UNIVERSIDAD DE INVESTIGACIÓN DE TECNOLOGÍA EXPERIMENTAL YACHAY

Escuela de Ciencias Matemáticas y Computacionales

TÍTULO: Gaussian Process Prior to Estimate the SIR Epidemic Model

Trabajo de integración curricular presentado como requisito para la obtención del título de Matemático

Autor:

Velasco Ramírez Byron Andrés

Tutor:

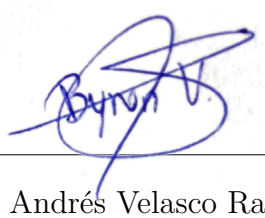
Ph.D. Infante Quirpa Saba Rafael

Urcuquí, Marzo del 2024

Autoría

Yo, **VELASCO RAMÍREZ BYRON ANDRÉS**, con cédula de identidad 0605023639, declaro que las ideas, juicios, valoraciones, interpretaciones, consultas bibliográficas, definiciones y conceptualizaciones expuestas en el presente trabajo; así cómo, los procedimientos y herramientas utilizadas en la investigación, son de absoluta responsabilidad de el/la autor/a del trabajo de integración curricular. Así mismo, me acojo a los reglamentos internos de la Universidad de Investigación de Tecnología Experimental Yachay.

Urcuquí, Marzo del 2024.



Byron Andrés Velasco Ramírez

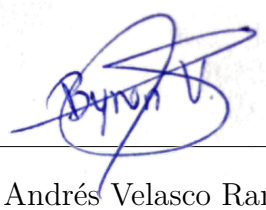
CI: 0605023639

Autorización de publicación

Yo, **VELASCO RAMÍREZ BYRON ANDRÉS**, con cédula de identidad 0605023639, cedo a la Universidad de Investigación de Tecnología Experimental Yachay, los derechos de publicación de la presente obra, sin que deba haber un reconocimiento económico por este concepto. Declaro además que el texto del presente trabajo de titulación no podrá ser cedido a ninguna empresa editorial para su publicación u otros fines, sin contar previamente con la autorización escrita de la Universidad.

Asimismo, autorizo a la Universidad que realice la digitalización y publicación de este trabajo de integración curricular en el repositorio virtual, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación

Urcuquí, Marzo del 2024.



Byron Andrés Velasco Ramírez

CI: 0605023639

Dedication

To my parents Byron and María.
To my sisters Ericka and Lussette.
To my grandparents César and Norma.
To my dear friend Julio César.
To all my friends who trusted me.

Byron Andrés Velasco Ramírez

Acknowledgment

I want to wholeheartedly thank my mom for being the most caring person for my well-being and my future. Without her support, I couldn't have come this far. Despite the difficulties encountered along the way, she has always been there to support me and encourage me to keep going. All this effort is thanks to you, Mom, to show you the great person you've raised since I was a baby.

I want to thank Julio César, who is a great motivator in life, for giving me advice on how to move forward with my projects, my work, and, above all, for improving my mom's days, who is the reason for my existence.

I want to thank my dad, who, despite the distance, has always been attentive to my studies, helping me solve academic problems and providing all possible support.

I want to thank my mentor, Professor Saba Infante, for his patience in my development, for his concern about my progress, and, above all, for his help in completing this work.

I want to thank all my loved ones. To my best friend Pily, my "hija" Kathy, and my "Amigos Mágicos" for brightening my days with their antics and making my days better. Without forgetting my family, who are the spark of joy in my life, who take away all the stress with their family conversations and board games, making me forget all the problems that exist in my life.

Byron Andrés Velasco Ramírez

Resumen

En los últimos años, ha habido una actividad significativa en el desarrollo y la aplicación de algoritmos computacionales eficientes para estimar estados y parámetros en el modelo estocástico SIR de epidemias. Estos modelos nos ayudan a comprender la realidad porque la cuantifican. Las poblaciones en estudio se dividen en estados o categorías. Las tasas de transferencia entre estados se expresan matemáticamente como derivadas con respecto al tiempo basadas en los tamaños de los estados utilizando sistemas de ecuaciones diferenciales ordinarias o ecuaciones diferenciales estocásticas. El objetivo principal de este trabajo es modelar la dinámica de la propagación del brote de la enfermedad por coronavirus 2019 y estimar la curva de tendencia del número reproductivo efectivo. Modelar la propagación de la epidemia facilita la inferencia estadística de los datos y ayuda a planificar estrategias de contingencia para la prevención en la población. La metodología utilizada para estimar los estados y parámetros del modelo estocástico SIR implica aplicar el algoritmo Euler-Maruyama, la aproximación Difussion Bridge, el filtro de Kalman y el proceso Gaussiano. Ilustramos la metodología utilizando estados epidémicos simulados y datos recopilados por la Secretaría Nacional de Gestión de Riesgos y Emergencias. Demostramos cómo los procesos o estados no observados pueden inferirse simultáneamente con los parámetros subyacentes. Entre las principales contribuciones de este trabajo se encuentra proponer estimaciones para el número de personas infectadas, susceptibles y recuperadas y proporcionar una herramienta de monitoreo en tiempo real para el número de casos acumulados.

Palabras Clave:
Modelo epidémico SIR, Euler-Maruyama, Diffusion Bridge, Filtro de Kalman, Proceso Gaussiano.

Abstract

In recent years, there has been significant activity in the development and application of efficient computational algorithms for estimating states and parameters in the stochastic SIR epidemic model. These models help us to understand reality because they quantify it. The populations under study are divided into states or categories. The transfer rates between states are mathematically expressed as derivatives with respect to time-based on the sizes of the states using systems of ordinary differential equations or stochastic differential equations. The main objective of this work is to model the dynamics of the spread of the 2019 coronavirus disease outbreak and estimate the trend curve of the effective reproductive number. Modeling the epidemic's spread facilitates statistical inference of the data and helps plan contingency strategies for population prevention. The methodology used to estimate the states and parameters of the stochastic SIR model involves applying the Euler-Maruyama algorithm, the Diffusion Bridge approximation, the Kalman filter, and the Gaussian process. We illustrate the methodology using simulated epidemic states and data collected by the Secretaría Nacional de Gestión de Riesgos y Emergencias. We show how unobserved processes or states can be inferred simultaneously with the underlying parameters. Among the main contributions of this work are proposing estimates for the number of infected, susceptible, and recovered individuals and providing a real-time monitoring tool for the number of cumulative cases.

Keywords:
SIR epidemic model, Euler-Maruyama, Diffusion Bridge, Kalman Filter, Gaussian Process.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Epidemiology is a discipline of biology that deals with studying public health in the population. Understanding the dynamics of infectious disease outbreaks can reduce their impact on society and the economy. Epidemiology can be well-explained through mathematical models. The fundamental reason is that phenomena observed at the population level are generally complex and difficult to interpret. Modeling epidemics allows us to understand the conditions under which an outbreak occurs and how it spreads. It also enables us to interpret data, test hypotheses, discover patterns, predict epidemics, determine the duration, evaluate intervention strategies to protect the population, and understand the properties of disease dynamics.

In these models, each individual is classified based on disease states and attributes. For example, the states in the SIR model are Susceptible, Infectious, or Recovered. The dynamics of disease transmission are modeled through a system of differential equations that describes the flow of interaction among individuals between states or categories. As the population mixes, the disease spreads, and infected individuals evolve over time through each stage of the disease. Models defined by a differential equation are the most appropriate because they allow for reasonable assumptions about the probabilities of a person getting infected or recovered.

In this study we propose four methods to estimate the curve that models COVID-19 outbreak in Ecuador, considering the stochastic SIR epidemic model: two numerical methods known as the Euler-Maruyama and Diffusion Bridge algorithms, a third method based on the Kalman filter, and finally, we propose a non-parametric method based on a

Gaussian process.

In recent years, the application of Gaussian processes has gained significant attention in epidemiological research due to their ability to capture and model complex, non-linear relationships in data. This is of paramount importance when dealing with infectious disease outbreaks like COVID-19, where various factors, such as government interventions, behavioral changes, and external influences, can lead to intricate and unpredictable dynamics. The uniqueness of Gaussian processes lies in their capacity to offer a flexible framework that can accommodate the inherent uncertainties and complexities of real-world data.

The Gaussian process methodology introduces a valuable tool for estimating the underlying epidemic curve in the context of COVID-19 in Ecuador. Unlike traditional methods, Gaussian processes do not rely on predetermined parametric assumptions about the disease's behavior. Instead, they allow us to infer the epidemic's trajectory directly from the observed data, adapting to the evolving nature of the pandemic. This adaptability is particularly crucial in the case of COVID-19, where new variants and changing behaviors can lead to significant shifts in the transmission dynamics.

Furthermore, Gaussian processes provide a powerful framework for uncertainty quantification. This is essential when making predictions and policy recommendations based on epidemic models. By incorporating uncertainty estimates, decision-makers can better assess the range of possible outcomes and tailor intervention strategies accordingly. The Gaussian process approach, when combined with other traditional modeling methods, offers a comprehensive and holistic view of the COVID-19 epidemic in Ecuador, thereby enhancing our ability to make informed decisions for public health and policy planning.

In addition to their flexibility in capturing complex dynamics and adapting to evolving situations, Gaussian processes are particularly valuable in this context for their ability to capture uncertainty and provide probabilistic estimates. In the realm of COVID-19 modeling, where data can be noisy and variable, the Gaussian process methodology offers a powerful means to quantify the range of possible outcomes and account for uncertainties in our predictions. This is essential for robust decision-making and for understanding the limits of our models in the face of dynamic and evolving pandemics.

The following sections will delve into the details of our approach, highlighting the specific methodologies and techniques employed to harness the potential of Gaussian pro-

cesses in modeling the COVID-19 epidemic, and we will demonstrate how these methods contribute to a more comprehensive understanding of disease dynamics and improved decision-making for the benefit of the population of Ecuador.

1.1 Background

Modern epidemiological mathematical models have their roots in early work, with significant contributions from pioneers such as McKendrick [1], who laid the foundation for the field by introducing methods to determine the probability of an epidemic reaching a certain size before extinguishing. McKendrick's work included the formalization of mathematical models that considered population dynamics over time. This marked the inception of models designed to understand and predict the spread of infectious diseases. Subsequently, Kermack and McKendrick [2] derived an equation for determining the size of an epidemic, considering a threshold based on population density. Their work provided valuable insights into the dynamics of disease propagation.

As the field of epidemiology continued to evolve, Bartlett [3] introduced a stochastic version of the McKendrick model, which allowed for a more nuanced exploration of the uncertainties inherent in infectious disease outbreaks. In the years that followed, a range of impactful models emerged, including deterministic models proposed by Anderson [4], Andersson [5], and Daley [6], among others. These models have played a crucial role in shaping our understanding of the dynamics of infectious diseases within populations.

The emergence of the COVID-19 pandemic in recent years has underscored the critical importance of mathematical modeling in epidemiology. The pandemic has presented unique challenges, demanding sophisticated modeling approaches to adapt to rapidly changing conditions. The ongoing nature of the pandemic, the uncertainties associated with the disease's behavior, and the interplay of various environmental, social, economic, and biological factors have made the modeling of COVID-19 exceptionally complex.

One of the primary challenges in modeling infectious diseases like COVID-19 is the inherent difficulty of studying infected individuals under controlled conditions, as repeated observations are often unfeasible. Moreover, complete data on the epidemic is rarely available. Even when such data is accessible, critical information such as the precise exposure

time and the unique contextual factors influencing disease spread remain elusive. Selecting relevant factors to include in the model, while considering the limitations of partially observed data, is another intricate aspect of epidemiological modeling. Furthermore, the dynamism of disease transmission necessitates the consideration of inherent randomness, which traditional models often struggle to address effectively.

In this context, the use of Gaussian processes emerges as a promising solution. These processes not only offer the flexibility needed to capture the uncertainties and complexities in the data but also adapt to evolving situations. By incorporating Gaussian processes into the modeling framework, we aim to enhance the precision of predictions, provide more accurate estimates, and offer a powerful means to quantify the uncertainty that characterizes infectious disease outbreaks. The expected utility of Gaussian processes lies in their ability to address the challenges posed by the COVID-19 pandemic, thereby contributing to more informed decision-making and more effective public health strategies in Ecuador and beyond.

1.2 Problem statement

The problem addressed in this study lies in the need to accurately understand and model the dynamics of the COVID-19 epidemic in Ecuador. This understanding is essential for making informed public health decisions and implementing effective strategies to control and prevent the spread of the disease.

The inherent complexity of infectious diseases, particularly COVID-19, is one of the primary causes of the problem. The dynamics of disease spread are subject to a range of changing factors such as virus variants, government interventions, and population behavior. These factors make it challenging to predict and accurately model the epidemic's evolution.

The relevance of addressing this problem is undeniable, as it has a direct impact on the health and well-being of Ecuador's population. The ability to accurately predict disease spread and its potential scenarios provides crucial information for decision-making, resource allocation, and intervention strategy implementation. Furthermore, the COVID-19 pandemic has placed significant strain on the country's healthcare systems and economy, underscoring the importance of effectively addressing this problem.

The proposed solution involves the application of Gaussian processes in modeling the COVID-19 epidemic in Ecuador. These processes offer a flexible and powerful methodology that can adapt to the complexities and inherent uncertainties in epidemic data. By using Gaussian processes, the aim is to improve prediction accuracy and the ability to quantify uncertainty in models, thereby enabling more evidence-based public health decisions and disease control policies.

In this study, four methods are proposed for estimating the curve that models the spread of COVID-19 in Ecuador. These methods are based on the stochastic SIR model and employ numerical approaches, such as the Euler-Maruyama and Diffusion Bridge algorithms, as well as the Kalman filter method. Additionally, a non-parametric approach based on Gaussian processes is introduced. The inclusion of this latter method aims to address the need to capture uncertainty in the data and provide probabilistic estimates that support evidence-based decision-making and adaptability to changes in the epidemiological situation.

The resolution of this problem will lead to a deeper understanding of the dynamics of the COVID-19 epidemic in Ecuador and, consequently, contribute to the improvement of intervention strategies, disease control, and, ultimately, the well-being of the Ecuadorian population.

1.2.1 Author's contribution

This study plays a crucial role in addressing a critical and complex issue in the field of epidemiology during the COVID-19 pandemic in Ecuador. Its contribution lies in the application of innovative methods and the implementation of Gaussian processes in epidemic modeling, significantly expanding the toolkit available for understanding and predicting disease spread. By proposing and developing four different approaches, the study demonstrates a commitment to continuous improvement and the pursuit of more precise and adaptable solutions. The implementation of these methods will not only enrich the scientific understanding of the epidemic but also provide health authorities and decision-makers with a more robust and flexible framework for addressing real-time epidemiological challenges. The valuable contribution of this study translates into a significant step towards effective pandemic management and control in Ecuador.

1.3 Objectives

1.3.1 General Objective

To develop a comprehensive epidemiological modeling approach to estimate the model parameters and predict the spread of the COVID-19 epidemic in Ecuador using Gaussian process methods, in order to provide a solid foundation for public health decision-making and the effective implementation of disease control strategies.

1.3.2 Specific Objectives

To achieve the main goal, we will accomplish the following objectives:

- To make a literature review on diffusion processes in pandemic modeling.
- To analyze and process the COVID-19 cases dataset in Ecuador.
- To propose a stochastic diffusion model to model the pandemic.
- To implement the algorithms: Euler-Maruyama, Diffusion Bridge, Kalman Filter, and Gaussian Process.
- To estimate the solution states of the stochastic SIR model.
- To compare the results obtained with the real data.
- To assess the effectiveness of Gaussian processes as a tool for inference in the SIR model based on observed epidemiological data.

Chapter 2

Theoretical Framework

Below are key concepts that will help you understand the study topic.

2.1 SIR model

The SIR model is a widely used mathematical tool in epidemiology for studying the spread of infectious diseases within a population. It was developed by Kermack and McKendrick in 1927 and is based on dividing the population into three main compartments: Susceptibles (S), Infectious (I), and Recovered (R).

- **Susceptible (S):** Represents individuals who are vulnerable to infection because they have not yet been exposed to the pathogen or have not developed immunity.
- **Infectious (I):** Includes individuals who are currently infected and can transmit the disease to susceptibles.
- **Recovered (R):** Represents individuals who have recovered from the disease and have developed immunity, making them unable to be infected again.

The SIR model is described by a set of differential equations that govern the dynamics of these groups over time. These equations consider infection, recovery, and transmission rates and are essential for understanding how an epidemic evolves within a given population.

The SIR model, proposed by Kermack and McKendrick in 1927, has been fundamental in epidemiology and has served as the basis for the development of more complex models for the study of infectious diseases. Its simplicity makes it a valuable tool for understanding

key concepts in disease spread, such as the basic reproduction number (R0), and the importance of control measures in epidemic mitigation.

Let $X_S(t)$, $Y_I(t)$, and $Z_R(t)$ random integer numbers in the population denote susceptible, infectious, and removed, respectively. Let $N(t) = (X_S(t), Y_I(t), Z_R(t))$ a continuous-time Markov chain with events and rates:

$$\begin{aligned}(X_S, Y_I, Z_R) &\to (X_S - 1, Y_I + 1, Z_R) \quad \textit{at rate} \quad \frac{\beta}{N} X_S Y_I \\ (X_S, Y_I, Z_R) &\to (X_S, Y_I - 1, Z_R + 1) \quad \textit{at rate} \quad \gamma Y_I\end{aligned}$$

where the population size is:

$$N = X_S + Y_I + Z_R$$

Note that the vector $N(t)$ is a Markov chain that can be treated as a diffusion process using a stochastic differential equation, or can be treated as a Gaussian process, and is modeled with a Kalman filter if the process is linear, or a Taylor approximation if it is non-linear.

