Transmission Dynamics and Control of Cholera in Africa: A Mathematical Modelling Approach

EBENEZER OLAYINKA ADENIYI

A THESIS SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF ART

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS
YORK UNIVERSITY
TORONTO, ONTARIO

NOVEMBER, 2024

Abstract

Background: Cholera, caused by *Vibrio cholerae*, is a global health threat, with outbreaks surging since 2021, particularly in Africa. In 2024, over 13 African countries faced outbreaks worsened by climatic events, poverty, and weak healthcare systems. A shortage of vaccines further complicates control efforts.
Objective: This study uses data science, machine learning, and modelling to analyze cholera dynamics, identify outbreak drivers, and propose targeted interventions.
Methods: A compartmental model with Bayesian estimation analyzed cholera data from eight African countries. Sensitivity analysis identified key transmission parameters, and hierarchical clustering grouped countries by outbreak characteristics.
Results: Average R_0 was 2.0, ranging from 1.41 (Zimbabwe) to 2.80 (Mozambique). Factors like infection rate and human shedding increased R_0, while recovery rate reduced it. Clustering identified three outbreak drivers: natural disasters, conflict, and sanitation issues.
Conclusion: Tailored, data-driven interventions are critical for effective cholera management across diverse contexts.

Acknowledgements

I express my deepest gratitude to my thesis advisor, Prof. Jude Kong, for his invaluable guidance and unwavering support throughout this research. His expertise in mathematical modelling and artificial intelligence, along with his leadership at the Artificial Intelligence and Mathematical Modelling Lab (AIMMLab) at the University of Toronto, has been instrumental in shaping this work and inspiring my academic growth.

I am profoundly thankful to Dr. Sarafa Adewale Iyaniwura, Dr Han Qing and Dr. Andrew Omame for their mentorship, insights, and encouragement, which greatly enhanced the depth of this study.

My sincere appreciation goes to the Department of Mathematics and Statistics at York University and the Department of Mathematics at the University of Ibadan for their resources and foundational support.

Finally, I acknowledge the AIMMLab team at the University of Toronto for their collaborative spirit and dedication to advancing AI and mathematical methodologies, which has been a constant source of motivation. Thank you to everyone who supported me in making this milestone possible.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Biological Background of Cholera

Cholera is classified as an ancient disease that still hunts mankind since its discovery in 1854, and it is reported to have originated in India [4]. Cholera is an infection characterized by severe diarrheal disease caused by Vibrio cholerae, mainly found in aquatic medium. This section will discuss the biological aspects of infection, providing a comprehensive overview of the bacterium, its life cycle, and pathogenesis

1.1.1 The Causative Agent: Vibrio cholerae

Vibrio cholerae is a gram-negative, curved, rod-shaped bacterium from the family Vibrionaceae. It is depicted as a comma-shaped organism with one polar flagellum conferring motility, thus able to swim through aquatic environments and the human intestinal tract. V. cholerae is an oxidase-positive facultative anaerobe: its metabolism occurs either with or without oxygen [55].

The structure of the Gram-negative cell wall includes an outer membrane, a thin layer of peptidoglycan, and an inner cytoplasmic membrane. The outer membrane contains lipopolysaccharides, essential in the bacterium's interaction with the host's immune system. The LPS includes O-antigen, differentiating among the over 200 serogroups of V. cholerae. However, two serogroups in this bacterium remain the cause of most cholera outbreaks: $O1$ and $O139$ [51].

The $O1$ serogroup of V. cholera has been responsible for most cholera pandemics, which can be further divided into two biotypes: the classical biotype, responsible for the first six pandemics, and the El Tor Biotype, which is attributed to causing the recent(7th) outbreaks, displacing the former. The $O139$ serogroup is attributed to South Asia, identified in the early 1990s [21].

1.2 Epidemiology – Epidemic model and basic reproduction ratio (R_o)

1.2.1 Overview

Although Epidemiology originated more than a century ago, it has grown significantly. There have been over 102 definitions since 1978 due to developments in the use of terms such as "populations," "study," "disease," "health," and "distribution" [22]. Various scholars have attempted to define epidemiology in the quest for a uniform definition. For example,Lilienfeld [35] constructed one based on 23 existing definitions, but inconsistencies emerged. Evans [20] analyzed Lilienfeld's definition and defined it as the quantitative analysis of disease processes in population groups ~~and~~ their prevention and control. Both definitions have influenced the understanding and calculation of the frequency of disease occurrence.

In addition, epidemiology also involves the study of risk factors associated with a particular disease or health condition. This includes identifying and analyzing various environmental, genetic, and lifestyle factors that may contribute to the development or progression of a disease. This information can also be used to develop targeted interventions and strategies to reduce the disease burden within a population. Epidemiology also plays a vital role in monitoring and evaluating the effectiveness of public health interventions and policies. By tracking changes in disease patterns and trends over time, epidemiologists can determine the impact of various interventions and make adjustments as necessary. Furthermore, epidemiology also plays a critical role in disease surveillance and outbreak investigations [54]. By quickly identifying and tracking outbreaks of infectious diseases, epidemiologists can help contain and control the spread of disease and protect public health.

1.2.2 Epidemiology: The Mathematical View

The mathematical view of epidemiology involves using concepts from the field to give insight into the behaviour of occasionally occurring diseases. For many centuries, epidemics have come and disappeared after claiming so many lives, impacting how we live, and so much more. The goal of an epidemiologist is first to understand the causes of a disease, then to predict its course, and finally, to create methods to contain it. The first work using data to study infectious disease was by John Grant (1620-1674) in his book "Natural and Political Observations Made upon the Bills of Mortality." Weekly records were taken of the numbers and causes of death in London during his time. This record spans from 1592 to 1603, which later provided the data John Grant used to analyze the various causes of death, and he gave a method for calculating comparative risks of dying from various diseases [8]. In as much as John Grant was the first to use data to study infectious disease, the first epidemiology model is associated with the work of Daniel Bernoulli (1700-1782) on inoculation against smallpox [37].

Further epidemiological developments were made in the 19th century by figures such as Florence Nightingale and John Snow, who used statistical methods and mapping tech-

niques to investigate the spread of disease [12]. Their work laid the foundation for modern epidemiology and continues to influence the field today.

It is also worth noting that epidemiology has evolved to include the study of infectious diseases, chronic diseases, and injuries and the broader determinants of health, such as social and environmental factors. The definitions and concepts used in epidemiology have also evolved, with more recent definitions emphasizing quantitative methods and focusing on understanding the determinants and distribution of health and disease in populations [23].

The mathematical view of epidemiology has continued to evolve with the development of new statistical and computational methods. For example, in the 20th century, the development of the Kermack-McKendrick mathematical model of infectious disease transmission (Kermack and McKendrick, 1927) provided a framework for understanding how diseases spread through populations. Anderson and May (1991) further refined this model in their book "Infectious Diseases of Humans: Dynamics and Control," which introduced new mathematical techniques for understanding the dynamics of infectious disease epidemics.

In recent years, the field of epidemiology has been revolutionized by the availability of large-scale data sets and advanced computational methods. For example, using electronic health records (EHRs) and other data sources has made it possible to conduct population-level studies of disease patterns and risk factors on a much larger scale than was previously possible [6]. Additionally, the development of machine learning and artificial intelligence techniques has allowed for the development of more sophisticated models for predicting the course of epidemics and identifying potential intervention strategies [36].

The mathematical view of epidemiology has played a crucial role in studying infectious diseases throughout history. From the early work of John Grant and Daniel Bernoulli to the modern use of advanced computational methods, the field continues to evolve and provide new insights into the spreading and controlling of infectious diseases.

1.2.3 Compartmental Model

In formulating the mathematical model that depicts the spread of a communicable disease, it is almost impossible to incorporate all physical parameters affecting the behaviour of the disease. Instead of this, assumptions are made to incorporate only the essential features of the system into the model.
In 1906, Hamer proposed the spread of infection by dividing the human population into two distinct compartments: susceptible and infective individuals [26]. He introduced the most used mass action law for the rate of new infections crucial today in modelling infectious diseases. The basic most common compartmental models to describe infectious disease transmission are by Kermack and McKendrick (1927,1932,1933) in a sequence of papers.

The Kermack-McKendrick 1927 epidemic model is given as

$$\begin{aligned} v(t) &= -x'(t) \\ x'(t) &= -x(t)\left[\int_0^t A(s)v(t-s)ds + A(t)y_0\right] \\ z'(t) &= \int_0^t C(s)v(t-s)ds + C(t)y_0 \\ y'(t) &= \int_0^t B(s)v(t-s)ds + B(t)y_0 \end{aligned} \tag{1.1}$$

$v(t)$ denotes the number of persons in the unit area who became infected at t. $X(t)$ is the number of susceptibles, $y(t)$ is the number of infectious individual, $z(t)$ is the number of recovered individuals and Φ is the recovery rate.

$$B(s) = \exp^{-\int_0^t \Phi(s)ds}, A(s) = \phi(s)B(s).$$

He assumed there were no disease deaths, and the population size remained constant. He also was able to derive a final size relation in the form

$$\log \frac{1-\frac{y_0}{N}}{1-p} = pN\int_0^\infty A(s)ds$$

with
$N :=$ Total population size
$p :=$ Disease attack ratio

$$p = 1 - \frac{x_\infty}{N}$$

when
$S(t) = x(t)$, $A(s) = B(s) = \exp^{-\gamma s}$, $I(t) = \frac{N}{a}y(t)$
If the model is modified, with "a" being the disease incidence rate, it will transform into the following

$$\begin{aligned} \frac{dS(t)}{dt} &= -aS(t)\frac{I}{N} \\ \frac{dI(t)}{dt} &= aS(t)\frac{I}{N} - \gamma I(t) \\ \frac{dR(t)}{dt} &= \gamma I(t) \end{aligned} \tag{1.2}$$

The above is the simple Kermack-McKendrick model.

1.2.4 Basic Reproduction Number

The basic reproductive number, denoted as R_0, plays a role in studying infectious diseases. It represents the average number of cases from a single typical infection within a susceptible

population. It's worth noting that R_0 is a figure, not a rate measured in terms of time [61]. Calculating the reproduction number involves tracking the secondary cases of one infected individual entering a fully susceptible population. However, if there are classifications or subpopulations beyond just infected and susceptible groups, it's crucial to monitor the secondary infections within these subgroups separately. An approach to grasp the concept of the reproduction number is through this equation;

$$R_0 \propto \left(\frac{infection}{contact}\right) \cdot \left(\frac{contact}{time}\right) \cdot \left(\frac{time}{infection}\right).$$

This equation illustrates how the number of infections a person's contacts, infection duration and transmission time to others are interconnected. In epidemiology, R_0 serves as a measure to assess disease outbreak potential and intervention effectiveness. When R_0 surpasses 1, it indicates disease transmission and an outbreak occurrence. If the value of R_0 is below one, it suggests that the disease spread is limited and may eventually fade away. When the R_0 drops below 1, it indicates that the measures have effectively managed to control the disease transmission. It's crucial to note that the value of R_0 is not fixed and can fluctuate over time and among groups of people. Factors such as population density, genetic diversity and immunity levels can impact the value of R_0 [13]. For example, populated regions have higher contact rates, leading to a higher R_0. Similarly, the R_0 tends to be higher in communities with lower immunity levels since there is increased susceptibility to infections.

The basic reproduction number is a fundamental concept in the study of infectious diseases and is used to assess the potential for an outbreak and the effectiveness of interventions. It is a dimensionless number that can change over time and in different populations. Understanding the basic reproduction number is essential for controlling and preventing the spread of infectious diseases.

1.2.5 Computing the Basic Reproduction Number

The basic reproduction number can be computed using various means, with Statistical and Mathematical methods dominating. The two key approaches in mathematics are the Jacobian method and the next-generation matrix, with the latter being more widely used. Both methods will give the same result of R_0.

Next Generation Matrix

One method of calculating R_0 is using the next-generation matrix, introduced by Diekmann et al. in 1990 [14]. The next generation matrix, represented as $\mathbb{G}$, is composed of two parts: $\mathcal{F}$ and $\mathcal{V}^{-1}$. The matrix F represents the new infections that occur in the population, while V represents the transfer of infection from one compartment to another. The matrix $\mathcal{V}^{-1}$ is the inverse of the Jacobian matrix of V concerning the parameters, and $\mathcal{F}$ is the Jacobian matrix of F. The disease-free equilibrium state, represented as x_0, is also considered when calculating $\Re_0$.

It is important to note that calculating $\Re_0$ using the next-generation matrix is a straightforward extension of the theory when there are multiple infected individuals. However, if there are subpopulations or distinct compartments within the population, it is necessary to calculate $\Re_0$ separately for each sub-population or compartment.

It is also important to note that $\Re_0$ is the spectral radius of the matrix $\underline{\mathbb{G}}$, which is the dominant eigenvalue. This has been shown in various studies such as van den Driessche and Watmough 2002 [56] and [15].

1.3 Statistical Analysis

Analyzing data statistically is essential for spotting patterns and making decisions based on the data relationships and insights drawn from them. Statistical tests can be broadly categorized into **Parametric** and **Non-Parametric Tests**, each type suited to different data characteristics.

1.3.1 Parametric Tests

Parametric tests assume that the data follows a specific distribution, typically a normal distribution and that certain assumptions are met, such as homogeneity of variance and the normality test. When the conditions are met properly for these tests to work well, they tend to be more effective and offer outcomes.

- **Independent t-test**: Compares the means of two independent groups (e.g., comparing test scores between two classes).
- **Paired t-test**: Compares the means of two related groups (e.g., blood pressure measurements before and after treatment within the same individuals).
- **One-Way ANOVA**: Compares the means of three or more independent groups (e.g., comparing weight loss across three different diet plans).
- **Repeated Measures ANOVA**: Used when comparing means across multiple conditions or time points within the same group.
- **Pearson Correlation**: Measures the linear relationship between two continuous variables (e.g., hours studied and test scores).
- **Linear Regression**: Predicts the value of a dependent variable based on one or more independent variables, assuming a linear relationship.

Non-Parametric Tests

Non-parametric tests do not require the assumption of normality or equal variances and are suitable for ordinal data, skewed distributions, or data with outliers. They are useful when parametric test assumptions are not met.

- **Mann-Whitney U Test**: The non-parametric alternative to the independent t-test compares the medians of two independent groups.
- **Wilcoxon Signed-Rank Test**: Non-parametric alternative to the paired t-test; compares the medians of two related groups.
- **Kruskal-Wallis Test**: Non-parametric alternative to One-Way ANOVA; used to compare the medians of three or more independent groups.
- **Friedman Test**: Non-parametric alternative to Repeated Measures ANOVA; compares medians across multiple conditions or time points within the same group.
- **Spearman Correlation**: Non-parametric alternative to Pearson correlation; assesses the relationship between two ranked or ordinal variables.
- **Chi-Square Test of Independence**: Tests relationships between two categorical variables (e.g., the relationship between gender and product preference).

1.3.2 Choosing Between Parametric and Non-Parametric Tests

The decision to use parametric or non-parametric tests depends on the data's characteristics and whether the assumptions for parametric tests are met. Parametric tests are preferable when data is normally distributed with equal variances, as they provide greater statistical power. However, non-parametric tests are a reliable alternative if these assumptions are violated.

1.3.3 Testing Assumptions: Homogeneity of Variance and Normality

To decide between parametric and non-parametric tests, testing for **homogeneity of variance** and **normality** is essential.

Testing for Homogeneity of Variance

Homogeneity of variance (also known as homoscedasticity) assumes that the variances across different groups are equal. This assumption is crucial for tests like ANOVA and t-tests.

- **Levene's Test**:

  - *Purpose*: Checks if variances across groups are equal.
  - *Interpreting Results*: A p-value greater than 0.05 suggests that variances are equal (homogeneity is met). A p-value less than 0.05 indicates significant differences in variances, violating the homogeneity assumption.

- **Bartlett's Test**:
  - *Purpose*: Another test for homogeneity of variances, suitable for normally distributed data.
  - *Interpreting Results*: A p-value greater than 0.05 suggests equal variances, while a p-value less than 0.05 indicates unequal variances.

- **Visual Inspection**:
  - *Boxplots*: Visualize the spread of values across groups. Similar spreads indicate homogeneity.
  - *Residual Plots*: Check residuals after fitting a model; random distribution supports homogeneity.

Testing for Normality

Normality tests assess whether the data follows a normal distribution, a key assumption for parametric tests.

- **Shapiro-Wilk Test**:
  - *Purpose*: Tests normality, especially useful for small samples.
  - *Interpreting Results*: A p-value greater than 0.05 suggests the data is normally distributed. A p-value less than 0.05 indicates deviation from normality.

- **Kolmogorov-Smirnov Test**:
  - *Purpose*: Compares sample distribution with a normal distribution, suitable for larger samples.
  - *Interpreting Results*: A p-value greater than 0.05 indicates normality, while a p-value less than 0.05 suggests non-normality.

- **Visual Inspection**:
  - *Histogram*: Check if the data forms a bell-shaped curve.
  - *Q-Q Plot*: Compares quantiles of data with those of a normal distribution; points along the diagonal suggest normality.

- **Skewness and Kurtosis**:

  - *Purpose*: Measures the data's asymmetry and "tailedness".
  - *Interpreting Results*: Skewness close to zero and kurtosis near three suggest a normal distribution.

