

[54] Cohen J. A, et al. Mechanistic modeling of SARS-CoV-2 immune memory, variants, and vaccines. In *medRxiv*, 2021.

[55] Collobert R and et al. . Natural language processing, automatic speech recognition, and machine translation. *Speech and Language Processing*, pages 895–949, 2011.

[56] Cooper I, et al. A SIR model assumption for the spread of COVID-19 in different communities. *Chaos, Solitons & Fractals*, 139:110057, Oct 2020.

[57] Croccolo F. Spreading of infections on random graphs: A percolation-type model for COVID-19. *Chaos, Solitons, and Fractals*, 139:110077, 2020.

[58] CTV News . 'Significant' COVID-19 variant outbreak in alberta announced by Hinshaw Saturday, by Adam Lachacz. `https://edmonton.ctvnews.ca/significant-covid-19-variant-outbreak-in-alberta-announced-by-hinshaw-saturday-1.5373506?cache=ihcaobeag%3FclipId%3D375756`. April 3, 2021.

[59] Cui S, et al. Hierarchical gaussian processes and mixtures of experts to model COVID-19 patient trajectories. *bioRxiv*, 2021.

[60] Cunniffe N, et al. Observability, identifiability and epidemiology – a survey, 2023.

[61] da Silva Francisco Jr R, et al. Pervasive transmission of E484K and emergence of VUI-NP13L with evidence of SARS-CoV-2 co-infection events by two different lineages in Rio Grande do Sul, Brazil. *Virus Research*, 296:198345, 2021.

[62] Dehning J, et al. Model-based and model-free characterization of epidemic outbreaks. *medRxiv*, 2020.

[63] Dela A, et al. Multi-method global sensitivity analysis of mathematical models. *Journal of Theoretical Biology*, 546(111159), 2022.

[64] Devlin J, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018.

[65] Diestel R. *Graph theory*. Graduate Texts in Mathematics Series ; Volume 173. Springer-Verlag, Berlin, Germany, sixth edition. edition, 2017.

[66] EASA . Amendment to the airspace requirements on ads-b and mode s. https://www.easa.europa.eu/newsroom-and-events/news/amendment-airspace-requirements-ads-b-and-mode-s, 2022. Accessed: 2022-07-18.

[67] Escamilla J, et al. Prediction of cases of infection and deaths caused by COVID-19 in mexico through the construction of probabilistic models under health conditions in 2020. pages 9–21, 2021.

[68] Evans T. S. Clique graphs and overlapping communities. *Journal of Statistical Mechanics: Theory and Experiment*, 2010(12):P12037, dec 2010.

[69] Fath B. D and JÃÿrgensen S. E. *Encyclopedia of Ecology*. Elsevier, Amsterdam, 2018.

[70] Foncea P, et al. Replacing quarantine of COVID-19 contacts with periodic testing is also effective in mitigating the risk of transmission. *Scientific Reports*, 12, 2021.

[71] Fonseca V, et al. Growth functions as a tool to model SARS-CoV-2 pandemic trajectory and related-deaths worldwide. In *medRxiv*, 2021.

[72] Fox S. J, et al. Optimizing the number of models included in outbreak forecasting ensembles. In *medRxiv*, 2024.

[73] Fraser C, et al. Factors that make an infectious disease outbreak controllable. *PNAS*, 101(16):6146–6151, 2004.

[74] Furstova J, et al. Towards bayesian evaluation of seroprevalence studies. 2021.

[75] Gallardo D, et al. Parametric quantile regression models for fitting double bounded response with application to COVID-19 mortality rate data. *Mathematics*, 2021.

[76] Gelman A, et al. *Bayesian Data Analysis*. CRC Press, 3rd edition, 2013.

[77] Ghaffari A and Saadati R. $*$-fuzzy measure model for COVID-19 disease. *Advances in Difference Equations*, 2021, 2021.

[78] Ghosh S and Ghosh S. A mathematical model for COVID-19 considering waning immunity, vaccination and control measures. *Sci Rep*, 13(1):3610, 2023.

[79] Ghostine R, et al. An extended seir model with vaccination for forecasting the COVID-19 pandemic in saudi arabia using an ensemble kalman filter. volume 9, page 636, 2021.