The stochastic SIR epidemic model is a continuous-time discrete-space Markov Chain. We have the following considerations:

- Closed and fixed population with $N + a$ individuals, homogeneously mixed and no latent period,
- $X_S + Y_I + Z_R = N + a$, where N and a are constant. Then $Z_R = N + a - X_S - Y_I$,
- $X_S(0) = N$ and $Y_I(0) = a$,
- All individuals are equally susceptible and infectious,
- An infected individual remains infected, before being removed, for a random period with mean $\frac{1}{\mu}$,
- A susceptible individual after contact with an infected individual immediately becomes infectious,
- A recovered individual acquires immunity to the disease,

- The epidemic ends when $Y_I(t_0) = 0$ for some time $t_0 > 0$.

At this point, the inference problem is to estimate the average infection rate β, the average recovery rate γ (both are considered positive), and R_0 known as the basic reproduction number (the average number of new infections caused by a single infected at an earlier stage of the epidemic). This quantity gives us information about the final size of the epidemic, thus, a large outbreak can occur if and only if $R_0 > 1$. There are several approaches for estimating R_0, depending on assumptions and data limitations; most methods are based on deterministic models which cannot accommodate uncertainty regarding parameter estimation.

The movement of individuals from one compartment to another is represented through the ordinary differential equations below. For the remainder of this paper, we use X, Y, and Z to denote the number of individuals in the S, I, and R compartments, respectively.

$$\begin{aligned}\frac{dX}{dt} &= -\frac{\beta XY}{N}\\ \frac{dY}{dt} &= \frac{\beta XY}{N} - \gamma Y\\ \frac{dZ}{dt} &= \gamma Y\end{aligned}$$

That is, susceptible individuals become infected at a rate that is proportional to the percentage of infected individuals multiplied by β, the infection rate, and the number of susceptible individuals. Infectious individuals recover at a rate of γ multiplied by the number of infected individuals. Individuals begin in the X state as susceptible individuals, possibly become infectious and move to the Y state, and finally possibly recover in the Z state. An outbreak occurs if the rate of change of infectious individuals is positive $\frac{dY}{dt} > 0$, That is, an outbreak occurs if the rate of new infections is greater than the rate of recovery. Formally, an outbreak occurs if:

$$R_0 = \frac{\beta}{\gamma} > 1$$

Now, we consider the stochastic model which is discussed in this paper. The epidemic is completely determined by $\{(X(t), Y(t)); t > 0\}$ which is a continuous-time Markov Chain

on the state space:

$$\mathbf{E} = \{(i, j); 0 \leq i \leq N, 0 \leq j \leq (N - i) + a\}$$

The transition probabilities from time t to $t + h$ are given in the Table 2.1:

Current state	*Transition*	*Next state*	*Probability*
(i, l)	$\longrightarrow$	$(i - 1, l + 1)$	$\frac{\beta}{N} ilh + o(h)$
(i, l)	$\longrightarrow$	$(i, l - 1)$	$\mu lh + o(h)$
(i, l)	$\longrightarrow$	(i, l)	$1 - \left(\frac{\beta}{N} ilh + \mu lh\right) + o(h)$

Table 2.1: Transition probabilities

For $(i, l) \in E$, we define the transition matrix:

$$p_{il}(t) = p\left(S(t) = i, I(t) = l\right) = p\left(I(t) = l | S(t) = i\right) p\left(S(t) = i\right)$$

The changes are given by $dS(t) = S(t + h) - S(t)$, and $dI(t) = I(t + h) - I(t)$. The transition probabilities given in the Table 2.1 satisfy the Kolmogorov forward equations:

$$\frac{\partial p_{(i,l)}(t)}{\partial t} = \frac{\beta(i + 1)(l - 1)}{N} p_{(i+1,l-1)}(t) + \mu(l + 1) p_{(i,l+1)}(t) + \left(1 - \left(\frac{\beta}{N} il + \mu l\right)\right) p_{(i,l)}(t)$$

for $(i, l) \in E$, with $p_{il}(t) = 0$ if $(i, l) \notin E$ and $p_{Na}(0) = 1$. The Markov process given previously leads to a stochastic model that allows us to understand the dynamics of the epidemic.

The Markov chain defined by in the Table 2.1 is well approximated by the solution $X(t)$ of the SDE

$$dX(t) = \mu\left(X(t), \theta\right) dt + \sqrt{\Sigma\left(X(t), \theta\right)} dB_t \tag{2.1}$$

where B_t denotes independent Brownian movements, $\mu = \mu(X(t), \theta)$ is called the drift vector (describes the trend of the stochastic process), and $\Sigma = \Sigma(X(t), \theta)$, is the diffusion matrix (determines the variability around the trend).

When the drift and diffusion processes are sufficiently regular functions, the transition density satisfies the forward Kolmogorov or Fokker-Planck-Kolmogorov equation, so-called Kloeden and Platen:

$$\frac{\partial p_\theta(t;x_0,x)}{\partial t} = -\sum_{i=1}^{d} \frac{\partial\left(\mu_i(x,\theta)p_\theta(t;x_0,x)\right)}{\partial x_i} + \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} \frac{\partial^2}{\partial x_i \partial x_j}\left(\left(\Sigma(x,\theta)Q\Sigma^T(x,\theta)\right)_{ij}\right) p_\theta(t;x_0,x) \quad (2.2)$$

where Q is the diffusion matrix of Brownian motion, and the diffusion process by the equation is given, that stationary and ergodic is considered. Instead of using S and I, we normalise the process by the transformations

$$x(t) = \frac{S(t)}{N} \quad and \quad y(t) = \frac{I(t)}{N}, \quad (2.3)$$

Now, a Fokker-Planck-Kolmogorov SDE for the bivariate stochastic process is obtained, for the SIR epidemic model; that is, in the equation (2.2) $d = 2$, $\mathbf{x} = (x = x_1, y = x_2)^T$:

$$\frac{\partial p_\theta(t;x_0,x)}{\partial t} = -\frac{\partial\left(\mu_1(x,\theta)p_\theta(t;x_0,x)\right)}{\partial x} - \frac{\partial\left(\mu_2(x,t)p_\theta(x,t)\right)}{\partial y} + \frac{1}{2}\frac{\partial^2\left(\Sigma_{11}(x,\theta)p_\theta(t;x_0,x)\right)}{\partial x^2} + \frac{1}{2}\frac{\partial^2\left(\Sigma_{22}(x,\theta)p_\theta(t;x_0,x)\right)}{\partial y^2} + \frac{\partial^2\left(\Sigma_{21}(x,\theta)p_\theta(t;x_0,x)\right)}{\partial x \partial y} \quad (2.4)$$

That is:

$$\begin{pmatrix} dx \\ dy \end{pmatrix} = \begin{pmatrix} -\beta xy \\ \beta xy - \gamma y \end{pmatrix} dt + \frac{1}{\sqrt{N}} \begin{pmatrix} \sqrt{\beta xy} & 0 \\ -\sqrt{\beta xy} & \sqrt{\gamma y} \end{pmatrix} \begin{pmatrix} dB_1 \\ dB_2 \end{pmatrix} \quad (2.5)$$

where B_1 and B_2 are two standard independent Brownian motions. Also:

$$\mu_1(\cdot) = -\beta xy \quad , \quad \mu_2(\cdot) = \beta xy - \gamma y \quad , \quad \Sigma_{11}(\cdot) = \sqrt{\frac{\beta xy}{N}}$$

$$\Sigma_{12}(\cdot) = 0 \quad , \quad \Sigma_{21}(\cdot) = -\sqrt{\frac{\beta xy}{N}} \quad and \quad \Sigma_{22}(\cdot) = \sqrt{\frac{\gamma y}{N}}.$$

Additionally, in the diffusion process defined in the equation (2.1) the data are partially known, this means that:

$$Z_t = (X_t, Y_t)^T$$

where X_t defines the unobservable part and Y_t the observable part of the system. Also, X_t and Y_t have dimensions d_1 and d_2 respectively such that Z_t has dimension $d = d_1 + d_2$. Since the process Y_t is subject to measurement error such that we actually observe:

$$Y_t = f(X_t) + \zeta_t \quad , \quad \zeta_t \sim N\left(0, \sigma_\zeta^2\right), \quad \zeta_t = diag\left(\sigma_{\varsigma_1}^2, \ldots, \varsigma_{\zeta_{d_1}}^2\right)$$

where $f(\cdot)$ is a known real-valued vector non-linear function, and ζ_t represents an independent noise process. The model object of study becomes:

$$\begin{aligned} dX_t &= \mu(X_t, \beta, \gamma)\, dt + \sqrt{\Sigma(X_t, \beta, \gamma)} dB_t \\ Y_t &= f(X_t) + \zeta_t \quad , \quad \zeta_t \sim N\left(0, \sigma_\zeta^2\right) \end{aligned} \tag{2.6}$$

When the $\{Y_t, t \geq 0\}$ are conditionally independent given $\{X_t, t \geq 0\}$ then (2.6) is a state-space model (or a hidden Markov model). The unknown parameters are:

$$\Theta = \left(\beta, \gamma, \sigma_{\varsigma_1}^2, \ldots, \varsigma_{\zeta_{d_1}}^2\right)$$

These models explain how data and a disease transmission model are related as they consist of two components: an unobserved, time-continuous state process, which operates on the population level and describes the dynamics of disease spread, and an observation model, which describes how the data collected at discrete points in time is connected to the transmission model.

The likelihood for the partially observed dynamic system given in equation (2.6) is as follows: let $Y_{1:T} = (Y_1, \ldots, Y_T)$ denote the random variable counting the observations at time t_n in each of the two states $X_n = (S(t_n), I(t_n))$ at that time t_n. Our goal is to infer the latent SIR population dynamic $X_{0:T} = (X_0, \ldots, X_T)$ and the rate parameters vector Θ over time grid $0 = t_0 < t_1 < \ldots < t_n = T$. Let $p(X_0)$ and $p(\Theta)$ denote the prior densities for the initial compartment states and the SIR parameters respectively. The joint density of the states and the observations is defined in terms of the transition density $f_{X_t|X_{t-1}}(x_t|x_{t-1}, \Theta)$, the observation density $g_{Y_t|X_t}(y_t|x_t, \Theta)$, and the initial density $f_{X_0}(x_0, \Theta)$ as:

$$f_{X_{0:T}, Y_{1:T}}(x_{0:T}, y_{1:T}, \Theta) = f_{X_0}(x_0, \Theta) \prod_{t=1}^{T} f_{X_t|X_{t-1}}(x_t|x_{t-1}, \Theta)\, g_{Y_t|X_t}(y_t|x_t, \Theta)$$

Then the posterior distribution for the population trajectory $X_{0:T}$ and parameters Θ given observed $Y_{1:T}$ is:

$$p\left(X_{0:T}, \Theta | Y_{1:T}\right) \propto p\left(Y_{1:T} | X_{0:T}, \Theta\right) p\left(X_{1:T} | X_0, \Theta\right) p\left(\Theta\right) p\left(X_0\right)$$

where:

$$p\left(X_{1:T} | X_0, \Theta\right) = \prod_{t=1}^{T} f_{X_t|X_{t-1}}\left(x_t | x_{t-1}, \Theta\right) \quad \textit{and} \quad p\left(Y_{1:T} | X_{0:T}, \Theta\right) = \prod_{t=1}^{T} g_{Y_t|X_t}\left(y_t | x_t, \Theta\right)$$

Note that the SIR transition density becomes intractable as population size N is large, this makes the process of inference complicated in large populations. The general objective of this work is to estimate the solution states $X_{1:T}$ of the given system in equation (2.6) and the unknown parameters Θ.

2.2 Stochastic Differential Equations (SDE)

Stochastic Differential Equations (SDEs) constitute an essential field in the modeling of dynamic phenomena in the presence of uncertainty. Unlike ordinary differential equations, SDEs incorporate the component of stochastic noise, stemming from random sources, into the mathematical descriptions of dynamic systems. These equations are fundamental in a wide range of disciplines, from physics and biology to economics and engineering, where randomness plays a critical role in the behavior of systems.

SDEs can be expressed in general form as:

$$dX(t) = F(X(t), t)\ dt + G(X(t), t)\ dW(t)$$

where:

- $X(t)$ represents the state of the system at time t.
- $F(X(t), t)$ is the deterministic part that describes the expected evolution of the system.
- $G(X(t), t)$ is the stochastic part that models random fluctuations.
- $dW(t)$ denotes a Wiener process, which is a source of stochastic noise.

SDEs provide a powerful tool for analyzing systems affected by uncertainty and variability and are especially useful in situations where experimental data are limited or noisy.

A key reference illustrating the importance of Stochastic Differential Equations can be found in [7]. This text delves deeply into the concepts and applications of SDEs in the context of physics and natural sciences, highlighting their relevance in understanding physical and chemical processes subject to random fluctuations.

2.3 Markov Chains

Markov chains, named in honor of the Russian mathematician Andrei Markov, are a widely used probabilistic model for describing systems that evolve in discrete steps over time, where the future depends solely on the current state and not on previous states. These chains are fundamental in a wide range of applications in science, engineering, economics, and other disciplines.

A Markov chain is characterized by the following key elements:

- **State Space:** It represents all possible states that the system can take at each time step.
- **Transition Matrix:** This matrix defines the transition probabilities between states. Each entry in the matrix indicates the probability of moving from one state to another in a single step.
- **Markov Property:** The Markov property states that the probability of transitioning to a future state depends only on the current state and not on the previous history of the system.
- **Stationary State:** Some Markov chains reach a stationary state where the probabilities of being in each state no longer change over time.

Markov chains are used to model a variety of phenomena, from weather forecasting to financial risk assessment. A common example is the "Random Walk Problem," where an individual moves randomly along a timeline, and Markov chains are used to analyze the probability of their future location.

A valuable resource that explores Markov chains and their applications in detail can be found [8]. This text provides a solid foundation in the theory and practice of Markov chains and is widely used in academic courses on stochastic processes.

2.4 Monte Carlo Algorithms

Monte Carlo algorithms are a class of computational techniques used to solve numerical problems by generating random samples and computing statistics on these samples. These algorithms are named after the famous Monte Carlo casino in Monaco, known for its randomness and games of chance, as they extensively rely on random numbers in their implementation.

One of the most well-known Monte Carlo algorithms is the Monte Carlo Method for estimating integrals. The main idea behind this method is to randomly sample points within a region of interest and calculate a numerical approximation of the integral as the average of function evaluations at these points multiplied by the total area of the region.

The Monte Carlo Method algorithm can be summarized in the following steps:

1. Define the region of interest containing the function to be integrated.
2. Generate a set of uniformly distributed random points within this region.
3. Evaluate the function at each of these points.
4. Calculate the average of the function evaluations and multiply it by the total area of the region to obtain an estimate of the integral.

The power of Monte Carlo algorithms lies in their ability to tackle complex, high-dimensional problems such as simulating physical systems, evaluating financial risks, optimization, and solving integral equations, among others.

A widely used reference resource for understanding Monte Carlo algorithms and their applications is [9]. This book provides a detailed introduction to concepts and techniques related to Monte Carlo algorithms, including the Monte Carlo Method for estimating integrals, along with examples and applications in statistics and computational sciences.

2.5 Monte Carlo Markov Chain Algorithms

Monte Carlo Markov Chain (MCMC) algorithms are computational techniques used for efficiently and approximately sampling probability distributions, especially in situations where obtaining samples directly is challenging. MCMC algorithms are based on Markov chains, which are sequences of states that evolve according to specific transition probabilities.

Two of the most commonly used MCMC algorithms are the Metropolis-Hastings (MH) algorithm and Gibbs sampling. The MH and Gibbs sampling algorithms are essential in Bayesian statistics and find applications in a wide variety of fields, including machine learning, statistical modeling, and inference in complex systems.

A widely used reference resource for understanding Monte Carlo Markov Chain algorithms, including Metropolis-Hastings and Gibbs sampling, is [10]. This book provides a comprehensive introduction to Bayesian statistics and MCMC methods, with detailed examples and applications.

2.5.1 Metropolis-Hasting Algorithm (MH)

The Metropolis-Hastings algorithm is used to sample from a desired probability distribution, even when obtaining direct samples from this distribution is not straightforward. Here are the steps of the MH algorithm:

1. **Initialization:** Begin with an arbitrary initial value for the variable you want to sample.

2. **Proposal of a New State:** Generate a new candidate for the next state of the Markov chain from a proposal probability distribution (e.g., a normal distribution centered around the current state).

3. **Acceptance or Rejection:** Calculate the acceptance ratio, which is the probability of accepting the new state based on its relative probability compared to the current state and the proposal probability. If the new state is more probable than the current state, it is accepted with high probability. If it is less probable, it is accepted with

a probability equal to the acceptance ratio. If it is rejected, the current state is repeated in the chain.

4. **Iteration:** Repeat steps 2 and 3 a sufficient number of times to obtain a representative sample of the distribution of interest.

2.5.2 Gibbs Sampling

Gibbs sampling is an MCMC algorithm used to sample from multivariate joint distributions by breaking the sampling process into simpler conditional sub-processes. Here are the steps of the Gibbs sampling algorithm:

1. **Initialization:** Start with an initial value for all the variables you want to sample.
2. **Iteration:** For each variable in turn, sample a new value from its conditional distribution given the updated information from the other variables. This sampling is based on the conditional distributions of each variable as a function of the others.
3. **Repetition:** Repeat step 2 for each variable a sufficient number of times to obtain a representative sample from the joint distribution.

2.6 Kalman filter

The Kalman filter is an algorithm used to estimate unobserved states of a system from noisy observations. It is derived in two stages: the prediction stage and the update stage.

2.6.1 System model

Let's assume we have a linear and dynamic system that can be described by two equations:

1. State model:

$$x_k = A \cdot x_{k-1} + w_k$$

where x_k is the true state at time k, A is the state transition matrix, and w_k is the process noise at time k.

