Comparison of the 2022 Monkeypox (Mpox) Outbreak Using Mathematical Modeling and Time Series Clustering

A thesis

presented to

the faculty of the Department of Mathematics

East Tennessee State University

In partial fulfillment

of the requirements for the degree

Master of Science in Mathematical Sciences

by

Mark-Daniels Tamakloe

May 2023

Michele Lynn Joyner, Ph.D., Chair

Jeff Randall Knisley, Ph.D.

Ariel Cintron-Arias, Ph.D.

Keywords: Mathematical Model, Monkeypox virus, Bootstrapping, Clustering.

ABSTRACT

Comparison of the 2022 Monkeypox (Mpox) Outbreak Using Mathematical Modeling and Time Series Clustering

by

Mark-Daniels Tamakloe

Monkeypox virus (MPXV) is the causative agent of monkeypox (mpox), a rare viral disease that affects humans [1]. It is primarily found in Africa and is transmitted to humans through contact with sick animals, particularly rodents and monkeys, or through human-to-human transmission [2]. From the beginning of May 2022, cases of mpox have been recorded from non-endemic nations, and the illness has continued to be reported in other endemic nations. Majority of confirmed cases have been recorded in Europe and North America. In this thesis, we compare the spread of the outbreak across the top ten countries using a combination of two different techniques. First, we look at the similarity of the outbreak from a mathematical modeling point of view using a simple SIR model to describe the dynamics of the spread and compare parameters of the model among most prevalent countries. Using the model as the general trend of the outbreak, we then look at the spread from a clustering perspective, grouping countries based on a time-series clustering technique.

DEDICATION

To my beautiful wife and mother, whose steadfast love, support, and sacrifice have served as the bedrock of my scholastic path. You have been my rock and my inspiration, giving me the strength and support I needed to overcome the obstacles and disappointments that have come my way. My thesis is dedicated to you, with deep thanks and respect for everything you have done for me. Your love and support have been crucial to my achievement, and I will be eternally thankful.

ACKNOWLEDGMENTS

I would like to begin by expressing my utmost gratitude to God Almighty, whose unwavering grace, guidance, and blessings have been the foundation of my life and academic journey. Without His divine mercy and assistance, this accomplishment would not have been possible.

I would also like to thank my thesis advisor, Dr. Michele Joyner, for her outstanding guidance, support, and mentorship throughout my research journey. Her insights and feedback have significantly improved the quality of my work. I am honored to have had the opportunity to work with such an exceptional advisor and scholar. Dr. Joyner has not only been a mentor but also a role model, and I will always cherish the lessons and experiences I have gained from her.

I am grateful to the members of my thesis committee, Dr. Jeff Knisley, and Dr. Ariel Cintron-Arias, for their time, effort, and valuable feedback on my thesis. Their constructive criticism has been instrumental in shaping my research.

I would like to express my appreciation to my family and friends, who have been a constant source of love, support, and encouragement. Their steadfast belief in me has been a driving force behind my success.

Finally, I would also like to acknowledge the staff and faculty of the Department of Mathematics and Statistics, whose dedication and hard work have provided me with an excellent academic environment in which to learn and grow.

Once again, I express my heartfelt thanks to God Almighty, and to all the individuals who have supported and encouraged me throughout my master's thesis journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Monkeypox virus (MPXV) is a zoonotic virus that causes an infectious disease in humans and non-human primates [1]. While it is a rare disease, outbreaks of monkeypox (mpox) have been reported in Central and West Africa, as well as in the United States, and some parts of Europe [2]. The disease was first discovered in 1958 when outbreaks occurred among monkeys kept for research purposes [1]. Since then, sporadic outbreaks of monkeypox have been reported in Central and West African countries, with human cases first reported in 1970 in the Democratic Republic of Congo (DRC) [3]. The incubation period is typically 3 - 17 days and during this period, a person does not have symptoms and may feel fine [5]. The symptoms of mpox are similar to those of smallpox and include fever, headache, muscle aches, and a rash that typically starts on the face and then spreads to other parts of the body [6]. In severe cases, mpox can cause complications such as pneumonia, sepsis, and even death [4]. It can initially look like pimples or blisters and may be itchy. The rash will go through several stages, including scabs, before healing [4]. There is currently no particular therapy or vaccination for the mpox virus, and treatment consists mostly of supportive care. Preventive actions such as avoiding contact with diseased animals and exercising excellent hygiene, on the other hand, can help lower the chance of transmission [8].

In recent years, there have been several outbreaks of mpox in African countries, but in 2022, the disease was reported in several countries outside of Africa, including the United States and several European countries. According to the World Health Organization (WHO), the top ten most prevalent mpox countries as of November, 2022 are United States, Brazil, Spain, France, Colombia, United Kingdom, Peru, Mexico, Germany, and Canada [43]. Figure 1 shows a comparison of the total number of infections from March 2022 through November 2022 for the top ten countries whiles Figures 2 and 3 show the progression of the outbreak across the most prevalent months.

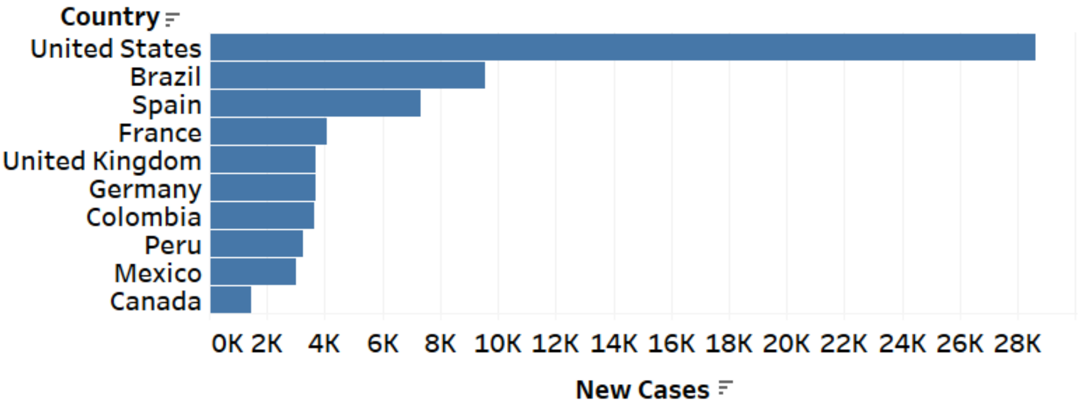

Figure 1: Total Cases of Monkeypox Outbreak Per Country based on WHO data [43].

Figure 1 shows that the country with the highest number of mpox cases as at the end of November 2022 was the United States ($n = 28,379$), followed by Brazil ($n = 9162$), Spain ($n = 7317$), France ($n = 4094$), the United Kingdom ($n = 3698$), Germany ($n = 3662$), Colombia ($n = 3298$), Peru ($n = 3048$), Mexico ($n = 2654$), and Canada ($n = 1437$) [16]. Also, the number of cases in the United States far exceeds all other countries in the top ten whiles Brazil and Spain are really close in terms of the number of cases. France, United Kingdom, Germany, Colombia, Peru, Mexico all have almost equal number of cases with Canada having the least number of mpox cases.

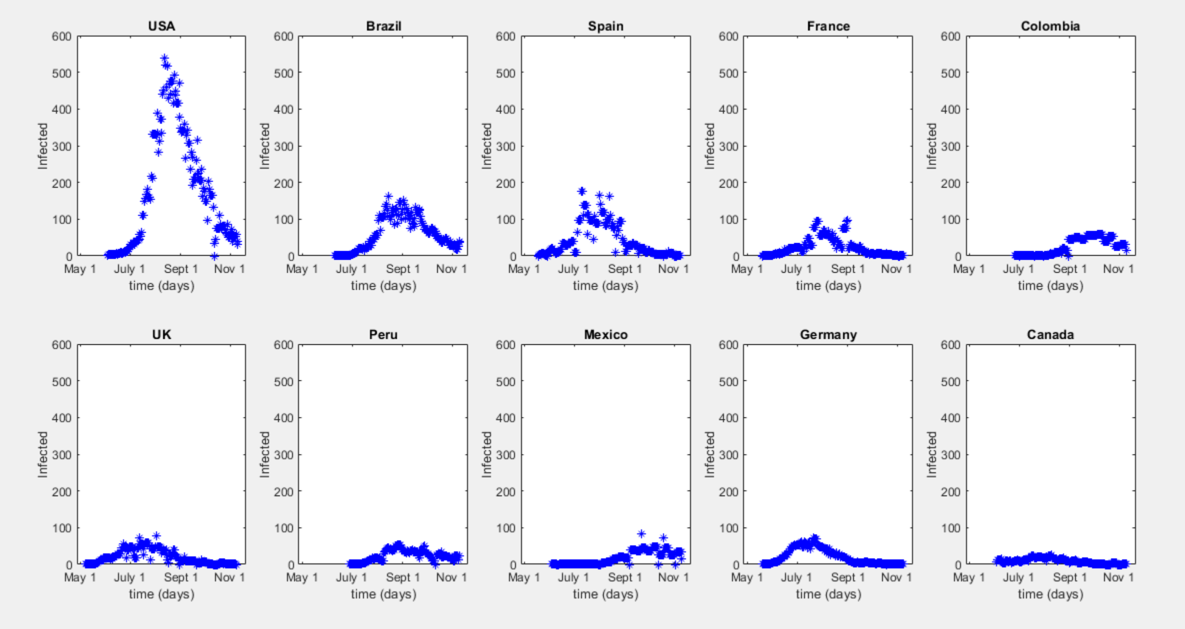

Figure 2: Time Series Progression for monkeypox in the top 10 plotted on the same scale [43].

From Figure 2 above, it is easy to tell there is an outbreak but difficult to discern any similarities in the outbreak when plotted on the same scale. Also, the initial start date of the outbreak seems to be different across each country with some variation in peak levels of the infection. In Figure 3, we plot the outbreak on their own respective scales.