[80] Giordano G, et al. Modeling vaccination rollouts, SARS-CoV-2 variants and the requirement for non-pharmaceutical interventions in italy. *Nature Medicine*, 27:993 – 998, 2021.

[81] Goodman J and Weare J. Ensemble samplers with affine invariance. *Communications in applied mathematics and computational science*, 5(1):65–80, 2010.

[82] Griette Q, et al. A robust phenomenological approach to investigate covid-19 data for france. *Mathematics in Applied Sciences and Engineering*, 2021.

[83] Hackenberger B. K. Bayes or not bayes, is this the question? *Croatian Medical Journal*, 60(1):50+, 2023/3/8/ 2019.

[84] Hellewell J, et al. Feasibility of controlling covid-19 outbreaks by isolation of cases and contacts. *The Lancet Global Health*, 8(4):e488–e496, 2020.

[85] Herrera M. Exploring the roles of local mobility patterns, socioeconomic conditions, and lockdown policies in shaping the patterns of COVID-19 spread. *ArXiv*, abs/2103.02701, 2021.

[86] Hethcote H. W. Qualitative analyses of communicable disease models. *Mathematical biosciences*, 28(3):335–356, 1976.

[87] Hinshaw D. `https://twitter.com/CMOH_Alberta/status/1378421652556455937?ref_src=twsrc%5Etfw%7Ctwcamp%5Etweetembed%7Ctwterm%5E1378421652556455937%7Ctwgr%5E%7Ctwcon%5Es1_&ref_url=https%3A%2F%2Fedmonton.ctvnews.ca%2Fsignificant-covid-19-variant-outbreak-in-alberta-announced-by-hinshaw-saturday-1.5373506%3Fcache%3Dihcaobeag3FclipId3D375756`. April 3, 2021.

[88] Hogan C, et al. Rapid increase in SARS-CoV-2 P.1 lineage leading to codominance with B.1.1.7 lineage, British Columbia, Canada, January-April 2021. *Emerging Infectious Diseases*, 27(11):2802–2809, 2021.

[89] Høiby N. Pandemics: past, present, future. *APMIS*, 129(7):352–371, 2021.

[90] Hong H. G and Li Y. Estimation of time-varying reproduction numbers underlying epidemiological processes: A new statistical tool for the COVID-19 pandemic. *PLoS One*, 15(7):e0236464, 2020.

[91] Huo X, et al. Estimating asymptomatic, undetected and total cases for the COVID-19 outbreak in wuhan: a mathematical modeling study. *BMC Infectious Diseases*, 21(1):476, 2021.

[92] Ibrahim A, et al. Modeling the effect of population density on controlling Covid-19 initial spread with the use of matlab numerical methods and stringency index model. *2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, pages 612–617, 2020.

[93] ICAO . Data sets - aircrafts. `http://www.lsv.fr/~sirangel/teaching/dataset/`, 2022. Accessed: 2022-06-10.

[94] Ivorra B, et al. Mathematical modeling of the spread of the coronavirus disease 2019 (COVID-19) taking into account the undetected infections. the case of china. *Communications in Nonlinear Science and Numerical Simulation*, 88:105303, Sep 2020.

[95] Jena R. M, et al. Time-fractional order mathematical modeling of COVID-19 with vaccination using non-singular kernel functions. *Pramana*, 101, 2023.

[96] JosephI K, et al. Comparative model profiles of Covid-19 occurrence in nigeria. *International Journal of Mathematics Trends and Technology*, 68:297–310, 2020.

[97] Julien Arino,StÃľphanie Portet, Nicolas Bajeux, Adriana-Stefania Ciupeanu . Investigation of global and local covid-19 importation risks, 2020.

[98] Jurafsky D and Martin J. H. *Speech and Language Processing*. Pearson Education Limited, 2014.

[99] Kamara A. A, et al. Predicting required COVID-19 vaccine coverage and its impact in sierra leone using mathematical models. 2021.

[100] Karaulov V, et al. Mathematical model of generalized assessment of the rating of similar objects based on statistical data from the standpoint of epidemiological safety (on the example of the incidence of COVID-19 in the regions of the volga federal district). *Issues of Risk Analysis*, 2021.

[101] KEN POLE . CanadaâĂŹs ads-b mandate raises collective concerns among industry. `https://skiesmag.com/news/canadas-ads-b-mandate-raises-collective-concerns-among-industry/`, 2022. Accessed: 2022-07-18.