2. Observation Model:

$$z_k = H \cdot x_k + v_k$$

where z_k is the observation at time k, H is the observation matrix, and v_k is the observation noise at time k.

The step-by-step derivation of the Kalman filter equations is as follows:

Step 1: Prediction (a priori): Prediction is the estimation of the state at time k based on the information available at $k-1$. In this step, the a priori state and covariance are predicted.

1. State Prediction:

$$\hat{x}_k^- = A \cdot \hat{x}_{k-1}$$

where $\hat{x}_k^-$ is the a priori state estimate at time k, and $\hat{x}_{k-1}$ is the state estimate at time $k-1$.

2. Covariance Prediction:

$$P_k^- = A \cdot P_{k-1} \cdot A^T + Q$$

where P_k^- is the a priori covariance at time k, P_{k-1} is the covariance at time $k-1$, and Q is the process noise covariance.

Step 2: Update (a posteriori): The update combines the prediction information with the observations at time k to obtain a more accurate a posteriori estimate.

1. Residual (Innovation):

$$y_k = z_k - H \cdot \hat{x}_k^-$$

where y_k is the residual or innovation at time k, z_k is the observation at time k, $\hat{x}_k^-$ is the a priori state estimate at time k, and H is the observation matrix.

2. Kalman Gain:

$$K_k = P_k^- \cdot H^T \cdot (H \cdot P_k^- \cdot H^T + R)^{-1}$$

where K_k is the Kalman gain at time k, P_k^- is the a priori covariance at time k, H is the observation matrix, and R is the observation noise covariance.

3. Updated State Estimate:

$$\hat{x}_k = \hat{x}_k^- + K_k \cdot y_k$$

where $\hat{x}_k$ is the a posteriori state estimate at time k, $\hat{x}_k^-$ is the a priori state estimate at time k, K_k is the Kalman gain, and y_k is the residual.

4. Covariance Update:

$$P_k = (I - K_k \cdot H) \cdot P_k^-$$

where P_k is the a posteriori covariance at time k, P_k^- is the a priori covariance at time k, K_k is the Kalman gain, H is the observation matrix, and I is the identity matrix.

The Kalman filter allows for optimal state estimation, taking into account noisy observations and the system's dynamics.

2.7 Gaussian processes

Gaussian processes (GPs) have emerged as a powerful tool in epidemiological modeling, allowing for uncertainty capture and the adaptability necessary to address ever-evolving infectious diseases, such as COVID-19. Recent studies in epidemiology have demonstrated the value of GPs in modeling epidemic dynamics, providing robust and probabilistic estimates of disease spread [11].

Formally, a Gaussian process is defined as a collection of random variables, where any subset of them follows a joint Gaussian distribution. The Gaussian process is characterized by its mean function and its covariance function. In the context of epidemic modeling, GPs enable the representation of disease spread dynamics over time and the capture of uncertainty in estimates.

2.7.1 Mean and covariance function

The mean function, denoted as $\mu(t)$, describes the expected value of the process function at a specific moment, i.e., $\mu(t) = \mathbb{E}[f(t)]$, where $f(t)$ represents the process at time t. The covariance function, denoted as $K(t, t')$, describes how different observations of the process are correlated over time.

The covariance function $K(t, t')$ allows for capturing how observations at different times interact and influence each other. The choice of a specific covariance function plays a critical role in the Gaussian process's ability to capture relevant patterns and features in epidemiological data.

2.7.2 Bayesian inference with Gaussian processes

Bayesian inference is a central approach in the application of Gaussian processes in epidemiology. The goal is to estimate the parameters of the Gaussian process from the observed data, taking uncertainty into account. For this purpose, the Bayesian theorem is used, which relates the posterior distribution of the parameters to the likelihood function and prior distributions.

Let's assume we have observations y from the Gaussian process and we are interested in the process parameters, represented by the vector θ. Bayesian inference allows us to calculate the posterior distribution of θ given the observed evidence y.

This posterior distribution is expressed as:

$$p(\theta|y) = \frac{p(y|\theta) \cdot p(\theta)}{p(y)}$$

Here, $p(\theta|y)$ is the posterior distribution, $p(y|\theta)$ is the likelihood of the data given the parameters, $p(\theta)$ is the prior distribution of the parameters, and $p(y)$ is the marginal likelihood of the data.

Bayesian inference with Gaussian processes enables us not only to estimate the process parameters but also to quantify the uncertainty in these estimates. This is particularly relevant in epidemiological modeling, where inherent uncertainty in disease spread can be high and needs to be considered in public health decision-making.

Chapter 3

State of the Art

In recent years, there has been a significant increase in theoretical development and applications of methodologies for analyzing data from infectious disease outbreaks. In most cases, models governed by a system of differential equations are used. This increase is mainly due to the appearance of the novel coronavirus (COVID-19) at the end of 2019, which originated in Wuhan, China, and has spread to all countries worldwide. Markov chain Monte Carlo (MCMC) methods have been one of the main tools used to analyze infectious disease models in recent years. The implementation of MCMC methods allows for greater flexibility in building a model. Likelihood functions that were previously intractable can now be considered, leading to increasingly detailed inference. Pioneering works in this line of research include [12], [13], and [14]. Recent works of note include [15], [16], [17], [18], [19], [20], [21], among others.

Alternative approaches have been developed to overcome the intractability of models dealing with epidemiological data. These include the Sequential Monte Carlo (SMC) algorithm, also known as particle filters ([22], [23]), and the approximate Bayesian computation (ABC) method ([24], [19]). SMC methods can be used to update the posterior distribution of parameters and the state of the epidemic as the disease progresses. ABC algorithms are a popular tool for analyzing epidemic data due to their simplicity in simulating samples. However, these methods do not solve the problem of updating epidemic process estimates as the disease progresses over time and new data becomes available.

Currently, many extensions of infectious disease models are being studied, with an emphasis on non-parametric methods, which increase flexibility. This is achieved using

Gaussian Processes in [25] and [19]. Other important developments can be found in [26] and [27].

In "Calibration and prediction for the inexact SIR model" by Yan Wan, et al [28], the study proposes calibration and prediction methods for the SIR model, acknowledging heteroscedastic observation errors. It introduces two predictors: a calibrated one and another corrected for discrepancy, integrating a calibrated SIR model with a Gaussian Process-based predictor. It employs bootstrap resampling and numerical assessment, demonstrating that the new predictors outperform existing ones by enhancing the accuracy of discrepancy-corrected prediction by at least 49.95%. Additionally, it introduces a weighted least squares estimator that accounts for the heteroscedasticity of observation errors. It shows that the calibrated SIR model has lower variance than the discrepancy-corrected one, highlighting the inaccuracy of the SIR model and the strong relationship between observation errors and cases.

In "A Gaussian-process approximation to a spatial SIR process using moment closures and emulators" by Parker Trostle, et al [29], the study presents an innovative approach to model disease spread in spatial locations using Gaussian Processes. It extends the SIR model to incorporate spatial aspects and develops a moment closure approximation to simplify parameter estimation. The resulting differential equations are addressed using a low-rank emulator, and a hierarchical model is applied to estimate actual infections from noisy data. Results underscore the effectiveness of combining moment closure approaches with emulators, addressing computational challenges. Areas for improvement are identified, such as the need to investigate more specific conditions for the moment closure approximation and simplify the method's implementation for broader adoption. It is suggested that this approach could be extended to other Markovian processes in various research domains.

In "Bayesian non-parametric inference for stochastic epidemic models using Gaussian Processes" by Xiaoguang Xu, et al [25], the study introduces an innovative approach in epidemiological modeling by using non-parametric Bayesian methods with Gaussian Processes (GPs) to estimate infection dynamics in epidemics. Demonstrating its efficacy with real and simulated data, it emphasizes the accuracy in capturing infection dynamics and its adaptability across various contexts. The feasibility of non-parametric inference in epidemiological models through GPs is highlighted, yet there is an acknowledgment of the

need for further exploration to broaden its applicability to other scenarios, particularly those with more complex infection structures. The discussion includes the selection of covariance functions, emphasizing their impact on the smoothness of estimates without significantly affecting the outcomes. Additionally, it highlights the computational challenge, underscoring the need for improvements to handle larger datasets within suitable time frames.

In "SIR-SI model with a Gaussian transmission rate: Understanding the dynamics of dengue outbreaks in Lima, Peru" by Max Ramírez, et al [30], the study focused on understanding the dynamics of dengue transmission in three districts of Lima, Peru, affected by a recent outbreak. Weekly data on dengue cases in the districts of Comas, Lurigancho, and Puente Piedra were utilized along with temperature data to investigate transmission dynamics. The susceptible-infected-recovered in humans and susceptible-infected in the vector (SIR-SI) model was applied, adjusted through an infection rate modeled by a Gaussian function. The results revealed that this adjusted model effectively captured the behavior of the dengue outbreak in the analyzed districts, displaying a strong dependence on meteorological, cultural, and demographic variables. The influence of climate on transmission was highlighted, establishing a robust qualitative relationship between the transmission rate and the effective reproduction number, Rt. These findings underscore the importance of considering external factors beyond the classic SIR-SI model to comprehend disease spread.

In "Enhanced Gaussian process regression-based forecasting model for COVID-19 outbreak and significance of IoT for its detection" by Shwet Ketu and Pramod Kumar Mishra [31], the study proposes the use of a Multi-Task Gaussian Process (MTGP) regression model to predict the global outbreak of COVID-19. This prediction aims to assist countries in planning preventive measures against the disease's spread. The proposed model is compared with other prediction models such as Linear Regression, Random Forest Regression, Support Vector Regression, and Long Short-Term Memory. Model accuracy is assessed using performance metrics like Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE) for various prediction horizons, ranging from 1 day to 15 days. The results demonstrate that the proposed model outperforms other models in terms of MAPE and RMSE across all prediction horizons, showcasing its suitability and

accuracy. Additionally, the study discusses the significant role of the Internet of Things (IoT) in COVID-19 detection and prevention, exploring potential IoT-based solutions to minimize the disease's impact.

In "Forecasting seasonal influenza with a state-space SIR model" by Dave Osthus, et al [32], the study focused on predicting the intensity and timing of seasonal flu in the United States using a Probabilistic State-Space Model (DBSSM) based on a deterministic mathematical model (SIR). They emphasized the importance of carefully specifying the prior distribution in the model, as the results critically depend on this specification. This prior distribution enabled them to leverage known relationships between the latent initial conditions of the SIR model and public health surveillance data. They compared their approach with other alternatives, demonstrating significant advantages in terms of prediction accuracy. They also proposed incorporating multiple disease surveillance systems and stressed the need for standardized and meaningful forecast metrics to compare competing models. They highlighted that clearly defining the predicted event and accuracy measures is essential, emphasizing the importance of direct comparisons between different forecasting approaches.

In "Forecasting COVID-19 cases based on a parameter-varying stochastic SIR model" by J Hespanha, et al [33], the study focused on predicting the evolution of COVID-19 using a time-varying parameter SIR model, employing time series of new cases and deaths. Despite uncertainties within the model, they managed to generate reliable forecasts, validated against a broad dataset. They emphasize the significance of confidence intervals, noting that while the forecasts might not always be precise, these intervals typically encompass future measurements.

In "Gaussian process approximations for fast inference from infectious disease data" by Elizabeth Buckingham-Jeffery, et al [34], the study focused on comparing Gaussian process approximations with stochastic models of infectious diseases, presenting a framework to evaluate the accuracy of these approximations in rapid inference from outbreak data. Researchers developed a flexible framework to derive and quantify the accuracy of these Gaussian process approximations for SIR and SEIR models. They highlighted the capability of these approximations to make swift maximum likelihood inferences using estimates of infected populations. They also demonstrated the feasibility of inferring unobserved pro-

cesses simultaneously with underlying parameters. The results aim to encourage wider use of Gaussian process approximations in infectious disease epidemiology, exploring multiple approaches and assessing errors through simulated and real data. Additionally, they aim to address future methodological challenges, including accuracy in case identification, the need for real-time analysis methods, and the ability to make future predictions in uncertain environments.

In "When and How to Lift the Lock down? Global COVID-19 Scenario Analysis and Policy Assessment using Compartmental Gaussian Processes" by Zhaozhi Qian, et al [35], the study focused on developing a predictive model for COVID-19 lockdown policies globally, using a two-layer Gaussian Process approach. The first layer tailors specific models for each country and policy, while the second layer shares data across countries and focuses on each nation's characteristics and policy indicators. The study compared COVID-19 mortality projections with other models and assessed lockdown strategies, highlighting their impact on mortality. The results underscore the importance of basing government decisions on predictive models in times of crisis. It aims to inform governments and public health about the impact of policies and social behavior on global health, contributing to the global effort in managing this crisis.

Chapter 4

Methodology

4.1 Phases of Problem Solving

4.1.1 Description of the Problem

The COVID-19 pandemic, caused by the SARS-CoV-2 virus, has posed an unprecedented global challenge in public health and epidemiology. Ecuador, like many other countries, has experienced the spread of this disease across its territories, leading to a critical need to understand and predict the virus's spread for informed decision-making and effective mitigation strategies. The Susceptible-Infectious-Recovered (SIR) epidemic model has been widely used to model the spread of infectious diseases, including COVID-19.

The application of the SIR model in epidemiology has been a valuable tool for understanding the spread of infectious diseases. However, the accurate estimation of its parameters, such as the infection rate and recovery rate, is essential to make the model more precise and useful in public health decision-making [36]. In the context of COVID-19 in Ecuador, the precise estimation of these parameters is crucial for planning effective responses and assessing the impact of interventions such as social distancing and vaccination.

As mentioned before, the SIR model is a simple model used to simulate the dynamics of an infectious disease. In this context, numerical and stochastic methods are applied to model the dynamics of COVID-19 infection in Ecuador. Implementing approaches like Euler-Maruyama, Diffusion Bridge, Kalman filter, and Gaussian processes helps simulate the infection dynamics by solving the proposed SDEs (Stochastic Differential Equations) in the simple SIR model. Simulating contagious diseases, such as COVID-19, which led to

a pandemic, is an extremely important resource for implementing safety measures and risk planning in the affected area.

As previously established, the SIR model has several limitations that affect the true dynamics of the disease. For example, in real life, it's almost impossible to have a closed population where there's no migration, births, or deaths. Additionally, it overlooks the variability in infection and recovery rates. In the case of COVID-19, different variants of the virus have been observed, each with varying infection and recovery rates.

Despite these limitations, we will simulate the dynamics of the COVID-19 virus in Ecuador based on real infection data and by proposing initial estimates of infection and recovery rates for the virus. With the algorithms presented in this section, they can be further used in more complex epidemiological models, such as the SEIR model, SIS model, among others.

The challenge lies in the complexity of COVID-19 epidemiological data, which includes multiple sources of uncertainty, such as data quality, variability in case detection, and population dynamics. Directly estimating the parameters of the SIR model from this data can be problematic and underestimate uncertainty [37]. Therefore, there is a need to address this problem using more advanced approaches.

4.1.2 Analysis of the Problem

In recent years, there has been a growing interest in the use of Gaussian processes to address parameter estimation in epidemiological models, including the SIR model. Gaussian processes are machine learning techniques that can effectively model uncertainty and variability in data [11]

In the context of COVID-19 in Ecuador, the application of Gaussian processes offers several advantages. Firstly, it allows for a more realistic incorporation of uncertainty into the estimation of SIR model parameters. This is especially valuable in situations where data is limited or noisy.

Secondly, Gaussian processes can capture spatial and temporal relationships in disease spread. Since the spread of COVID-19 can vary in different regions of Ecuador and over time, this spatial-temporal modeling capability is essential for a better understanding of disease dynamics.

However, it is important to note that the application of Gaussian processes in epidemiology also presents challenges, such as the proper selection of hyperparameters and the integration of heterogeneous data sources.

4.1.3 Algorithm Design

Euler-Maruyama

The problem with the estimation of the parameters is that the model (2.6) is formulated in continuous time, while the sampling data are naturally available only at discrete time frequencies. A numerical solution of the stochastic differential equation:

$$dX_t = \mu\left(X_t, \beta, \gamma\right) dt + \sqrt{\Sigma\left(X_t, \beta, \gamma\right)} dB_t \quad in \quad [0, T], \quad X_0 = x_0 \tag{4.1}$$

is a stochastic process that solves the equivalent equation when the differentials are replaced by difference approximations. Take a partition of $[0, T]$;

$$0 = t_0 < t_1 < \ldots, t_k < t_{k+1} < \ldots < t_n = T, \quad \Delta t = t_{i+1} - t_i = \frac{T}{k}$$

So that:

$$t_{i+1} = t_i + \Delta t = i\Delta t, \quad \Delta B_i = \Delta B_{t_i} = \left(B_{t_i + \Delta t} - B_{t_i}\right)$$

In the method of Euler-Murayama, which is the simplest one,

$$dX_{t_i} \approx \Delta X_{t_i} = X_{t_{i+1}} - X_{t_i} \quad , \quad dB_{t_i} \approx \Delta B_i$$

Therefore:

$$X_{i+1} = X_i + \mu\left(X_i, \theta\right) \Delta t + \sqrt{\Sigma\left(X_i, \theta\right)} \Delta B_i$$

Since $\Delta B_i \sim N\left(0, \Delta t\right)$ and $\eta \sim N(0, 1)$, then $\sqrt{\Delta t}\eta \sim N\left(0, \Delta t\right)$. To generate a trajectory or realization it is necessary to generate random values of η_i.

$$X_{i+1} = X_i + \mu\left(X_i, \theta\right) \Delta t + \sqrt{\Sigma\left(X_i, \theta\right) \Delta t} \eta_i \quad , \quad i = 0, 1, \ldots, k-1 \quad , \quad X_0 = x_0$$

Rewriting the expression dX_t in terms of the random variables $S(t)$ and $I(t)$,if we assume that ΔB_t^1 and ΔB_t^2 are independent Wiener processes, the Euler–Maruyama discrete version would be:

$$S_{t_{k+1}} = S_{t_k} - \beta S_{t_k} I_{t_k} \Delta t + \sqrt{\frac{\beta S_{t_k} I_{t_k}}{N}} \Delta t \Delta B_i^1, \quad S_{t_0} = s_{t_0}$$
$$I_{t_{k+1}} = I_{t_k} + \left(\beta S_{t_k} I_{t_k} - \gamma I_{t_k}\right) \Delta t + \left(\sqrt{\frac{\beta S_{t_k} I_{t_k}}{N}} \Delta B_i^1 + \sqrt{\gamma I_{t_k}} \Delta B_i^2\right), \quad I_{t_0} = i_{t_0}$$

with independent Brownian increments $\Delta B_i^1 = \eta_i^1 \sqrt{\Delta t_i}$, $\Delta B_i^2 = \eta_i^2 \sqrt{\Delta t_i}$ where η_i^1 and η_i^2 are independent draws from the standard normal distribution.