When Assumptions Are Violated

- **If Normality Is Violated**: Use a non-parametric test (e.g., Mann-Whitney U test instead of a t-test).
- **If Homogeneity of Variance Is Violated**: Use **Welch's t-test** (for two groups) or **Welch's ANOVA** (for multiple groups), as they are robust to unequal variances.

By conducting these tests for normality and homogeneity of variance, researchers can select the appropriate statistical test, ensuring robust and accurate results for their analysis.

Chapter 2

Literature Review

2.1 Overview of Cholera

Cholera is a major infection in regions that do not have access to safe drinking water, good hygiene, and adequate sanitation. As Cholera is transmitted through the fecal-oral route [52], it is caused by a bacterium called Vibrio Cholera found in water outlets exposed to human waste(feces). Vibrio cholera produces cholera toxin (CT) to cause severe diarrhea, which is the most prominent symptom of the infection, while others are mild [16]. Diarrhea is the primary symptom that can turn to dehydration and can lead to death if not quickly treated [57]. An exposed person starts to exhibit symptoms 12 hours to 5 days into the incubation period [31]. Without proper treatment, the body of the infected becomes too acidic, increasing the likelihood of death within 24 hours. According to research, patients can recover with immunity, which can endure for a long time based on various circumstances [27]. Cholera is transmitted mostly indirectly to humans, with its primary mode of transmission from drinking contaminated water, food and direct contact with the feces of an infected person [47, 48]. There is also the direct transmission of the infection, which can be human-to-human transmission. All this contributes to the persistence of the bacterium in the environment [45] and its seasonal epidemic [7]. The first cholera pandemic happened in the late 18th century in London in 1849. The British physician John Snow was the first to map out the cause of the outbreak of drinking contaminated water with human excreta [5]. Ever since the first outbreak of cholera, there have been seven pandemics of cholera affecting a lot of countries, and it has claimed a lot of lives over this course [17]. Countries affected have no access to clean water, hospitals, basic amenities or war-prone countries [2].

2.1.1 Transmission Dynamics of Cholera

Cholera is an infection caused by the intake of food or water contaminated with the bacterium Vibrio cholerae. If not treated well and on time, it can lead to death. Bacterial transmission is indirect because it involves contact with a human in an infected reservoir. The in-host life cycle and outside-the-host life cycle of the bacteria are discussed.

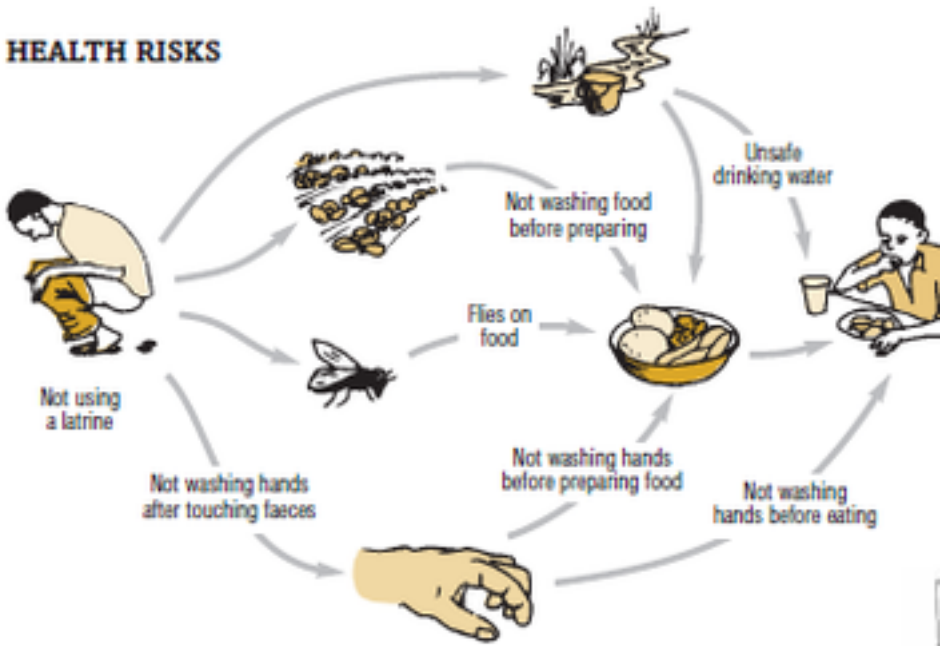


Figure 2.1: Outside the host cholera transmission [29]

The transmission may occur due to the bacterium through the feces of human beings by landing flies on food, poor hand hygiene after touching feces, water sources, not washing hands before preparing food, unsafe drinking water, and not washing hands before eating.

Cholera bacteria first enter the body through the mouth and reach the stomach. Upon ingestion, the bacteria must survive the low-pH environment of the stomach and then colonize the small intestine [44]. Using a single flagellum, V. cholerae propels itself through the mucus layer of the intestine, responding to host signals like bile, mucins, pH, and oxygen availability [63]. This induces the expression of virulence factors, including toxin-coregulated pilus (TCP) and cholera toxin (CT). TCP helps colonize the epithelial surface while CT binds to GM1 gangliosides on host cells and induces the formation of cyclic AMP (cAMP). This high amount of cAMP inhibits the normal transport of ions, resulting in a high efflux of electrolytes and water into the intestinal lumen, which gives the characteristic watery diarrhea of cholera.

The transmission cycle of V. cholerae then continues by producing microcolonies and biofilm components at high cell density. The bacteria employ quorum sensing to regulate their gene expression. The bacteria then detach from the epithelial surface and migrate to the intestinal lumen through the down-regulation of genes whose products are involved in attachment to the host cells and upregulate those needed for transition to the aquatic environment. V. cholerae is excreted in diarrheal stools and vomits, contaminating the environment [28]. Contaminated drinking water or food can infect new hosts, perpetuating the cycle. The spread of cholera is worsened by poor sanitation, inadequate hand hygiene, and unsafe drinking water [59]. Implementing comprehensive public health measures targeting

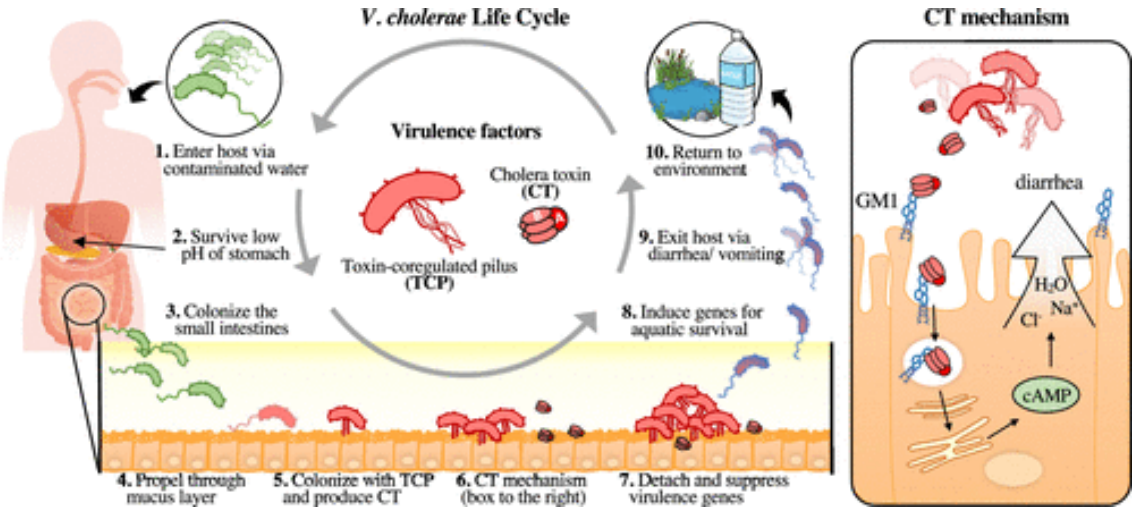


Figure 2.2: Cholera In-Host life cycle [11]

these areas is crucial to breaking the transmission cycle and controlling cholera outbreaks [33].

2.2 Cholera Outbreaks in Africa

Since the start of the current or seventh pandemic of cholera in 1961, about 2.9 million cases and 95000 deaths have occurred annually [50]. A large fraction of the cases and deaths are from Africa and Asia [1]. More than 80% of the world cases come from Africa, which can be attributed to unstable refugee situations, natural disasters, and lack of clean water and sanitation in the cholera-endemic part of Africa [38]. As of February 2023, approximately 25 nations have reported cholera cases. The regions most affected are South Asia and Africa. Since the start of 2023, over 14 African countries, including Nigeria, Cameroon, the Democratic Republic of the Congo, South Sudan, Somalia, Ethiopia, Kenya, Tanzania, Zambia, Malawi, Mozambique, Zimbabwe, South Africa, and Eswatini, have reported cholera cases. Additionally, instances have been recorded in several other nations across the Americas, the Mediterranean, the Pacific, and Asia. By August 2023, there were $547,626$ confirmed cases of cholera and $4,927$ documented fatalities. [18]. Natural disasters such as the 2000 floods in Mozambique and the volcanic eruption in the Democratic Republic of the Congo in 2002 were seen to cause new outbreaks of this disease in this region [43]. An unusual relationship was found between an increase in cholera cases and EL Nino-sensitive regions, highlighting an additional $50,000$ cases in East Africa during this event [39].
WHO suggests an effective control strategy for this disease that combines surveillance, water, sanitation and hygiene, social mobilization, treatment, and oral cholera vaccines [58]. This is an approach put in place by the public or private health sector to curb its increase. According to the WHO, Malawi is among the most affected, having over $59,000$ cases and 1770 deaths between the years 2022 and 2023. The country has witnessed a polio outbreak and the devastating effects of tropical storms during the last few years. Despite all this, the country

has realized WASH projects in affected areas besides improving the surveillance system for early case detection and treatment; this saw cases drop significantly within the first quarter of 2023. However, despite African countries incorporating the WASH(Water, Sanitation, and Hygiene) strategy in taming the disease, cases have increased across the continent.

2.3 Epidemiological Models

The cholera outbreak in Africa seems to be never-ending, which is alarming. However, addressing the specific factors driving this issue can lead to a solution. Mathematical models can help best inform public policies that can be used to reduce disease transmission through simulations and predictions about future behaviour [9]. The dynamics of a disease will determine the type of mathematical model that can be used to study its transmission dynamics. A classic SIR model by Kermack McKendrick 1927 [32] is the basic framework for studying diseases and understanding their dynamics. In the case of cholera, its transmission is mainly indirect, which means humans contract the bacterium, causing cholera from a medium harbouring it. Capasso and Fontana first described a model describing the dynamics of cholera in 1979 [10]. All Cholera models used by epidemiologists are based on this.

To better understand the transmission process of cholera in China, [53] developed a mathematical model known as the SIB (Susceptible-Infected-Concentration level of the bacteria) model. They use a SIB model in which Holling's type II function represents the infection term, and the recovered population is salient because of the formula $N = S + I + R$. Their analysis revealed that improved environmental sanitation and the provision of clean water are more effective strategies than vaccination. Since the two closely similar models have distinct dynamical behaviours, [62] provided two models to determine the influence of awareness/unawareness. Their findings emphasize the significance of model validation as a requirement for adoption. Similar research was conducted to determine the role of immigration in the spread of cholera and to evaluate the efficacy of different management strategies. An expanded $SIRB$ deterministic epidemiological model for cholera was created and rigorously examined [46]. It was discovered that the model had two equilibria: a distinct endemic equilibrium (EE) and a disease-free equilibrium (DFE). Based on the basic reproductive number (R_0)—the number of secondary infections that arise from introducing a single infected individual into a population—it was discovered that the local stability of the DFE and EE depends on this threshold. Specifically, the DFE is stable when $R_0 < 1$, while the EE is stable when $R_0 > 1$.

According to [3], vaccination and medication-assisted therapy are adequate to eradicate cholera, notwithstanding the likelihood of human-to-human transmission. A cholera epidemiology model, modified by [24], is presented and analyzed. The possibility of disease transmission from person to person is included in the expanded model, along with preventative and control strategies. Disease-free equilibrium state (DFE) and endemic equilibrium state (EE) are established by conducting equilibrium analysis for the extended model in two epidemic and endemic equilibrium scenarios. They determine the local asymptotical stability and calculate the basic reproduction numbers for both models. The outcomes are then used

to compare the models at the DFE states in terms of how control affects the extended model. Furthermore, the extended model's endemic equilibrium state (EE) is examined and discovered to be locally asymptotically stable when the fundamental reproduction number is less than 1. This demonstrates that effective preventative and control strategies are necessary to eradicate cholera in a community.

To effectively control the dynamics of cholera, we need to understand its transmission and determine effective control strategies by comprehending the parameters that influence its transmission [24]. One way to do this is by estimating the parameters best fitting real data to a model describing its transmission. Although there is the problem of the non-existence of a cholera model that best fits real data [40], we will use a simple iSIRB model with an immunological threshold.

Chapter 3

Methodology

3.1 Study Area and Data Sources

3.1.1 Geographic Scope of the Study

This research dwells on the most affected African countries with cholera from 2022 to 2024. The focus is on countries like Mozambique, Zambia, and Zimbabwe, which generally recorded heavy outbreaks. They would offer very useful learning sites about the dynamics and nature of the disease and the intervention measures needed to be engraved in the public health system. Other African countries that recorded cases of cholera during this period that will be considered in this study are the Comoros, Cameroon, Somalia, Malawi, Mozambique and Sudan, which are also to be looked into to learn from an even larger range of regional variations and transmission patterns of the epidemic.

3.1.2 Data Collection Methods

For this study, the dataset is collected from the dashboard of Cholera and Acute Watery Diarrhea's weekly reporting of cases and deaths from the World Health Organization [60]. This dataset is one of the critical sources of information that provides complete and updated details regarding the number of cases, fatalities, and geographic spread of cholera outbreaks in the areas under study.

3.1.3 Data Quality and Preprocessing

The data quality would be as good as the predictions made by the models. To that end, rigorous preprocessing of the collected data will be done, which shall consist of a few salient steps:

- **Data Cleaning**: This step includes checking missing values, inconsistencies, and outliers in data and their respective treatments to ensure the data is clean and complete.

Combining and integrating data from different sources will yield a unified, comprehensive dataset that makes analysis easier.

In this case, the datasets are of high integrity, so little or no preprocessing stage is necessary.

3.2 Model and Model Analysiss

3.2.1 Mathematical model

We used a compartmental indirect-susceptible-infected-recovered-bacteria (iSIRB) model [30, 41] to study cholera dynamics in African countries. Our model has four compartments: three for the human population and the remaining for bacteria concentration (see Figure 3.1 and Table 3.1). For the human population, we have the susceptible compartment (S), which includes members of the population with the potential to come in contact with water from a reservoir contaminated with vibro cholera bacteria and become infected. The infected

Table 3.1: Model variables and description.

Variable	Description/Values	Unit
S	Susceptible population	Persons
I	Infected population (value= 1)	Persons
R	Recovered population	Persons
B	Concentration of V. Cholera in the reservoir (value= 10^6)	Cell liter^{-1}
N	Total population(value= 10^8)	Persons

compartment (I) includes the members of the population who are currently infected with the bacteria. The bloodstream bacteria concentration level for these individuals is above the minimum infection dose required to overcome the immune system and make an individual sick. The recovered compartment (R) includes those who have recovered from the bacteria infection. The last compartment in our model is used to track the concentration of vibro cholera bacteria (B) in water reservoir, through which members of the population become infected.

Although cholera can be transmitted from person to person through in-person contact, the possibility is low [25]. The majority of cholera transmission in Africa occurs through contaminated water reservoirs [19]. As a result, our model focuses on the indirect route of cholera transmission, which is represented by the 'indirect' in iSIRB. We assume that susceptible individuals in the population acquire cholera when they consume water from a contaminated reservoir. Since our model describes the dynamics of cholera in a country, and not every person in the country would be exposed to contaminated water (or bacteria-infested Water sources) at the same time, we assumed only a fraction of the population of each country is susceptible during the cholera outbreaks that we considered, and estimated this fraction using Bayesian inference. The model assumes that the concentration of bacteria

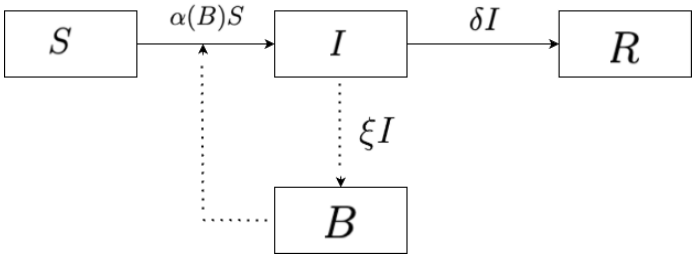


Figure 3.1: **Model schematic diagram.** The human population is divided into three compartments. Susceptible population (S): individuals with the potential to come in contact with water from a reservoir contaminated with vibro-cholera bacteria and become infected. Infected compartment (I): those who are currently infected with the bacteria, and recovered compartment (R): those who have recovered from the bacteria infection. The last compartment in our model describes the bacteria concentration in contaminated water reservoirs, denoted by B. The solid arrows show the flow of individuals in the population from one compartment to another, while the dashed arrows show the shedding of vibro cholera bacteria by infected individuals into water reservoirs and how susceptible individuals acquire cholera indirectly through contaminated reservoirs.

in the reservoir is directly proportional to the concentration of bacteria consumed in the water obtained from the reservoir. As a result, an individual will only be infected with cholera when the concentration of bacteria in the reservoir exceeds a certain threshold, referred to as the minimum infectious dose. We assume infected individuals shed vibrio cholera bacteria into reservoirs through fecal contamination [49]. This is when an infected person's feces enter back into a reservoir, thus increasing the bacteria concentration level in the environment.