[102] Khedhiri S. Statistical modeling of COVID-19 deaths with excess zero counts. *Epidemiologic Methods*, 10, 2021.

[103] Ko Y, et al. Quantifying the effects of non-pharmaceutical and pharmaceutical interventions against COVID-19 epidemic in the republic of korea: Mathematical model-based approach considering age groups and the delta variant. In *medRxiv*, 2021.

[104] Korber B, et al. Tracking changes in SARS-CoV-2 spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell*, 182:812–827.e19, 2020.

[105] Kraay A, et al. Modeling the use of SARS-CoV-2 vaccination to safely relax non-pharmaceutical interventions. In *medRxiv*, 2021.

[106] Krackhardt D. Graph theoretical dimensions of informal organizations, computational organization theory. *Computational Organizational Theory, K. Carley, and M. Prietula, Eds. Hillsdale, NJ: Lawrence Erlbaum Associaties, Inc*, pages 89–111, 1994.

[107] Kucharski A. J, et al. Effectiveness of isolation, testing, contact tracing, and physical distancing on reducing transmission of SARS-CoV-2 in different settings: a mathematical modelling study. *The Lancet Infectious Diseases*, 20(10):1151–1160, Oct 2020.

[108] Kucharski A. J, et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *The Lancet Infectious Diseases*, 20(5):553–558, May 2020.

[109] Kumar J. J, et al. Study of the trend pattern of COVID-19 using spline-based time series model: a bayesian paradigm. *Japanese Journal of Statistics and Data Science*, 5:363 – 377, 2021.

[110] Kwasny S and Faisal K. Overcoming limitations of rule-based systems: An example of a hybrid deterministic parser. volume 252, pages 48–57, 01 1990.

[111] Lahcene B. Probability distributions related to modeling epidemic spread data COVID-19 status and developments. volume 5, pages 134–144, 2021.

[112] Lakman I. A, et al. CovidâĂŚ19 mathematical forecasting in the russian federation. volume 3, pages 288–294, 2020.

[113] Layton A. T and Sadria M. Understanding the dynamics of SARS-CoV-2 variants of concern in Ontario, Canada: a modeling study. *Scientific Reports*, 12(1):2114, 2022.

[114] Levin S. A. Community equilibria and stability, and an extension of the competitive exclusion principle. *The American Naturalist*, 104:413 – 423, 1970.

[115] Li J and Ma M. Multiple outbreaks resulting from asymptomatic infection in spreading of COVID-19. 2021.

[116] Liu Z, et al. Estimating parameters of two-level individual-level models of the COVID-19 epidemic using ensemble learning classifiers. In *Frontiers of Physics*, volume 8, 2021.

[117] Lobato F, et al. Mathematical modelling of the second wave of COVID-19 infections using deterministic and stochastic sidr models. *Nonlinear Dynamics*, 106:1359–1373, 2021.

[118] Long J. COVID-19 real-time tracker and analytical report. *arXiv: Applications*, 2020.

[119] Lotfi M, et al. Innate immune response against COVID-19: The first report on the theory of COVID-19 treatment by a combined method of mathematics and medicine. 2021.

[120] Lu M and Ishwaran H. A competing risk compartmental model for COVID-19 transmission and vaccination effects. *Statistics in Medicine*, 2021.

[121] Mandal S, et al. IndiaâĂŹs pragmatic vaccination strategy against COVID-19: a mathematical modelling-based analysis. *BMJ Open*, 11, 2021.

[122] Mandal S, et al. Combining serology with case-detection, to allow the easing of restrictions against SARS-CoV-2: a modelling-based study in india. *Scientific Reports*, 11, 2021.

[123] Manning C. D, et al. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[124] Manning C. D, et al. The stanford corenlp natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, 2014.

[125] Marino S, et al. A methodology for performing global uncertainty and sensitivity analysis in systems biology. *Journal of Theoretical Biology*, 254(1):178–196, 2008.

[126] Marino S, et al. A methodology for performing global uncertainty and sensitivity analysis in systems biology. *Journal of theoretical biology*, 254(1):178–196, 2008.

[127] Martcheva M. A non-autonomous multi-strain sis epidemic model. *Journal of Biological Dynamics*, 3(2-3):235–251, 2009. PMID: 22880832.