The path is simulated through a recursive application of

$$X_{t_{k+1}} | X_{t_k} \sim N\left(X_{t_k} + \mu_k \Delta t, \Sigma\left(X_i, \theta\right) \Delta t\right)$$

The error satisfies:

$$\mathbb{E}\left(\left|X\left(t_i\right) - X_i\right|^2 \middle| X\left(t_{i-1}\right) - X_{i-1}\right) = o\left(\Delta t^2\right) \quad \textit{(one step)}$$

and

$$\mathbb{E}\left(\left|X\left(t_i\right) - X_i\right|^2 \middle| X(0) - X_0\right) = o\left(\Delta t\right)$$

Order of weak convergence equal to 1. The joint density of this approximation is:

$$p\left(X_{t_{1:T}} | X_0, \theta\right) \propto \prod_{k=0}^{T} N\left(X_{t_k} + \mu_k \Delta t, \Sigma\left(X_i, \theta\right) \Delta t\right)$$

Diffusion Bridge Approximation

Suppose the process given in (2.6), and that a discrete time realization of X_t is generated conditional in x_0 and Y_T. Let partition $[0, T]$ as

$$0 = t_0 < t_1 < \ldots < t_{N-1} < t_N = T, \quad \Delta t = \frac{T}{m}$$

The continuous-time conditioned process is then approximated by the discrete-time

diffusion bridge, with latent values

$$x_{(0,T]} = (x_{t_1}, \ldots, x_{t_m} = x_T)^T$$

having the posterior density

$$p\left(x_{(0,T]}|x_0, y_T, \Theta\right) \propto p\left(y_T|x_T, \sigma_\zeta^2\right) \prod_{k=1}^{m-1} p\left(x_{t_{k+1}}|x_{t_k}, \Theta\right)$$

where

$$p\left(x_{t_{k+1}}|x_{t_k}, \Theta\right) \sim N\left(X_{t_k} + \mu_k \Delta t, \Sigma\left(X_i, \theta\right)\Delta t\right)$$

is the transition density under the Euler–Maruyama approximation,

$$p\left(y_T|x_T, \sigma_\zeta^2\right) \sim N\left(f(x_T), \sigma_\zeta^2\right)$$

For know x_T, [38] derives a linear Gaussian approximation of $p\left(x_{t_{k+1}}|x_{t_k}, \Theta\right)$ Extensions are considered in [39]. Then the joint distribution $p\left(X_{t_{k+1}}, Y_T|X_{t_k}\right)$ is approximated by:

$$\left(\begin{array}{c} X_{t_{k+1}} \\ Y_T \end{array}\right) \Bigg| X_{t_k} \sim N\left(\left(\begin{array}{c} x_{t_k} + \mu_k \Delta t \\ F'\left(x_{t_k} + \mu_k \Delta t\right) \end{array}\right), \left(\begin{array}{cc} \Sigma_k \Delta t & \Sigma_k F \Delta t \\ F'\Sigma_k \Delta t & F'\Sigma_k F \Delta k + \sigma_\zeta^2 \end{array}\right)\right)$$

where $f(X_T) = F'X_T$ is a constant $d \times d$ matrix, $\mu_k = \mu(x_{t_k})$, $\Sigma_k = \Sigma(x_{t_k})$ and $\Delta k = T - t_k$. Conditioning on $Y_T = y_T$, is obtained:

$$\mu_{MDB}\left(x_{t_k}\right) = \mu_k + \Sigma_k F\left(F'\Sigma_k F \Delta_k + \sigma_\zeta^2\right)^{-1}\left(y_T - F'\left(x_{t_k} + \mu_k \Delta t\right)\right)$$

and

$$\Sigma_{MDB}\left(x_{t_k}\right) = \Sigma_k - \Sigma_k F\left(F'\Sigma_k F \Delta_k + \sigma_\zeta^2\right)^{-1} F'\Sigma_k \Delta t.$$

In the case of no measurement error and observation of all components then x_T is known, and

$$\mu_{MDB}\left(x_{t_k}\right) = \frac{x_T - x_{t_k}}{T - t_{t_k}} \quad , \quad \Sigma_{MDB}\left(x_{t_k}\right) = \frac{T - t_{k+1}}{T - t_k}\Sigma(x_{t_k})$$

Kalman Filter

Suppose that the joint distribution of the susceptible and infectious population can be approximated by a bivariate normal distribution. In [34] is shown that for the SIR model, we obtain a set of 5 ODEs for the mean, variance, and covariance of the susceptible and infectious populations. To make an inference with the SIR model, suppose that, $x(t)$ and $y(t)$ represent the number of susceptible and infectious people respectively at time t. We consider priors

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} \sim GP\left(\mu(t), C(t,t)\right)$$

where

$$\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} = \begin{pmatrix} \mathbb{E}\left(x\left(t\right)\right) \\ \mathbb{E}\left(y\left(t\right)\right) \end{pmatrix}, \quad C(t,t) = \begin{pmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{pmatrix} = \begin{pmatrix} \mathbb{V}ar\left(x\left(t\right)\right) & \mathbb{C}ov\left(x\left(t\right), y\left(t\right)\right) \\ \mathbb{C}ov\left(y\left(t\right), x\left(t\right)\right) & \mathbb{V}ar\left(y\left(t\right)\right) \end{pmatrix}$$

This process satisfies the following system of ordinary differential equations

$$\begin{aligned}
\frac{d\mu_x}{dt} &= -\frac{\beta}{N}\left(\mu_x\mu_y + C_{xy}\right) \\
\frac{d\mu_y}{dt} &= \frac{\beta}{N}\left(\mu_x\mu_y + C_{xy}\right) - \gamma\mu_y \\
\frac{dC_{xx}}{dt} &= \frac{\beta}{N}\left(\mu_x\mu_y + C_{xy} - 2\mu_x C_{xy} - 2\mu_y C_{xx}\right) \\
\frac{dC_{xy}}{dt} &= \frac{\beta}{N}\left(\mu_x\left(C_{xy} - C_{yy}\right) + \mu_y\left(C_{xx} - C_{xy}\right) - \mu_x\mu_y - C_{xy}\right) - \gamma C_{xy} \\
\frac{dC_{yy}}{dt} &= \frac{\beta}{N}\left(2\mu_x C_{yy} + 2\mu_y C_{xy} + \mu_x\mu_y + C_{xy}\right) - \gamma\left(2K_{yy} - \mu_y\right)
\end{aligned}$$

These equations are obtained by applying the assumption of normality to the moment equations of the Markov chain. We now show the approximation of a linear stochastic process proposed in [40], [41], and [34]. Here, assuming that the susceptible population evolves deterministically, and the infective individuals are normally distributed, gives the following set of three ODEs for the evolution of the deterministic susceptible population, $s(t)$, and the mean, $\mu_y = \mathbb{E}\left(y\left(t\right)\right)$, and the variance $C_{yy} = \mathbb{V}ar\left(y\left(t\right)\right)$ of the infectious

population:

$$\begin{aligned}
\frac{ds}{dt} &= -\frac{\beta}{N} s\mu_y \\
\frac{d\mu_y}{dt} &= \frac{\beta}{N} s\mu_y - \gamma\mu_y \\
\frac{dC_{yy}}{dt} &= \frac{\beta}{N}\left(2sC_{yy} + s\mu_y\right) - \gamma\left(2C_{yy} - \mu_y\right)
\end{aligned}$$

For inferential purposes, it is preferable that the susceptible population be random, and then can consider an approximation of a linear stochastic differential equation. Let a finite set of noisy observations $y_{1:N} = (y_1, \ldots, y_N)$ of a hidden state $\{X_t, t \geq 0\}$; and if assumed that the time evolution of X_t is described by an (Itô) stochastic differential equation (SDE):

$$\begin{aligned}
dx_t &= F(t,\theta)x(t)dt + L(t,\theta)dB_t \\
y_k &= H_k x(t_k) + r_k \quad , \quad r_k \sim N\left(0, R_k\right)
\end{aligned} \tag{4.2}$$

where $x(t_k)$ is the state at time t_k, $\theta \in \Theta \subseteq \mathbb{R}^d$ is the vector of parameters to be estimated, $F(t,\theta) = A(t)x + b$, $F : [0,\infty) \to \mathbb{R}^n$ is a linear dynamic model function, $L : [0,\infty) \times \Theta \to \mathbb{R}^{n\times s}$ is a linear matrix valued function, B_t is $s-$dimension Brownian motion with diffusion matrix $Q \in \mathbb{R}^{s\times s}$, $y_k \in \mathbb{R}^m$ is the measurement at time t_k, $H : \mathbb{R}^n \to \mathbb{R}^m$ is the measurement model function. Archambeau et al. (2007) showed that the Gaussian process solution $p(x) \sim GP\left(\mu(t), C(t,t)\right)$ satisfying

$$\begin{aligned}
\frac{dm_t}{dt} &= Am_t + b \\
\frac{dC_t}{dt} &= AC_t + C_tA^T + \Sigma_t, \quad \Sigma_t = L(t,\theta)QL^T(t,\theta)
\end{aligned} \tag{4.3}$$

For SIR model, [34] showed the following results:

$$\begin{pmatrix} \frac{dS(t)}{dt} \\ \frac{dI(t)}{dt} \end{pmatrix} = \begin{pmatrix} -\frac{\beta}{N}\left(s(t)I(t) + S(t)i(t) - s(t)i(t)\right)dt \\ \frac{\beta}{N}\left(s(t)I(t) + S(t)i(t) - s(t)i(t)\right) - \gamma i(t) \end{pmatrix} dt + \begin{pmatrix} \frac{\beta}{N}s(t)i(t) & -\frac{\beta}{N}s(t)i(t) \\ -\frac{\beta}{N}s(t)i(t) & \frac{\beta}{N}s(t)i(t) + \gamma i(t) \end{pmatrix}^{\frac{1}{2}} dB_t$$

where: $S(t) = s(t) + \tilde{S}(t)$, $I(t) = i(t) + \tilde{I}(t)$, $\tilde{S}(t)$, $\tilde{I}(t)$ are assumed to be small in the approximation, ignoring quadratic terms and considering only linear terms gives

$$A(t) = \begin{pmatrix} -\frac{\beta}{N} i(t) & -\frac{\beta}{N} s(t) \\ \frac{\beta}{N} i(t) & \frac{\beta}{N} s(t) - \gamma \end{pmatrix} \quad , \quad b(t) = \begin{pmatrix} \frac{\beta}{N} s(t) i(t) \\ -\frac{\beta}{N} s(t) i(t) \end{pmatrix}$$

and

$$\Sigma(t) = \begin{pmatrix} \frac{\beta}{N} s(t) i(t) & -\frac{\beta}{N} s(t) i(t) \\ -\frac{\beta}{N} s(t) i(t) & \frac{\beta}{N} s(t) i(t) + \gamma i(t) \end{pmatrix}$$

Now we propose to make an approach using the Kalman filter (KF). For a review of the KF see, e.g., [42], [43], [44], which originally appeared in [45]. The KF algorithm provides a recursive efficient computation of dynamic states from which the mean of the squared error is minimized. For the derivation of the filtering steps for the KF algorithm see [46]

Algorithm 1 The Kalman Filter

---

1. Initialize the mean m_0 and covariance C_0
2. For $K = 1, 2, \ldots,$ perform the following:
   (a) Prediction step:

$$\frac{dm_k^-}{dt} = Am_k^-(t) + b$$
$$\frac{dC_t^-}{dt} = AC_t^- + C_t^- A^T + \Sigma_t$$

   where $m_k^-(t_{k-1}) = m_{k-1}$ and $C_k^-(t_{k-1}) = C_{k-1}$, and the prediction result is given as $m_k^- = m_k^-(t_k)$, $C_k^- = C_k^-(t_k)$.

3. Update step:

$$
\begin{aligned}
S_k &= H_k C_k^- C_k^T + R_k \\
K_k &= C_k^- H_k^T S_k^{-1} \\
m_k &= m_k^- + K_k \left(y_k - H_k m_k^- \right) \\
C_k &= C_k^- + K_k S_k K_k^T
\end{aligned}
$$

---

where m_k^- is a prior state estimate, m_k is a posterior state estimate, C_k^- is a prior estimate error covariance, C_k is a posterior estimate error covariance and $\Sigma_t = L(t,\theta) Q L^T(t,\theta)$.

Gaussian Process

A Gaussian Process (GP) is a collection of random variables, where any subset of these variables follows a joint Gaussian distribution. A GP is completely specified by its mean function and the covariance and variance function. We define mean function $\mu(x)$ and the covariance function $C(x,x)$ of a real process $f(x)$ as

$$
\mu(x) = \mathbb{E}\left(f(x)\right), \quad C\left(x, x^T\right) = \mathbb{E}\left\{ \left[f(x) - \mu(x)\right] \left[f(x^{'}) - \mu(x^{'})\right] \right\}
$$

where the mean function $\mu(x) : \mathbb{R} \to \mathbb{R}$, and the covariance function $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ will write the GP as

$$
f(x) \sim GP\left(\mu(x), k(x, x^{'})\right)
$$

Let $f(x)$ a GP prior and consider that the observations are given by

$$
(T, Y) = \{(t_1, y_1) \ldots, (t_n, y_n)\}
$$

having likelihood $N\left(y; x_T, \Lambda\right)$ give rise to a posterior $GPs\left(x; \mu^s, k^s\right)$ with

$$
\mu_t^{post} = \mu_{prior} + k_{lT}\left(k_{TT} + \Lambda\right)^{-1}\left(y - \mu_T\right) \quad and \quad k_{uv}^{post} = k_{uv} - k_{uT}\left(k_{TT} + \Lambda\right)^{-1} k_{Tv}
$$

GPs provide a distribution of overfitted functions and associated gradients. Incorporating priors (prior distributions) into the parameters of the GP model and the SDE represents a powerful and highly versatile approach to parameter estimation in Bayesian contexts. Gaussian processes are commonly used to model nonlinear relationships and provide smooth and continuous predictions in various applications, such as regression and prediction. Introducing priors into the parameters of these models allows the incorporation of prior information or beliefs about parameter values, enhancing parameter estimation and prediction by accounting for uncertainty in the data. This Bayesian approach not only offers greater modeling flexibility but also provides enhanced statistical inference robustness by leveraging the advantages of Bayesian statistics in handling uncertainty coherently and efficiently.. These GP approaches have similar computational complexity and can run up to two orders of magnitude faster than numerical integration. This plays a similar role to numerical integration but without the corresponding high computational cost. GPs are closed under linear maps, in particular, the joint distribution over x and its derivative is:

$$p\left[\begin{pmatrix} x \\ \dot{x} \end{pmatrix}\right] = GP\left[\begin{pmatrix} x \\ \dot{x} \end{pmatrix}; \begin{pmatrix} \mu \\ \mu^{\partial} \end{pmatrix}, \begin{pmatrix} k & k^{\partial} \\ \partial_k & \partial_{k^{\partial}} \end{pmatrix}\right]$$

where:

$$\mu^{\partial} = \frac{\partial \mu(t)}{\partial t} \quad , \quad k^{\partial} = \frac{\partial k(t,t^{'})}{\partial t^{'}} \quad , \quad \partial_k = \frac{\partial k(t,t^{'})}{\partial t} \quad , \quad \partial_{k^{\partial}} = \frac{\partial^2 k(t,t^{'})}{\partial t \partial t^{'}}$$

We consider continuous time dynamical systems in which the motions of d states $\mathbf{x}(t) = (x_1(t), \ldots, x_d(t))$ are represented by $d-$dimensional Itȯ process governed by the SDE

$$dx_t = \mu\left(x_t, \theta\right) dt + \sqrt{\Sigma\left(x_t, \theta\right)} dB_t \tag{4.4}$$

where θ is a vector of parameters of the SDE. For notational convenience, let $\mathbf{X} = (x(t_1), \ldots, x(t_T))$ and $k-$th state sequence $x_k = (x_k(t_1), \ldots, x_k(t_T))$, and given noisy observations model of $\mathbf{X}$, then the task is to infer a posterior distribution over the parameters θ. Let T observations $\mathbf{y}(t) = [y(t_1), \ldots, y(t_T)]$ are obtained from the states $\mathbf{X}$ in terms of:

$$y(t) = x(t) + \epsilon_t \quad , \quad \epsilon_t \sim N\left(0, \sigma^2\right) \tag{4.5}$$

This leads to an observation model:

$$p_{OBS}\left(Y|X\right) = \prod_{t=1}^{T} p_{OBS}\left(y(t)|x(t)\right) \quad , \quad p_{OBS}\left(y(t)|x(t)\right) \sim N\left(x(t), \sigma^2\mathbf{I}\right)$$

We propose the following model over states X, their derivatives $\dot{X}$, observations Y and parameters, where the joint distribution is given by:

$$p\left(Y, X, \dot{X}, \phi^+.\theta\right) \propto p\left(\phi^+\right) p\left(\theta\right) p_{GP}\left(Y|\dot{X}, \phi^+\right) p_{SDE}\left(\dot{X}|X, \theta\right) p\left(X|\phi^+\right)$$

where $\phi^+ = (x_0, \phi)$. To generate data from this model the procedure is as follows:

1. Generate $\phi^+ \sim p_1(.)$ and $\theta \sim p_2(.)$,
2. Generate $X|\phi^+ \sim p_{GP}\left(X|\phi^+\right)$,
3. Generate $\dot{X}|X, \theta \sim p_{SDE}\left(\dot{X}|X, \theta\right)$, and
4. Generate $Y|\dot{X}, \phi^+ \sim p_{GP}\left(Y|\dot{X}, \phi^+\right)$.