The model differential equations are given by

$$\begin{aligned} \frac{dS}{dt} &= -\alpha(B)S, \\ \frac{dI}{dt} &= \alpha(B)S - \mu I - \gamma I, \\ \frac{dR}{dt} &= \gamma I, \\ \frac{dB}{dt} &= rB\left(1 - \frac{B}{K}\right) + \xi I, \end{aligned} \tag{3.1}$$

where $\alpha(B)$ is the cholera transmission incidence rate, given by

$$\alpha(B) = \begin{cases} 0 & B < c, \\ \dfrac{\eta(B-c)}{(B-c)+H} & B \geq c. \end{cases} \tag{3.2}$$

Here, c is the minimum infectious dose of vibro cholera bacteria required to trigger an infection [41]. We observed from (3.2) that there is no cholera infection in the population when

$B < c$, and the spread of the disease only occurs when the reservoir bacteria concentration is equal to greater than the threshold (i.e., $B \geq c$).

Table 3.2: Model parameters and description.

Parameter	Description	Values/Source	Unit
r	Bacteria intrinsic growth rate	Estimated	Week^{-1}
K	Bacteria carrying capacity	Fixed	Cell liter^{-1}
H	Half-saturation constant for incidence rate	Fixed	Cell liter^{-1}
η	Maximum rate of infection	Estimated	Week^{-1}
ξ	Bacteria shed rate	Estimated	$\text{Cell liter}^{-1}\,\text{Week}^{-1}\,\text{Person}^{-1}$
μ	Cholera-induced death rate	0.105 [42]	Week^{-1}
c	Minimum Infectious Dose	Fixed	Cell liter^{-1}
γ	Recovery rate	Estimated	Week^{-1}

We modelled the disease incidence rate when $B \geq c$ using the type-II Holling function (see Eq. 3.2), where η is the maximum infection rate and H is the half-saturation constant, which specifies the concentration of bacteria that gives half the maximum infection rate. In Eq. (3.1), μ is the cholera-induced death rate, γ is the recovery rate, and ξ is the rate at which infected individuals shed vibrio cholera bacteria into the water reservoir. We assume that the bacteria concentration grows in the reservoirs with an intrinsic growth rate r and a carrying capacity K. See Table 3.2 for a detailed description of model parameters and their units.

3.2.2 Equilibrium Points

Henceforth, since $N = S + I + R$, the recovered class can be suppressed, and the equations for S and I will be used to determine the equilibrium points. We set the derivatives to zero and solve the resulting algebraic equations:

Disease-Free Equilibrium (DFE)

For the disease-free equilibrium, we consider $I = 0$:

$$\begin{cases} 0 = -\alpha(B)S \\ 0 = \alpha(B)S - (\gamma + \mu)I = 0 \implies I = 0 \\ 0 = \xi I + rB\left(1 - \frac{B}{K}\right) = 0 \implies B = 0 \text{ or } B = K \end{cases}$$

This leads to the disease-free equilibrium points:

$$(S, I, B) = (S_0, 0, 0) \quad \text{for } K < c$$

$$(S, I, B) = (S_0, 0, K) \quad \text{for } K \geq c$$

Since the model is an epidemic model, there is no endemic equilibrium point.

3.2.3 Basic reproduction number

The most important summary measure of the transmissibility of pathogens is the basic reproduction number (R_0). Put another way, given an entirely susceptible population, the basic reproduction number, R_0, is defined as the average number of secondary infections caused by a single infected individual. It is the fundamental measure of the potential of an outbreak for any infectious disease. The model's disease-free equilibrium point is $E_0 = (S_0, 0, K) = (N, 0, K)$. The next-generation matrix is one of the most popular methods for calculating the basic reproduction number. It is used here. Let

$$f := \begin{bmatrix} \alpha(B)S \\ \xi l \end{bmatrix}$$

These are the introduction of the infection in the I and B compartments. While

$$v := \begin{bmatrix} \mu I + \gamma I \\ -rB\left(1 - \frac{B}{K}\right) \end{bmatrix}$$

Is the rate of transfer of individuals and bacteria out of I and B. The Jacobian matrix of both matrices at the disease-free equilibrium point (E_0) are

$$F := \begin{bmatrix} 0 & \frac{aN}{K-c+H} + \frac{acN}{(K-c+H)^2} \\ \xi & 0 \end{bmatrix} \quad V^{:} = \begin{bmatrix} \mu+\gamma & 0 \\ 0 & -r \end{bmatrix} and \quad V^{-1} := \begin{bmatrix} \mu+\gamma & 0 \\ 0 & -\frac{1}{r} \end{bmatrix}$$

Then

$$FV^{-1} := \begin{bmatrix} 0 & -\frac{aNH}{(K-c+H)^2 r} \\ \frac{\xi}{\mu+\gamma} & 0 \end{bmatrix}$$

The basic reproduction number is the spectral radius of the FV^{-1}, the largest positive eigenvalues, which is

$$R_0^2 := \frac{\xi a N H}{(\mu+\gamma)(K-c+H)^2 r} \tag{3.3}$$

3.3 Parameter Estimation of Epidemiological Models

Parameter estimation in an epidemiological model is critical in understanding and dealing with outbreaks. During the process, we derive invaluable information about the complicated dynamics of the outbreak. We can, therefore, advise policymakers on the most appropriate measures to reduce the impact and control of such outbreaks.

There are two major methods for parameter estimation: frequentist and Bayesian approaches. The former is associated with the long-run probabilities, such as the probability of

getting a given dataset under a null hypothesis. The Bayesian approach deals with the probability of a hypothesis given a particular dataset and includes prior information in the analysis. On the other hand, the frequentist approach is dependent only on the observed data. The Bayesian approach allows for calculating the probability of some particular hypothesis being true. In frequentist analysis, the P-value represents the probability of obtaining another dataset as extreme as that collected. The results from a Bayesian approach are usually more intuitively interpretable than frequentist results, which often lead to misinterpretation. For this project, the estimation of parameters will be done using the Bayesian approach.

3.3.1 Monte Carlo Sampling

Sampling is the selection of a subset or a fraction of a total population to make inferences about the whole group. How can we sample a large dataset and be sure of the integrity of one's selection to reflect the whole representation? Different sampling methods have been developed with different assumptions which can be employed for sampling.

Monte Carlo sampling is a process of sampling from a probability distribution aimed at approximating some quantity of interest with accuracy. Consider a random variable with a pdf denoted by $f(x)$. The idea here comes down to generating samples $x_1, x_2, \ldots, x_n$ from the said distribution so that the distribution of samples approximates the actual distribution, $f(x)$.

If D is the observed data, and θ denotes the model parameters, the knowledge of the prior distribution $P(\theta)$, which is a good guess of the parameters and the likelihood of the observed data $P(\theta|D)$ given some specific values of the parameters is crucial in getting the posterior distribution of the parameters $P(\theta|D)$.

The prior distribution and the likelihood of having a joint distribution are defined as

$$P(D, \theta) = P(D|\theta)P(\theta)$$

Also, after observing the data, the posterior distribution uses Baye's theorem to update the parameters by combining the prior distribution with the information provided by the data.

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{\int P(D|\theta)P(\theta)d\theta}$$

The expected value of the posterior of a function $f(\theta)$ can now be evaluated as

$$E[f(\theta|D)] = \frac{\int f(\theta)P(D|\theta)P(\theta)}{\int P(D|\theta)P(\theta)d\theta}$$

Markov Chain

A Markov chain is a memoryless random walk having the future state depending only on the current state, not the preceding states. A countable set S with a stochastic process defined

on it as $X = \{X_n : n \geq 0\}$ is called a Markov chain if it satisfies the following conditions

$$P\{X_{n+1} = j | X_0, ..., X_n\} = P\{X_{n+1} = j | X_n\} \tag{3.4}$$

$$P\{X_{n+1} = j | X_n = i\} = P_{ij} \tag{3.5}$$

The first condition is called the Markov property, with the second describing how the transition probabilities p_{ij} do not depend on the time parameter n.

Markov Chain Monte Carlo (MCMC) is classified as a class of algorithms that uses the Monte Carlo sampling method following the Markov chain to sample from a probability distribution.

3.3.2 NUTS Algorithm

The No-U-Turn Sampler (NUTS) is an advanced MCMC algorithm that solves the problem of that of the Hamiltonian Monte Carlo(HMC))sampling algorithm. The HMC incorporated the Hamiltonian Mechanics in its sampling methods, which makes it more robust than traditional ones in solving complex high-dimensional probability distribution by introducing momentum variables. A major flaw is the manual inputting of the step size and the number of leapfrog steps, which leads to long computation times.

NUTS generalizes HMC, which automates the tuning of parameters, mainly at the number of leapfrog steps. The recursive construction of a binary tree makes it possible for NUTS to adaptively extend the length of the trajectory until some stopping criterion is met. The No-U-Turn criterion is probably the real innovation of NUTS, which prevents the trajectory from doubling back on itself and thus guarantees that the sampler will not spend computational effort on redundant paths. This makes the algorithm efficient and more user-friendly, reducing the need for manual intervention in parameter setting.

Moreover, NUTS has included an adaptive tuning of the step size during the warm-up phase for an optimal balance between exploration and acceptance rates. This adaptive tuning is useful when dealing with high-dimensional spaces where geometry may vary significantly in a posterior distribution. NUTS automatically adjusts the length of the trajectory and the step size so that the sampler efficiently explores the complex posterior landscapes without the risk of getting caught in local modes or due to poor mixing. This recommends NUTS as an ideal choice to obtain Bayesian inference in models with high-dimensional or highly correlated parameter spaces since it provides robust and reliable sampling even in very hard cases.

We will not detail the statistical theory underlying this sampling method. Instead, we will concentrate on the practical implementation using the STAN probabilistic programming language that we will interface from the R software environment. In this way, we will be able to make proper use of both the advanced capabilities of STAN in Bayesian inference and another complex statistical modelling and extensive R ecosystem in data manipulation, analysis, and visualization. The combination of R and STAN provides a very powerful and flexible framework for effectively running probabilistic models and analyses.

The R-hat statistic (commonly called the Gelman Rubin diagnostic or potential scale reduction factor) serves as a metric for evaluating the convergence of Markov Chain Monte Carlo (MCMC) simulations within analysis processes. During MCMC execution, we create chains to sample from the distribution. The R hat metric aids in determining whether these chains have reached a distribution. An R-hat value nearing one often indicates a good convergence that implies well-mixed chains and a high likelihood sampling from the target posterior. Values greater than 1 indicate that the chains have not yet converged, implying the need for more iterations or further adjustments. Monitoring R-hat is essential for ensuring reliable parameter estimates from MCMC samples.

Chapter 4

Results

In this chapter, we will share the findings of our research focusing on the estimated parameters of the cholera transmission model. These estimated parameters offer insights into how cholera outbreaks have progressed dynamically, particularly between 2022 and 2024, in African nations significantly impacted by the disease. The parameters include the maximum infection rate, recovery rate, human shedding rate of bacteria, intrinsic growth rate and basic reproduction number. Given the transmission of the disease, only individuals in contact with the bacterial reservoir are vulnerable to infection. This implies that only a portion of the population is at risk due to their interaction with the source. Both initial susceptibility and initial infection cases will be calculated. The R-square error will determine if the fit is the best data set.

R-squared (R^2) measures how well a model explains the variance in the dependent variable. It ranges from 0 to 1, where a value of 1 means the model perfectly explains all the variability in the data, and 0 means it explains none of it.

The formula for R-squared is:

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}}$$

Where:
$SS_{residual}$ - (Residual Sum of Squares) represents the variance not explained by the model, i.e., the differences between the observed and predicted values.
SS_{total} (Total Sum of Squares) is the total variance in the observed data, calculated as the differences between the observed values and their overall mean. Specifically, it shows the proportion of the total variance in the observed data that is accounted for by the model's predictions.

For example, an R^2 of 0.75 means that 75% of the variance in the observed data can be explained by the independent variables in the model. In other words, the model's predictions capture 75% of the changes or fluctuations in the dependent variable.

The remaining 25% of the variance is unexplained, meaning it is due to factors not included in the model or inherent randomness in the data. This unexplained variance reflects

the difference between the observed values and the predicted values, which is captured by the residual sum of squares ($SS_{residual}$).

The process of estimating parameters during a cholera outbreak using the NUTS algorithm through STAN probabilistic programming language offers an examination of epidemic dynamics within a specific country. While not all parameters are subject to estimation, certain factors like carrying capacity for bacterial growth (K-assumed), cholera-induced death rate (μ), and minimum infection dose ($C - assumed$) remain constant across all countries. The acceptable range for the R-square error will be $85\% - 100\%$.

4.0.1 Cameroon

Cameroon has been facing outbreaks of this disease since the cholera outbreak began in this area on October 25, 2021. Between October 25, 2021, and October 12, 2023, the WHO reported a total of 20, 933 cases of cholera, with 2, 050 cases confirmed through lab tests and a total of 492 deaths, resulting in a CFR (Case Fatality Rate) of 2.4%. The outbreak spread across nine regions, with active transmission occurring in the central coastal and Southwest regions. Patients affected by the outbreak ranged from as young as two months to as old as years old. The male-to-female ratio was calculated to be 1.3. These epidemiological figures highlight the scale and seriousness of the outbreak. It Emphasizes the need for continued public health interventions and monitoring.

The recent outbreak began around March 6^{th}, 2023 peaked on May 15^{th}, then declined by July that same year. On average, the highest infection rate parameter, denoted as a, is around 0.93 with a deviation of 0.05. The 95% credible interval spans from 0.83 to 1.00, suggesting some variability in this factor that may reflect differences in outbreak severity or varying interpretations of data.

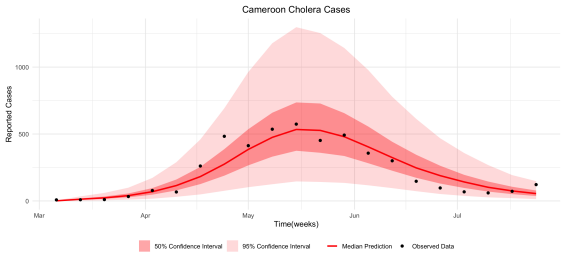
The estimated recovery rate parameter, labelled γ, averages 0.96 with a credible interval ranging from 0.88 to 1.

Regarding the bacteria growth rate parameter denoted as r, infection rates across the population vary, with an average of 0.0.92 and a credible interval ranging from 0.64 to 1.24. The parameter ξ is well defined, but it still exhibits some variability that impacts outbreak dynamics; its mean value is 128.32, with a standard deviation of about 4.75 and a credible interval spanning from 119.28 to 137.67.

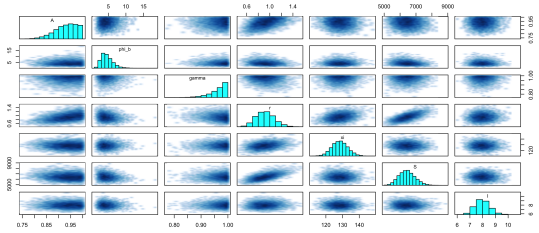
The initial susceptible population has an average estimate of approximately 6438.53 individuals with a credible interval between roughly 5507 and up to about 7519 people. Highlighting significant uncertainty surrounding this vulnerable initial population. At the start of the outbreak, the average number of individuals infected was 8, with a range of 7 to 9, giving us a solid estimate of its size. All the parameters have an R-hat of 1. The sample sizes were sufficient, showing that our estimates are trustworthy and that we've reached a good understanding. This thorough examination will show how the outbreak spreads and how people recover, offering insights for healthcare planners and intervention strategies.

Based on our calculations, the basic reproduction number stands at 1.573501, which tells us how many people are infected on average by one infected individual. The R-square error

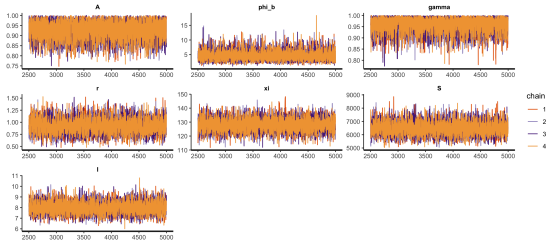
is 0.8662042, which depicts a good fit for the observed data.