By plotting the graphs of the top ten countries on their respective scales, we can see that all the countries have fairly defined peaks although some of the peaks are wider than others indicating that the outbreak might have stayed at its peak for longer in some countries than others. For example, USA and Germany seem to have some similarities in the shapes of the peaks which indicates that the outbreak rises steadily, attains a maximum at the peak and then falls steadily again, whiles Brazil

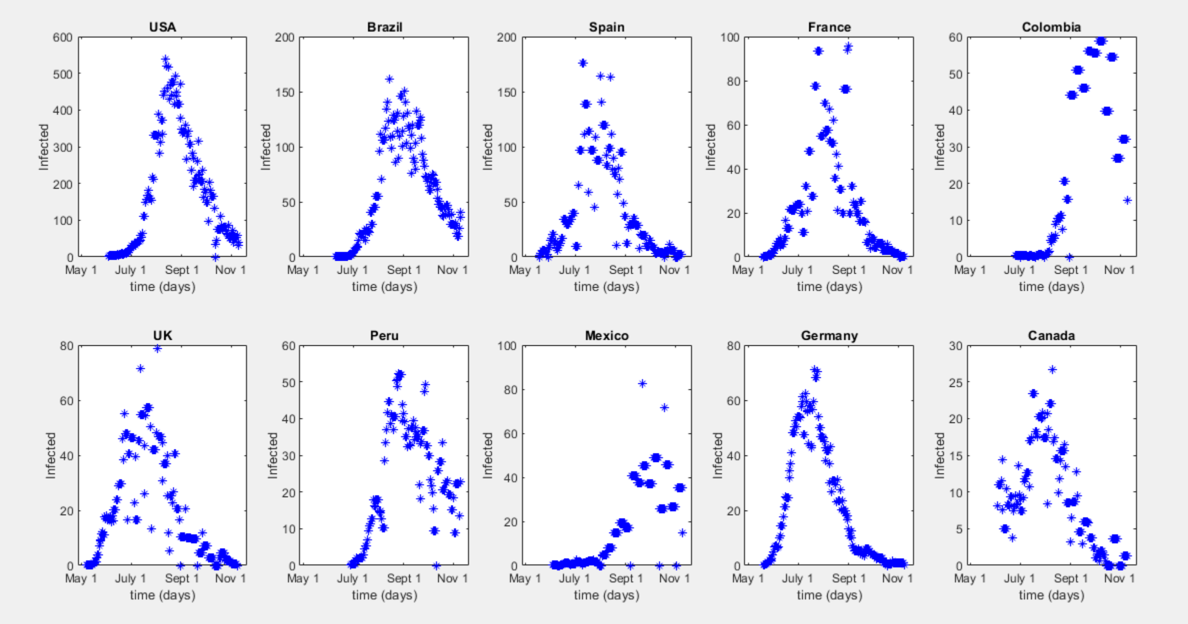

Figure 3: Time Series Progression for monkeypox in the top 10 plotted on different scales [43].

and UK also have shapes that shows that the outbreak lasted longer at their peaks.

The aim of this thesis is to compare the spread of the mpox across the top ten countries by combining two different approaches. First, we look at the similarity of the spread by using an epidemiological model to describe the dynamics of the spread and compare parameters of the model among the most prevalent countries. Then, we use time series data to identify the trend of the disease and group countries based on clustering techniques.

This thesis is organized into five chapters. In Chapter 2, we discuss mathematical modeling using the United States mpox data as an example. In this section, we develop both a SIR and SEIR model for mpox. We then set up inverse problem for parameter estimation and compare the two models using the Akaike Information

Criterion (AIC). We then finally display modeling results for all ten countries. In Chapter 3, we introduce the bootsrapping method to compute bootstrap intervals for the model parameters and then compare the results for the top ten countries. In Chapter 4, time series clustering is introduced and implemented to group the countries into clusters based on the similarity of the outbreak between countries. We then conclude in Chapter 5 with a summary and future work.

2 MATHEMATICAL MODELING FOR U.S. DATA

Mathematical modeling is a powerful technique that helps scientists, engineers, and researchers to simplify and abstractly describe and evaluate complicated systems. Mathematical models can help researchers in understanding the behavior of complex systems, forecasting their future behavior, and identifying potential interventions or solutions to issues by employing mathematical equations, algorithms, and other analytical tools. It has numerous applications in a wide range of disciplines, including physics, engineering, biology, economics, finance, ecology, and social sciences, among many others [7]. In general, mathematical models can be divided into two broad categories: deterministic models and stochastic models [9]. Deterministic models are based on a set of equations that describe the behavior of a system with certainty, while stochastic models incorporate randomness or uncertainty into the equations [10]. The study of infectious disease transmission and the evaluation of the effects of treatments like immunization, quarantine, and social isolation are both done in epidemiology using mathematical models. The formal definition of the term *mathematical modeling* as used in this thesis is given in Definition 2.1:

Definition 2.1. *Mathematical modeling is the process of using mathematical equations and tools to represent and analyze real-world systems and phenomena. It involves creating a simplified representation of a system, often with the use of mathematical symbols and equations, and using it to make predictions or test hypotheses about the behavior of the system* [11].

There are a group of models, epidemiological models, used to model the transmission of infectious diseases in the population. In this thesis, we consider two epidemiological models which are the SIR (Susceptible-Infected-Recovered) and the SEIR (Susceptible-Exposed-Infected-Recovered) model[12]. The SEIR model seems most appropriate for this thesis because of the incubation period of the mpox virus; however, the SIR is a simpler model.

2.1 The SIR Model

The SIR model is a mathematical model that is used to study and forecast the spread of infectious diseases[13]. At any given time t, the population is divided into three groups under the model: susceptible (S), infected (I), and recovered (R). The model assumes that the rate of infection is proportional to the number of susceptible and infected persons, and that the rate of recovery is related to the number of infected individuals. The model also assumes that once a person recovers from the sickness, they are immune and cannot be infected again. The dynamics of the model are shown in Figure 4.

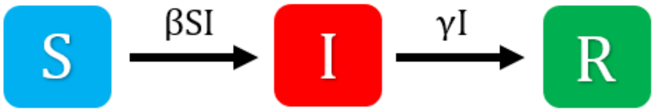

Figure 4: Compartmental diagram of the SIR epidemic model

The set of differential equations that describes the SIR model are given by

$$\begin{aligned}\frac{dS}{dt} &= -\beta SI \\ \frac{dI}{dt} &= \beta SI - \gamma I \\ \frac{dR}{dt} &= \gamma I\end{aligned} \tag{1}$$

where β represents the effective transmission rate, γ is the recovery rate and the total population is given by $N = S + I + R$, which is constant over time because $\frac{dN}{dt} = 0$. The variables S, I, and R in epidemiological models such as the SIR or SEIR models describe the number of persons in different disease states and are time dependent. Sometimes, the equation includes standardizing of the β parameter by the size of the population N, but we instead account for the value of N when comparing the effective reproduction number, R_0, between countries in Section 3.

2.2 The SEIR Model

The SEIR model is a derivative of the SIR model in that it has an additional compartment called the exposed compartment (E) which is included to capture the latency period of the infection [14]. This model assumes a latent period which means there is a period of time during which an infected individual is asymptomatic but still infectious. The dynamics of the SEIR model are shown in Figure 5.

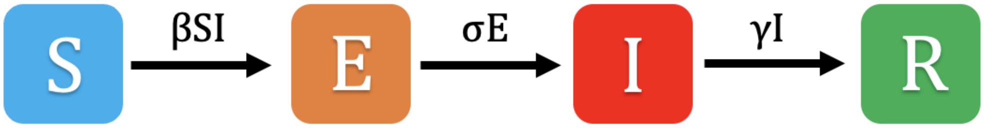

Figure 5: Compartmental diagram of the SEIR epidemic model

The set of ordinary differential equations that describe the SEIR model are given by

$$\begin{aligned}\frac{dS}{dt} &= -\beta SI \\ \frac{dE}{dt} &= \beta SI - \sigma E \\ \frac{dI}{dt} &= -\sigma E - \gamma I \\ \frac{dR}{dt} &= \gamma I\end{aligned} \tag{2}$$

where β is the effective transmission rate, σ is the incubation rate, γ is the recovery rate and the total population $N = S + E + I + R$, where N is constant and S, E, I, and R are functions of time.

2.3 Inverse Problems

Inverse problems are a type of mathematical problem in which the goal is to estimate the parameters of a mathematical or statistical model of a physical system based on observations of that system [15]. In other words, instead of predicting the output of a system given the parameters, inverse problems entail determining the parameters of a system given the observed outputs [18, 21]. In this thesis, we use the Ordinary Least-Squares (OLS) method to estimate the parameters of the SIR model [22]. Here, we assume constant variance of the Y variable where our statistical model is of the form

$$Y_j = f(t_j;\theta_0) + \epsilon_j, \quad j = 1, 2, ..., N$$

where Y_j is a random variable for the observation system at time t_j, $f(t_j;\theta_0)$ is the observed part of the solution of our statistical model with θ_0 considered to be the 'true' model parameters, ϵ_j is the measurement error with the assumption that ϵ_j is independent and identically distributed with mean zero and constant variance [23]. The OLS estimate for the parameter θ is given by

$$\theta_{OLS} = \theta_{OLS}^N(Y) = \arg\min_{\theta\in\Omega_0} \sum_{j=1}^{N} |Y_j - f(t_j;\theta)|^2.$$

In other words θ_{OLS} is estimated by minimizing

$$J(\theta) = \sum_{j=1}^{N} |y_j - f(t_j;\theta)|^2$$

where y_j is the data used for the estimation [18]. In this thesis, we wanted to estimate the parameters β which is the effective transmission rate of the virus, t_0 which is the initial time of the first infected case, and S_0 which is the initial susceptible population. Note that we assume the average rate of recovery γ is constant across this limited outbreak as the virus has not had significant time to mutate. We set $\gamma = \frac{1}{22}$ per day, where 22 days is the mean of the infection period for the mpox. However, we also assume an initial infected individual at unknown time t_0. Although cases are reported, we are assuming not all cases are reported and therefore the first infected case may occur at an unknown time. We further assume that at the initial stages of the outbreak, the entire population is not at risk, so we wanted to estimate the initial susceptible population S_0. We used the *fminsearch* algorithm in MATLAB which employs the Nelder-Mead simplex method, an optimization algorithm to estimate our parameters.

The Nelder-Mead simplex technique is a derivative-free optimization procedure that iteratively searches for the minimum of a function using a geometric approach with a simplex (a set of vertices)[19]. The parameter estimation problem in the context of the SIR and SEIR models is frequently described as a convex optimization problem. Because the models are linear combinations of nonlinear components, the parameters act as coefficients in these combinations. As a result, the objective function in the optimization problem becomes convex. Convexity assures that the Nelder-Mead algorithm, a more advanced version of the "bisection" approach, converges to a unique minimum that is guaranteed to be within the convex polytope's interior. The convexity of the objective function and parameter space is critical because it ensures the reliability and global optimality of the Nelder-Mead algorithm estimates.