Prior on latent state $p\left(X|\phi^+\right)$ is as follows; the GP prior assumes that the states are a prior independent, then the prior on the latent state is given by:

$$p_{GP}\left(x_k|\phi^+\right) \sim GP\left(\mu(t), k(t, t^{'})\right)$$

Also, we need to approximate $p_{SDE}\left(\dot{X}|X, \theta\right)$, to achieve this objective, we proposed to use a numerical procedure based on an ordinary differential equation (ODE), the generalization to the case of a stochastic differential equation (SDE) is immediate just add Gaussian random noise to the equation.

The temporal evolution of the ODE can be modeled by assigning a probability distribution over the solution. An ODE initial value problem is to find a function $x(t) : \mathbb{R} \to \mathbb{R}^d$, such that the ODE

$$\dot{x} = \frac{\partial x}{dt} = f(x, t), \quad \textit{holds for} \quad t \in T = [t_0, t_T], \quad \textit{and} \quad x(t_0) = x_0$$

We assume that a unique solution exists. Runge-Kutta methods (RKs) see [47], are carefully designed linear extrapolation methods operating on small contiguous subintervals

$[t_n, t_n + h] \subset T$ of length h. These methods collect observations of approximate gradients of the solved ODE, by evaluating the vector field f at an estimated solution, which is a linear combination of previously collected observations. In an interval $[t_0, t_0 + h]$, the RK method is to evaluate:

$$y_i = f\left(\hat{x}_i, t_0 + hc_i\right), \quad i = 1, \ldots, s$$

where:

$$\hat{x}_i = x_0 + h \sum_{j=1}^{i-1} \omega_{ij} y_j$$

then returns a simple prediction for the solution of the IVP at $t_0 + h$, as $\hat{x}(t_0 + h) = x_0 + h \sum_{i=1}^{s} b_i y_i$. In compact form:

$$y_i = f\left(x_0 + h \sum_{j=1}^{i-1} \omega_{ij} y_j, t_0 + hc_i\right), \quad i = 1, \ldots, s \quad \hat{x}(t_0 + h) = x_0 + h \sum_{i=1}^{s} b_i y_i$$

Runge-Kutta methods can be constructed naturally from a Gaussian process over $x(t)$, where the y_i are treated as observations of $\hat{x}\left(t_0 + hc_i\right)$ and the $\hat{x}_i$ are subsequent posterior estimates, see [48].

A recursive algorithm analogous to RK methods can be constructed see [49], and [47], by setting the prior mean to the constant $\mu(t) = x_0$, then recursively estimating $\hat{x}_i = \mu^i\left(t_0 + hc_i\right)$ in some from the current posterior over x. We know that $y_i = f\left(\hat{x}_i, t_0 + hc_i\right)$, and $y_i | x \sim N\left(\dot{x}\left(t_0 + hc_i\right), \Lambda\right)$, this result allows us to obtain a recursive algorithm:

$$\hat{x}\left(t_0 + hc_i\right) = x_0 + \sum_{j=1}^{i-1} \sum_{l=1}^{i-1} k^{\partial}\left(t_0 + hc_i, t_0 + hc_l\right) \left(\partial_{k^{\partial}} + \Lambda\right)^{-1}_{lj} y_j = x_0 + h \sum_{j=1}^{i-1} \omega_{ij} y_j$$

The prediction is the posterior mean at the point:

$$\hat{x}\left(t_0 + h\right) = x_0 + \sum_{i=1}^{s} \sum_{j=1}^{s} k^{\partial}\left(t_0 + hc_i, t_0 + hc_l\right) \left(\partial_{k^{\partial}} + \Lambda\right)^{-1}_{lj} y_j = x_0 + h \sum_{i=1}^{s} b_i y_i$$

To obtain $p\left(Y | \dot{X}, \phi^{+}\right)$ it is carried out using an implicit integration:

$$p_{GP}\left(Y | \dot{X}\right) = \prod_{k=1}^{T} p_{GP}\left(y_k | \dot{x}_k\right) \quad , \quad p_{GP}\left(y_k | \dot{x}_k\right) = \int p_{OBS}\left(y_k | x_k\right) p_{GP}\left(x_k | \dot{x}_k\right) dx_k \sim GP\left(\mu_k^{y|\dot{x}}, \Sigma_k^{y|\dot{x}}\right)$$

To simulate samples we can use either Monte Carlo Markov Chain (MCMC), Gibbs sampling, or Metropolis-Hasting. We present an approach for generate from these conditional distributions, the algorithm is summarized:

- Initialize ϕ^0, θ^0 at random and draw $X^0 \sim p_{GP}\left(X|Y,\phi^0\right)$
- For $i = 1 : L$ do
  1. Sample $\theta^i, \phi^i|X^{i-1} \sim p\left(\theta,\phi|X^{i-1},y\right)$,
  2. Sample $X^i \sim p\left(X|\theta^i,\phi^i,Y\right)$.

To simulate samples from parameters:

1. Set $\theta^{i,0} = \theta^{i-1}$, $\phi^{i,0} = \phi^{i-1}$
2. For $j = 1 : L_p$ do
   (a) Sample $\phi^{i,j} \sim p\left(\phi|X^{i-1},\theta^{i,j-1},Y\right)$,
   (b) Sample $\theta^{i,j} \sim p\left(\theta|X^{i-1},\phi^{i,j},Y\right)$,
3. Set $\theta^i = \theta^{i,L_p}$ $\phi^i = \phi^{i,L_p}$.

To simulate samples from $p\left(X|\theta,\phi,Y\right)$ we can use either Metropolis-Hasting, with $X|\theta,\phi,Y \sim p_{GP}\left(X|\phi,Y\right)$ as the proposal see [50].

4.1.4 Implementation

The first phase of the methodology involved the collection of epidemiological data on COVID-19 in Ecuador. Weekly data on confirmed cases was obtained from the official website of the Ministerio de Salud Pública. This data is essential to feed the SIR model and evaluate its performance.

The collected data underwent a rigorous data preparation process. This included data cleaning to remove duplicate or inconsistent entries, interpolation to fill in possible missing data, and data transformation when necessary to achieve normalization.

The SIR model, based on a system of ordinary differential equations describing the dynamics of disease spread, was implemented. Initial conditions of the model were configured, considering the total population and the initial number of infected individuals,

assuming that the estimation is always done at the beginning of the epidemic, meaning the population of recovered individuals is zero.

To incorporate the proposed algorithms into the model, the approach was adapted to the specific needs of epidemic parameter estimation.

The Euler-Maruyama approximation was used to make a first approximation to the dynamic of COVID-19 in Ecuador. The Diffusion Bridge was used to simulate the dynamic of the virus inter-time (at a continuous time), interpolating the real data of the infected population. Kalman filter was used to simulate the dynamic of the virus, through an updating step of the covariance of the error of the simulations generated in previous iterations. Gaussian processes were used to model uncertainty in the parameters of the SIR model, including the infection rate and recovery rate. Appropriate hyperparameters for the Gaussian process were selected, and iterations were performed to fit the model parameters to the observed data.

Comparisons were made with real COVID-19 data in Ecuador, and the model's ability to predict disease spread over a specified time horizon was calculated. The results obtained through the implementation of the model were interpreted in light of observed trends in the spread of COVID-19 in Ecuador. Projected scenarios were examined, and the utility of Gaussian processes in capturing uncertainty in epidemic parameter estimation was assessed.

The limitations of the methodology were acknowledged, including the availability of limited data at certain times, the assumption of homogeneity in virus transmission throughout the population, and the possibility of changes in public health policies affecting disease dynamics.

4.1.5 Testing

When applying the proposed methods, graphs and tables were generated to allow for a comparison of the quality of the generated simulations with actual observations of infected individuals. To quantify the quality of the fit, the calculation of mean squared error and the coefficient of determination were performed. Similarly, several configurations were tested to visualize the impact on predictions, such as setting different values for the beta and gamma parameters, hyperparameters of the Gaussian process, Kalman filter matrices, proposed data variance, among others. This aids in understanding the robustness of the

model. Additionally, uncertainty in predictions was assessed and how it changes when performing sensitivity analysis.

4.2 Model Proposal

The aim of this study is to estimate the SIR epidemic model using the algorithms detailed in Section 4.1.3. This will allow us to assess the effectiveness of these algorithms in making predictions about the dynamics of the COVID-19 virus. In other words, we will evaluate how well these predictions fit the actual data, how results vary with different initial parameter values, and the applicability of these algorithms to more complex epidemiological models, such as the SEIR model, the SIS model, or other models.

4.2.1 Observed data description

The data analyzed for this study were collected from the official website of the Secretaría Nacional de Gestión de Riesgos y Emergencias (https://www.gestionderiesgos.gob.ec/informes-de-situacion-covid-19-desde-el-13-de-marzo-del-2020/). The data represent confirmed COVID-19 cases at the national level, collected on a weekly basis from March 13, 2020, until week 37 of 2023 (a total of 189 weeks). For the total population size (N), it is assumed to be the population of Ecuador at the beginning of 2020, which was approximately 17,600,000 inhabitants. It's important to note that one of the theoretical assumptions is that the model is studied in a closed population, so it is assumed that from 2020 until week 37 of 2023, there was no immigration, emigration, births, or deaths.

4.3 Analysis Method

The results will be analyzed using a graphical method as we aim to visualize the infection dynamics compared to actual observed data. The accuracy of the fit will be quantified by calculating the mean squared error and the coefficient of determination. Alongside the infection dynamics, we will graphically observe the dynamics of the parameters beta and gamma through iterations when applying Gaussian processes. Estimates of these parameters will be calculated using the mean of all values produced until they reach their

equilibrium point (when their dynamics remain nearly constant). Additionally, maximum likelihood estimation will be used to provide an initial estimate of the SIR model parameters, compared to an initial random simulation of the infected population.

As will be seen in section 5, a study of three scenarios based on observed data is conducted: the first scenario includes all the data (referred to as the "full group"), the second scenario is formed by considering data from week 1 to week 90 (referred to as the "first group"), and the third scenario is formed by considering data from week 91 to week 189. Therefore, the analysis of the results is conducted based on these three scenarios (except in Euler-Maruyama approximation).

4.4 Experimental Setup

The application of the proposed algorithms is carried out with the assistance of the R software, for which we utilize libraries such as "deSolve", "stats", "MASS", "mvtnorm" and "ggplot2". These libraries are focused on solving ordinary differential equations (ODEs), handling multivariate distributions (for the use of Gaussian processes), and creating custom graphics.

4.4.1 Parameters

For each proposed scenario, the first step is to solve the ODE of the SIR model using the "deSolve" package in R. With this initial approximation, parameter optimization for beta and gamma is performed through Maximum Likelihood Estimation (MLE) to be used as initial parameters in each scenario, resulting in the following values:

- **First scenario:** beta = 0.2, gamma = 0.07
- **Second scenario:** beta = 0.3, gamma = 0.1
- **Third scenario:** beta = 0.5, gamma = 0.11

4.4.2 Euler-Maruyama Approximation

The application of this method is carried out for four scenarios with different parameters. Since this method indicates the dynamics of the S, I, and R populations, given initial values

of beta and gamma, only the initial data of these populations are needed. So, we use the following data:

- **First scenario:** N = 1000, S0 = 999, I0 = 1, R0 = 0, beta = 0.5, gamma = 0.1
- **Second scenario:** N = 1000, S0 = 999, I0 = 1, R0 = 0, beta = 5, gamma = 0.01
- **Third scenario:** N = 1000, S0 = 999, I0 = 1, R0 = 0, beta = 0.17, gamma = 0.1
- **Fourth scenario:** N = 1000, S0 = 999, I0 = 1, R0 = 0, beta = 0.1, gamma = 0.1

In all scenarios, it's used a total time (T) of 100 and a time-step (dt) equal to 0.1.

4.4.3 Diffusion Bridge Approximation

The application of this method is carried out for the three scenarios of observed data described before, using the values of beta and gamma found with MLE and the observed data from each group. The values of the other parameters for this algorithm are as follows:

- **Total population:** $N = 17,600,000$
- **Initial susceptible population:** $S0 = N - Y[1]$, where $Y[1]$ is the first real observation from the group.
- **Initial infected population:** $I0 = Y[1]$, where $Y[1]$ is the first real observation from the group.
- **Initial recovered population:** $R0 = 0$ (start of the epidemic).
- **Study time:** $times = length(Y)$, where $length(Y)$ is the number of observations in the group.

4.4.4 Kalman filter

The application of this method is carried out for the three scenarios of observed data described above, using the values of beta and gamma found using MLE and the observed data from each group. The values of the other parameters for this algorithm are as follows:

- **Total population:** $N = 17,600,000$

- **Initial susceptible population:** $S0 = N - Y[1]$, where $Y[1]$ is the first real observation from the group.
- **Initial infected population:** $I0 = Y[1]$, where $Y[1]$ is the first real observation from the group.
- **Initial recovered population:** $R0 = 0$ (start of the epidemic).
- **Study time:** $times = length(Y)$, where $length(Y)$ is the number of observations in the group.
- **Proposed variance:** $obs\_variance = var(Y)$, where $var(Y)$ is the variance of the observed data. This variance will be used for the proposal of the matrix R.
- **State transition matrix:** it is constructed using the differential equations of the SIR model, with dt = 1:

$$A = \begin{bmatrix} 1 - \beta * dt & \beta * dt & 0 \\ 0 & 1 - \gamma * dt & \gamma * dt \\ 0 & 0 & 1 \end{bmatrix}$$

  This matrix helps predict the state of the system at the next time step from the current state.
- **Process covariance matrix:** represents the covariance of the process and describes the uncertainty in the model predictions. In our case, we set Q as a diagonal matrix of dimension $3x3$, where its entire diagonal has the values of the variance of the differences in the observed data (constant matrix as the initial proposal).
- **Observation matrix:** specifies how the state of the system relates to the observations. In our case, since we only have the values of the infected population, H is a matrix of dimension $nx3$, where n is the total number of observations in the group, and its second column has values of 1, while the other columns have a value of 0. This indicates that the states are entirely dependent on the observed infected data.
- **Observation covariance matrix:** represents the covariance of the observations and describes the uncertainty in the measurements made. In our case, R is a diagonal

matrix of dimension nxn, where n is the total number of observations in the group, and its entire diagonal has the value of the variance of the observations.

- **Estimation error covariance matrix:** represents the covariance of the error in estimating the latent states of the system. This matrix is updated over time, depending on how well the current latent state fits, and is used for the next iteration. In our case, as the initial value of P, we set P as a diagonal matrix of dimension $3x3$, where its entire diagonal has a large uncertainty value ($1e6$) to reflect high uncertainty.

4.4.5 Gaussian process

The application of this method is carried out for the three scenarios of observed data described above, using the values of beta and gamma found using MLE and the observed data from each group. The values of the other parameters for this algorithm are as follows:

Hiperparameters

The key hyperparameters in our model are ”length” and ”standard deviation.” The ”length” determines the spatial scale over which the Gaussian process can vary, while the ”standard deviation” controls the degree of variation around the predicted mean.

In the experimentation phase, we assigned a value of 10 to ”length” to allow the Gaussian process to capture patterns over a broader spatial interval. This choice was supported by our consideration that the spread of diseases may be influenced by geographic or demographic factors over long distances.

Regarding the ”standard deviation,” we opted for a value of 0.1 to indicate a moderate level of uncertainty. This reflects our assumption that, while there may be variations, we do not expect extremely abrupt changes in the spread of the disease.

To determine these values optimally, we employed optimization techniques. We used the ”L-BFGS-B” optimization method to find values that minimize the sum of squared errors between the model predictions and the observed data. Additionally, we implemented constraints on the allowed values to ensure solutions that are realistic and consistent with the epidemiological context.

This iterative process of hyperparameter tuning and optimization allowed us to find

combinations that best fit the observed data and reflect our expectations about the behavior of the underlying process. The final choice of hyperparameters is supported not only by numerical optimization but also by problem interpretation and empirical validation of the model predictions.

Mean function and Covariance matrix

We set the mean function to be $m(t) = 0$ for simplicity, since we expect the infected population to reach its threshold and then decrease until it dissapears.

For the covariance matrix, we use the optimized hiperparameters as follows

$$K(t, t') = \sigma^2 \exp\left(-\frac{(t-t')^2}{2l^2}\right)$$

where σ is the optimized standard deviation, l is the optimized length and t and t' are two points in time.

Other parameters

- **Total population:** $N = 17,600,000$
- **Initial susceptible population:** $S0 = N - Y[1]$, where $Y[1]$ is the first real observation from the group.
- **Initial infected population:** $I0 = Y[1]$, where $Y[1]$ is the first real observation from the group.
- **Initial recovered population:** $R0 = 0$ (start of the epidemic).
- **Study time:** $times = length(Y)$, where $length(Y)$ is the number of observations in the group.
- **Number of iterations:** $num\_iterations = 10000$
- **Proposal beta:** we choose a normal distribution, with mean its previous value, and standard deviation equal to the standard deviation proposed.
- **Proposal gamma:** we choose a normal distribution, with mean its previous value, and standard deviation equal to the standard deviation proposed.