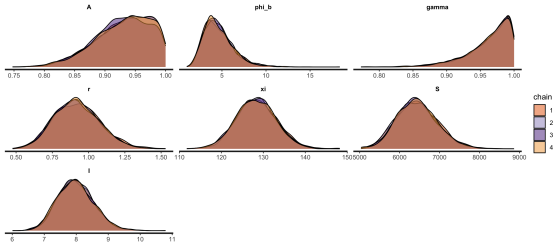
(a) Cameroon Cases data fitting



(b) Pairs plot



(c) Trace plot



(d) Density plot

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
A	0.93	0.00	0.05	0.83	0.90	0.94	0.97	1.00	8529	1
phi_b	4.62	0.02	1.69	2.11	3.42	4.35	5.53	8.63	7576	1
gamma	0.96	0.00	0.03	0.88	0.95	0.97	0.99	1.00	8999	1
r	0.92	0.00	0.16	0.64	0.81	0.91	1.02	1.24	4548	1
xi	128.32	0.06	4.75	119.28	125.09	128.26	131.43	137.67	6570	1
S	6438.53	6.92	512.41	5507.58	6079.83	6417.12	6768.66	7519.03	5488	1
I	7.98	0.01	0.56	6.95	7.59	7.96	8.35	9.15	8444	1

(e) Estimated Parameter and variables

Figure 4.1: Data Fitting and Parameter Estimation of Cameroon

4.0.2 Comoros

The cholera outbreak in Comoros, first reported in February at 2, 2024, has reached enormous proportions. The nation had recorded $10,142$ cases as of June $30, 2024$. This has primarily been reported in Ndzuwani, where $8,942$ cases have been registered. Ngazidja and Mwali have also been affected and reported 625 and 575, respectively. Sadly, the outbreak claimed 147 lives, yielding a CFR of 1.4%.. Using STAN with the Half-saturation pathogen density(H) being $10^{5.4}$, the following are the result obtained.

The maximum infection rate parameter, a, has a mean value of 0.64 with a standard deviation of 0.05. The 95% credible interval ranges from 0.55 to 0.74. It thus covers part of the variability in outbreak severity that may be related to differences in data interpretation or even real epidemiological factors forming the basis for the estimates.

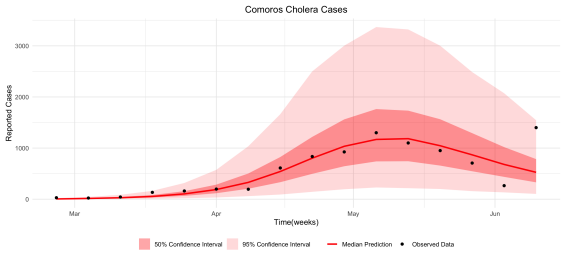
The recovery rate parameter, γ, has a mean of 0.66, with a credible interval ranging from 0.46 to 0.91. In future, this will be a key parameter to learn precisely how quickly people recover from the disease, as this can be integral in understanding the general dynamics of the outbreak. Also, the average of the human shredding rate is 22.56.

Parameters like the bacterial growth rate,r, show very little variability, with the mean at 0.66 and credible intervals ranging from 0.46 to 0.91. Thus, an estimate for this factor has a mean of 0.66 and a standard deviation of 0.12, with a credible interval from 18 to 30.

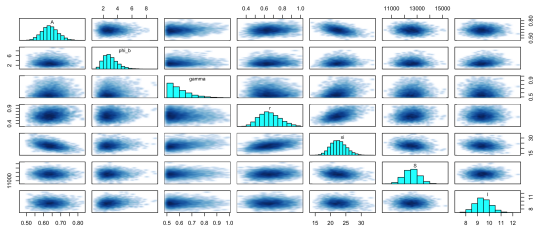
This gives the initial susceptible population,S, a mean estimate of 12564.48 with a standard deviation of 593 and a credible interval of 11429 to $12,558$, which indicates the estimate of the initial population at risk. The initial infected population has a mean of 9, with a credible interval from 8 to 10, giving a reliable estimate of the outbreak's initial size.

All the Rhat values for the parameters are 1, and the effective sample sizes are high enough to ensure that the convergence of the estimates is good and, hence, reliable. These results are relevant to planning in public health and intervention strategies, especially for offering insight into the transmission and recovery processes involved in this outbreak. The basic reproduction number is estimated to be 1.966067, wherein one infected person will most likely transfer his disease to an average of 2.447295 others, which states that it can still develop into an outbreak if it is not controlled properly.

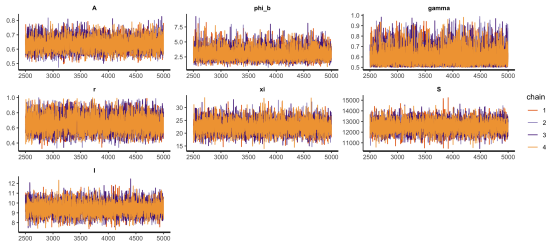
The R-squared value is 0.6480183, which indicates that approximately 64.8% of the variance in the dependent variable is explained by the model. This suggests a moderate fit, where the model captures a significant portion of the data's variability, but about 35.2% of the variance remains unexplained.



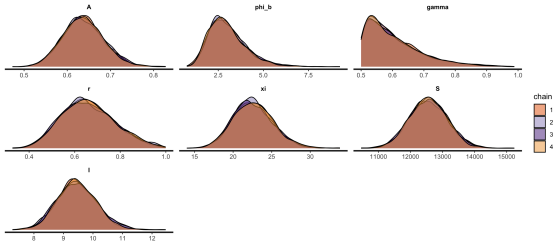
(a) Comoros Cases data fitting



(b) Pairs plot



(c) Trace plot



(d) Density plot

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
A	0.64	0.00	0.05	0.55	0.61	0.64	0.67	0.74	10714	1
phi_b	3.02	0.01	1.03	1.47	2.28	2.86	3.58	5.45	10022	1
gamma	0.60	0.00	0.09	0.50	0.54	0.58	0.65	0.83	10330	1
r	0.66	0.00	0.12	0.46	0.58	0.65	0.74	0.91	8156	1
xi	22.56	0.03	2.59	17.83	20.76	22.44	24.25	27.97	8978	1
S	12564.48	5.44	593.36	11429.30	12161.04	12561.56	12965.36	13745.92	11881	1
I	9.46	0.01	0.67	8.20	9.00	9.43	9.91	10.83	12558	1

(e) Estimated Parameter and variables

Figure 4.2: Data Fitting and Parameter Estimation of Comoros

4.0.3 Malawi

There have been cholera outbreaks in twenty-nine districts, including Machinga, since March 2022. Up to April 30, 2024, a total of 59,361 cases and 1,772 deaths have been reported, resulting in a case fatality rate of 3.0%. The analysis period from November 18, 2022, to June 4, 2023, was chosen as it captures the epidemic phase. STAN was utilized for data fitting and parameter estimation with the aim of enhancing insights into outbreak dynamics and refining response strategies.

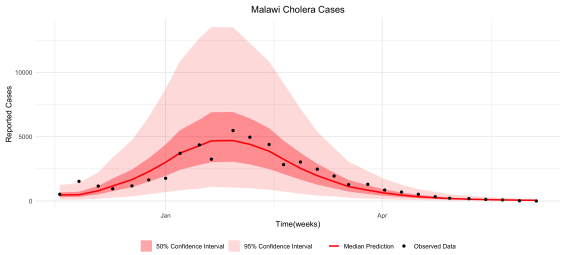
The average maximum infection rate parameter (A) is estimated at 0.83 with a deviation of 0.07 and a credible interval of 95% ranging from 0.71 to 0.97. This suggests some degree of variability in the severity of outbreaks due to factors like public health measures or population density.

Regarding the recovery rate parameter (γ), its mean value is calculated at 0.43 with an interval spanning from 0.15 to 0.87, portraying a wide spectrum of possible recovery scenarios within the population. The growth rate of bacteria denoted as r is tightly determined, with an average of 0.1 and a narrow credible range, indicating an estimation with high confidence. The parameter ξ averages 1.75 with a standard deviation of 0.51 and a 95% credible range between 0.99 and 2.98.

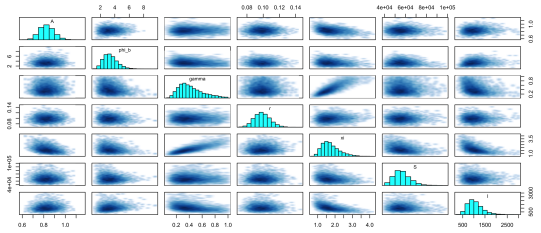
For the susceptible population denoted as S, the estimated mean stands at 56665 with a standard deviation of 91.77. The credible interval ranges from 43429 to 74960, indicating uncertainty surrounding the initial population at risk. The mean for the infected population is 1054.5, with a credible interval ranging from 521 to 1887.

All Rhat values return as 1, and the effective sample sizes are satisfactory; thus, we can trust that the parameter estimates have converged well. The calculated basic reproduction number based on these estimates is 2.494212. These findings help comprehend the dynamics of the cholera outbreak and offer valuable insights for devising effective public health strategies and interventions.

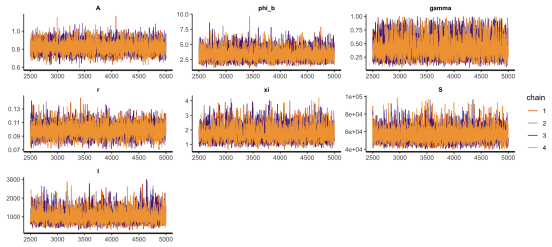
The R-squared value is 0.8298686, meaning that the model's predictions explains approximately 84.2% of the variance in the observed data. This indicates a strong fit between the model and the actual data, as the model can accurately predict most changes in the observed values.



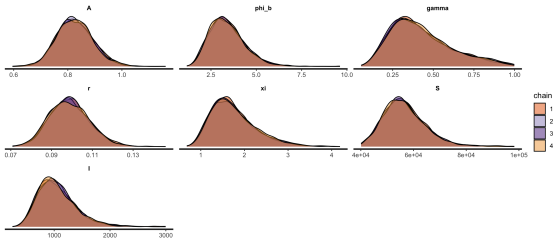
(a) Malawi Cases data fitting



(b) Pairs plot



(c) Trace plot



(d) Density plot

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
A	0.83	0.00	0.06	0.71	0.78	0.83	0.87	0.96	7045	1
phi_b	3.19	0.01	0.91	1.76	2.54	3.08	3.70	5.29	6021	1
gamma	0.61	0.00	0.20	0.26	0.46	0.60	0.76	0.97	4112	1
r	0.10	0.00	0.01	0.08	0.09	0.10	0.10	0.12	7655	1
xi	2.16	0.01	0.56	1.22	1.75	2.12	2.53	3.34	3526	1
S	56280.66	97.77	8120.61	43188.64	50492.28	55300.01	60930.63	74846.72	6899	1
I	936.27	4.37	308.70	476.69	717.12	888.43	1103.96	1657.37	4987	1

(e) Estimated Parameter and variables

Figure 4.3: Data Fitting and Parameter Estimation of Malawi

4.0.4 Mozambique

Beginning on September 14, 2022, the cholera outbreak hit the country starting in the Niassa province. There have been 7,294 reported cases as of April 28, 2024, with 12 deaths, at a CFR of 0.2 percent. Since the onset of the outbreak in September 2022, there have been cumulative 48,181 cases and 174 deaths for an overall case fatality rate of 0.4 percent in the country. The cases have been declining over the last three weeks of April, with eight provinces and 24 districts still reporting active cases. As of June 23, 2024, the cumulative number of cases stood at 8,024, with 18 deaths, thus accounting for a CFR of 0.2%. From February 6, 2023, to June 19, 2023 shall therefore be used to study the dynamics of cholera in this region for a better understanding of the outbreak behaviour and possible intervention strategies.

The parameter describing the maximum infection rate,"a", is such that its average is 0.72, a standard deviation of 0.05, and the 95% credible interval ranges from 0.69 to 0.83. The range indicates some variability in the outbreak's intensity, perhaps due to different environmental or social factors.

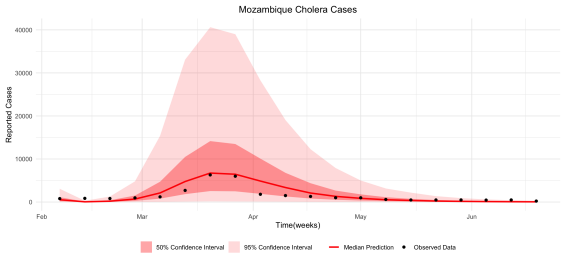
The estimated average for the recovery rate parameter,γ, is 0.71, and its credible interval goes as low as 0.41 and as high as 0.98, indicating a wide range of recovery rates among the patients.

The bacterial growth rate parameter,r, has a mean of 0.72, with a narrow credible interval ranging from 0.58 to 0.87; it indicates little variation in the pattern of bacterial growth among the population.

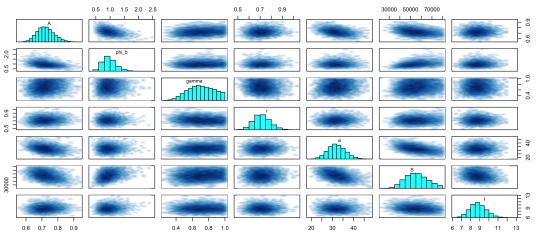
The mean estimate is $55,546$ for the initial susceptible population S, and the credible interval has a large range of $(37122, 76132)$, indicating high uncertainty for the population at risk. The estimate of the initial infection is 8, with a credible interval of 7 to 10, which will be a very robust measure of the size of the initial outbreak.

All the Rhat values are 1, and all the effective sample sizes are large enough; therefore, one can confidently believe in the convergence of the parameter estimates. With the estimated parameters, the basic reproduction number is 2.804156. Such estimates will be instrumental in enlightening the spread and recovery dynamics of the epidemic, which will help formulate public health interventions and strategies.

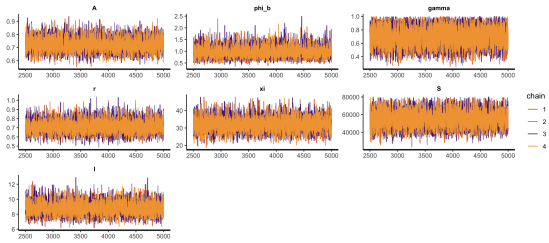
The R-squared value is -1.009007, near -1. This suggests that the model performs significantly worse than a simple model that predicts the mean of the observed data for all observations. A value near -1 indicates that the model's predictions are systematically poor and far from the observed values. Instead of capturing the data's variance, the model introduces additional errors, implying that the model might be unsuitable for the data or incorrectly specified.



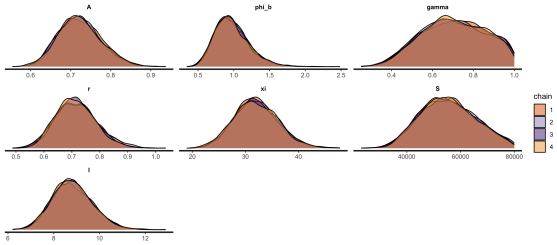
(a) Mozambique Cases data fitting



(b) Pairs plot



(c) Trace plot



(d) Density plot

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
A	0.72	0.00	0.05	0.63	0.69	0.72	0.76	0.83	6025	1
phi_b	0.97	0.00	0.24	0.58	0.80	0.95	1.12	1.51	7045	1
gamma	0.71	0.00	0.16	0.41	0.59	0.71	0.83	0.98	7378	1
r	0.72	0.00	0.07	0.58	0.66	0.71	0.76	0.87	8977	1
xi	32.03	0.05	4.15	24.32	29.17	31.83	34.78	40.68	6928	1
S	55546.95	146.20	10206.32	37122.22	48090.53	55108.25	62697.10	76132.92	4873	1
I	8.81	0.01	0.89	7.22	8.20	8.77	9.39	10.67	9026	1

(e) Estimated Parameter and variables

Figure 4.4: Data Fitting and Parameter Estimation of Mozambique

4.0.5 Somalia

In Somalia, an outbreak of Cholera has killed nine people, while 474 cases were reported from 7 to 13 January 2024 at a CFR of 1.9%, above the WHO emergency threshold. Most of the cases were from Hirshabelle state, particularly from Belet Weyne, Bulo Burto, Jalalaqsi, and Jowhar, following flooding in the latter part of 2023, because of contaminated water sources. Meanwhile, there were 352 cases reported specifically among displaced communities in Daynille, Banadir region. Other reports came from Hargeysa and Wajaale in Somaliland. In 2023, there were more than $18,304$ cases, among them a large proportion of children under 5, with 46 reported deaths. This outbreak was attributed to poor access to safe water, sanitation, and health care. The time period taken for analysis of the epidemic was between January 23, 2023, and August 14, 2023.

The maximum infection rate parameter, a, had a mean value of 0.65 with a standard deviation of 0.05 and a 95% credible interval of 0.55 to 0.76. This range reflects the moderate variability of infection rate, which may result from differing environmental conditions and/or intervention measures.

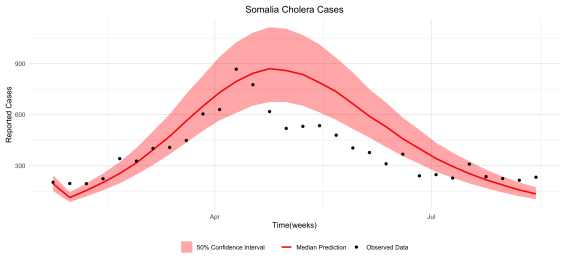
The recovery rate parameter, γ, has a mean of 0.92, with a credible interval ranging from 0.74 to 1.00, indicating that the recovery process for those affected is relatively homogeneous.