Table 1 gives the estimated parameters for US data, while Figure 6 shows the graph of the data with optimal parameters. From Figure 6, we see that the model fits the US mpox data well. Although the model "averages" the data well, we note that it does not capture the peak of the infection. This can be problematic as the peak of the infection is determined by the actual data, and our model here would suggest these peaks are outliers. This can be further witnessed by the higher residuals in Figure 7 near the peak of the infection. This is one of the problems which might be encountered with this approach. However, this averaging behavior is found in the models for almost all countries as we show in Figure 8 in Section 2.5. Figure 7 displays the residuals (ϵ_j) as a function of time which suggests that the constant variance assumption is satisfied. The assumption of constant variance is closely related to the assumption of normally distributed residuals with a mean of zero. This assumption suggests that the predicted values will be equal to the true values on average, with any difference owing to random noise or measurement error [18, 20]. If the assumption is correct, the residual plot should show a random scatter of points around a horizontal line at zero as seen in Figure 7 and there will be no fan-like pattern in the data. This suggests that the residuals are symmetrically distributed around zero and that the errors have no systematic bias or trend.

Table 1: Parameter Estimates

Parameter	**Description**	**Estimate**
β (persons/day)	Effective Transmission Rate	1.0740×10^{-4}
t_0 (days)	Time First Infected	May 31, 2022
S_0 (persons)	Susceptible Population	1389
R_0	Basic Reproduction Number	3.28

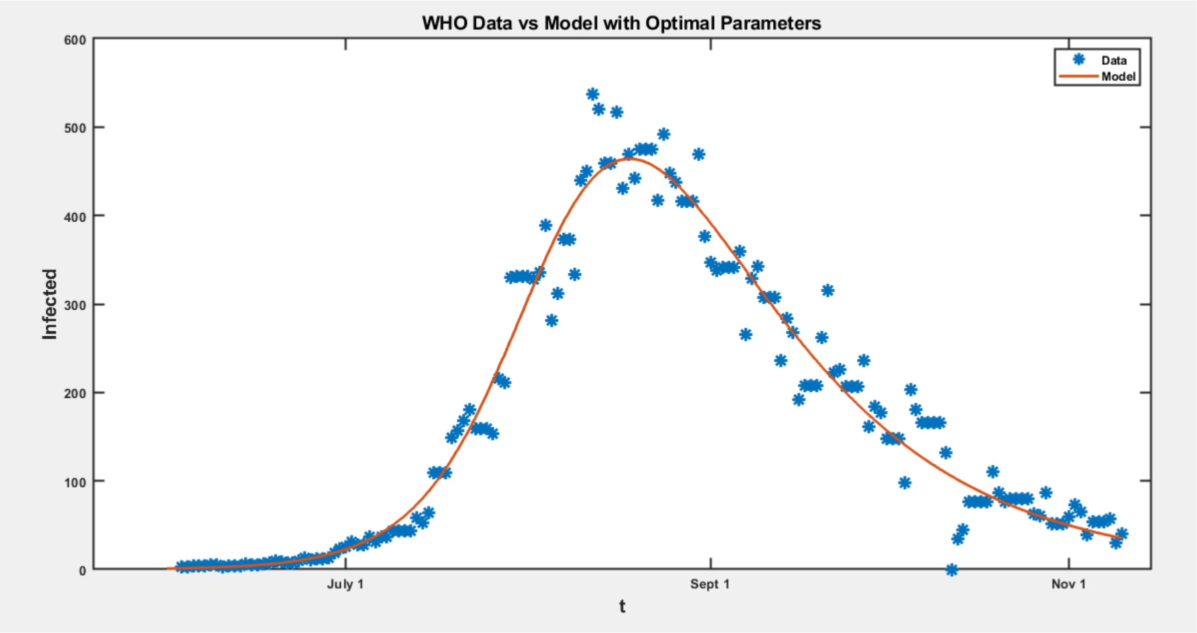

Figure 6: Graph of US mpox data with Optimal Parameters

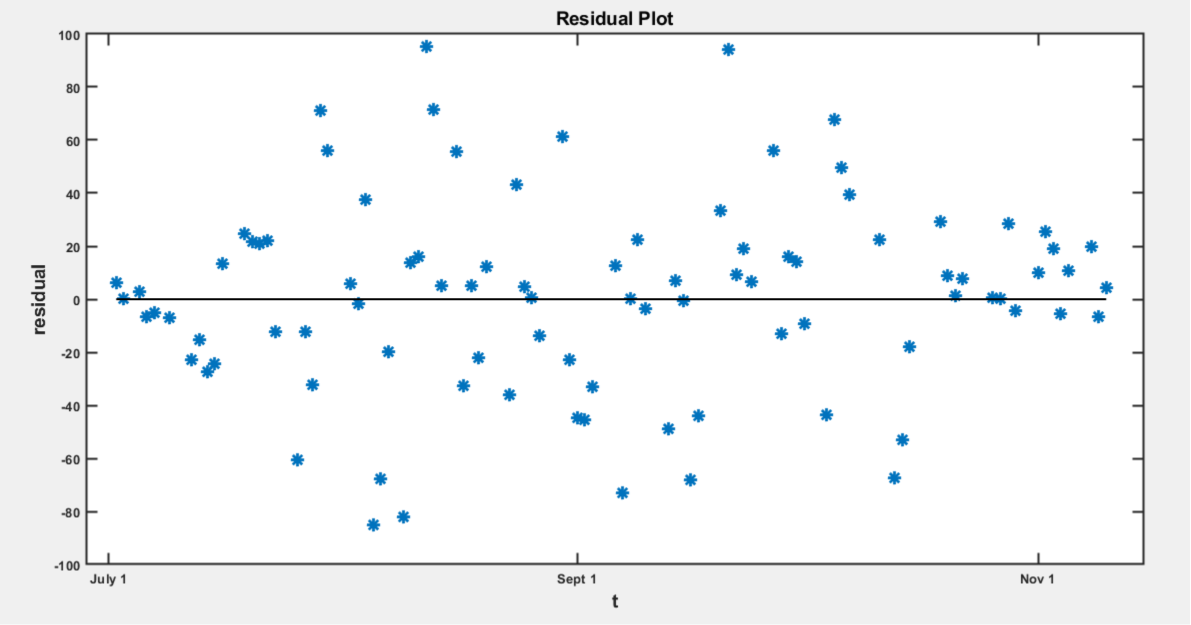

Figure 7: Residual Plot of US mpox data with Optimal Parameters

2.4 Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) is a statistical metric used to assess and select amongst alternative models based on their goodness of fit and complexity. First proposed in 1974 by Japanese statistician Hirotugu Akaike, the AIC is based on the principle that a good model should fit the data well, but not be too complex [24]. It is defined as

$$AIC = -2\ln\bigg(\mathcal{L}(\hat{\theta}_{MLE}|y)\bigg)+2\kappa_\theta$$

where log-likelihood is the maximum value of the likelihood function for the model, and κ_θ is the number of parameters in the model. Since we are using the OLS method to estimate parameters of our model, then we use the AIC under a constant variance statistical model which is given in [24] by

$$AIC_{OLS} = N\ln\left(\frac{\sum_{j=1}^{N}[y_j - f(t_j;\theta_0)]^2}{N}\right) + 2(\kappa_\theta + 1)$$

The AIC penalizes models with a larger number of parameters, because increasing the number of parameters tends to improve the fit to the data but at the expense of increased complexity. The AIC value is a relative indicator of model quality, with lower values suggesting better models. When two or more models are compared, the model with the lowest AIC value is often favored because it provides the greatest balance between goodness of fit and model complexity.

The weights, also known as the Akaike weights help us judge how much more likely the best model is compared to the next best model [24]. It is defined as

$$w_i(AIC) = \frac{\exp(-\frac{1}{2}\Delta_i(AIC))}{\sum_{k=1}^{K}\exp(-\frac{1}{2}\Delta_k(AIC))}$$

where K is the number of models being compared. The sum of the AIC weights for all the models being compared is equal to 1, so the higher the weight, the better the model. Akaike weights are useful when comparing multiple models because they provide a way to account for the uncertainty in model selection. Instead of just selecting the model with the lowest AIC score, one can use the weights to evaluate the relative likelihood of each model being the best model. In our work, we compared the two epidemiological models, SIR and SEIR to determine which is the best using the AIC.

We observed that the AIC for the SEIR model was 1110.65 with a weight of 0.185 and the AIC for the SIR model was 1107.68 with a weight of 0.815. In view of this, we chose the SIR model as the best model for this research given that it had a lower AIC and a higher weight.

2.5 Initial Model Results for All Countries

In this research, we aimed at estimating the effective transmission rate (β), the initial time of the first infected case (t_0), the initial susceptible population S_0 and then calculating the basic reproduction number R_0 for the top ten mpox countries using the formula $R_0 = \frac{\beta S_0}{\gamma}$ [25] where γ is fixed at $\frac{1}{22}$. In Table 2 below, we give the results for β, S_0, t_0 and R_0 for the top ten monkeypox countries. The β coefficient represents the effective transmission rate of the infection in each country, with higher values indicating a faster rate of transmission. The S_0 value represents the estimated size of the susceptible population at the beginning of the epidemic, and the t_0 value represents the estimated time at which the first infection occurred in each country.

Table 2: Model Results for the Top 10 Mpox Countries

Country	$\beta(\times 10^{-4})$ **persons/day**	S_0 **persons**	t_0 **(2022)** **days**	R_0
United States	1.0740	1389	May 31	3.28
Brazil	2.3356	508	May 27	2.61
Spain	3.6154	384	May 13	3.05
France	5.7474	219	May 20	2.77
United Kingdom	5.3102	208	April 17	2.43
Germany	8.0127	182	May 12	3.21
Colombia	6.9501	195	July 20	2.98
Peru	6.2642	181	June 12	2.49
Mexico	5.4011	211	July 10	2.51
Canada	13.3056	81	May 13	2.37

Based on the values in the table above, it can be seen that Canada has the highest estimated value for the beta coefficient, indicating a relatively higher rate of transmission of the infection in this country compared to the other countries in the table. The United States has the lowest estimated value for the beta coefficient, indicating a relatively slower rate of transmission. The initial susceptible populations (S_0) vary widely among the countries, with Canada having the smallest estimated susceptible population, while the United States has the largest. The initial time of the first infection (t_0) also varies among the countries, with most countries experiencing their first infection in May 2022. In the table, the R_0 numbers represent the basic reproductive number for each of the top ten mpox countries. R_0 is a critical epidemiological measure that sheds light on the dynamics of infectious disease transmission. It calculates the average number of new infections caused by a single infected person in a completely vulnerable community [26]. The mpox outbreak is expected to continue in a country if R_0 has a value > 1 and to end if R_0 is < 1 [27]. The basic reproduction number (R_0) is an important indication for understanding infectious disease transmission dynamics. When the R_0 values in the table are analyzed, it is clear that countries such as Colombia, United States, Spain, and Germany have higher R_0 values, indicating a higher level of infectiousness and perhaps greater obstacles in preventing the disease's spread. Brazil, France, Mexico, and the UK have moderate R_0 values, indicating a slightly reduced but still significant amount of infectiousness. Peru, and Canada has the lowest R_0 value, implying a lower level of infectiousness and, maybe, better disease control. When interpreting the implications of R_0 values, it is critical to consider other factors such as population density, healthcare infrastructure, and public health programs.