Chapter 5

Results and Discussion

In this chapter, We will show the results and discussions. In order to do this, we divide the studies into the following sections: results and discussions of the observed data, Euler-Maruyama approximation, Diffusion Bridge approximation, Kalman filter, and Gaussian process.

5.1 Observed data

Figure 5.1 shows the observed data

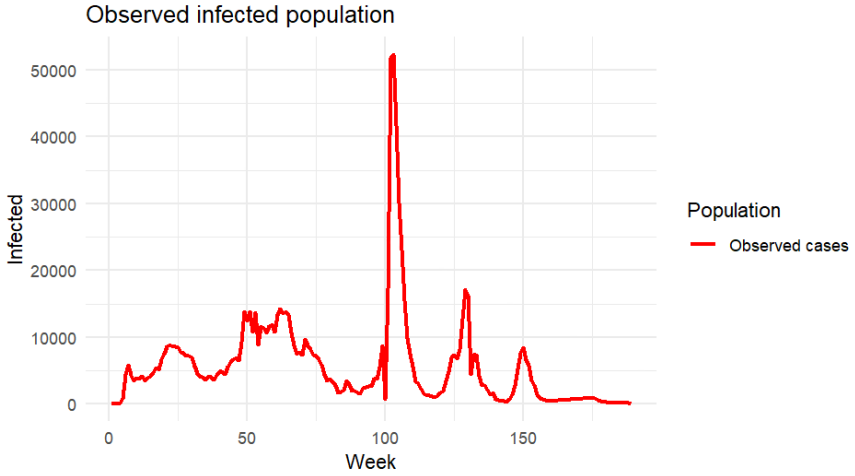


Figure 5.1: Observed Data (complete group)

In this figure, we can observe the weekly dynamics of the COVID-19 virus at the national level from March 13, 2020, to week 37 of year 2023. Around week 100, a peak of rapid growth and decline can be seen. This is because historically, during these weeks

(around December 2021), the Omicron variant of the virus emerged, which had a high infection rate but a very short infection period. For this reason, the data was divided into two groups: the first group spans from week 1 to week 90, and the second group spans from week 91 to week 189. This division is made considering that graphically, the infection rate and recovery rate differ in each group.

It should be noted that there were more variants of the COVID-19 virus, but a significant difference in the infected population is only observed with the Omicron variant. The graphs of these new data are shown below:

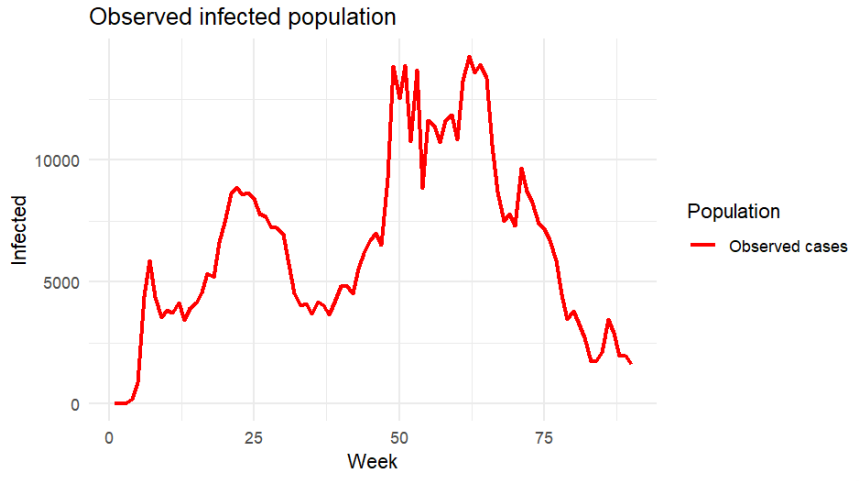


Figure 5.2: Observed Data (group 1)

Figure 5.2 displays the observations of infected individuals in Group 1 (from week 1 to week 90). It can be observed that the highest peak occurs around week 60, with a value of approximately 15,000. This illustrates the initial dynamics of the disease when there was no vaccine available, resulting in a rising number of infected individuals. However, as the vaccination campaign began (around week 60), the population of infected individuals started to decrease significantly.

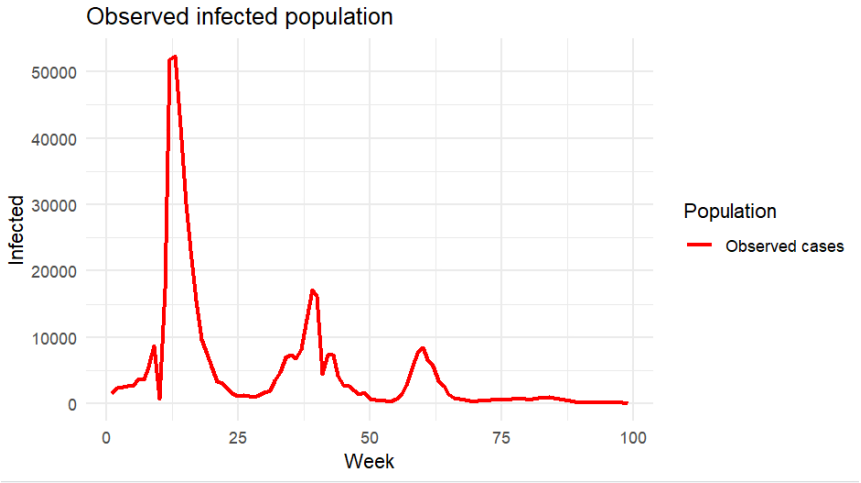


Figure 5.3: Observed Data (group 2)

Figure 5.3 depicts the observations of infected individuals in Group 2 (from week 91 to week 189). It can be observed that the highest peak occurs around week 12 (week 102 in the complete data), with a value of approximately 53,000. This illustrates the initial dynamics of the disease, which marked the onset of this highly contagious variant, resulting in a rapid increase in the population of infected individuals, but with a short duration, leading to a swift decline in the population of infected individuals.

The graph of observed infections demonstrates that the biosecurity measures implemented in the country kept infection peaks low (except during the onset of the Omicron variant). For example, measures such as border closures and travel restrictions, temporarily closing air, land, and sea borders, preventing entry and exit from the country; curfews and weekend movement restrictions, which significantly reduced social interaction; mandatory mask-wearing in both outdoor (initially) and indoor settings (where social distancing couldn't be maintained); the rapid distribution of vaccines, with the population receiving up to four doses; among other measures, had an impact on containing the spread of the COVID-19 virus.

5.2 Euler-Maruyama Approximation

As defined earlier, the Euler-Maruyama method allows us to understand and model the spread of infectious diseases in the SIR model, addressing the dynamics of the disease in a population over a specific period of time. This algorithm demonstrates that the dynamics

of susceptible, infected, and recovered populations depend entirely on the parameters beta (infection rate) and gamma (recovery rate).

To illustrate the effects of the Euler-Maruyama method, four scenarios have been created, each showcasing different dynamics in the susceptible, infected, and recovered populations.

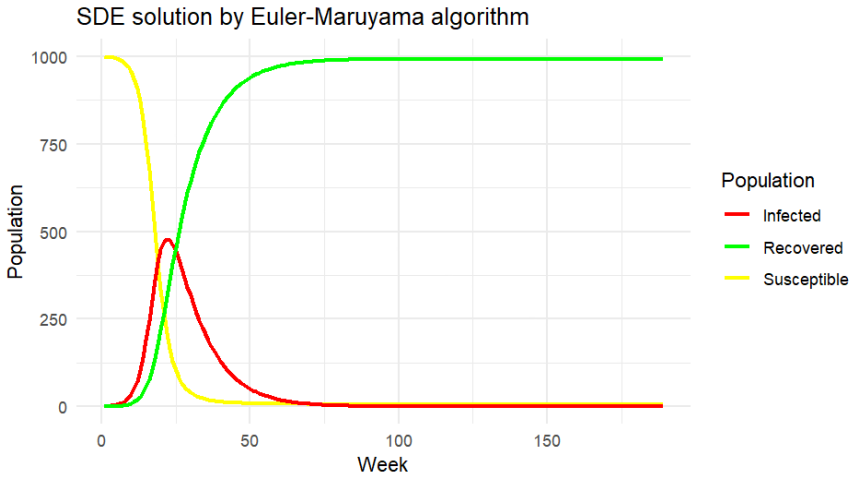


Figure 5.4: Euler-Maruyama Approximation (Scenario 1)

Figure 5.4 depicts the Euler-Maruyama approximation with a total population of $N = 1000$ individuals. The infection rate (beta) is 0.5, the recovery rate (gamma) is 0.1, and the initial population is $S = 999$, $I = 1$, $R = 0$.

This scenario shows a dynamic where the disease infects the entire population, concluding the infection when there are no more susceptible individuals to infect, and the entire infected population starts recovering at a rate gamma until the entire population is recovered. This scenario demonstrates that $R0 = beta/gamma = 5 > 1$, making the disease a large-scale epidemic.

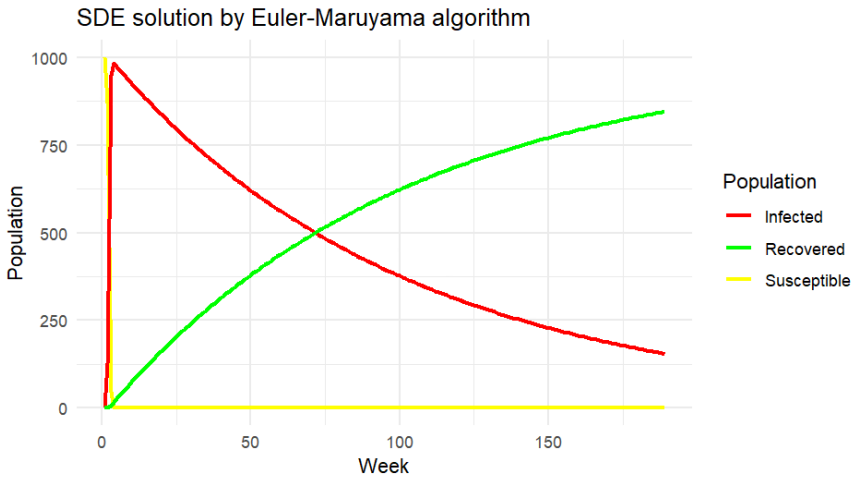


Figure 5.5: Euler-Maruyama Approximation (Scenario 2)

Figure 5.5 depicts the Euler-Maruyama approximation with a total population of $N = 1000$ individuals. The infection rate (beta) is 5, the recovery rate (gamma) is 0.01, and the initial population is $S = 999$, $I = 1$, $R = 0$.

This scenario shows a dynamic where the disease infects the entire population at a high speed due to its high infection rate, and its recovery is slower because the recovery rate is smaller. This scenario demonstrates that $R0 = beta/gamma = 500 > 1$, making the disease a large-scale epidemic.

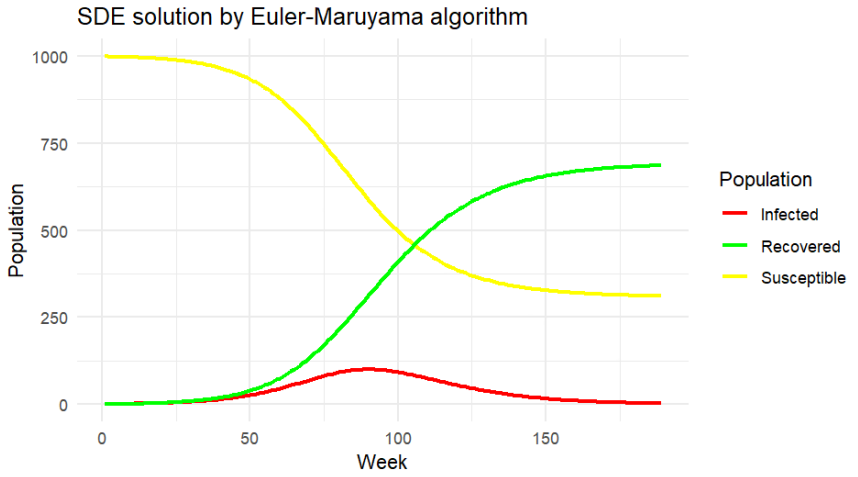


Figure 5.6: Euler-Maruyama Approximation (Scenario 3)

Figure 5.6 shows the Euler-Maruyama approximation with a total population of $N = 1000$ individuals. The infection rate (beta) is 0.17, the recovery rate (gamma) is 0.1, and the initial population is $S = 999$, $I = 1$, $R = 0$.

This scenario illustrates a dynamic in which the disease infects more than half of the population but does not completely infect the susceptible population. This scenario demonstrates that $R0 = beta/gamma = 1.7 > 1$, making the disease a large-scale epidemic but eventually coming to an end without infecting the entire susceptible population.

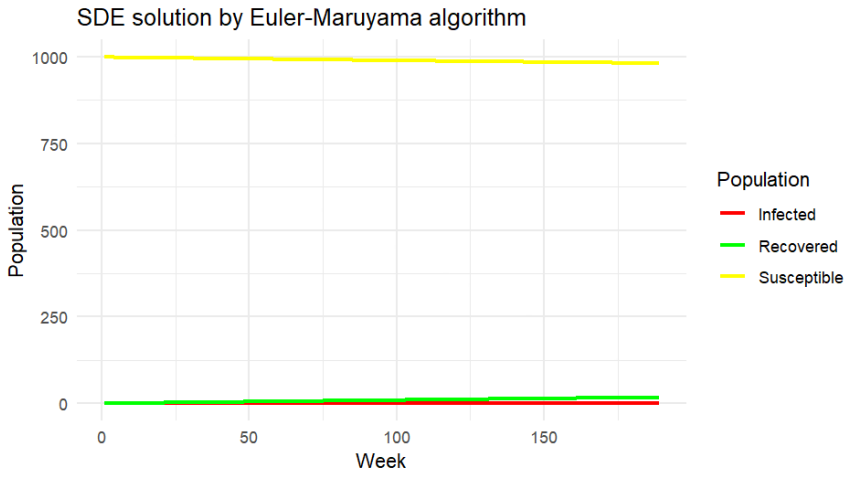


Figure 5.7: Euler-Maruyama Approximation (Scenario 4)

Figure 5.7 shows the Euler-Maruyama approximation with a total population of $N = 1000$ individuals. The infection rate (beta) is 0.1, the recovery rate (gamma) is 0.1, and the initial population is $S = 999$, $I = 1$, $R = 0$.

This scenario illustrates a dynamic in which the disease affects a few susceptibles and quickly comes to an end. This scenario demonstrates that $R0 = beta/gamma = 1 \leq 1$, meaning the disease does not become a large-scale epidemic. This is because the infection rate is equal to the recovery rate, causing the disease to extinguish immediately.

In all four scenarios presented, the following can be observed: the dynamics depend entirely on the infection rate (beta) and the recovery rate (gamma). The susceptible population is always decreasing, based on the assumption that once an infected individual recovers, they acquire immunity to the disease, which prevents them from returning to the susceptible population. The recovered population is increasing, following the same assumption as for susceptibles. This population can be easily determined by subtracting the total population from the sum of susceptibles and infected individuals (this keeps the total population constant). Even when the susceptible population reaches zero, there is still dynamic interaction between the infected and recovered populations. The dynamics

of all three populations only end when the infected population reaches zero, indicating the disease has been extinguished.

Additionally, one can identify the maximum peaks of the disease and the time required for the number of recovered individuals to reach a significant value. These results are important for understanding disease dynamics and the effectiveness of control measures. The graph provides valuable information about disease spread in the population under specified conditions and parameters. These findings can be useful for assessing the effectiveness of control measures such as quarantine or vaccination, identifying the critical point at which an infection peak is reached, and understanding how parameter variations affect disease dynamics.

Despite the advantages of this method, it's essential to consider that this study is based on simplifications of the SIR model and specific assumptions, which come with several limitations. These limitations include not accounting for geographic dynamics, assuming constant parameters throughout the simulation, and neglecting seasonal variation or external fluctuations, among others.

5.3 Diffusion Bridge Approximation

As mentioned earlier, the Diffusion Bridge method simulates inter-time data in such a way that the simulated data interpolates with the observed data. This approach allows for a more precise modeling and estimation of disease dynamics, taking into account the uncertainty in the observed data. In this case, we have three scenarios.

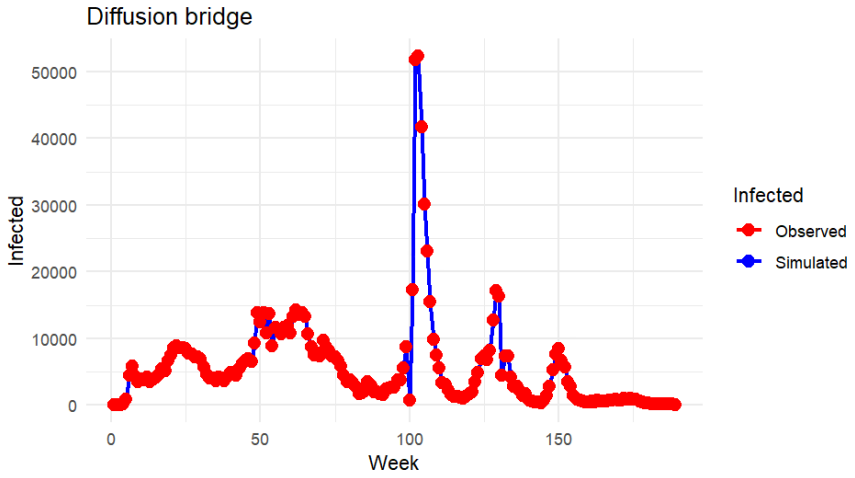


Figure 5.8: Diffusion Bridge Approximation (Scenario 1)

Figure 5.8 shows the Diffusion Bridge approximation for the complete group of infected individuals (189 weeks). The real total population is used ($N = 17,600,000$). The infection rate (beta) is 0.2, and the recovery rate (gamma) is 0.07, with initial populations of $S = 17,599,999$, $I = 1$, and $R = 0$. The chosen variance is the variance of the observed data.