The estimate for the mean of the bacterial growth rate parameter, r, is 0.24, with a very narrow 95 % credible interval of 0.19 to 0.29, which suggests very high precision and very low variability in this estimate. The parameter ξ has a mean of 13.27, a standard deviation of 1.78, and a credible interval of 9.89 to 16.84. This range captures some variability in conditions that impact the epidemic's progression.

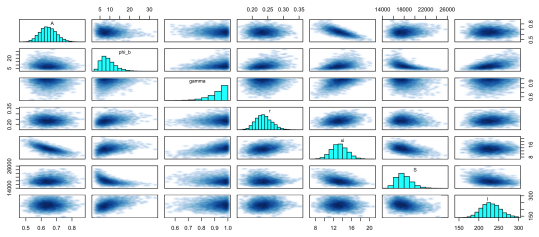
The mean estimate for the initial susceptible population, S, is 17963, with a credible interval from 15644 to 21350. This shows a great deal of uncertainty in the initial population at risk. For the initial infected population, I, the mean is 230, and the credible interval is from 186 to 281, again providing a reliable estimate of the initial outbreak size.

All Rhat values are 1, and all the effective sample sizes are also very large, indicating good convergence of the parameter estimates and that one should have no qualms about relying on them. These are extremely important parameter estimates for understanding how the dynamics of this cholera outbreak unfolded and for devising a robust public health response. With the estimated parameter, the basic reproduction number is 1.681665.

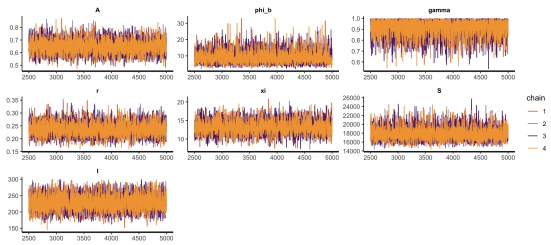
The R-squared value is -0.04707184, indicating that the model explains none of the variance in the observed data and is performing slightly worse than a model that simply predicts the mean of the observed data for all observations. This negative R-squared value suggests that the model introduces more error into its predictions than a basic model using the mean would. While not drastically poor, this still implies that the model is unsuitable for explaining the dataset's relationship and requires further refinement or reconsideration.



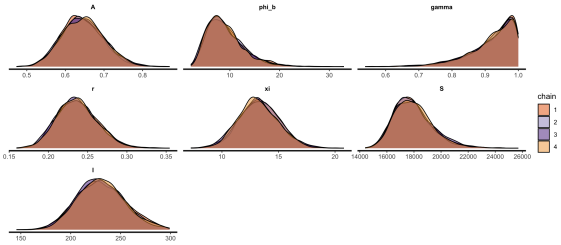
(a) Somalia Cases data fitting



(b) Pairs plot



(c) Trace plot



(d) Density plot

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
A	0.65	0.00	0.05	0.55	0.61	0.64	0.68	0.76	4730	1
phi_b	9.15	0.06	3.78	3.79	6.42	8.43	11.15	18.25	4348	1
gamma	0.92	0.00	0.07	0.74	0.89	0.94	0.97	1.00	7145	1
r	0.24	0.00	0.03	0.19	0.22	0.24	0.25	0.29	7068	1
xi	13.27	0.03	1.78	9.89	12.06	13.22	14.44	16.84	3945	1
S	17963.82	22.65	1446.14	15644.23	16928.83	17787.10	18771.47	21350.12	4076	1
I	230.43	0.36	23.99	186.12	213.48	229.49	246.32	281.10	4441	1

(e) Estimated Parameter and variables

Figure 4.5: Data Fitting and Parameter Estimation of Somalia

4.0.6 Sudan

The cholera outbreak in Sudan started on July 19, 2023, with suspected cases emerging in South Kordofan. The official outbreak declaration occurred on September 26, 2023, following identifying and confirming the vibrio cholerae bacterium. Since then, the disease has spread to 46 localities across nine states, including Gedaref, Red Sea Aj Jazirah, and White Nile. By the end of December 23, 2023, there were 8,267 suspected cases recorded, with a total of 224 reported deaths resulting in an overall case fatality rate of around 2.7%. The rapid spread within this timeframe is attributed to variations in water shortages during dry periods, and contamination from flooding during rainy seasons are significant contributing factors. These challenges form a foundation for addressing public health issues within the region. Various humanitarian initiatives have been launched, such as vaccination drives and efforts

to enhance water sanitation and hygiene conditions (WASH). The analysis covers the period from August 7, 2023, to May 27, 2024.

The average value of parameter "a", representing the maximum infection rate, is approximately 0.99, with a standard deviation of 0.01 and a credible interval of 0.96 to 0.1 at a confidence level of 95%.

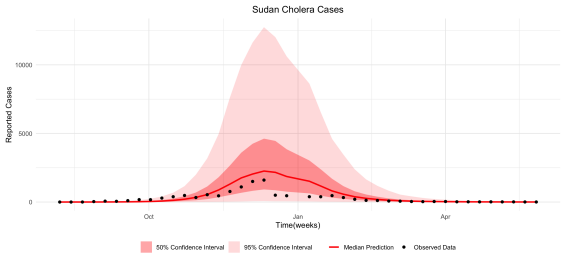
The recovery rate parameter γ has a mean value of around 0.88; however, the credible interval varies significantly from 0.66 to 1.00.

For the growth rate parameter r, the average value is approximately 0.02, with a very narrow credible interval precisely between 0.02. This high precision implies consistent bacterial growth dynamics across different scenarios.

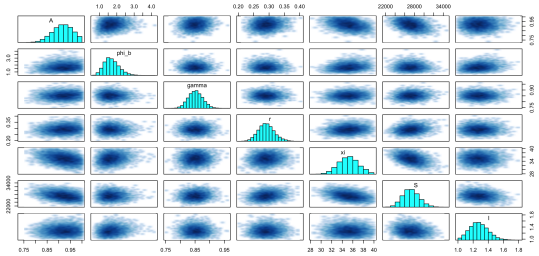
Additionally, an estimate for epidemiological factors, ξ, has an average of about 14.85, a standard deviation of 0.14, and a credible interval spanning from 14.46 to 15. This wide range reflects variations in conditions that impact outbreak development. The estimated number of people at risk is around 28315, with a broad uncertainty range from 25112 to 31748. The average value for the initial infected is 1, with a narrow range from 1 to 1, indicating a highly precise estimation of the initial outbreak scale.

All Rhat values are at 1. The sample sizes are sufficiently large to ensure that all parameter estimates are trustworthy and stable. These findings help us comprehend how cholera spreads and devise effective public health measures and strategies. The basic reproduction number derived from these calculations is 3.093062.

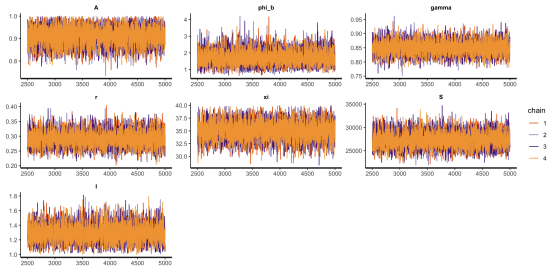
The R-squared value is 0.8422973, indicating that the model explains approximately 84.2% of the observed data variance. This suggests a good fit, as the model successfully captures a substantial portion of the variability in the data. The remaining 15.8% of the variance is unexplained, potentially due to factors not included in the model or random variations. Overall, this R-squared value reflects that the model is performing well in explaining the relationship between the independent and dependent variables, making it a reliable fit for the data.



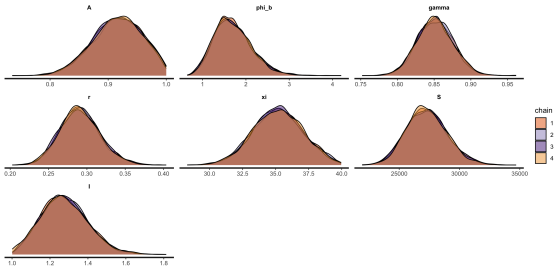
(a) Sudan Cases data fitting



(b) Pairs plot



(c) Trace plot



(d) Density plot

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
A	0.92	0.00	0.04	0.83	0.89	0.92	0.95	0.99	6760	1
phi_b	1.70	0.00	0.44	0.97	1.40	1.66	1.97	2.67	8367	1
gamma	0.85	0.00	0.02	0.80	0.83	0.85	0.87	0.90	8453	1
r	0.29	0.00	0.03	0.24	0.27	0.29	0.31	0.35	7964	1
xi	35.18	0.02	1.88	31.47	33.88	35.20	36.48	38.86	6546	1
S	27291.60	20.91	1700.47	24152.35	26117.01	27232.63	28394.07	30856.54	6615	1
I	1.28	0.00	0.12	1.06	1.19	1.27	1.36	1.53	8400	1

(e) Estimated Parameter and variables

Figure 4.6: Data Fitting and Parameter Estimation of Sudan

4.0.7 Zambia

The current cholera outbreak in Zambia started on 20 January 2023 in Lusaka Province, particularly in the cholera-prone areas in the peri-urban of Lusaka. Currently, from October 2023 to 6 and May 2024, there has been a report of cholera cases in nine provinces, while outbreaks have been confirmed in seven. Already, 40 districts have reported local transmission, bringing about a cumulative total of 23,221 cases with 740 deaths, equivalent to a case fatality rate of 3.2%. The period taken for this overview is from 20 November 2023 to 3 June 2024 to learn lessons on the transmission dynamics for the outbreak under review.

The maximum infection rate parameter, a, is estimated to have a mean value of 0.71 with a standard deviation of 0.05 and a 95% credible interval of 0.62 to 0.81. This would

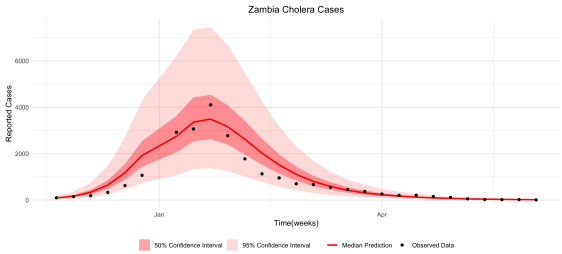
probably tell of a medium amount of variability in infection rate, possibly due to the effects of public health intervention, environmental conditions, and population susceptibility.

The recovery rate parameter,γ, has a mean of 0.67 and a standard deviation of 0.09, with a credible interval of 0.67-1.

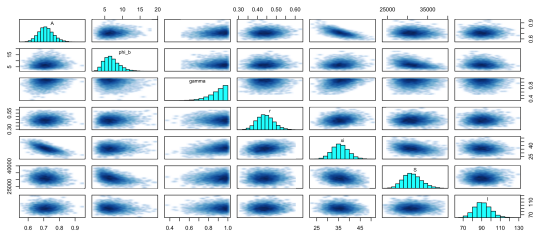
The bacterial growth rate parameter is estimated to have a mean of 0.44, with a standard deviation of 0.04 and a credible interval extending from 0.36 to 0.53, indicating some variability in bacterial proliferation rates. The parameter ξ has a mean of 35.77, with a standard deviation of 3.73 and a credible interval from 28.69 to 43.39.

The mean estimate for the initial susceptible population is 31178, with a credible interval from 27080 to 36200. The mean estimate of the initial infected population is 90, with a credible interval from 74 to 109.

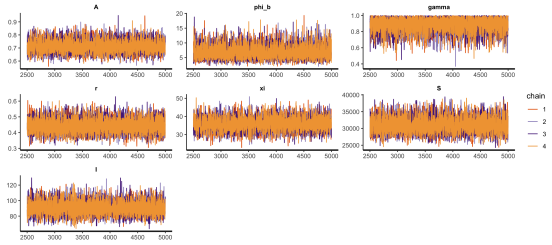
All the parameters have Rhat 1, indicating good convergence and, hence, the reliability of the estimates. Effective sample sizes are also adequate, reiterating the strength behind these parameter estimates. These insights are critical to understanding the spread of the cholera outbreak and guiding effective public health responses and strategies. The basic reproduction number within this timeline is 2.133918.



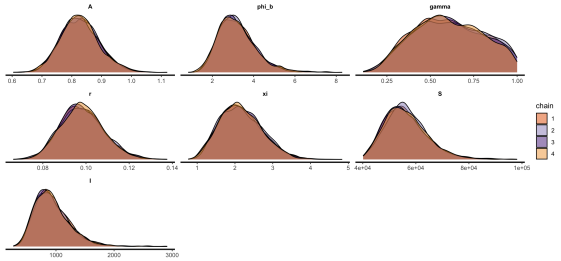
(a) Zambia Cases data fitting



(b) Pairs plot



(c) Trace plot



(d) Density plot

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
A	0.71	0.00	0.05	0.62	0.68	0.71	0.74	0.81	6672	1
phi_b	6.95	0.03	2.23	3.50	5.35	6.64	8.20	12.15	6711	1
gamma	0.89	0.00	0.09	0.67	0.84	0.91	0.96	1.00	8888	1
r	0.44	0.00	0.04	0.36	0.41	0.44	0.47	0.53	8864	1
xi	35.77	0.05	3.73	28.69	33.19	35.67	38.22	43.39	6084	1
S	31178.15	28.61	2309.05	27080.98	29550.17	31030.27	32653.11	36200.33	6515	1
I	90.92	0.09	8.75	74.75	84.85	90.52	96.67	109.23	8503	1

(e) Estimated Parameter and variables

Figure 4.7: Data Fitting and Parameter Estimation of Zambia

4.0.8 Zimbabwe

The very first outbreak of cholera in the country dates back to February 12, 2023. As of June 30, 2024, this outbreak has already claimed a cumulative 34,549 cases and 718 deaths, thereby translating to a case fatality rate of 2.0%. The disease has swept through areas beyond the 17 known cholera hotspot districts, indicating that there is considerable geographical expansion of areas affected by the disease and the period used for the in-depth review of the development and transmission structure of the outbreak ranged from October 2, 2023, to June 3, 2024.

The maximum infection rate parameter, a, was estimated to have a mean of 0.59, a standard deviation of 0.05, and a 95% credible interval ranging from 0.49 to 0.70. This range approximates moderate variability in infection rate and may be due to population

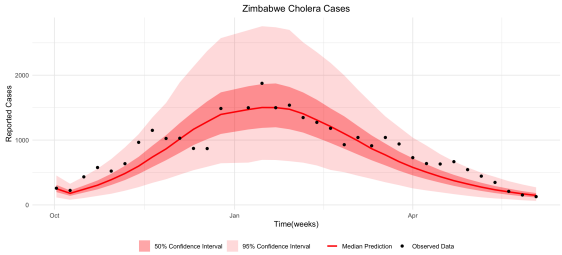
density, environmental conditions, or mitigating interventions.

The bacterial growth rate parameter,r, has an estimated mean of 0.47, with a standard deviation of 0.05 and a credible interval ranging from 0.39 to 0.57, indicating some variability in bacterial growth rates. The parameter ξ, accounting for other epidemiological factors, has a mean value of 19.36, with a standard deviation of 2.57, and a credible interval from 14.91 to 24.93, again showing variability in factors influencing the progression of the epidemic.

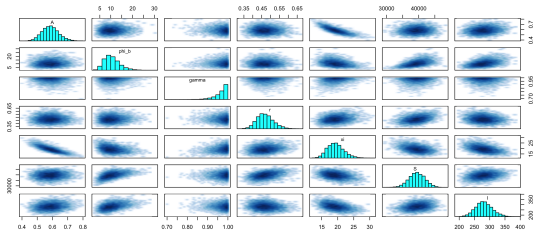
For the susceptible population, this is 39425.66, with a very wide credible interval from 34497.53 to 44242.92, showing considerable uncertainty about the population at risk. The mean for the initial infected population is 281.10, with a credible interval from 229.80 to 338.89, thus representing an effective and reliable estimate of the initial outbreak size.

All the Rhat values are 1, ensuring good convergence and estimates' reliability. In addition, the effective sample sizes are also adequate for the robustness of these parameter estimates with a basic reproduction number of 1.41065. These findings on the details of the cholera outbreak are very important in designing effective public health interventions and strategies.

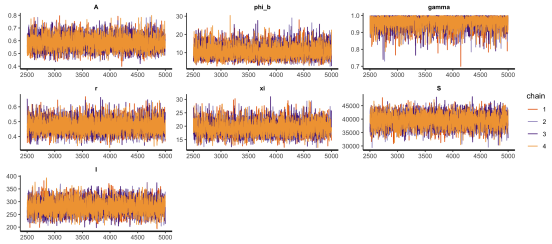
The R-squared value is 0.8122124, indicating a good fit. The model explains approximately 81.2% of the variance in the observed data, showing that it effectively captures the relationship between the variables.