Figure 8 shows a subplot comparing various country models. Each subplot depicts a different country, such as the United States, Brazil, Spain, France, the United Kingdom, Germany, Colombia, Peru, Mexico, and Canada. The subplots show the data for each country and includes a line plot of the model for the corresponding country. This visualization enables a thorough comparison of the model outputs for the top 10 countries, providing analysis and insights into the mpox outbreak. We note that for each country, the model appears to capture the trend of the outbreak well. We see the maxima for most of the model are again well-below the maxima for the data, but the data is “real” indicators of maxima values. This again shows a constraint of this approach.

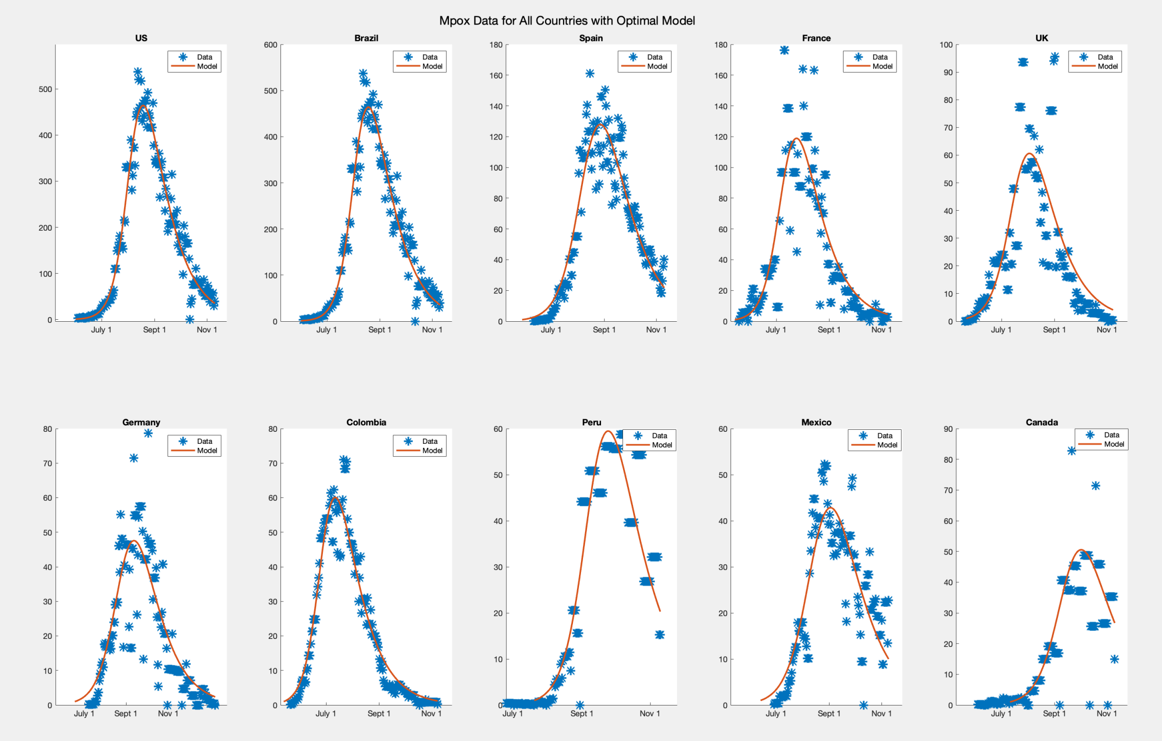

Figure 8: Model Solutions with Data for the Top 10 Countries[43]

Figure 9 shows a graph of the comparison of parameter values for β and S_0 for the top 10 countries. The graph employs dual y-axes, with the left y-axis representing the β parameter and the right y-axis representing S_0 , allowing the two sets of data to be plotted separately. The x-axis shows the names of the countries arranged in order of prevalence as given in Figure 1.

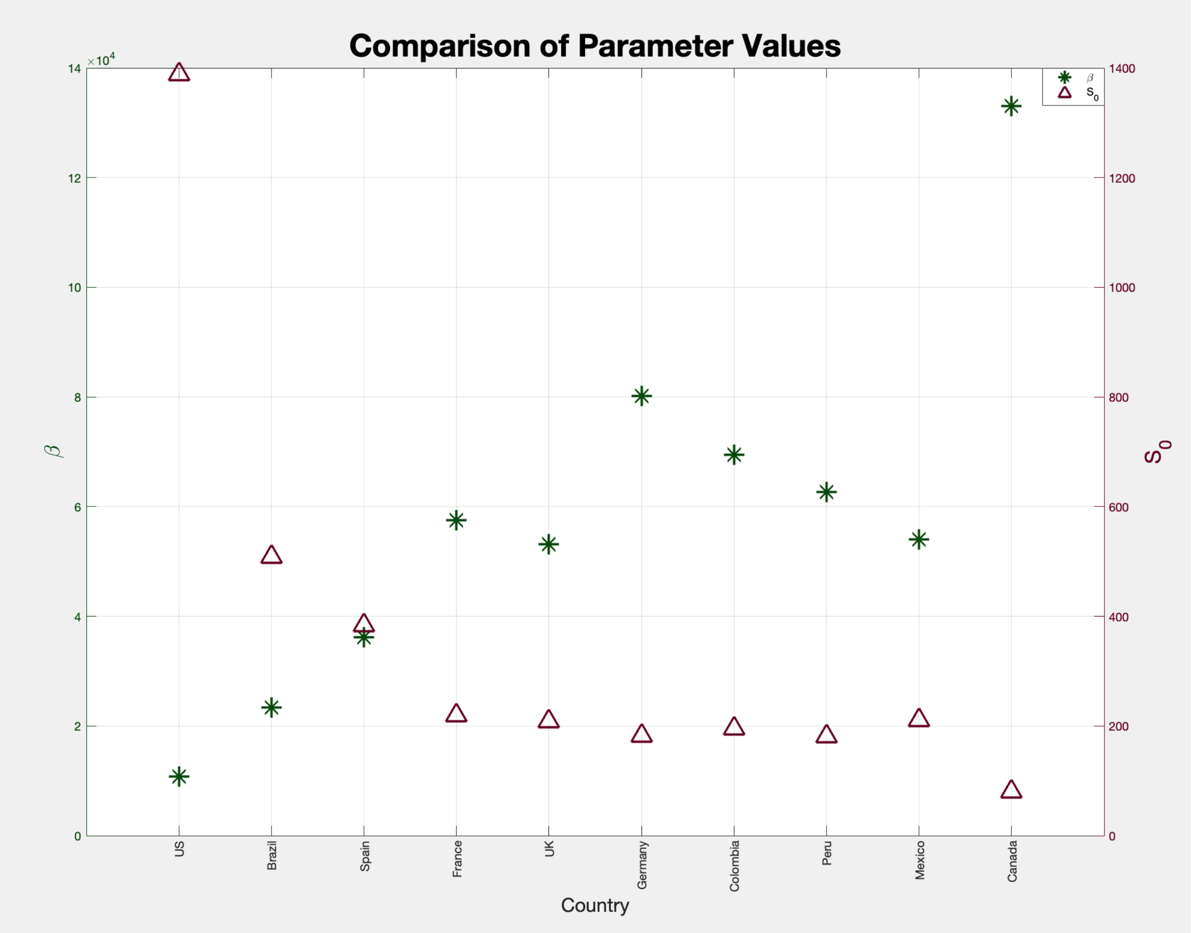

Figure 9: Comparison of Parameter Values For β and S_0[43]

The values of β are represented by green asterisks, while the values S_0 are represented by red triangles. The graph visualizes the relationship between these parameter values and their associated values across countries. It offers a clear comparison, indicating any patterns or differences. There seems to be a relationship between the size of S_0 and the size of β and this could be due to the population influence on β. However, it was difficult to tell which countries cluster together based on the parameter values. For example, by looking at the US, we see a very small value for β but a bigger value for S_0. On the other hand for Canada, we see a very small value for S_0 and a very big value for β. This makes it difficult to tell how similar or different the outbreaks are across the countries, especially given the widespread range of values. Incorporating the size of the initial susceptible population (S_0) in the calculation of R_0 (as shown in Table 2), we are now able to determine some similarities in outbreaks across countries.

We now point out some of the limitations of the OLS algorithm in our parameter estimation and then further investigate the potential uncertainty in the parameter estimates in the next section.

The OLS method is sensitive to optimization algorithm [28]. It relies on optimization methods such as Nelder Mead, which can fail spectacularly if there isn't a unique minimum. This sensitivity to the optimization process may cause instability or inaccuracy in the outcomes.

For example, Nelder Mead is widely used for convex problems and works well in low-dimensional cases with a few parameters. However, the reliability and effectiveness of OLS can be jeopardized when applied to non-convex problems with a larger parameter space[29].

Finally, OLS can sometimes provide curve fits that indicate maximum values that are far lower than the observed data [30]. This can be problematic, particularly when substantial data supports greater maximum values that are not outliers. Such disparities between model predictions and actual data may cast doubt on the dependability and defensibility on the OLS method.

3 UNCERTAINTY ANALYSIS

Uncertainty analysis is a systematic method for assessing and quantifying the uncertainties involved in measurements, computations, models, or any other process that has built-in variability or errors [32]. The results or conclusions drawn from these methods are intended to have a certain level of reliability or credibility. In this thesis, we used the bootstrapping technique in computing confidence intervals for our parameter estimates.

Bootstrapping is a statistical resampling technique used to estimate the sampling distribution of a statistic or to draw conclusions about population parameters. It entails repeatedly sampling by using the residuals from the original dataset with replacement to create several simulated datasets [24]. This resampling procedure generates simulated datasets that closely mimic the original data, allowing for the estimate of uncertainty metrics such as confidence intervals and standard errors. Bootstrapping is a non-parametric strategy that does not rely on specific distributional assumptions. It is especially effective when traditional statistical approaches are constrained by small sample sizes or breaches of distributional assumptions. It offers a versatile and reliable tool for estimating uncertainty and assessing the dependability of statistical outcomes [31].