We can observe the dynamics of the infected population spanning all the weeks in question. The Diffusion Bridge method simulated infected data between time points (between each week) in a way that matches (interpolates) the observed infected data points.

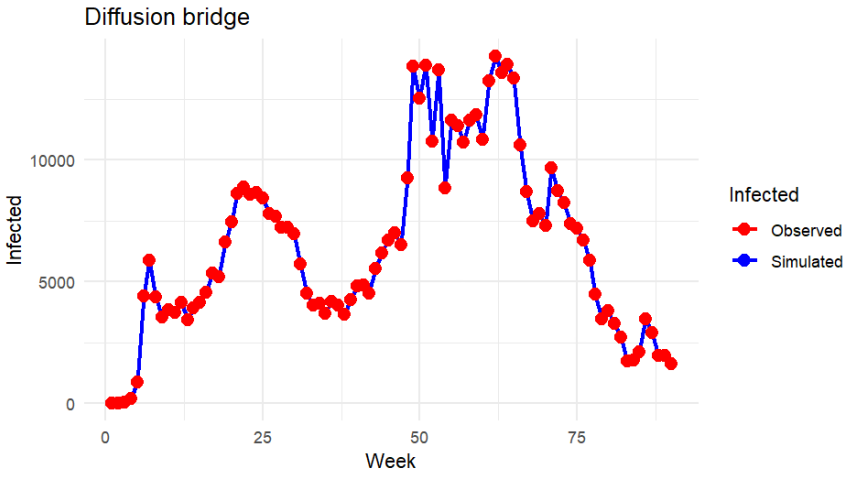


Figure 5.9: Diffusion Bridge Approximation (Scenario 2)

Figure 5.9 displays the Diffusion Bridge approximation for the first group of infected individuals (from week 1 to week 90). We used the actual total population ($N = 17,600,000$). The infection rate (beta) is 0.3, and the recovery rate is 0.1, with initial populations of

$S = 17,599,999$, $I = 1$, and $R = 0$. The chosen variance is the variance of the observed data.

We can observe the dynamics of the infected population within the first group of infections (from week 1 to week 90). The Diffusion Bridge method simulated infected data between time points (between each week) in a way that matches (interpolates) the observed infected data points.

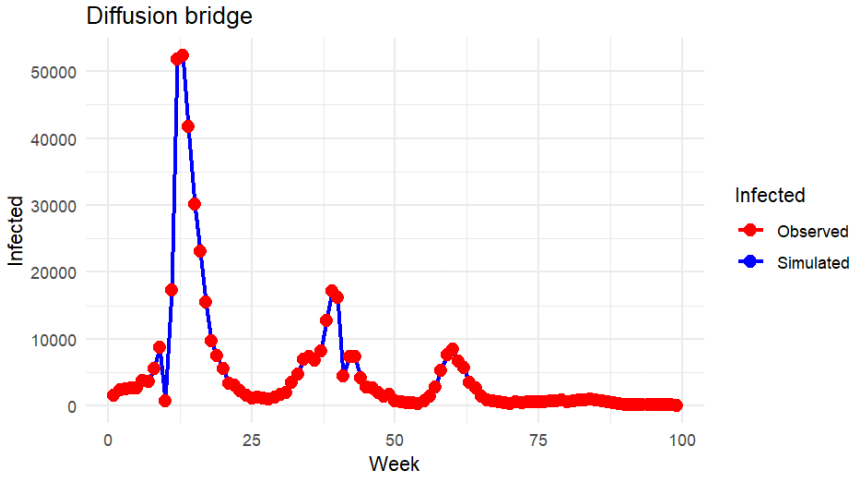


Figure 5.10: Diffusion Bridge Approximation (Scenario 3)

Figure 5.10 shows the Diffusion Bridge approximation for the second group of infected individuals (from week 91 to week 189). We used the actual total population ($N = 17,600,000$). The infection rate (beta) is 0.5, and the recovery rate is 0.11, with initial populations of $S = 17,598,439$, $I = 1,561$, and $R = 0$. The chosen variance is the variance of the observed data.

We can observe the dynamics of the infected population within the second group of infections (from week 91 to week 189). The Diffusion Bridge method simulated infected data between time points (between each week) in a way that matches (interpolates) the observed infected data points.

In all three cases presented, the application of the Diffusion Bridge method shows a good approximation of the trajectory followed by the real observed data. The simulated trajectories are consistent with the observations and capture the inherent variability in the disease's spread. In all three cases, it can be observed that the simulation's trajectory interpolates the points of the observed infected individuals.

In the three scenarios presented, it can be observed that the use of the Diffusion Bridge method creates a good approximation in simulating the infected population, interpolating each of the real observations. With this method, an approximation of the inter-time infection behavior can be obtained, meaning the dynamics of latent observations since their dynamics are a continuous process. However, in reality, data cannot be collected at every infinitesimal time step. In fact, the data collected in this study have a one-week time step, which introduces significant variability in the data and results in a loss of information.

The results of this method are valuable for assessing the accuracy of estimates and the capability of the Diffusion Bridge method to model disease dynamics, identify possible gaps or significant deviations between estimates and observed data, and understand the associated uncertainty in the estimates of the infected population.

Similar to the Euler-Maruyama method, this method is subject to various limitations due to simplifications of the SIR model. These limitations include the dependence on model parameters (beta and gamma), the assumption of a probability distribution for estimation, and the need for high-quality observed data and proper error modeling.

5.4 Kalman filter

The implementation of the Kalman filter allows us to simulate the population of infected individuals with great accuracy in its propagation dynamics, taking into account both the observed data and the associated uncertainty. In this case, we have three scenarios.

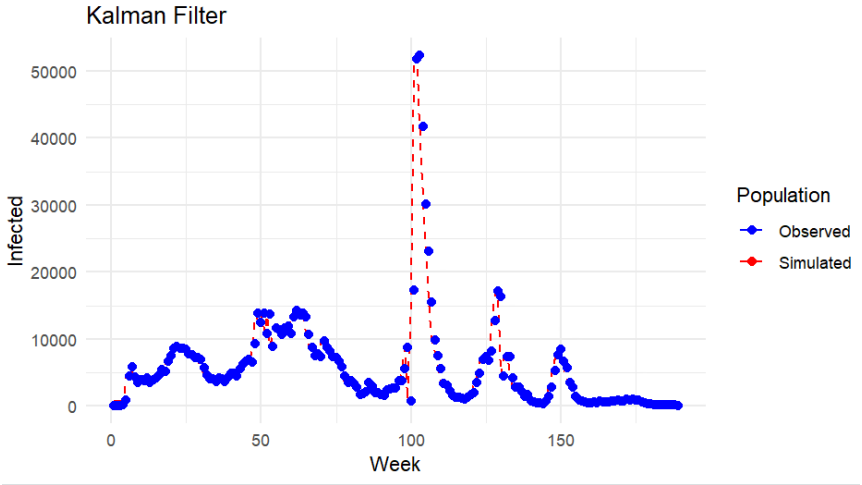


Figure 5.11: Kalman Filter (Scenario 1)

Figure 5.11 shows the Kalman filter for the complete group of infected individuals (189 weeks). The actual total population is considered ($N = 17600000$). The infection rate (beta) is 0.2, and the recovery rate is 0.07, with initial populations of $S = 17599999$, $I = 1$, and $R = 0$. The variance is chosen as the variance of the observed data.

The dynamics of the infected population over all the weeks in question are observed. The Kalman filter method simulated the dynamics of the observed data, coming very close to passing through the actual points of infected individuals.

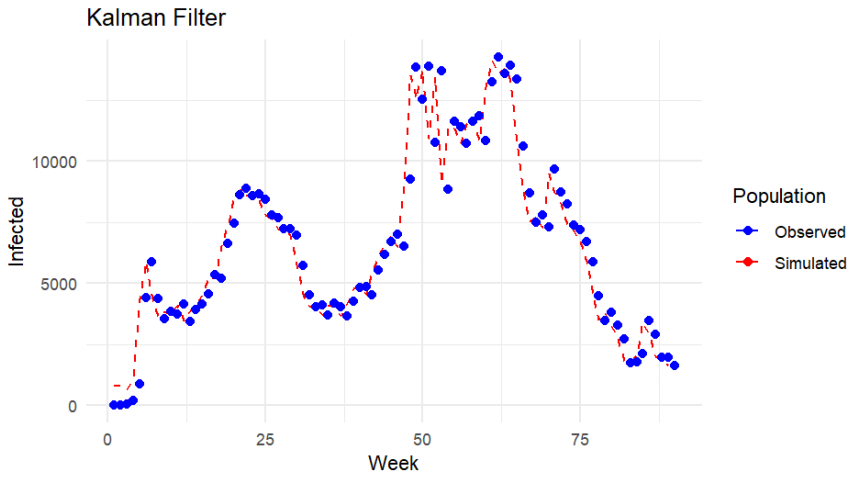


Figure 5.12: Kalman Filter (Scenario 2)

Figure 5.12 shows the Kalman filter for the first group of infected individuals (from week 1 to week 90). The actual total population is considered ($N = 17600000$). The infection rate (beta) is 0.3, and the recovery rate is 0.1, with initial populations of $S = 17599999$, $I = 1$, and $R = 0$. The variance is chosen as the variance of the observed data.

The dynamics of the infected population covering the first group of infection (from week 1 to week 90) are observed. The Kalman filter method simulated the dynamics of the observed data, coming very close to passing through the actual points of infected individuals.

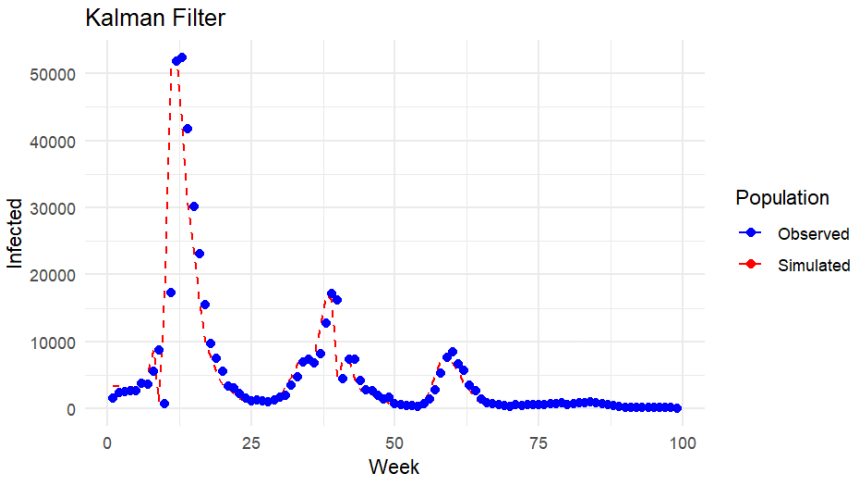


Figure 5.13: Kalman Filter (Scenario 3)

Figure 5.13 shows the Kalman filter for the second group of infected individuals (from week 91 to week 189). The actual total population is considered ($N = 17600000$). The infection rate (beta) is 0.5, and the recovery rate is 0.11, with initial populations of $S = 17598439$, $I = 1561$, and $R = 0$. The variance is chosen as the variance of the observed data.

The dynamics of the infected population covering the second group of infection (from week 91 to week 189) are observed. The Kalman filter method simulated the dynamics of the observed data, coming very close to passing through the actual points of infected individuals.

In all three cases presented, it is observed that the simulations created by the Kalman filter are consistent with the observed data and reflect the uncertainty in the spread of the disease.

In the three scenarios shown, the Kalman filter managed to create simulations that closely match reality. Through updates in the simulations and the covariance matrix of estimation error, an equilibrium (convergence) can be reached, resulting in simulations that closely align with reality. This takes into account the variability of the observed data and its relationship with the estimated SIR model.

The results of the simulations of infected individuals in the SIR model provide valuable insights into the spread of infectious diseases and the ability of the Kalman filter to model this dynamic. This is important for assessing the accuracy of the simulations and the

Kalman filter's ability to fit the observed data, understanding the associated uncertainty in the simulations and its impact on future projections, and evaluating how changes in model parameters affect the simulations and disease spread.

Limitations of this method include its dependence on model parameters (beta and gamma), the need for high-quality observed data and proper error modeling, and the assumption that the SIR model is an accurate representation of the disease in question.

5.5 Gaussian Process

The application of the Gaussian process for estimations in SDEs is a very powerful tool. Through this method, simulations of the infected population were successfully conducted to observe its dynamics, taking into account the observed data and associated uncertainty. Additionally, it allows us to track the evolution of the parameters beta and gamma over iterations until they reach an equilibrium. For the results of this method, the constant value of the standard deviation of phi is initially set as a first approximation to its value ($proposal\_sd = 0.1$). The following are three scenarios presented.

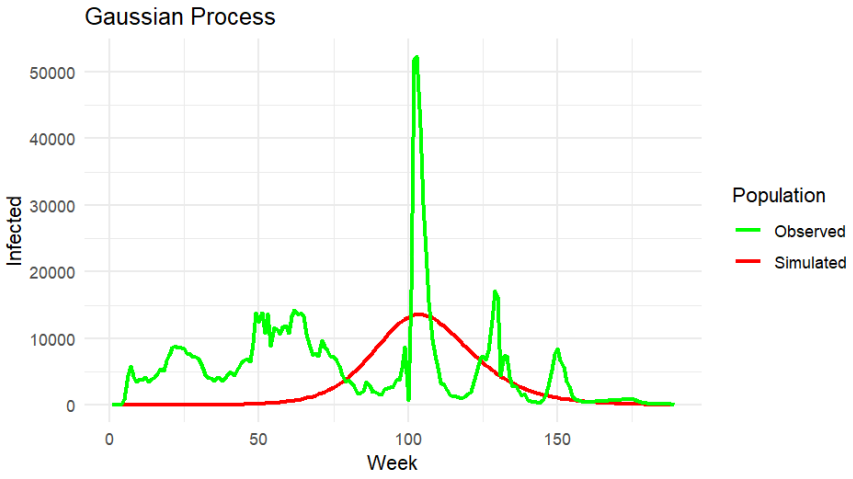


Figure 5.14: Gaussian Process (Scenario 1)

Figure 5.14 shows the Gaussian process applied to the entire group of infected individuals (189 weeks). The real total population is used ($N = 17,600,000$). The initial infection rate (beta) is 0.2, and the initial recovery rate is 0.07, with initial populations $S = 17,599,999$, $I = 1$, and $R = 0$.

The simulation displays the dynamics of the entire group of infected individuals. The simulation demonstrates significant growth around week 100 of the data. However, it needs further adjustment to match the true data better, which can be achieved by improving the initial parameter values or the proposed standard deviation of phi.

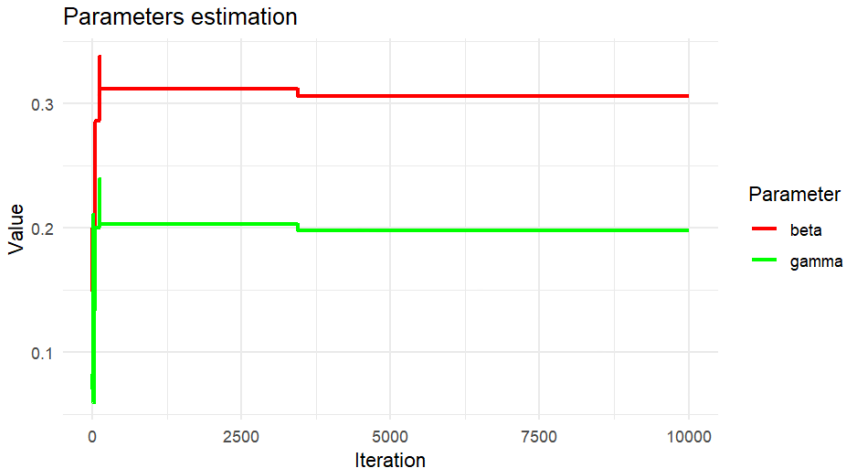


Figure 5.15: Beta and gamma evolution (Scenario 1)

mean_beta	mean_gamma
0.3072605	0.1992114

Table 5.1: Beta and gamma mean value (Scenario 1)

Figure 5.15 displays the evolution of beta and gamma over the iterations. The values of beta and gamma reach an equilibrium almost immediately, around 100 iterations. Table 5.1 presents the mean values of beta and gamma. According to this data, the initial values of 0.2 and 0.07, respectively, were a good initial assumption, with an increase of 0.11 in beta and 0.13 in gamma.