[128] Martcheva M. *An Introduction to Mathematical Epidemiology*. Springer, New York, 2015.

[129] Mbabazi F. K, et al. A mathematical model approach for prevention and intervention measures of the COVID-19 pandemic in uganda. *Asian Research Journal of Mathematics*, 2020.

[130] McKendrick A. G and Kermack W. O. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115, 1927.

[131] Micheletti A, et al. Mathematical models of the spread and consequences of the SARS-CoV-2 pandemics. *Journal of Mathematics in Industry*, 11, 2021.

[132] Miyamoto K. An analysis of the COVID-19 epidemic in japan using a logistic model. *Journal of Disaster Research*, 16:12–15, 2021.

[133] Mizumoto K and Chowell G. Transmission potential of the novel coronavirus (covid-19) onboard the diamond princess cruises ship, 2020. *Infectious Disease Modelling*, 5:264–270, 2020.

[134] M.K. A and WiliÅĎski A. Mathematical modeling and estimation for next wave of covid19 in poland. 2021.

[135] Mohamed H, et al. Estimation of the daily recovery cases in egypt for COVID-19 using power odd generalized exponential lomax distribution. *Annals of Data Science*, 9:71 – 99, 2021.

[136] Moriconi F. Extended mario wuethrich model for COVID-19: Italian and chinese data analysis. *Applied Mathematical Modelling*, 80, 2020.

[137] Mukandavire Z, et al. Quantifying early COVID-19 outbreak transmission in south africa and exploring vaccine efficacy scenarios. *PLoS ONE*, 15, 2020.

[138] Murakami D and Matsui T. Improved log-gaussian approximation for over-dispersed poisson regression: Application to spatial analysis of COVID-19. *PLoS ONE*, 17, 2021.

[139] National Collaborating Centre for Infectious Diseases . Updates on COVID-19 variants of concern (voc). `https://nccid.ca/covid-19-variants/`, 2022. Accessed: 2022-06-16.

[140] Negre C. F. A, et al. Eigenvector centrality for characterization of protein allosteric pathways. *Proceedings of the National Academy of Sciences*, 115(52):E12201–E12208, 2018.

[141] Newman M. E. J. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3), sep 2006.

[142] Newman M. E. J. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74:036104, Sep 2006.

[143] Newman M. E. *Networks: An Introduction*. Oxford University Press, 2010.

[144] Ng K.-S, et al. Implementation of the compulsory universal testing scheme in hong kong: Mathematical simulations of a household-based pooling approach. *Frontiers in Public Health*, 10, 2022.

[145] Ngonghala C. N, et al. Mathematical assessment of the impact of non-pharmaceutical interventions on curtailing the 2019 novel coronavirus. *Mathematical Biosciences*, 325:108364, 2020.

[146] Nishimoto Y and Inoue K. A mathematical model for repetitive behaviors of COVID-19. In *medRxiv*, 2021.

[147] Obasi V. C and Nwaka B. Transmission dynamics of coronavirus pandemic: Modeling and stability analysis. *Journal of Scientific Research and Reports*, 2021.

[148] of Encyclopaedia Britannica T. E. Black deathpandemic, medieval europe, 2023.

[149] Olive X, et al. Crowdsourced air traffic data from The OpenSky Network 2020, June 2022.

[150] organisation W. H. Hiv.

[151] Oud M. A. A, et al. A fractional order mathematical model for COVID-19 dynamics with quarantine, isolation, and environmental viral load. *Advances in Difference Equations*, 2021, 2021.

[152] OurAirport . Open data downloads. `https://ourairports.com/data/`, 2022. Accessed: 2022-06-10.

[153] Pan X. Quantitative analysis and prediction of the worldwide COVID-19 pandemic. In *IOP Conference Series: Earth and Environment*, volume 693, 2021.

[154] Pang B and Lee L. *Opinion mining and sentiment analysis*. Morgan and Claypool Publishers, 2008.

[155] Pathak A, et al. Statistical inferences: Based on exponentiated exponential model to assess novel corona virus (COVID-19) kerala patient data. *Annals of Data Science*, 9:101 – 119, 2021.

[156] Péni T, et al. Nonlinear model predictive control with logic constraints for COVID-19 management. *Nonlinear Dynamics*, 102(4):1965–1986, 2020.