In order to obtain parameter estimates, we considered a constant variance model and absolute error measurement datasets in this thesis. We followed the bootstrapping algorithm used in [24, p. 96-98] to compute bootstrap intervals for our estimated parameters. Our statistical model is of the form

$$y_j = f(t_j; \theta_0) + \epsilon_j, \quad j = 1, 2, ..., N$$

We first assume that $\hat{\theta}_{BOOT}$ is our bootstrapping estimate for the true parameter θ_0 and then follow the steps below.

1. Using Ordinary Least Square (OLS) approach, estimate the bootstrapping estimate $\hat{\theta}^0$ from the entire sample $\{y_j\}_{j=1}^N$.

2. We then define the standard residuals in terms of this estimate as

$$\overline{r}_j = \sqrt{\frac{N}{N - \kappa_0}} \left(y_j - f(t_j; \hat{\theta}^0) \right), j = 1, 2, 3, ..., N$$

   where κ_0 is the number of estimated parameters and N is the sample size.

3. By using random sampling, create a bootstrap sample of size N from the original data residuals $\{\overline{r}_1, \ldots, \overline{r}_N\}$ to form a bootstrapping sample $\{r_1^m, \ldots, r_N\}$

4. Now, we create bootstrap sample points

$$y_j^m = f(t_j; \hat{\theta}^0) + r_j^m, \quad j = 1, 2, ..., N$$

5. We now, obtain a new estimate $\hat{\theta}^{m+1}$ from the bootstrapping sample by using OLS.

6. Set $m = m + 1$ and repeat steps 3–5 until $m \geq M$ (e.g., here $M = 500$ as in our calculations below).

By following the bootstrapping algorithm as outlined in [24, p. 96-98], we used a constant variance data together with the Ordinary Least Squares (OLS) to compute the parameters for our models as given in Table 3 and Table 4.

Table 3 and Table 4 present median bootstrap parameter estimates for the effective transmission rate β, initial susceptible population S_0, the initial time of the first infected case t_0 of the mpox, and the calculated R_0 values along with their 90% bootstrap intervals respectively for the top ten mpox countries. These intervals were based upon 90% of the bootstrapping estimates being in the given interval and was calculated based on the 5% and 95% quantiles. These intervals provide ranges of plausible values for each parameter, accounting for the uncertainty in the estimates in Table 3.

Table 3: Median Parameter Estimates for the Top 10 Mpox Countries

Country	$\beta(\times 10^{-4})$ **persons/day**	S_0 **persons**	t_0 **(2022) days**	R_0
United States	1.029	1417	May 29	3.21
Brazil	2.233	521	May 25	2.56
Spain	2.999	422	May 3	2.79
France	4.155	262	May 2	2.40
United Kingdom	4.424	232	April 6	2.25
Germany	7.966	181	May 11	3.18
Colombia	6.662	198	July 17	2.90
Peru	5.744	190	June 8	2.40
Mexico	4.471	235	July 28	2.32
Canada	11.865	87	May 8	2.27

In Table 3, the R_0 values reveal information about the potential transmission dynamics of the mpox disease in each country. According to the provided estimations, the R_0 values range from 2.25 to 3.21 among countries. These numbers represent the average number of secondary infections induced by a single infected person in a susceptible population. These R_0 estimates can be used to better understand and evaluate the infectiousness and potential spread of mpox in different nations, assisting in the development of effective control and prevention efforts. The estimated values of β per day range from 1.029×10^{-4} to 11.865×10^{-4}. A higher β score underlines the necessity for more vigorous and urgent measures to contain the transmission of the virus. The initial start time of the epidemic (t_0) is given in days, and the estimated values range from April 6 to July 28, 2022.

Table 4: Median Parameter Estimates for the Top 10 Mpox Countries

Country	**90% CI for** β ($\times 10^{-4}$)	**90% CI for** S_0	**90% CI for** t_0 **(2022)**	**90% CI for** R_0
United States	$[0.973, 1.084]$	$[1390, 1450]$	[May 25, June 1]	$[3.10, 3.33]$
Brazil	$[2.074, 2.432]$	$[501, 539]$	[May 19, May 31]	$[2.46, 2.69]$
Spain	$[2.433, 3.833]$	$[383, 464]$	[April 18, May 17]	$[2.46, 3.23]$
France	$[2.790, 6.430]$	$[217, 327]$	[March 30, May 26]	$[2.00, 3.06]$
United Kingdom	$[3.673, 5.310]$	$[211, 257]$	[March 23, April 18]	$[2.07, 2.48]$
Germany	$[7.443, 8.618]$	$[176, 186]$	[May 8, May 15]	$[3.03, 3.35]$
Colombia	$[5.719, 7.780]$	$[187, 213]$	[July 10, July 25]	$[2.65, 3.25]$
Peru	$[4.970, 6.639]$	$[177, 205]$	[May 30, June 16]	$[2.24, 2.59]$
Mexico	$[1.789, 7.892]$	$[176, 429]$	[May 1, July 28]	$[1.71, 3.12]$
Canada	$[9.821, 14.805]$	$[78, 97]$	[April 28, May 19]	$[2.08, 2.55]$

From Table 4, the 90% bootstrap intervals indicate the level of uncertainty surrounding these estimates. These estimates provide useful information about the characteristics of mpox in each country. The variation in β, S_0, t_0, and R_0 values between countries indicates differences in transmission rates, population susceptibility, and disease dynamics. These parameter estimations and their accompanying intervals are critical for understanding and modeling the spread of mpox, allowing policymakers and public health professionals to implement effective disease management and mitigation methods. The estimated values and their bootstrap intervals offer insights into the parameters of interest and their likely values for each country. Figure 10 gives a plot of the variation in β values for the top 10 countries.

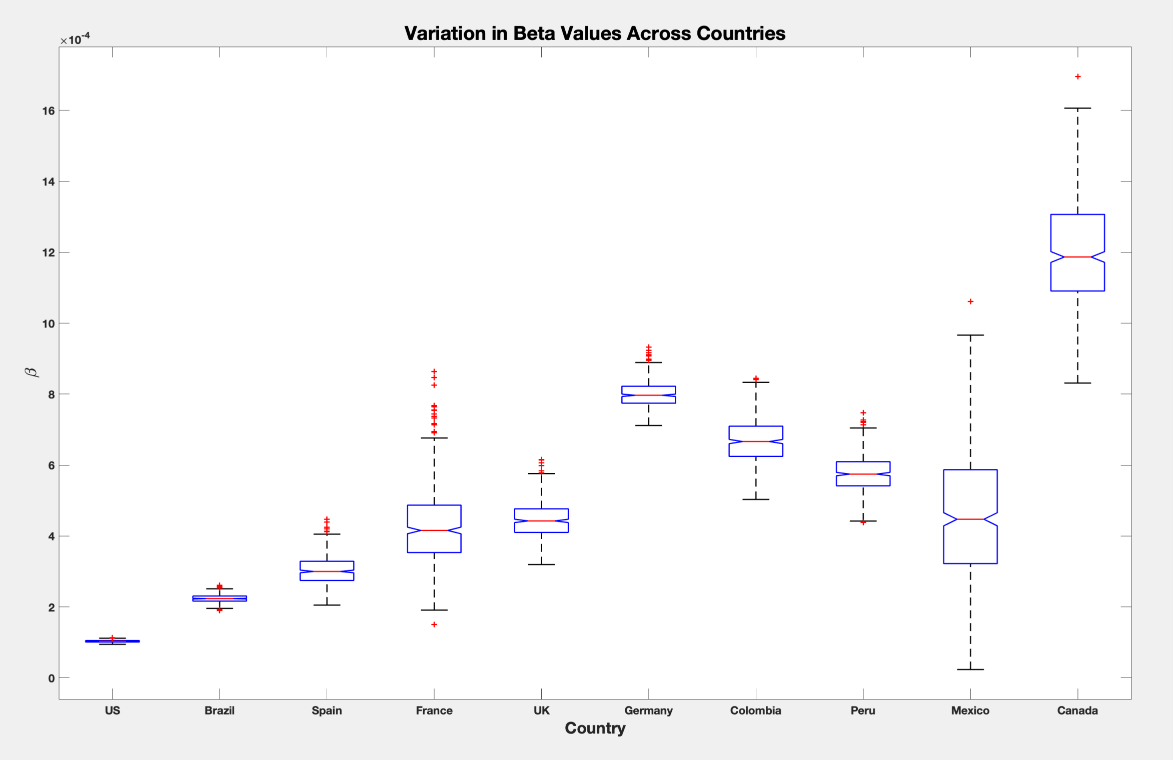

Figure 10: Boxplot Comparison For the Variation of β Across Countries[43]

From Figure 10, we visualize the estimated values for β using a boxplot. We again observe that Canada stands out with the highest value among the countries listed. This suggests that Canada has a relatively higher rate of transmission of the mpox infection compared to the other countries. Conversely, the United States exhibits the lowest estimated value for β, indicating a relatively slower rate of transmission. However, the reverse is true for the initial susceptible population as given in Figure

11 which shows a boxplot comparison of the values of S_0 across the top 10 mpox countries.

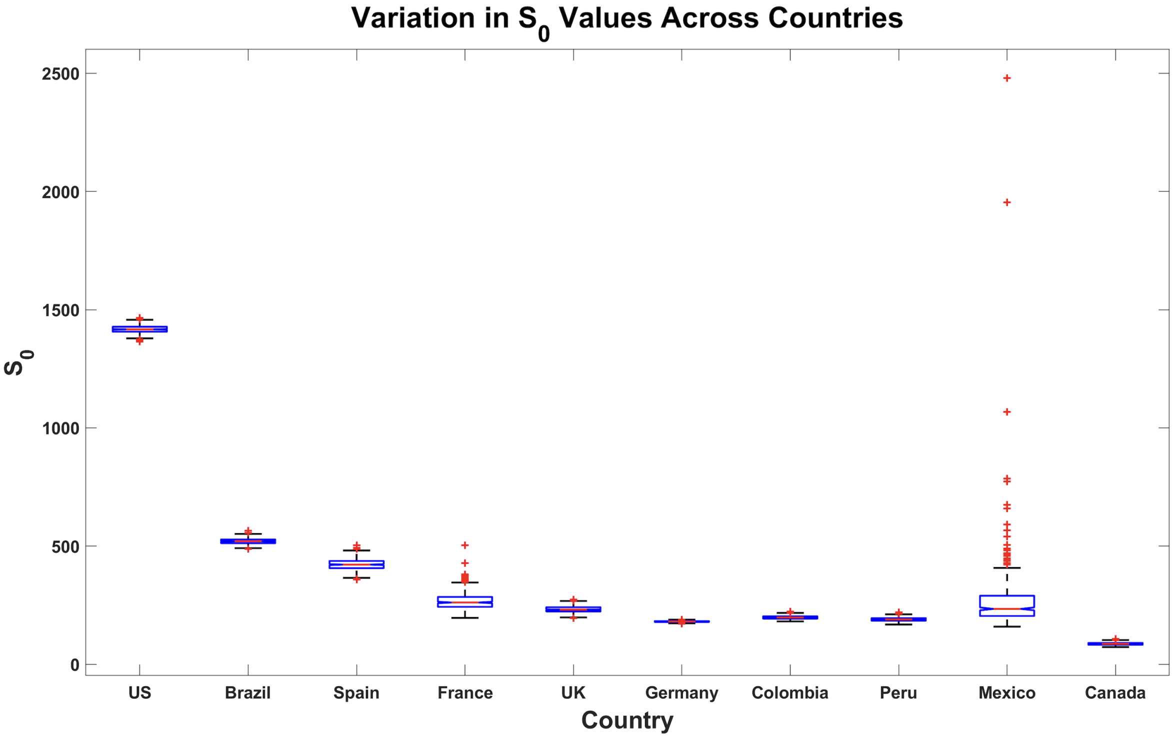

Figure 11: Boxplot Comparison For the Variation of S_0 Across Countries[43]