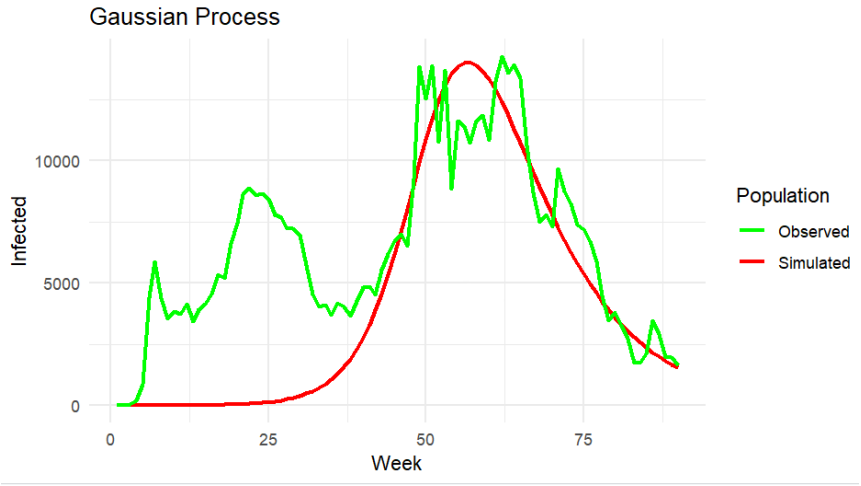


Figure 5.16: Gaussian Process (Scenario 2)

Figure 5.16 shows the Gaussian process for the first group of infected individuals (from week 1 to week 90). The actual total population is used ($N = 17,600,000$). The initial infection rate (beta) is 0.3, and the initial recovery rate (gamma) is 0.1, with initial populations $S = 17,599,999$, $I = 1$, and $R = 0$.

The dynamics of the simulation for the first group of infected individuals are observed. The simulation demonstrates a similar behavior to the observed data from around week 45 onwards, although it needs further adjustments to match the actual data. This can be achieved by improving the initial parameter values or the proposed standard deviation of phi.

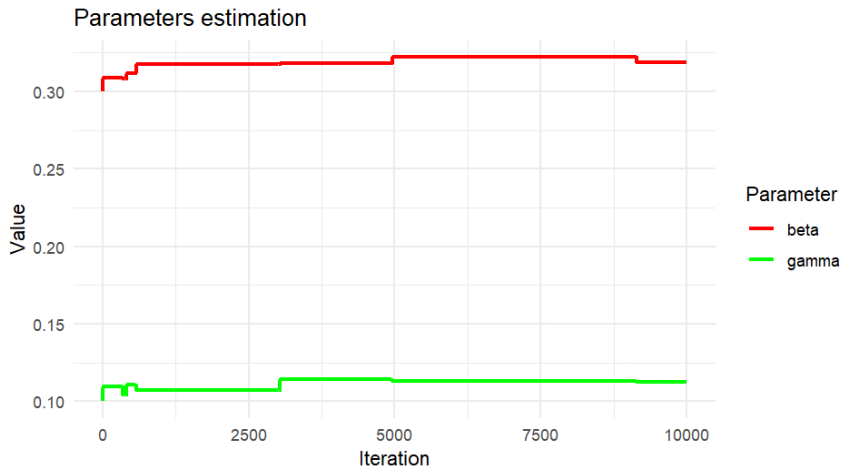


Figure 5.17: Beta and gamma evolution (Scenario 2)

mean_beta	mean_gamma
0.3193035	0.1115379

Table 5.2: Beta and gamma mean value (Scenario 2)

Figure 5.17 displays the evolution of beta and gamma throughout the iterations. The values of beta and gamma exhibit nearly constant behavior throughout the process. Table 5.2 shows the mean of beta and gamma. According to this data, the initial values of 0.3 and 0.1, respectively, were good initial assumptions, with an increase of 0.1 in both parameter values.

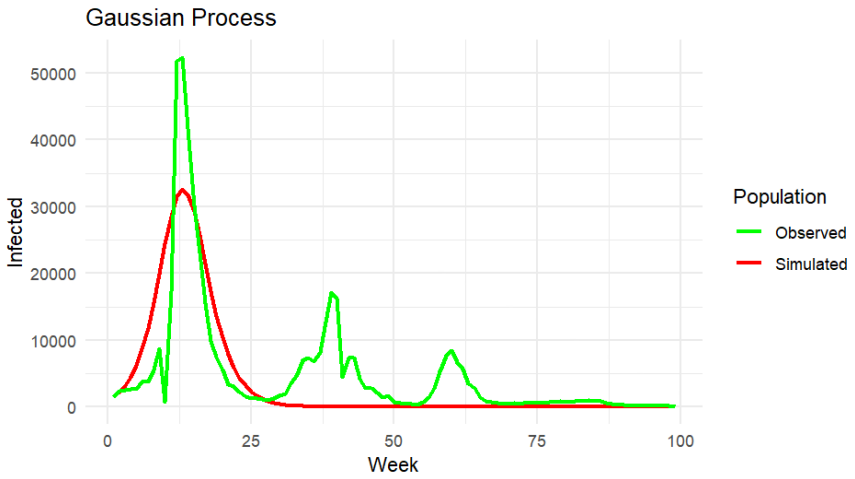


Figure 5.18: Gaussian Process (Scenario 3)

Figure 5.18 shows the Gaussian process applied to the second group of infected individuals (from week 91 to week 189). The real total population ($N = 17,600,000$) is considered. The initial infection rate (beta) is 0.5, the initial recovery rate (gamma) is 0.11, and the initial populations are $S = 17,598,439$, $I = 1,561$, and $R = 0$.

The dynamics of the simulation of the second group of infected individuals are observed. The simulation exhibits similar high growth behavior before week 115 (week 25 according to the graph), although it needs further adjustment to match the true data. This can be achieved by improving the initial parameter values or the proposed standard deviation of phi.

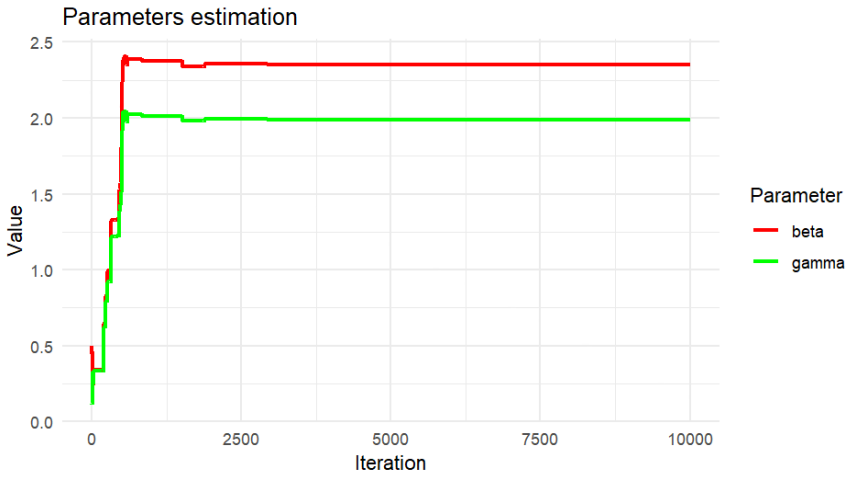


Figure 5.19: Beta and gamma evolution (Scenario 3)

mean_beta	mean_gamma
2.279166	1.930900

Table 5.3: Beta and gamma mean value (Scenario 3)

Figure 5.19 displays the evolution of beta and gamma over the iterations. Beta and gamma values reach their equilibrium after 500 iterations. The graph in Table 5.3 represents the mean of beta and gamma. According to these data, the initial values of 0.5 and 0.11, respectively, were not good initial assumptions, as there is an increase of 1.8 in the value of both parameters.

In the three scenarios presented, despite the simulations not fitting reality as closely, they show the trends of high peaks in each case and their decreasing behavior toward the end of the process. The dynamics of the parameters reveal the trend of their values until reaching an equilibrium point, which could provide an initial estimation of the parameters in the study population to improve the quality of the fit.

The results of the estimation in the simple SIR model using the Gaussian process provide a detailed view of the spread of infectious diseases. These findings can be valuable for assessing the accuracy of the estimates and the Gaussian process's ability to fit the observed data, understanding the associated uncertainty in the estimates and its impact on future projections, and evaluating how changes in model parameters affect the estimates and disease spread, as well as the trend of parameter values.

Like the previous methods, this method has limitations, including its dependence on

model parameters (beta and gamma), the need for high-quality observed data and proper error modeling, and the assumption that the SIR model is an accurate representation of the disease in question.

Chapter 6

Conclusions

In conclusion, the Euler-Maruyama approach implemented for the SIR model provides a valuable initial insight into the spread of infectious diseases in a population, such as COVID-19. The Diffusion Bridge approach offers a powerful tool for estimating the infected population in the SIR model. The results enable a deeper understanding of the spread of infectious diseases and the assessment of estimation quality. The use of the Kalman filter to perform simulations of infected individuals in the SIR model provides a valuable tool for understanding and estimating the spread of infectious diseases. The obtained results allow for a detailed assessment of the simulations and the evaluation of uncertainty in these projections. The use of the Gaussian process for estimation in the SIR model provides a valuable tool for understanding and estimating the spread of infectious diseases. The obtained results allow for a detailed assessment of estimations and the evaluation of uncertainty in these projections.

All of these results contribute to a better understanding of the dynamics of the disease in a given population and can be useful in making decisions related to public health and resource planning.

This study provides a solid insight into the SIR model applied to the spread of COVID-19 in Ecuador, as well as a deep understanding of the implementation of Gaussian processes in parameter estimation. Some of the future work identified includes:

- Using other epidemiological models such as the SEIR model (susceptible-exposed-infected-recovered) or stochastic compartmental models.
- Consideration of external factors such as the impact of control measures, virus vari-

ants, and other factors that allow for the evaluation of effective intervention strategies.

- Expanding the temporal focus of the research to make longer-term predictions about the evolution of the pandemic to improve long-term planning and decision-making.
- Optimization of control strategies, which would assess different combinations of public health measures and their impact on mitigating virus spread.
- Applying these methodologies to the study of other infectious diseases that exhibit similar dynamics, such as influenza, dengue, Ebola, to enhance understanding and response to future epidemiological threats.

In summary, future work should be focused on refining models, collecting additional data, and developing effective control strategies, contributing to preparedness and response to future health emergencies.

Bibliography

[1] A. McKendrick, "Applications of mathematics to medical problems," *Proceedings of the Edinburgh Mathematical Society*, vol. 44, pp. 98–130, 1925.

[2] W. Kermack and A. McKendrick, "A contribution to the mathematical theory of epidemics," *Proceedings of the Royal Society of London*, vol. 115, no. 772, pp. 700–721, 1927.

[3] M. Bartlett, "Some evolutionary stochastic processes," *Journal of the Royal Statistical Society*, vol. 11, no. 2, pp. 211–229, 1949.

[4] R. Anderson and R. May, *Infectious diseases of humans; dynamic and control.* Oxford University Press, 1991.

[5] H. Andersson and T. Britton, *Stochastic Epidemic Models and Their Statistical Analysis*, 1st ed. Springer, 2000.

[6] D. Daley and J. Gani, *Epidemic Modelling: an introduction.* Cambridge University Press, 1999.

[7] C. Gardiner, *Handbook of Stochastic Methods: For Physics, Chemistry and the Natural Sciences*, 2nd ed. Springer Berlin Heidelberg, 1985.

[8] G. Lawler, *Introduction to Stochastic Processes*, 2nd ed. Chapman Hall, 2006.

[9] C. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed. Springer New York, NY, 1999.

[10] H. S. D. D. A. V. A Gelman, J Carlin and D. Rubin, *Bayesian Data Analysis*, 3rd ed. CRC Press, 2013.

[11] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[12] G. Gibson, "Markov chain monte carlo methods for fitting spatiotemporal stochastic models in plant epidemiology," *Journal of the Royal Statistical Society Series C: Applied Statistics*, vol. 46, pp. 215–233, 1997.

[13] G. Gibson and E. Renshaw, "Estimating parameters in stochastic compartmental models using markov chain methods," *Mathematical Medicine and Biology: A Journal of the IMA*, vol. 15, pp. 19–40, 1998.

[14] P. O'Neill and G. Roberts, "Bayesian inference for partially observed stochastic epidemics," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 162, pp. 121–129, 1999.

[15] J. C. M Capistrán and J. Velasco-Hernández, "Towards uncertainty quantification and inference in the stochastic sir epidemic model," *Mathematical Biosciences*, vol. 240, pp. 250–259, 2012.

[16] S. B. H El Maroufy, T Kernane and A. Ouddadj, "Bayesian inference for nonlinear stochastic sir epidemic model," *Journal of Statistical Computation and Simulation*, vol. 86, pp. 2229–2240, 2014.

[17] J. D. M Li and B. Bolker, "Fitting mechanistic epidemic models to data: A comparison of simple markov chain monte carlo approaches," *Statistical Methods in Medical Research*, vol. 7, no. 27, 2018.

[18] F. C. L Tung and M. Suchard, "Direct likelihood-based inference for discretely observed stochastic compartmental models of infectious disease," *The Annals of Applied Statistics*, vol. 12, no. 3, pp. 1993–2021, 2018.

[19] P. N. T Kypraios and D. Prangle, "A tutorial introduction to bayesian inference for stochastic epidemic models using approximate bayesian computation," *Mathematical Biosciences*, vol. 287, no. 42-53, 2016.

[20] G. B. M Tang, G Dudas and V. Minin, "Fitting stochastic epidemic models to gene genealogies using linear noise approximation," *The Annals of Applied Statistics*, vol. 17, no. 1, pp. 1–22, 2019.

[21] L. Sánchez, S. Infante, and A. Hernández, "Epidemic models and estimation of the spread of sars-cov-2: Case study portoviejo-ecuador." *SmartTech-IC 2021*, vol. 1532, pp. 398–411, 2020.

[22] N. d. F. A Doucet and N. Gordon, *An Introduction to Sequential Monte Carlo Methods*. Springer, 2001.

[23] J. Welding, "Sequential monte carlo methods for epidemic data," Ph.D. dissertation, Lancaster University, 2020.

[24] N. D. J. D. J. M. M Baguelin, J Newton and J. Wood, "Control of equine influenza: scenario testing using a realistic metapopulation model of spread," *Journal of the Royal Society Interface*, vol. 7, no. 42, pp. 67–79, 2010.

[25] T. K. X Xu and P. O'Neill, "Bayesian non-parametric inference for stochastic epidemic models using gaussian processes," *Biostatistics*, vol. 4, no. 17, pp. 619–633, 2016.

[26] T. Britton, "Basic stochastic transmission models and their inference," *arXiv*, 2018.

[27] T. Britton and E. Pardoux, "Stochastic epidemic models with inference," *Lecture Notes in Mathematics()*, vol. 2255, 2019.

[28] G. L. Y Wang and J. Du, "Calibration and prediction for the inexact sir model," *Mathematical Biosciences and Engineering*, vol. 19, no. 3, pp. 2800–2818, 2022.

[29] B. R. P Trostle, J Guinness, "A gaussian-process approximation to a spatial sir process using moment closures and emulators," *stat.ME*, 2022.

[30] D. S. D. C. M. M. M Ramírez, J Bogado and C. Schaerer, "Sir-si model with a gaussian transmission rate: Understanding the dynamics of dengue outbreaks in lima, peru," *PLOS ONE*, 2023.

[31] S. Ketu and P. Mishra, "Enhanced gaussian process regression-based forecasting model for covid-19 outbreak and significance of iot for its detection," *Applied Intelligence*, no. 51, pp. 1492–1512, 2021.

[32] P. C. D. H. D Osthus, K Hickmann and S. D. Valle, "Forecasting seasonal influenza with a state-space sir model," *The Annals of Applied Statistics*, vol. 11, no. 1, pp. 202–224, 2017.

[33] R. C. M. E. J Hespanha, R Chinchilla and G. Yang, "Forecasting covid-19 cases based on a parameter-varying stochastic sir model," *Annual Reviews in Control*, vol. 51, pp. 460–476, 2021.

[34] V. I. E Buckingham-Jeffery and T. House, "Gaussian process approximations for fast inference from infectious disease data," *Mathematical Biosciences*, vol. 301, no. 111-120, 2018.

[35] A. A. Z Qian and M. Schaar, "When and how to lift the lockdown? global covid-19 scenario analysis and policy assessment using compartmental gaussian processes," *NeurIPS Proceedings*, 2020.

[36] M. Keeling and P. Rohani, *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, 2011.

[37] E. I. C Breto, D He and A. King, "Time series analysis via mechanistic models," *The Annals of Applied Statistics*, vol. 3, no. 1, pp. 319–348, 2009.

[38] G. Durham and A. Gallant, "Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes," *Journal of Business Economic Statistics*, vol. 20, no. 3, pp. 335–338, 2002.

[39] A. Golightly and D. Wilkinson, "Bayesian inference for nonlinear multivariate diffusion models observed with error," *Computational Statistics and Data Analysis*, vol. 52, pp. 1674–1693, 2008.

[40] V. Isham, "Assessing the variability of stochastic epidemics," *Mathematical Biosciences*, vol. 107, pp. 209–224, 1991.

[41] W. Tan and H. Hsu, "Some stochastic models of aids spread," *Statistics in Medicine*, vol. 8, pp. 121–136, 1989.

[42] A. Jazwinski, *Stochastic Processes and Filtering Theory*, 1st ed. Academic Press, 1970.

[43] X. L. Y Bar-Shalom and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation: Theory, Algorithms and Software*. Wiley-Interscience, 2001.

[44] M. Grewal and A. Andrews, *Kalman Filtering: Theory and Practice Using MATLAB*. Wiley Interscience, 2008.

[45] R. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, pp. 35–45, 1960.

[46] S. Särkkä, *Bayesian Filtering and Smoothing*. Cambridge University Press, 2013.

[47] D. D. M Schober and P. Hennig, "Probabilistic ode solvers with runge-kutta means," *Advances in Neural Information Processing Systems*, pp. 739–747, 2014.

[48] P. Hennig and S. Hauberg, "Probabilistic solutions to differential equations and their application to riemannian statistics," *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, vol. 33, pp. 347–355, 2014.

[49] J. Skilling, *Maximum Entropy and Bayesian Methods*, 1st ed. Springer, Dordrecht, 1991.

[50] Y. Wang and D. Barber, "Gaussian processes for bayesian estimation in ordinary differential equations," *Conference on International Conference on Machine Learning*, vol. 32, no. 2, pp. 1485–1493, 2014.

Appendices