[157] Perera R, et al. Spatial analysis of COVID-19 and socio-economic factors in sri lanka. In *Moratuwa Engineering Research Conference*, pages 444–449, 2021.

[158] Porter M. A, et al. Communities in networks. *CoRR*, abs/0902.3788, 2009.

[159] Radford A, et al. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018.

[160] Raffel C, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

[161] Rockett R. J, et al. Co-infection with SARS-CoV-2 Omicron and Delta variants revealed by genomic surveillance. *Nature Communications*, 13(2745), 2022.

[162] Roda W. C. Bayesian inference for dynamical systems. *Infectious Disease Modelling*, 5:221–232, 2020.

[163] Roda W. C, et al. Why is it difficult to accurately predict the covid-19 epidemic? *Infectious Disease Modelling*, 5:271–281, 2020.

[164] Rosvall M and Bergstrom C. T. Maps of random walks on complex networks reveal community structure. *Proceedings of the national academy of sciences*, 105(4):1118–1123, 2008.

[165] Rush A. M, et al. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, 2015.

[166] Russo L, et al. Tracing day-zero and forecasting the fade out of the COVID-19 outbreak in lombardy, italy: A compartmental modelling and numerical optimization approach. *medRxiv*, 2020.

[167] Saeed M, et al. An optimized decision support model for COVID-19 diagnostics based on complex fuzzy hypersoft mapping. *Mathematics*, 2022.

[168] Saltelli A, et al. *Global Sensitivity Analysis: The Primer*. John Wiley & Sons, Chichester, UK, 2008.

[169] Saraiva E, et al. A piecewise growth model for modeling the accumulated number of COVID-19 cases in the city of campo grande. *Revista Brasileira de Biometria*, 39:240, 2021.

[170] Sarma A. A, et al. Control-theoretic immune tradeoffs explain SARS-CoV-2 virulence and transmission variation. *bioRxiv*, 2021.

[171] Schäfer M, et al. Demonstration abstract: Opensky: A large-scale ADS-B sensor network for research. In *Proceedings of the 13th International Symposium on Information Processing in Sensor Networks*, IPSN '14, page 313âĂŞ314. IEEE Press, 2014.

[172] Scrucca L. A covindex based on a gam beta regression model with an application to the COVID-19 pandemic in italy. *Statistical Methods and Applications*, 31:881 – 900, 2021.

[173] Sebastiani F. Machine learning in information retrieval. *ACM Computing Surveys (CSUR)*, 34(1):1–47, 2002.

[174] Selgelid M. J. Ethics and infectious disease1. *Bioethics*, 19(3):272–289, 2005.

[175] Song H, et al. Forecast of the COVID-19 trend in india: A simple modelling approach. *Mathematical biosciences and engineering : MBE*, 18 6:9775–9786, 2021.

[176] Soubeyrand S, et al. COVID-19 mortality dynamics: The future modelled as a (mixture of) past(s). *PLoS ONE*, 15, 2020.

[177] Stavrou V and Gritzalis D. Introduction to social media investigation âĂŞ a hands-on approach, jennifer golbeck, elsevier publications, usa (2015). *Computers and Security*, 55:128–129, 2015.

[178] Sugiyanto S, et al. Stability analysis of mathematical modeling of interaction between target cells and COVID-19 infected cells. *Biology, Medicine, and Natural Product Chemistry*, 2021.

[179] Surowiec A and Warowny T. COVID-19 death risk estimation using var method. *EUROPEAN RESEARCH STUDIES JOURNAL*, 2021.

[180] Tognotti E. Lessons from the history of quarantine, from plague to influenza a. *Emerging Infectious Diseases*, 19(2):254–259, February 2013.

[181] Tomes N. "destroyer and teacher": Managing the masses during the 1918-1919 influenza pandemic. *Public Health Reports*, 125(Suppl 3):48–62, April 2010.

[182] Touvron H, et al. Llama: Open and efficient foundation language models, 2023.

[183] Traag V. A, et al. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1), March 2019.

[184] Traag V. A, et al. From louvain to leiden: guaranteeing well-connected communities. *CoRR*, abs/1810.08473, 2018.

[185] Tuite A. R, et al. Mathematical modelling of covid-19 transmission and mitigation strategies in the population of ontario, canada. *CMAJ*, 192(19):E497–E505, 2020.