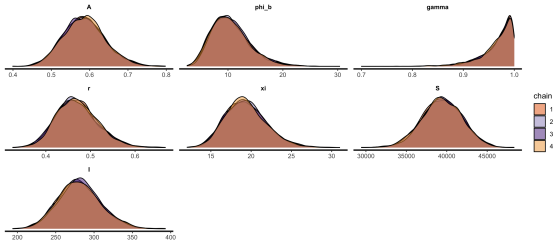
(a) Zimbabwe Cases data fitting



(b) Pairs plot



(c) Trace plot



(d) Density plot

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
A	0.59	0.00	0.05	0.49	0.55	0.58	0.62	0.70	4237	1
phi_b	10.53	0.05	3.51	4.97	8.01	10.11	12.56	18.69	5442	1
gamma	0.97	0.00	0.03	0.88	0.95	0.98	0.99	1.00	9219	1
r	0.47	0.00	0.05	0.39	0.44	0.47	0.50	0.57	6220	1
xi	19.36	0.04	2.57	14.91	17.54	19.18	20.95	24.93	3369	1
S	39425.66	36.46	2495.75	34497.53	37769.43	39434.24	41163.61	44242.92	4687	1
I	281.10	0.39	27.55	229.80	262.20	280.02	298.83	338.89	5059	1

(e) Estimated Parameter and variables

Figure 4.8: Data Fitting and Parameter Estimation of Zimabawe

Chapter 5

Sensitivity Analysis

Sensitivity analysis will be performed to determine the parameters' impact on the basic reproduction number $\mathcal{R}_0$, an important measure that determines the likelihood of spreading an infectious disease within the population. Sensitivity analysis will tell us the most critical parameters for controlling the spread of the disease. Also, we will consider the sensitivity analysis of the infection peak time and infection peak value.

We used the Partial Rank Correlation Coefficient (PRCC) method, which measures the connection between each parameter and the model's results while considering the impact of other parameters. PRCC is well suited for models because it can address nonlinear relationships and interactions among parameters.

The range of values is obtained from the range obtained from the STAN output. The lowest and greatest values are those across the countries. The range values of H, c, and K from [34] are also used.

Parameter	**Lower bound**	**Upper bound**	**Distribution**
ξ	0	700	U
η	0.1	1	U
S	1	36058	U
H	10^6	10^8	U
γ	0.15	1	U
r	2.4	100.1	U
c	10^5	10^7	U
K	10^5	10^7	U

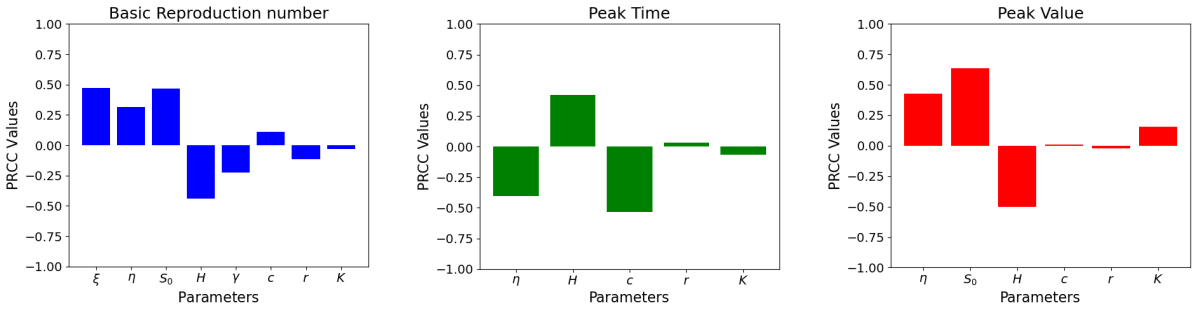


Figure 5.1: **Sensitivity analysis**. **Left panel**: Sensitivity analysis of the control reproduction number. **Middle panel**: Sensitivity analysis of the peak time of cholera cases. **Right panel**: Sensitivity analysis of the peak size of Cholera cases. Positive (negative) values of PRCC indicate a positive (negative) correlation between the quantities and the corresponding model parameter, while the magnitude reflects the measure of sensitivity.

The PRCC values were computed for every parameter using $10,000$ samples generated by Latin Hypercube Sampling (LHS), a technique that guarantees parameter space exploration. The findings reveal how much each parameter affects $\mathcal{R}_0$, the infection peak time and peak value. Also, bivariate distribution is considered for the parameters c and K to ensure the inequality $K > c$ for the infection to be epidemic. The PRCC values range from 1 to -1, indicating a positive or negative impact, respectively, while values near zero signify minimal influence.

The sensitivity analysis findings, as shown in the PRCC plot, emphasize the varying impact of factors on the reproduction number ($\mathcal{R}_0$). Factors like the population (S_0), minimum infection dose c, maximum infection rate a and human shredding rate ξ demonstrate strong positive relationships with $\mathcal{R}_0$, indicating that increases in these factors notably raise $\mathcal{R}_0$ and consequently boost disease spread potential. On the contrary, variables such as the Half-saturation bacteria density (H), recovery rate γ and growth rate r exhibit associations implying that higher values of these factors lead to a decrease in $\mathcal{R}_0$.

Infection peak time and peak value during the epidemic can also be affected by various model parameters. The sensitivity of the model parameters on the peak time and peak value was also conducted using PRCC. The result indicates the maximum infection rate (a), initial susceptibles and maximum carry carrying capacity of the bacteria positivity affect the peak value and the Half-saturation bacteria density negativity affect if with a strength of 0.5198 with other parameters having no or little effect on it. The parameter a and r affect the infection negatively with the strength of 0.3980 and 0.5337 while parameter r affects positivity with strength of 0.5337

These results underscore the significance of managing the maximum infection rate, initial susceptibles, and human shredding rates to effectively control and reduce disease transmission, as these play an essential role in the transmission of the infection.

Chapter 6

Statistical Analysis

In this section, we will explore the analysis used to assess the significance of a parameter's value or distribution across different regions in the country. Verifying normality and variance equality is essential to deciding whether to employ a nonparametric test.

6.1 Test of Normality

In this study, three different tests were used to determine whether the data distribution for each parameter follows a pattern or not; the Anderson Darling test, Kolmogorov Smirnov test, and Shapiro Wilk test are commonly applied in analysis to confirm the assumption of normality, an essential aspect for many parametric tests.

The findings from these examinations shown in the chart below enable us to assess whether parametric or non-parametric approaches are better suited for scrutiny based on the data distribution for each variable across various nations.

The graph shows the significance levels obtained from conducting three normality assessments—the Anderson Darling test, the Kolmogorov Smirnov test, and the Shapiro Wilk test—on a range of variables in countries including Cameroon, Comoros, Malawi, Mozambique, Somalia, Sudan, Zambia, and Zimbabwe. Each panel represents a country, with p-values plotted for parameters labelled as A, γ, I, r, S, and ξ. The bars, in colours, show the outcomes of each test for normality. Red for Anderson Darling test results, green for Kolmogorov Smirnov results, and blue for Shapiro Wilk results

When a p-value exceeds 0.05, it usually indicates that there isn't a deviation from normality in data for a parameter being analyzed here in countries based on what we see in this visualization, most p-values are clustered near zero, which implies that the normality assumption is not being met for parameters across these countries. However, there are exceptions, such as parameters in Cameroon, Comoros and Zimbabwe, where the Kolmogorov-Smirnov test reveals p-values suggesting that normality may hold in these scenarios. This examination assists in assessing the appropriateness of tests and guides modifications like opting for parametric tests for parameters that show deviations from the normal distribution.

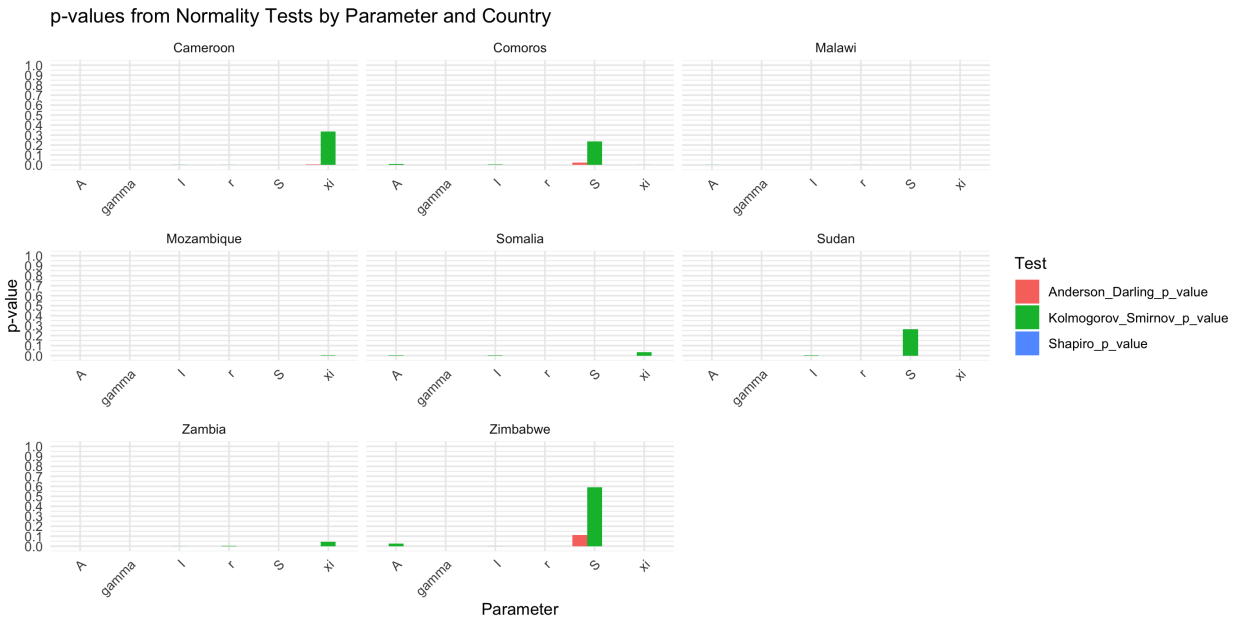


Figure 6.1: Test for Normality

	Parameter	Levene_p_value	Bartlett_p_value
	<chr>	<dbl>	<dbl>
1	A	0	0
2	I	0	0
3	S	0	0
4	gamma	0	0
5	r	0	0
6	xi	0	0

Figure 6.2: Test for Homogeneity of variance

6.2 Test of Homogeneity of Variance

When evaluating whether parametric tests are appropriate for analyzing the data at hand, it's crucial to check if the assumption of homogeneity of variances among groups is met. This assumption is crucial in tests like ANOVA and t-tests since unequal variances affect the outcomes' reliability. Two tests were carried out to verify this assumption. Levens Test and Bartletts Test. Levens Test is more reliable when dealing with deviations from normality. It is commonly used to check for equality in variances between groups, whereas Bartletts Test is more sensitive but assumes that data follows a normal distribution. The test results for each parameter are provided below to help you understand whether the assumption of homogeneity in variance holds.

The table displays the results of the homogeneity of variance tests—Levene's Test and Bartlett's Test—conducted for different parameters: A, I, S, γ, r, and ξ. Both tests returned p-values of 0 for all parameters, indicating significant results that suggest the assumption

of equal variances across groups is violated. When p-values are below the 0.05 threshold, as in this case, it suggests that variances are not homogeneous across the samples for each parameter. This outcome implies that traditional parametric tests that assume homogeneity of variance may not be suitable for analyzing these parameters, and adjustments may be required. Possible adjustments include using alternative tests that do not assume equal variances, such as Welch's ANOVA, or opting for non-parametric tests. This finding highlights the need for caution in interpreting results from parametric analyses on these parameters.

6.2.1 Kruskal-Wallis

Based on the results of the homogeneity of variance and normality tests, it was found that the assumptions for parametric testing were not met across several parameters. Levene's Test and Bartlett's Test both returned p-values of 0 for each parameter (A, I, S, γ, r, and ξ), indicating significant results that suggest the assumption of equal variances across groups is violated. Additionally, the normality tests (Anderson-Darling, Kolmogorov-Smirnov, and Shapiro-Wilk) showed that many parameters had p-values close to zero, suggesting that the data do not follow a normal distribution. These results indicate that parametric tests like ANOVA would not be appropriate for analyzing these parameters due to violations of both normality and homogeneity of variance.

Given these violations, the Kruskal-Wallis test, a non-parametric alternative to ANOVA, was applied to examine parameter differences across countries. The boxplots of parameters by country, along with their corresponding Kruskal-Wallis p-values, are shown in the figure. For each parameter (A, γ, I, r, S, and ξ), the Kruskal-Wallis test returned a p-value of less than 2×10^{-16}, indicating statistically significant differences across countries for all parameters. This suggests that the values of these parameters vary significantly by country, highlighting possible regional differences in the factors influencing the model.

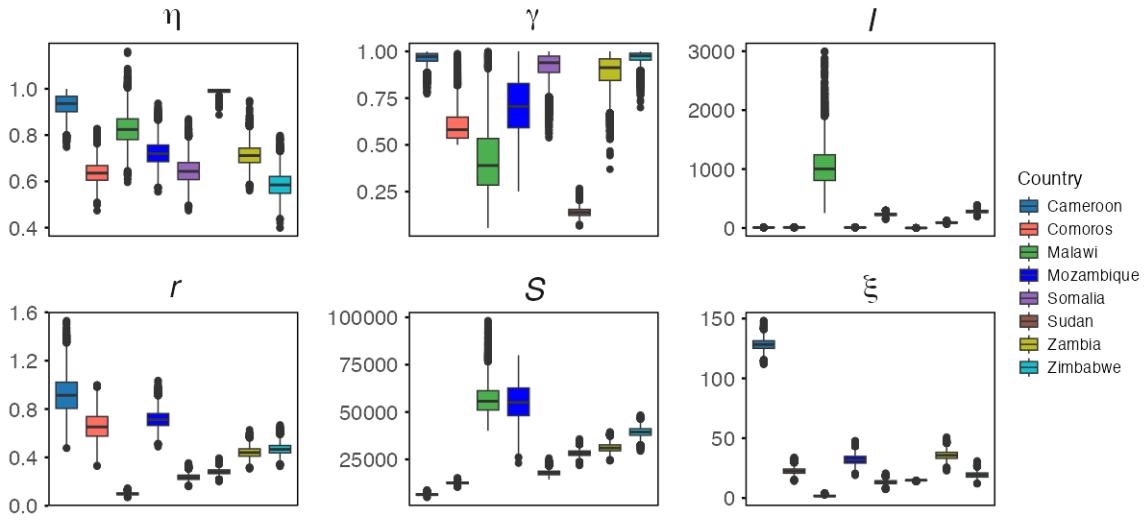


Figure 6.3: **Parameter Distribution and Statistical Test:** The Box plots display the distribution of parameters across countries with Kruskal-Wallis p-values. This is used to check if the distribution of the parameters is statistically significant. The results show a significant difference in parameter distributions between countries, indicating variations in cholera outbreak characteristics across Cameroon, Comoros, Malawi, Mozambique, Somalia, Sudan, Zambia, and Zimbabwe. The results highlight regional disparities in the infection dynamics in Africa.

These findings support the conclusion that there are substantial differences across countries regarding cholera transmission dynamics, recovery rates, bacterial growth rates, and other key parameters. This variation across regions may reflect differences in environmental factors, healthcare interventions, and population susceptibility, which are crucial for tailoring public health strategies to specific regions.

Chapter 7

Hierarchical clustering

Hierarchical clustering is a method in machine learning that categorizes data points into clusters without supervision based on their features or distances from each other. Unlike techniques that create separate groups of data points independently of each other, hierarchical clustering constructs a structure of clusters that can be visualized as a dendrogram. A tree-shaped graph displaying how clusters are formed at various stages. Various linkage criteria can be employed with this approach, such as linkage that calculates the distance between points in different clusters and single linkage that computes the shortest distance between clusters.

Hierarchical clustering was used with the cholera cases and estimated parameters to categorize countries with cholera dynamics characteristics into groups. The method creates a cluster hierarchy displayed as dendrograms to show the connections and resemblances between countries according to their cholera patterns.

Two approaches were utilized in the analysis: complete linkage and single linkage methods were employed to examine the data structure closely. The complete linkage method determines the distance between points in clusters to create compact and distinct clusters. In contrast to complete linkage, the linkage method looks at the minimum distance between points, leading to elongated clusters that might encompass chains of connected countries. The dendrograms visually represent the cluster formations based on each linkage method used in grouping countries according to their similarities in confirmed cases and estimated parameters.

For the complete linkage method, the dendrogram groups **Somalia** and **Cameroon** closely, with **Comoros** joining next, indicating similarities in cholera dynamics among these countries. Another cluster forms with **Zimbabwe**, **Sudan**, and **Zambia**, while **Malawi** and **Mozambique** are clustered separately. The single linkage dendrogram also shows **Somalia**, **Cameroon**, and **Comoros** as a close cluster, with similar connections observed for **Zimbabwe**, **Sudan**, and **Zambia**. However, **Malawi** and **Mozambique** are clustered separately, consistent with the complete linkage results.