We observe from Figure 11 the estimated values for S_0 and the variations using a boxplot. We observe that Mexico has the highest variation in S_0 with some outliers among the countries listed. The US has the highest value for S_0 as compared to Canada which has the smallest value for S_0. Next, we show in Figure 12 the variation in R_0 values across the top 10 mpox countries. We see that US has the least variation in R_0 as compared to Mexico which has the highest variation in R_0 suggesting that there is greater uncertainty and heterogeneity in the transmissibility of the mpox virus in Mexico than in the US.

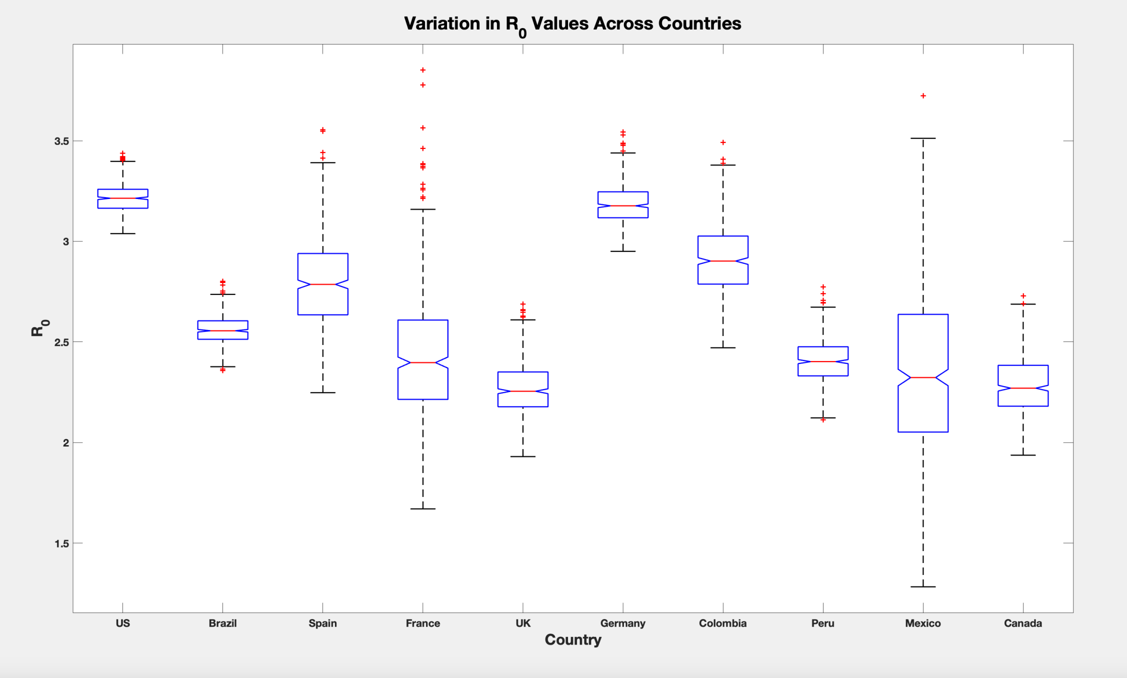

Figure 12: Boxplot Comparison For the Variation of R_0 Across Countries[43]

Directly comparing countries based on parameter values in the context of a comparative analysis of the mpox epidemic among prevalent countries can be difficult due to the inverse-like relationship between certain parameters, such as β and S_0. Even using the calculated values of R_0 which takes into account β and S_0, we still can not effectively determine which countries have trends that are similar across the course of the outbreak. A more effective technique to addressing this issue is to convert the parameter values into scaled trend data by dividing the model which already has the values of β in it by the total population. By doing so, we are getting the data for all the 10 countries on the same scale. We can then find patterns and commonalities among the prevalent countries by applying clustering techniques to this scaled data, allowing for a more accurate and insightful comparison of the outbreak among the countries.

4 TIME SERIES CLUSTERING

In this thesis, we decided to group all the top 10 mpox countries based on certain defined characteristics or differences exhibited by the countries. We used a technique called time series clustering to achieve this aim. Clustering helps identify which of the top 10 countries behave similarly or differently in terms of the spread and dynamics of the mpox.

Time series clustering is a technique that groups similar time series data into clusters based on similarities and trends [33]. It entails studying the data's temporal patterns and attributes in order to discover groups or clusters that display consistent behavior across time [33, 34]. Each time series is represented in time series clustering as a series of observations acquired at regular intervals. There are several techniques to cluster time series data, but for the purpose of this thesis, we look at the hierarchical clustering approach using Dynamic Time Warping as the distance metric on the top 10 mpox countries' model-simulated data.

Our original WHO data had some missing values and typically when dealing with time series data, a smoothing technique is applied to the data to distinguish the noise from the overall trend. For our purposes, the mathematical model already gives an average trend for the outbreak (recall Figure 8); therefore, we used the model to generate a smoothed estimate for the daily outbreak. Due to the "inverse-like" relationship between S_0 and β for some countries, we wanted to get the overall dynamics of the mpox, so we scaled the the data by dividing each country's model-simulated data by $N = S_0 + I_0 + R_0 = S_0 + 1$ where values of S_0 are given Table 2. We furthermore assumed 0 infected population until the respective date give by

t_0 for the country. Figure 12 shows the comparison of mpox trends across countries using scaled data. The plot compares the trends of mpox across different countries. Each country's data is plotted on the graph using the respective date and scaled data. The legend on the graph indicates the country names for each line. This visual comparison of the trends among the selected countries are based on their scaled data. We used a technique called dynamic time warping to access the similarity between these scaled trends and then hierarchical clustering to cluster the countries based on their similarities.

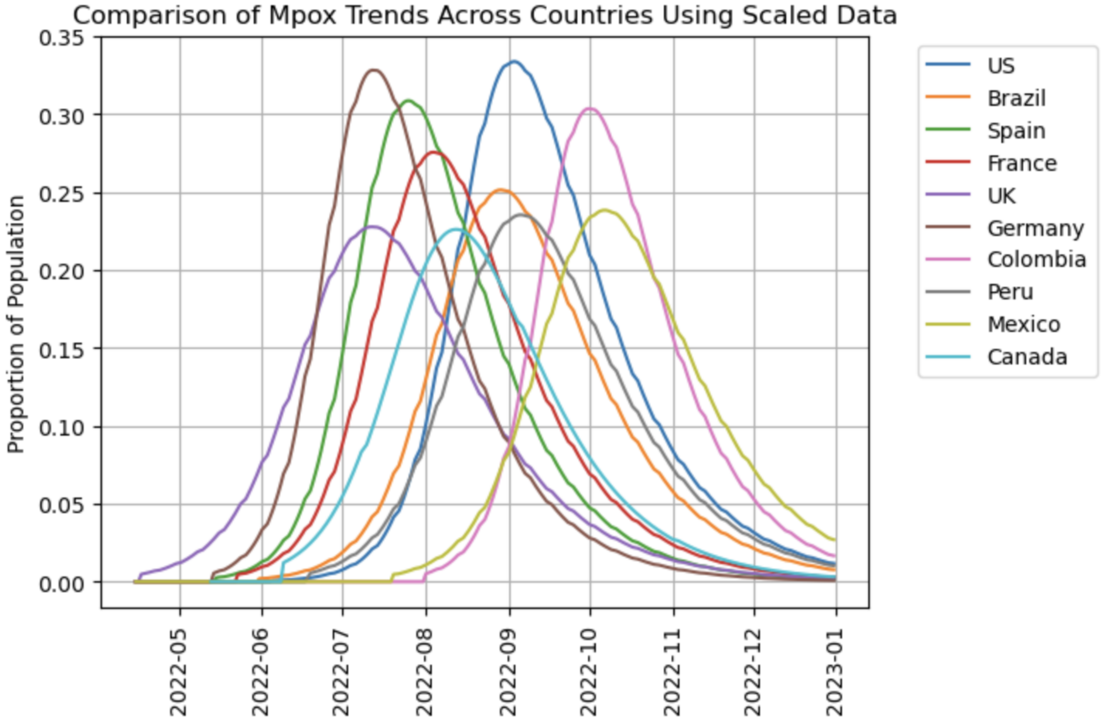

Figure 13: Comparison of Mpox Trends Across Countries[43]

It is critical to remember that the population sizes are now expressed relative to each country's overall population number (N). As a result, the y-scale in the graph indicates a percentage or proportion of the overall population, with infected individuals starting at $\frac{1}{N}$ for each country. This adjustment enables for consistent comparisons and interpretations across different populations while accounting for the varying sizes of the populations.

4.1 Dynamic Time Warping

Dynamic Time Warping (DTW) is a technique used to compare the similarity of two sequences that may differ in length or pace [35]. It is widely used in domains including speech recognition, gesture recognition, and time series analysis [36]. The core idea underlying Dynamic Time Warping is to warp the time axis of two sequences to obtain an ideal alignment. This enables for the comparison of corresponding sequence elements, even if their lengths or speeds differ. The best alignment is found by reducing the cumulative distance between the sequences' corresponding elements.

Consider the finite sequences $X = [X_1, X_2, X_3, ..., X_n]$ and $Y = [Y_1, Y_2, Y_3, ..., Y_m]$, where n and m are the lengths of the sequences, respectively. Dynamic Time Warping generates an $n \times m$ matrix, often known as the DTW matrix or accumulated cost matrix, abbreviated as D. Each coefficient $D_{i,j}$ of D indicates the distance between the numbers X_i and Y_j in the sequences X and Y, respectively. Dynamic programming is used to compute the DTW matrix iteratively [37].