[186] Turing A. M. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

[187] Vaidya N, et al. Modeling within-host dynamics of SARS-CoV-2 infection: A case study in ferrets. *Viruses*, 13, 2021.

[188] van de Schoot R, et al. Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1(1):1, 2021.

[189] van den Driessche P and Watmough J. A simple SIS epidemic model with a backward bifurcation. *Journal of Mathematical Biology*, 40:525–540, 2000.

[190] Vasconcelos G. L, et al. Situation of COVID-19 in brazil: An analysis via growth models as implemented in the modinterv system for monitoring the pandemic. In *medRxiv*, 2021.

[191] Vatteroni M. L, et al. Co-infection with SARS-CoV-2 omicron BA.1 and BA.2 subvariants in a non-vaccinated woman. *Lancet Microbe*, 2022.

[192] Verma P, et al. Theoretical and numerical analysis of fractional order mathematical model on recent COVID-19 using singular kernel. *Proceedings of the National Academy of Sciences, India Section A: Physical Sciences*, 93:219–232, 2023.

[193] Wallach H. M, et al. Right topic right time: Preemption in social recommender systems. In *International Conference on World Wide Web (WWW)*, pages 501–510, 2009.

[194] Wang S, et al. Backward bifurcation, basic reinfection number, and robustness of a seire epidemic model with reinfection. 2022.

[195] Wang X, et al. Complex systems analysis informs on the spread of COVID-19. *Epidemiologic Methods*, 10, 2021.

[196] Watson O. J, et al. Leveraging community mortality indicators to infer COVID-19 mortality and transmission dynamics in damascus, syria. *Nat Commun*, 12(1):2394, 2021.

[197] Watson O. J, et al. Global impact of the first year of COVID-19 vaccination: a mathematical modelling study. *The Lancet Infectious Diseases*, 22(9):1293–1302, Sep 2022. Epub 2022 Jun 23.

[198] Wiggins S. *Introduction to applied nonlinear dynamical Systems and Chaos*. Springer, New York, 1990.

[199] Wilkinson R. H. Homogeneous interpretable approximations to heterogeneous sir models. 2021.

[200] Williams D. E. Statistics of antibody binding to the spike protein explain the dependence of covid 19 infection risk on antibody concentration and affinity. *Scientific Reports*, 12, 2021.

[201] World Health Organization . Coronavirus disease (COVID-19) pandemic. `https://www.who.int/emergencies/diseases/novel-coronavirus-2019`, 2021. Accessed: 2022-05-25.

[202] Wu J, et al. The impact of public health interventions on delaying and mitigating against replacement by SARS-CoV-2 variants of concern. *Available at SSRN*, (February 4), 2021.

[203] Wu J. T and Cowling B. J. The use of mathematical models to inform influenza pandemic preparedness and response. *Experimental Biology and Medicine*, 236(8):955–961, 2011.

[204] Wu R, et al. Sparse topic modeling: Computational efficiency, near-optimal algorithms, and statistical inference. *Journal of the American Statistical Association*, 118:1849 – 1861, 2021.

[205] Xin Y, et al. A new generalized-x family for analyzing the COVID-19 data set: a case study. *Mathematical Problems in Engineering*, 2022.

[206] Xu T, et al. The impact of COVID-19 control measures on air quality in taiyuan, china: a gra-based and svr-based prediction. *Atmospheric Pollution Research*, 2021.

[207] Yadav S. K, et al. Modeling global COVID-19 dissemination data after the emergence of omicron variant using multipronged approaches. *Current Microbiology*, 79, 2022.

[208] Yang H, et al. Evaluating the trade-off between transmissibility and virulence of SARS-CoV-2 by mathematical modeling. In *medRxiv*, 2021.

[209] Yanuar F, et al. Modeling length of hospital stay for patients with COVID-19 in west sumatra using quantile regression approach. *CAUCHY*, 2021.

[210] Yuan P, et al. Efficacy of a "stay-at-home" policy on SARS-CoV-2 transmission in toronto, canada: a mathematical modelling study. *Canadian Medical Association Open Access Journal*, 10(2):E367–E378, 2022.

[211] Zhao S, et al. Modelling the association between COVID-19 transmissibility and D614G substitution in SARS-CoV-2 spike protein: using the surveillance data in california as an example. *Theoretical Biology and Medical Modelling*, 18, 2021.