To evaluate the clustering quality, several metrics were calculated:

- **Cophenetic Correlation**: The cophenetic correlation for complete linkage was 0.6548,

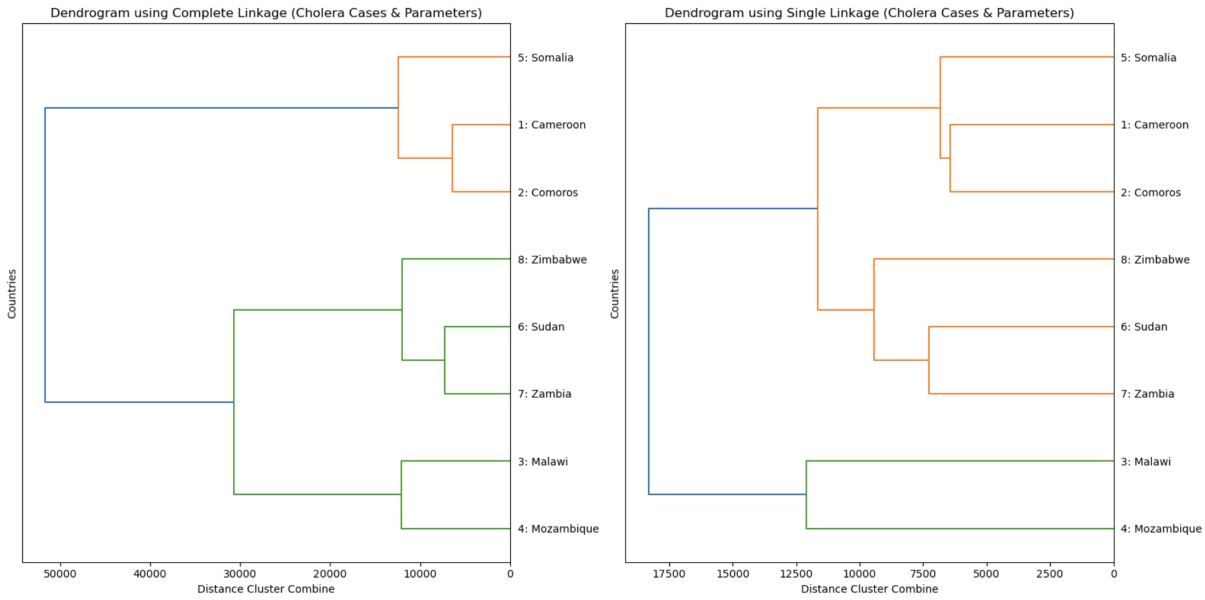


Figure 7.1: Hierarchical clustering

```
Cophenetic Correlation (Complete Linkage): 0.6547977958889282
Cophenetic Correlation (Single Linkage): 0.7691551833308149
Silhouette Score (Complete Linkage): 0.5275819626806111
Silhouette Score (Single Linkage): 0.3406698386672343
Davies-Bouldin Index (Complete Linkage): 0.48172150172403566
Davies-Bouldin Index (Single Linkage): 0.3174006928583663
```

Figure 7.2: Clustering Tests

and for single linkage, it was higher at 0.7692. This indicates that the single linkage method preserved the pairwise distances between original data points more faithfully in the dendrogram.

- **Silhouette Score**: The silhouette score for complete linkage was 0.5276, indicating a moderate level of cluster separation. The score for single linkage was lower at 0.3407, suggesting that the clusters formed were less well-defined with this method.

- **Davies-Bouldin Index**: For complete linkage, the Davies-Bouldin index was 0.4817, while for single linkage, it was higher at 0.3174. Lower values indicate better-defined clusters, so complete linkage demonstrated better cluster compactness and separation.

These results indicate that complete linkage produced more compact and well-separated clusters than single linkage. The clustering patterns and quality metrics provide valuable insights into regional similarities in cholera cases and transmission dynamics, supporting targeted interventions based on the identified clusters of countries.

Now, to know the true number of clusters in the Clustering, both the Elbow method and Silhouette score can be optimized to determine this.

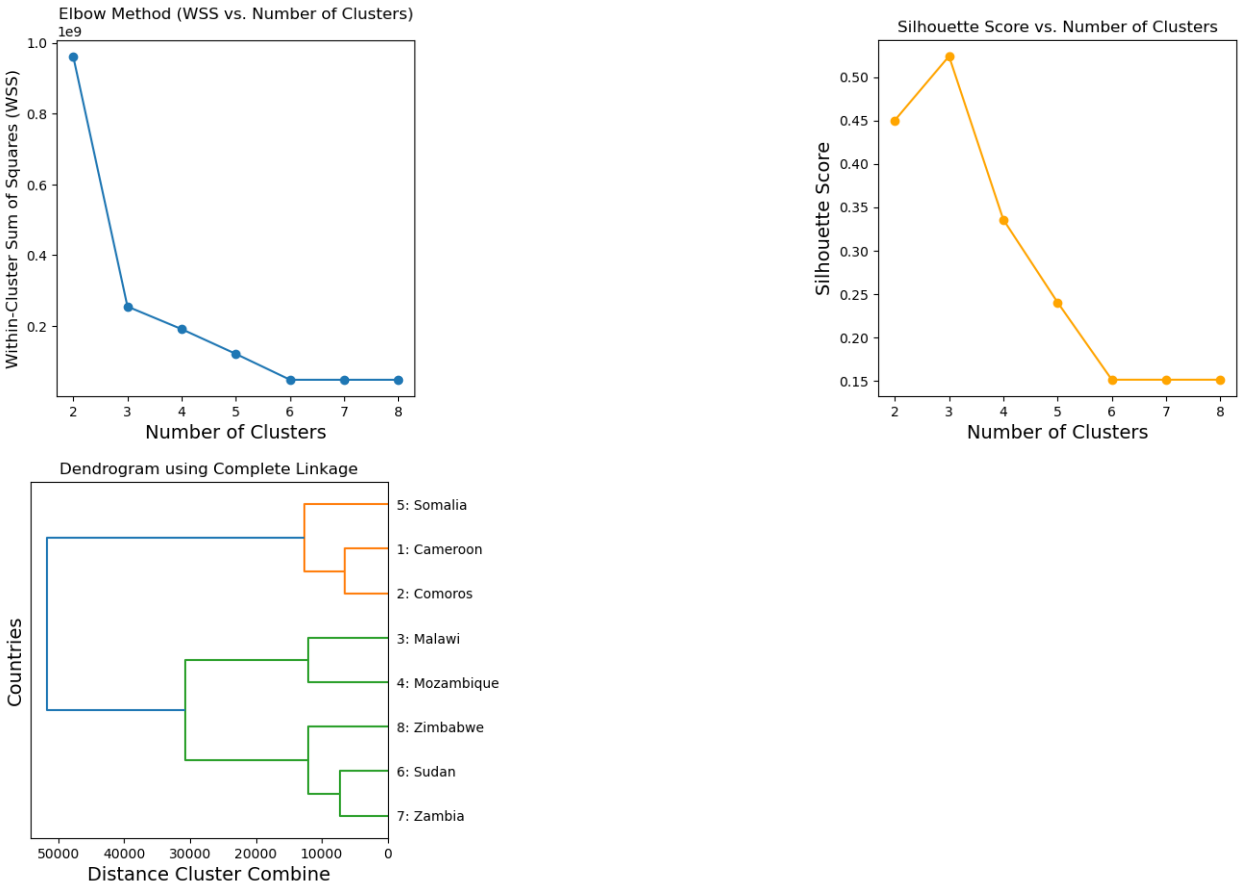


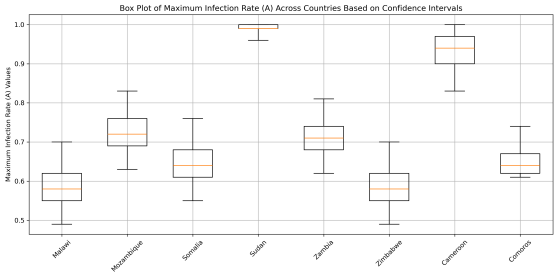
Figure 7.3: **Clustering Analysis: Left panel**: The Elbow Method is used to determine the appropriate number of clusters for a clustering, and it involves calculating the identify Within-Cluster the Sum 'elbow' of point Squares where (WSS) the for WSS various no potential longer clusters decreases in significant order with to each additional cluster. The optimal value is determined visually by the value at the elbow of the plot. In this case, it is three. **Center panel**: Silhouette score plot used to determine the optimal number of clusters. It involves calculating the Silhouette score for K number of clusters and determining the cluster with the highest score. In this study, we have three to be the optimal value. **Right panel**: Hierarchical clustering dendrogram obtained using the Complete Linkage method.

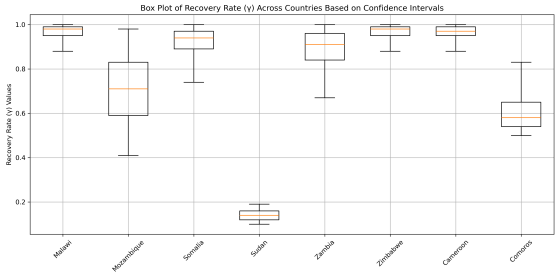
This confirms our previous assertion of three clusterings, and the best linage method is the complete linkage.

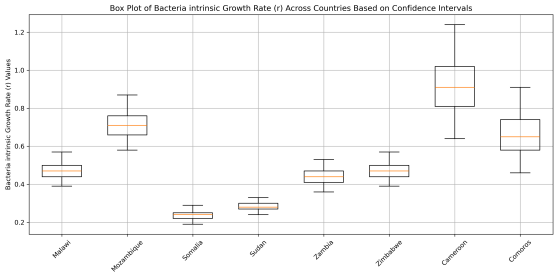
Chapter 8

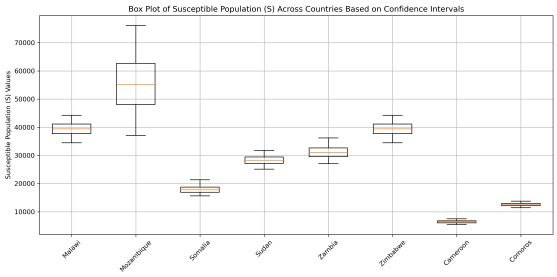
Discussion and Conclusion

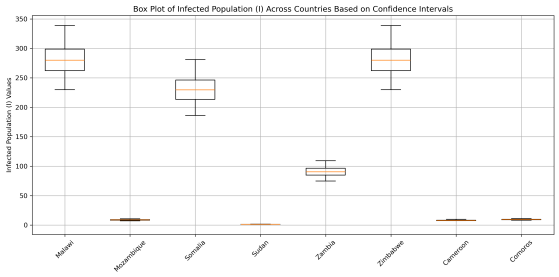
The differences in cholera outbreak patterns vary across Cameroon, Comoros, Malawi, Mozambique, Somalia, Sudan, Zambia, and Zimbabwe, offering insights into how the infection spreads in Africa.











Maximum Infection Rate (a)

There are differences in the maximum infection rate, A, among the countries under study. The maximum rate was found in Malawi at a rating of 0.83, which indicates a severe outbreak situation with a high transmission rate. The lowest value was that of Zimbabwe, 0.59, which indicated a less violent outbreak. Other noteworthy values were those of Cameroon, 0.93; Comoros, 0.64; Mozambique, 0.72; Somalia, 0.65; Sudan, 0.92; and Zambia, 0.83. These maximum infection rates differ from country to country. They are accounted for by differences in the intensity of outbreaks, possibly due to variations in public health infrastructure, intervention strategies, and population behaviour.

Recovery Rate γ

The recovery rates (γ) also had high variations among countries. In Comoros, the recovery rate ranged between 0.50 and 0.65, with a mean of 0.60, while in Mozambique, it ranged from as low as 0.41 to as high as 0.98 with a mean of 0.71. This means that health outcomes differ and might characterize varying levels of healthcare quality or accessibility. In sharp contrast, Zimbabwe exhibited a stable overall recovery rate of 0.97, ranging from 0.88 to 1.00. This could indicate that health responses have been consistent or the severity of the disease was less prominently manifested. Other countries also had very good recovery rates, including Cameroon with 0.96, Malawi with 0.61, Somalia with 0.92, Sudan with 0.85, and Zambia with 0.89.

Bacterial Growth Rate r

The growth rate of bacteria denoted as r was generally quite small and consistent across countries with minimal deviation. Mozambique had a rate of 0.72 ranging from 0.58 to 0.87, suggesting slightly more variability in bacterial growth than other regions. In contrast, Cameroon recorded a rate of 0.92, Comoros (0.66) and Malawi (0.1), Somalia (0.24), Sudan (0.28), and Zimbabwe (0.47), and Zambia at (0.44) for the bacteria growth rates. These low growth rates may imply differing environmental factors and strain characteristics influencing the outbreaks.

Initial Susceptible Population (S)

Regarding the susceptible population denoted as S, there was significant variability in the estimates reflecting differences in population exposure and potential spread of outbreaks. Zambia also showed high variability, with an average estimate of $31,178$. This wide range could stem from challenges in accurately reporting the at-risk population. In nations, the figures were as follows: Cameroon $(6,439)$, Comoros $(12,564)$; Malawi ($56,665$), Sudan (27291) Mozambique $(55,546)$, Somalia (17963) and Zimbabwe $(39,425)$.

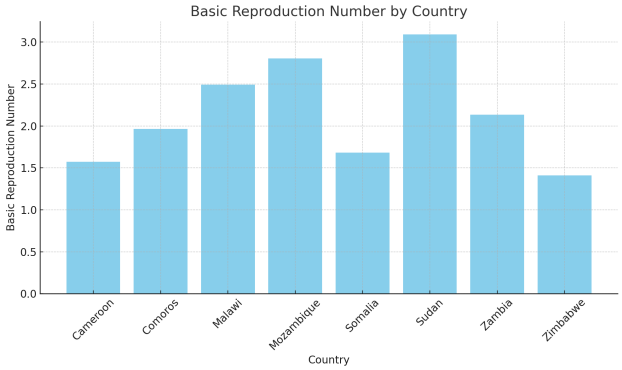


Figure 8.1: Basic Reproduction Number across countries

The Initial Infected population (I)

The estimated infected population ranges differently between the countries, which can be attributed to varying transmission dynamics and control measures. In Malawi, the infection levels were the highest, with an average of 936. Fluctuated greatly between 476 and 1657. Zambia and Somalia also had infection rates averaging 90 and 230, respectively, indicating spread but with moderate variability. On the other hand, Sudan had an average of 1, indicating well-controlled infections with minimal spread. Countries like Mozambique(8) and Zimbabwe(281), along with Comoros and Cameroon, have an average of 9 and 7, indicating a controlled spread of infections in these areas. The differences in infection rates point out how environmental factors, varying intervention approaches, and the characteristics of the strains impact the outbreaks in these regions overall.

The Shredding rate (ξ)

The shedding rate parameter, ξ, varies widely across different regions. Cameroon exhibited the highest shedding rate, with a mean of 128.32. Zambia followed with an average ξ of 35.77. Zimbabwe's shedding rate parameter averaged 19.36. Somalia had a mean ξ of 13.27.

For Comoros, the human shedding rate averaged 22.56. In Malawi, the shedding rate was lower, with an average of 1.75. Sudan's shedding rate parameter had an average of 14.85.

These differences in ξ values across regions highlight the influence of environmental and public health factors, localized interventions, and access to resources in managing cholera transmission dynamics.

Basic Reproduction Number (R_0)

The basic reproduction number, considered standard in measuring the potential of diseases to spread, had a fairly modest range. The country with the highest value was Mozambique, with an R_0 of 2.8, hence a high transmission potential warranting any urgent control measure. The lowest R_0 was for Zimbabwe at 1.41, thus with low transmission risk but still above the

threshold for epidemic potential. Other R_0 values were for Cameroon, 1.42; Comoros, 1.97; Malawi, 2.39; Sudan, 2.24; Somalia, 1.68; and Zambia, 2.24. These values underscore the differing levels of outbreak control required in each country.

8.1 Conclusion

The findings indicate that severe outbreaks, as indicated by the highest infection rate and reproduction number, are observed in Sudan and Malawi. This highlights the need for public health measures. Conversely, nations like Zimbabwe have low infection rates but still require ongoing monitoring and preventive actions to prevent outbreak escalation. The various parameters varied, emphasizing the importance of tailored public health responses based on a country's context and available resources. As a result, this comparative analysis enables targeted interventions for improved case management and resource allocation in these regions to manage and contain cholera outbreaks.

This research found that cholera outbreaks in countries are driven by factors unique to each region through hierarchical clustering of confirmed cases and estimated parameter analysis. Based on their cholera dynamics, interventions should be tailored to address the challenges faced by each country.

The clustering analysis identified three primary groups:

1. **Natural Disaster-Affected Countries** (Malawi, Mozambique): Countries in these regions often face calamities such as floods and cyclones that damage infrastructure and increase the chances of water pollution risks rising. The categorization has grouped these nations based on vulnerabilities, which indicates a mutual requirement for infrastructure and disaster management plans to curb cholera outbreaks after these occurrences. This classification underscores the significance of infrastructure and readiness measures to tackle hurdles.

2. **Conflict-Affected Countries** (Somalia, Sudan): The analysis of clusters also brought together countries affected by conflicts where damaged infrastructure restricts water availability and proper sanitation facilities. This group highlights the difficulties regions face where instability hinders public health initiatives. Nations within this group would see impacts from sanitation efforts, better healthcare accessibility and increased emphasis on hygiene support, especially in communities displaced and at risk of cholera outbreaks. Recognizing this group emphasizes the importance of intervention strategies designed for conflict regions.