The first step is to initialize the matrix with acceptable boundary conditions. This is often accomplished by specifying $D_{1,1}$ as the distance between the first elements of X and Y and filling the first row and first column with the calculated distances from the beginning point. The cumulative distance is then calculated for each element $D_{i,j}$

in the matrix as follows:

$$D_{i,j} = AC(X_i, Y_j) = (X_i - Y_j)^2 + \min\{D_{i-1,j}, D_{i,j-1}, D_{i-1,j-1}\}$$

where the term $AC(X_i, Y_j)$ is the accumulated cost at X_i, Y_j, the term $(X_i - Y_j)^2$ is taken as the Euclidean distance between X_i and Y_j, and the term $\min\{D_{i-1,j}, D_{i,j-1}, D_{i-1,j-1}\}$ is the cost of where it came from. Although the term $AC(X_i, Y_j)$ is not exactly the Euclidean distance, it incorporates the Euclidean distance as part of the accumulated cost calculation in the dynamic time warping algorithm. Recall the Euclidean distance is given by:

$$||X - Y||_2 = \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2 + \cdots + (X_n - Y_n)^2}$$

Further more, the warped path allows for horizontal moves $((i, j) \rightarrow (i, j+1))$, vertical moves $((i, j) \rightarrow (i + 1, j))$ and diagonal moves $((i, j) \rightarrow (i + 1, j + 1))$; therefore, internal moves are accounted for by examining the previous locations $D_{i-1,j}$, $D_{i,j-1}$ and $D_{i-1,j-1}$. After computing the DTW matrix, the best alignment path can be found by backtracking from the bottom-left cell to the top-right cell, taking the path with the shortest cumulative distance [38]. This alignment path represents the best mapping between X and Y elements.

Consider the following example to demonstrate the DTW algorithm. Suppose we have two time series sequences, X and Y with lengths $n = 5$ and $m = 7$ respectively: $X = [3, 1, 2, 2, 1]$, and $Y = [2, 0, 0, 3, 3, 1, 0]$ [39]. The sequences X and Y start at position 1, while X ends at position $n = 5$ and Y ends at position $m = 7$. X starts at $X_1 = 3$ whiles Y starts at $Y_1 = 2$ and X ends at $X_n = 1$, whiles Y ends at $Y_m = 0$. This therefore means that X_1 always pairs with Y_1 and X_n always pairs with Y_n.

We then form an accumulated cost matrix in which the ‘next’ move is the original ‘distance’ between components plus the minimum between the choices of moving from a specified adjacent point. It is easiest to visualize the warped path using an inverted cost matrix. In Table 5, we begin by aligning the Y sequence in an opposite order so that the first alignment is in the bottom corner position.

Table 5: Calculating Dynamic Time Warping Distances

0					
1					
3					
3					
0					
0					
2					
	3	1	2	2	1

The first component of the accumulated cost matrix is the component in the bottom left which is only the Euclidean distance component since it is the first move and is calculated as $AC(X_1 = 3, Y_1 = 2) = (X_1 - Y_1)^2 = (3 - 2)^2 = 1$ as shown in Table 6.

Table 6: Calculating Dynamic Time Warping Distances

0					
1					
3					
3					
0					
0					
2	1				
	3	1	2	2	1

Now, consider the other entries on the bottom boundary starting with $X_2 = 1$ and $Y_1 = 2$. For that entry, there is only a left component filled in already so we compute the (X_2, Y_1) entry and then add the already filled in component on the left of it. Thus, we get $AC(X_2 = 1, Y_1 = 2) = (X_2 - Y_1)^2 + AC(X_1, Y_1) = (1 - 2)^2 = 1 + 1 = 2$ as shown in Table 7.

Table 7: Calculating Dynamic Time Warping Distances

0					
1					
3					
3					
0					
0					
2	1	2			
	3	1	2	2	1

Similarly for $AC(X_i, Y_1 = 2), i = 3, 4, 5$, we have

$$AC(X_3 = 2, Y_1 = 2) = (2 - 2)^2 + AC(X_2, Y_1) = 0 + 2 = 2$$

$$AC(X_4 = 2, Y_1 = 2) = (2 - 2)^2 + AC(X_3, Y_1) = 0 + 2 = 2$$

$$AC(X_5 = 1, Y_1 = 2) = (1 - 2)^2 + AC(X_4, Y_1) = 1 + 2 = 3$$

which gives the completed buttom row shown in Table 8.

Table 8: Calculating Dynamic Time Warping Distances

0					
1					
3					
3					
0					
0					
2	1	2	2	2	3
	3	1	2	2	1

We repeat a similar procedure for the first column starting with the next to last entry in our inverted matrix i.e. $X_1 = 3$ and $Y_2 = 0$. For this entry, there is only a bottom component so,

$$AC(X_1 = 3, Y_2 = 0) = (X_1 - Y_2)^2 + AC(X_1, Y_1) = (3 - 0)^2 + 1 = 10$$

Similarly for $AC(X_3, Y_j), j = 3, 4, 5, 6, 7$, we have

$$AC(X_1 = 3, Y_3 = 0) = (3 - 0)^2 + AC(X_1, Y_2) = 9 + 10 = 19$$

$$AC(X_1 = 3, Y_4 = 3) = (3 - 3)^2 + AC(X_1, Y_3) = 0 + 19 = 19$$

$$AC(X_1 = 3, Y_5 = 3) = (3 - 3)^2 + AC(X_1, Y_4) = 0 + 19 = 19$$

$$AC(X_1 = 3, Y_6 = 1) = (3 - 1)^2 + AC(X_1, Y_5) = 4 + 19 = 23$$

$$AC(X_1 = 3, Y_7 = 0) = (3 - 0)^2 + AC(X_1, Y_6) = 9 + 23 = 32$$

The computed cells are shown in Table 9.

Table 9: Calculating Dynamic Time Warping Distances

0	32				
1	23				
3	19				
3	19				
0	19				
0	10				
2	1	2	2	2	3
	3	1	2	2	1

Once the boundaries are filled, the interior points can be filled. For the $AC(X_2 = 1, Y_2 = 0)$ entry, all previous entries are available; therefore we have

$$AC(X_2 = 1, Y_2 = 0) = (X_2 - Y_2)^2 + \min\{AC(X_1, Y_1), AC(X_1, Y_2), AC(X_2, Y_1)\}$$

$$= (1 - 0)^2 + \min\{10, 1, 2\}$$

$$= 1 + 1 = 2.$$

Similarly for the $X_3 = 2, Y_2 = 0$ entry, we have

$$\begin{aligned} AC(X_3 = 2, Y_2 = 0) &= (X_3 - Y_2)^2 + \min\{AC(X_2, Y_1), AC(X_2, Y_2), AC(X_3, Y_1)\} \\ &= (2 - 0)^2 + \min\{2, 2, 2\} \\ &= 4 + 2 = 6. \end{aligned}$$

Now, for the $X_2 = 1, Y_3 = 0$ entry, we have

$$\begin{aligned} AC(X_2 = 1, Y_3 = 0) &= (X_2 - Y_3)^2 + \min\{AC(X_1, Y_2), AC(X_1, Y_3), AC(X_2, Y_2)\} \\ &= (1 - 0)^2 + \min\{10, 19, 2\} \\ &= 1 + 2 = 3. \end{aligned}$$

We follow a similar procedure and fill the rest of the cells as shown in Table 10. We also plotted a heat map for the distance matrix in Python (see Figure 12) from the accumulated cost matrix in Table 10.

To obtain the warped path, we started in the bottom left and moved up, right or diagonal according to the lowest cost as illustrated by the arrows in Figure 13. We thus have the following alignment $X = [3, 1, 1, 2, 2, 1, 1]$ and $Y = [2, 0, 0, 3, 3, 1, 0]$. The best alignment of the two sequences is plotted in Figure 14.

Table 10: Calculating Dynamic Time Warping Distances

0	32	12	10	10	6
1	23	11	6	6	5
3	19	11	5	5	9
3	19	7	4	5	8
0	19	3	6	10	4
0	10	2	6	6	3
2	1	2	2	2	3
	3	1	2	2	1

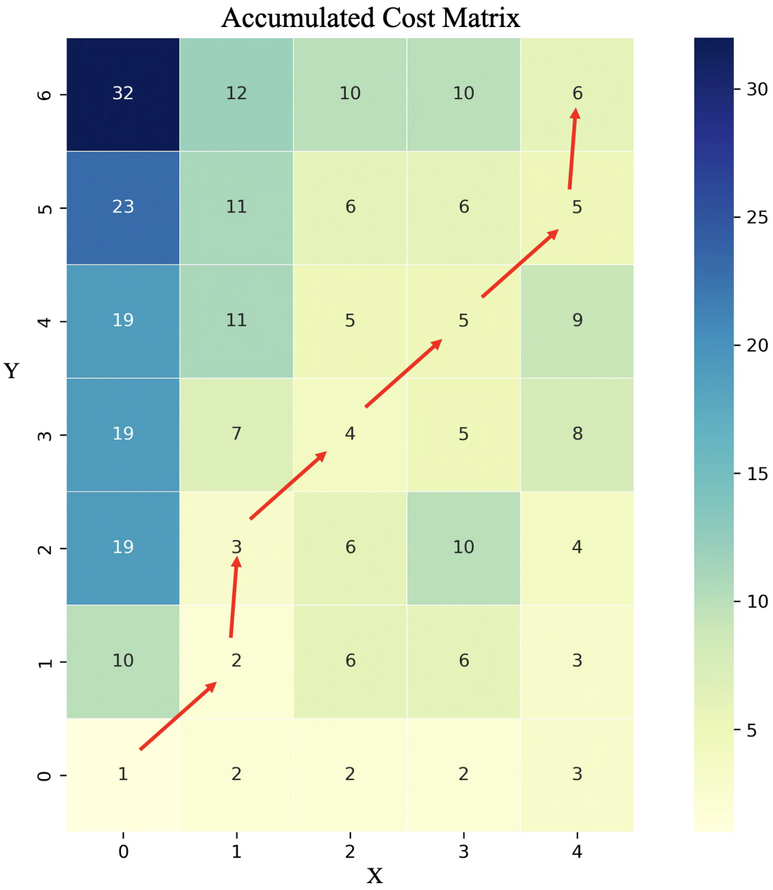

Figure 14: Accumulated Cost Matrix For X and Y[39]