3. **Countries with Chronic Sanitation Issues** (Cameroon, Comoros): This group of countries, clustered due to A collection of nations facing issues with access to clean water and sanitation facilities, often sees repeated outbreaks of cholera infections. Analyzing these countries as a group reveals lasting weaknesses that call for investments in upgrading water resources and sanitation services to reduce the risk of cholera out-

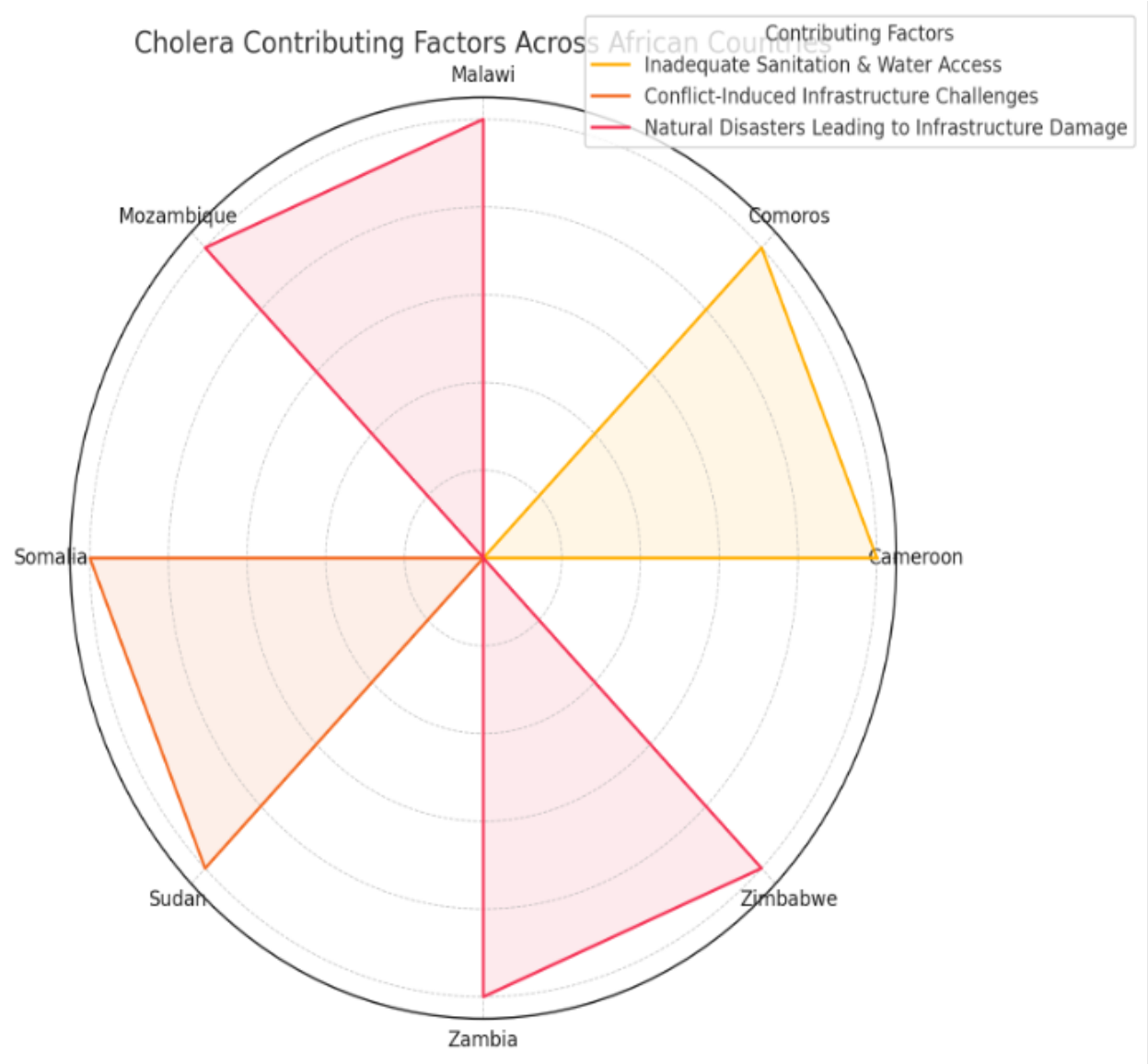


Figure 8.2: Spider Plot for Country's Cholera Contribution Factor

breaks over time effectively. This grouping implies that efforts in these areas should focus on enhancing water and sanitation infrastructure.tr

From the analysis, the dendrograms gave us a visual representation of patterns and relationships between countries that share common factors contributing to cholera outbreaks. By looking at metrics like correlation, silhouette score, and the Davies Bouldin index, results showed that using the linkage method helped create distinct clusters that shed light on how cholera dynamics vary across different regions.

The clustering analysis results show that the various socio-environmental factors that impact cholera outbreaks in parts of Africa are interconnected and intricate. The research emphasizes the importance of developing prevention and response plans for cholera by recognizing specific groups that share similar risks. Disaster readiness plays a crucial role in regions prone to natural calamities. Immediate interventions are vital in areas affected by conflicts, while countries dealing with persistent sanitation challenges require sustained infrastructure development efforts. Focusing on these aspects in every group can significantly enhance the public's health. Lessening the occurrence of cholera while also strengthening resilience in susceptible areas.

Bibliography

[1] M. Ali, A. L. Lopez, Y. A. You, Y.-E. Kim, B. Sah, B. Maskery, and J. Clemens. The global burden of cholera. Bulletin of the World Health Organization, 90(3):209–218, 2012.

[2] O. D. R. Annisha, Z. Li, N. M. D. Stenay Junior, and O. O. Donde. The perception and expectation of wash technology services in pointe-noire ville and tandou-boma, republic of congo through novel conventional-servqual-ahp model. Urban Water Journal, 19(2):119–129, 2022.

[3] A. A. Ayoade, M. Ibrahim, O. J. Peter, and F. A. Oguntolu. A mathematical model on cholera dynamics with prevention and control. Covenant Journal of Physical & Life Sciences (CJPL), 6(1):46–54, 2018.

[4] M. V. Ayyappan, P. Kishore, S. K. Panda, A. Kumar, D. Uchoi, R. K. Nadella, H. Priyadarshi, M. C. Obaiah, D. George, M. Hamza, S. K. Ramannathan, and C. N. Ravishankar. Emergence of multidrug resistant, ctx negative seventh pandemic vibrio cholerae o1 el tor sequence type (st) 69 in coastal water of kerala, india. Scientific Reports, 14(1):1–15, 2024.

[5] M. Azizi and F. Azizi. History of cholera outbreaks in iran during the 19th and 20th centuries. Middle East Journal of Digestive Diseases, 2(1):51–55, 2010.

[6] R. J. Biggar, M. Melbye, H. A. Eriksen, and M. Frisch. Electronic health records in epidemiology: opportunities and challenges. Epidemiology (Cambridge, Mass.), 29(3):331–338, 2018.

[7] M. E. Birmingham, L. A. Lee, N. Ndayimirije, S. Nkurikiye, B. S. Hersh, J. G. Wells, and M. S. Deming. Epidemic cholera in burundi: patterns of transmission in the great rift valley lake region. The Lancet, 349(9057):981–985, 1997.

[8] W. Boghurst. Loimographia: An Account of the Great Plague of London in the Year 1665. AMS Press, 1894.

[9] S. Bowong, J. J. Tewa, B. Momeni, and J. Kurths. Parameter estimation in dynamic models of infectious diseases. Mathematical Methods in the Applied Sciences, 39(14):4036–4054, 2016.

[10] V. Capasso and S. L. Paveri-Fontana. A mathematical model for the cholera epidemic in the european mediterranean region. Revue d'Épidémiologie et de Santé Publique, 27:121–132, 1979.

[11] D. Chac, C. N. Dunmire, J. Singh, and A. A. Weil. Update on environmental and host factors impacting the risk of vibrio cholerae infection. ACS Infectious Diseases, 7(5):1010–1019, 2021. PMID: 33844507.

[12] A. Champion, J. Hartley, A. Rogers, J. Phillips, and N. Summers. British Topography and Local History. David and Charles, Newton Abbot, UK, 1977.

[13] P. L. Delamater, E. J. Street, T. F. Leslie, Y. T. Yang, and K. H. Jacobsen. Complexity of the basic reproduction number (r0). Emerging Infectious Diseases, 25(1):1–4, 2019.

[14] O. Diekmann, J. A. P. Heesterbeek, and J. A. J. Metz. On the definition and the computation of the basic reproduction ratio r_0 in models for infectious diseases in heterogeneous populations. Journal of Mathematical Biology, 28:365–382, 1990.

[15] O. Diekmann, J. A. P. Heesterbeek, and J. A. J. Metz. On the definition and the computation of the basic reproduction ratio r0 in models for infectious diseases in heterogeneous populations. Journal of Mathematical Biology, 28(4):365–382, 1990.

[16] S. R. Dominguez, P. N. Doan, and F. Rivera-Chávez. The intersection between host–pathogen interactions and metabolism during vibrio cholerae infection. Current Opinion in Microbiology, 77:102421, 2024.

[17] M. Echenberg. Africa in the Time of Cholera: A History of Pandemics from 1817 to the Present, volume 114 of African Studies. Cambridge University Press, 2011.

[18] G. A. Erkyihun, N. Asamene, and A. Z. Woldegiorgis. The threat of cholera in africa. Zoonoses, 3(1):20230027, 2023.

[19] M. Eruaga and K. Davis. Evaluation of household water treatment technologies for cholera eradication in sub-saharan africa: Epidemiological and economic perspectives. Sustainability, 16(4):1422, 2024.

[20] A. S. Evans. Re: Definitions of epidemiology [letter]. American Journal of Epidemiology, 109:379–381, 1979.

[21] R. A. Finkelstein. Cholera, Vibrio cholerae O1 and O139, and Other Pathogenic Vibrios, chapter 24. University of Texas Medical Branch at Galveston, 4th edition, 1996.

[22] E. Frerot and A. Gagneur. The evolution of epidemiology definitions. International Journal of Environmental Research and Public Health, 14(5):535, 2017.

[23] M. Frérot, A. Lefebvre, S. Aho, P. Callier, K. Astruc, and L. S. Aho Glélé. What is epidemiology? changing definitions of epidemiology 1978-2017. PloS One, 13(12):e0208442, 2018.

[24] I. C.-H. Fung. Cholera transmission dynamic models for public health practitioners. Emerging Themes in Epidemiology, 11(1):1, Feb 2014.

[25] K. Goh, S. Teo, S. Lam, and M. Ling. Person-to-person transmission of cholera in a psychiatric hospital. Journal of Infection, 20(3):193–200, May 1990.

[26] W. Hamer. Epidemic disease in england—the evidence of variability and of persistence. The Lancet, 167:733–738, 1906.

[27] J. B. Harris. Cholera: Immunity and prospects in vaccine development. Journal of Infectious Diseases, 218(Suppl 3):141–146, 2018.

[28] J. B. Harris, R. C. LaRocque, F. Qadri, E. T. Ryan, and S. B. Calderwood. Cholera. The Lancet, 379(9835):2466–2476, 2012.

[29] Infonet-Biovision. Introduction to hygiene and sanitation, 2024. Accessed: 2024-07-28.

[30] R. I. Joh, H. Wang, H. Weiss, and J. S. Weitz. Dynamics of indirectly transmitted infectious diseases with immunological threshold. Bulletin of Mathematical Biology, 71:845–862, 2009.

[31] R. Kahn, C. M. Peak, J. Fernández-Gracia, A. Hill, A. Jambai, L. Ganda, M. C. Castro, and C. O. Buckee. Incubation periods impact the spatial predictability of cholera and Ebola outbreaks in Sierra Leone. Proceedings of the National Academy of Sciences of the United States of America, 117(9):5067–5073, 2020.

[32] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, 115(772):700–721, 1927.

[33] R. Kirby. Cholera and public health: The bacteriophage intervention. The Lancet Infectious Diseases, 9(7):421–422, 2009.

[34] J. D. Kong, W. Davis, X. Li, and H. Wang. Stability and sensitivity analysis of the isir model for indirectly transmitted infectious diseases with immunological threshold. SIAM Journal on Applied Mathematics, 74(5):1418–1441, 2014.

[35] A. M. Lilienfeld. Foundations of Epidemiology. Oxford University Press, New York, 1976.

[36] Y. Liu, Y. Chen, and Y. Liu. Machine learning in epidemiology. Journal of Epidemiology and Community Health, 74(6):477–485, 2020.

[37] R. M. May. Mathematical models and ecology: Past and future. In S. A. Levin, S. I. Rubinow, and L. J. Gross, editors, Mathematical Models in Biological Science, pages 73–87. Springer, New York, NY, 1977.

[38] M. N. Momba, P. O. Bessong, C. L. Obi, M. K. Serepa-Dlamini, and A. I. Okoh. Vibrio cholerae and cholera. Water and Health in Developing Countries, pages 115–136, 2017.

[39] S. M. Moore, A. S. Azman, B. F. Zaitchik, E. D. Mintz, J. M. Brunkard, D. Legros, A. Hill, H. S. McKay, and F. J. Luquero. El niño and the shifting geography of cholera in africa. Proceedings of the National Academy of Sciences, 114(17):4436–4441, 2017.

[40] Z. Mukandavire, D. L. Smith, J. G. Morris Jr, G. B. Nair, Y. You, H. M. Beckles, K. P. Geissler, F. J. Luquero, J. M. Brunkard, and D. Legros. Estimating the reproductive number of the Haiti cholera epidemic. Proceedings of the National Academy of Sciences, 108(21):8767–8772, 2011.

[41] K. M. Murphy, P. Travers, and M. Walport. Janeway's Immunobiology. Garland Science, New York, 7th edition, 2007.

[42] A. Mwasa and J. M. Tchuenche. Mathematical analysis of a cholera model with public health interventions. Bio Systems, 105(3):190–200, 2011.

[43] J. Naidoo, K. Patric, D. J. Surmon, P. Moodley, M. P. Nicol, and A. W. Sturm. Cholera in children in an endemic area. Journal of Tropical Pediatrics, 48(3):156–160, 2002.

[44] E. J. Nelson, J. B. Harris, J. G. Morris, S. B. Calderwood, and A. Camilli. Cholera transmission: The host, pathogen and bacteriophage dynamics. Nature Reviews Microbiology, 7(10):693–702, 2009.

[45] H. D. Nyamogoba and A. A. Obala. Combating diarrhoea in kenya: the impact of policy and programme strategies. Journal of Diarrhoeal Diseases Research, pages 31–37, 2002.

[46] M. O. Onuorah, F. A. Atiku, and H. Juuko. Mathematical model for prevention and control of cholera transmission in a variable population. Research in Mathematics, 9(1):2018779, 2022.

[47] G. Pande, B. Kwesiga, G. Bwire, P. Kalyebi, A. Riolexus, J. K. Matovu, and B. P. Zhu. Cholera outbreak caused by drinking contaminated water from a lakeshore water-collection site, kasese district, south-western uganda, june-july 2015. PloS One, 13(6):e0198431, 2018.

[48] M. D. Phelps, M. L. Perner, P. Keating, D. L. Smith, and A. S. Azman. Individuals' food choices affect the transmission of foodborne pathogens: implications for control of enteric infections in human populations. Epidemiology and Infection, 147:e162, 2019.

[49] Public Health Agency of Canada. For health professionals: Cholera, 2019. Accessed: 2024-12-06.

[50] X. Rodó, J. Ballester, D. Cayan, H. Melonie, N. Nicholls, G. Schumann, and N. C. Stenseth. Climate change and infectious diseases in europe: impact, projection and adaptation. The Lancet Planetary Health, 5(6):e416–e426, 2021.

[51] T. J. Silhavy, D. Kahne, and S. Walker. The bacterial cell envelope. Cold Spring Harbor Perspectives in Biology, 2(5):a000414, 2010.

[52] R. Stein and M. Chirilă. Foodborne Diseases. Elsevier, 2017.

[53] G. Sun, J. Xie, S. Huang, Z. Jin, M. Li, and L. Liu. Transmission dynamics of cholera: Mathematical modeling and control strategies. Communications in Nonlinear Science and Numerical Simulation, 45:235–244, 2017.

[54] S. B. Thacker, J. V. Bennett, T. F. Tsai, D. W. Fraser, J. E. McDade, C. C. Shepard, and T. C. Eickhoff. An outbreak in 1965 of severe respiratory illness caused by the legionnaires' disease bacterium. Journal of Infectious Diseases, 138(4):512–519, 1978.

[55] A. J. Van Alst, L. M. Demey, and V. J. DiRita. Vibrio cholerae requires oxidative respiration through the bd-i and cbb3 oxidases for intestinal proliferation. PLoS Pathogens, 18(5):e1010102, 2022.

[56] P. van den Driessche and J. Watmough. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. Mathematical Biosciences, 180:29–48, 2002.

[57] J. Wang. Mathematical models for cholera dynamics-a review. Microorganisms, 10(12):2358, 2022.

[58] World Health Organization. Water, sanitation and hygiene strategy 2018–2025. Geneva: World Health Organization, 2018.

[59] World Health Organization. Cholera, 2019. Retrieved from `https://www.who.int/news-room/fact-sheets/detail/cholera`.

[60] World Health Organization. Global cholera and acute watery diarrhea dashboard, 2024. Accessed: 2024-07-28.

[61] A. K. Yadav, S. Kumar, G. Singh, and N. K. Kansara. Demystifying r naught: Understanding what does it hide? Indian Journal of Community Medicine, 46(1):7–14, 2021.

[62] C. Yang, X. Wang, D. Gao, and J. Wang. Impact of awareness programs on cholera dynamics: Two modeling approaches. Bulletin of Mathematical Biology, 79(9):2109–2131, 2017.

[63] J. Zhu and J. J. Mekalanos. Quorum sensing-dependent biofilms enhance colonization in vibrio cholerae. Developmental Cell, 5(4):647–656, 2003.