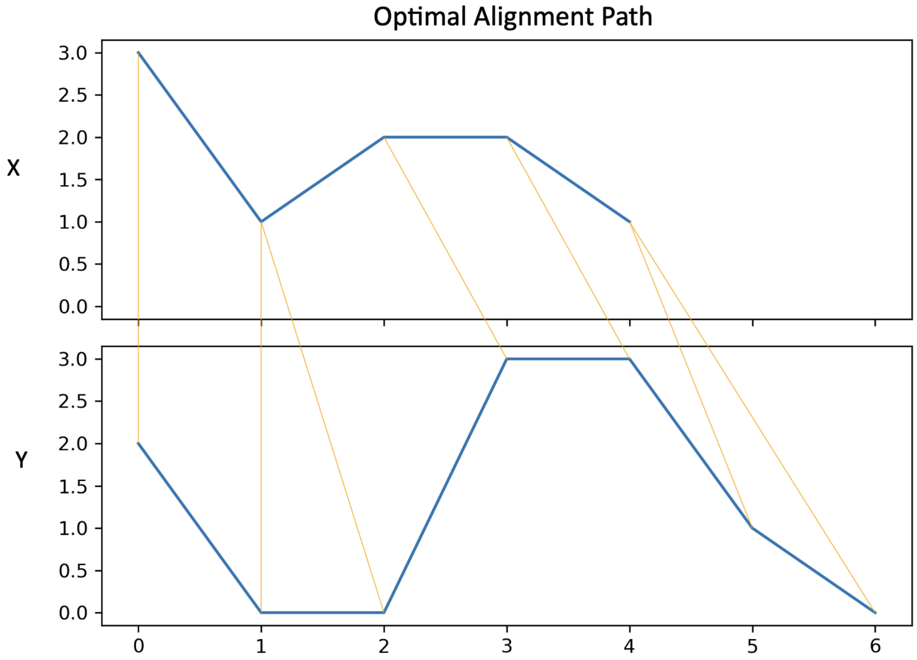

Figure 15: Optimal Path Alignment For X and Y [39]

In Figure 14, we observe that the point 3 in the X sequence aligns with the point 2 in the Y sequence, 1 in the X sequence aligns with 0 twice in the Y sequence, 2 in the X sequence aligns with 3 in the Y sequence and so on. After using Dynamic Time Warping to find the optimal alignment between X and Y, we can now calculate the Euclidean distance between this best optimal alignment as:

$$DTW(X,Y) = \sqrt{(3-2)^2 + (1-0)^2 + (1-0)^2 + (1-3)^2 + (2-3)^2 + (1-1)^2 + (1-0)^2}$$

$$= 2.45$$

4.2 Dynamic Time Warping on Mpox Model-Simulated Data

By following the algorithm of the DTW as illustrated, we constructed a 10 × 10 distance matrix for the top 10 mpox model-simulated data.

For hierarchical clustering implementation, a distance matrix between series is needed. The distance matrix is a square matrix that represents the pairwise distances between our dataset's series. Our Python method computes the DTW distance between each pair of series by iterating through the series indices and performing the DTW algorithm, omitting self-comparisons. The distances obtained are then recorded in a distance matrix. This distance matrix is an important input for hierarchical clustering, allowing the series' similarity patterns and hierarchical relationships to be identified. Table 11 shows the output of the distance matrix.

Table 11: Dynamic Time Warping Distance Measure for the Top Ten Mpox Countries

Country	US	Brazil	Spain	France	UK	Germany	Colombia	Peru	Mexico	Canada
US	0.00	0.34	0.09	0.22	0.47	0.07	0.10	0.42	0.41	0.47
Brazil	0.34	0.00	0.21	0.08	0.08	0.29	0.18	0.05	0.08	0.09
Spain	0.09	0.21	0.00	0.11	0.32	0.06	0.09	0.29	0.31	0.33
France	0.22	0.08	0.11	0.00	0.18	0.18	0.11	0.15	0.19	0.19
UK	0.47	0.08	0.32	0.18	0.00	0.41	0.30	0.05	0.16	0.02
Germany	0.07	0.29	0.06	0.18	0.41	0.00	0.12	0.37	0.40	0.42
Colombia	0.10	0.18	0.09	0.11	0.30	0.12	0.00	0.26	0.24	0.31
Peru	0.42	0.05	0.29	0.15	0.05	0.37	0.26	0.00	0.06	0.04
Mexico	0.41	0.08	0.31	0.19	0.16	0.40	0.24	0.06	0.00	0.14
Canada	0.47	0.09	0.33	0.19	0.02	0.42	0.31	0.04	0.14	0.00

The table displays the pairwise distances between the entire outbreak in the various countries. Each cell in the table indicates the measurement of distance or dissimilarity between two countries. DTW considers the local distances between corresponding elements of the time series and determines the best alignment between them [35]. The distances between countries in this table demonstrate their resemblance or dissimilarity based on specific criteria. A lower distance number indicates greater similarity, whereas a higher distance shows greater dissimilarity [40]. Looking at the first row, for example, we can see that the distance between the United States and Brazil is 0.34, indicating a moderate amount of dissimilarity. The gap between the UK and Canada, on the other hand, is only 0.02, indicating a higher degree of similarity. Figure 15 shows the warp path between UK and Canada.

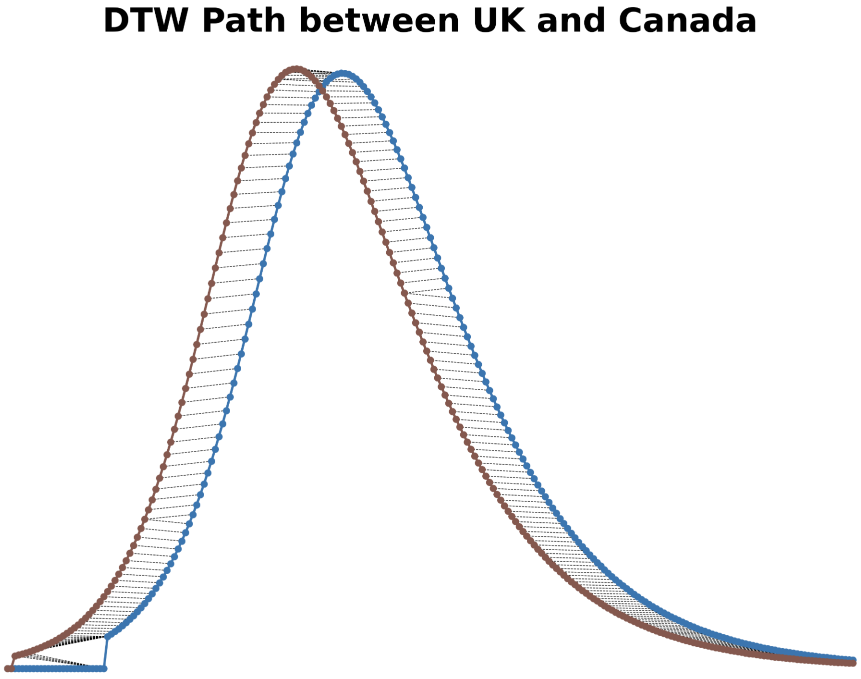

Figure 16: Warp Path Between UK and Canada

Figure 15 gives the optimal alignment path between UK and Canada using Dynamic Time Warping to visualize the correspond points and connecting lines.

4.3 Hierarchical Clustering

Hierarchical clustering is a cluster analysis method that aims to create a hierarchy of groups [41]. It is widely utilized in statistics and data mining. There are two types of hierarchical clustering strategies: agglomerative and divisive. Each observation begins in its own cluster in agglomerative clustering, and pairs of clusters are combined as one advances up the hierarchy. All observations begin in one cluster in divisive clustering, and splits are performed iteratively as one proceeds down the hierarchy. Hierarchical clustering produces a tree-based representation of the objects, known as a dendrogram. The complete linkage method is a distance measure used in hierarchical clustering to determine the dissimilarity across clusters. It computes the distance between two clusters as the maximum distance between any two data points from each cluster [42]. We will give a simple example of how the complete linkage metric works in forming clusters by the following steps.

1. Find the cluster pair with the shortest maximum distance between any two of their members. In other words, find the cluster pair with the closest link.

2. Combine the two clusters to form a new cluster.

3. Calculate the maximum distance between any pair of nations in the merged cluster and the other clusters to update the distance matrix.

4. Repeat the process until there is nothing to merge again in the distance matrix.

5. Make a dendrogram. Represent the clusters and their mergers as a dendrogram, displaying the clusters' hierarchical relationship.

By using the distance matrix in Table 11, we will show how the complete linkage method works to form clusters. According to the distances in the table, the shortest distance is 0.02, which occurs between the UK and Canada. As a result, the UK and Canada would form the first cluster. Then we update the distance matrix as follows

Table 12: Dynamic Time Warping Distance Measure for the Top Ten Mpox Countries

Country	UK/Canada	US	Brazil	Spain	France	Germany	Colombia	Peru	Mexico
UK/Canada	0.00								
US	0.47	0.00							
Brazil	0.09	0.34	0.00						
Spain	0.33	0.09	0.21	0.00					
France	0.19	0.22	0.08	0.11	0.00				
Germany	0.42	0.07	0.29	0.06	0.18	0.00			
Colombia	0.31	0.10	0.18	0.09	0.11	0.12	0.00		
Peru	0.05	0.42	0.05	0.29	0.15	0.12	0.26	0.00	
Mexico	0.16	0.41	0.08	0.31	0.19	0.40	0.24	0.06	0.00

From Table 12, we left out the elements of the upper diagonal since the distance matrix is symmetric. Using the complete linkage method, we compute the distances between the pair UK/Canada and the rest of the countries. To do this, we find from Table 11 the distance between US and UK which was 0.47. We then find the distance between US and Canada which was 0.47. We choose the maximum of the two which in this case will still be 0.47. Similarly, to find the distance between Brazil and the UK/Canada pair, we first find the distance between Brazil and UK which was 0.08. We then find the distance between Brazil and Canada which was 0.09 according to Table 11. We finally find the maximum of the two distances which turns out to be 0.09. We follow the same approach to find the distances between UK/Canada and the rest of the countries. Then we maintain the distances in other cells as they are in Table 11. With the updated distance matrix in Table 12, we repeat the process by finding the shortest distance in the matrix. This happens to be 0.05 and this occurs between the pair UK/Canada and Peru. This therefore means that UK/Canada and Peru will form another cluster as shown in Table 13.

Table 13: Dynamic Time Warping Distance Measure for the Top Ten Mpox Countries

Country	UK/Canada/Peru	US	Brazil	Spain	France	Germany	Colombia	Mexico
UK/Canada/Peru	0.00							
US	0.47	0.00						
Brazil	0.09	0.34	0.00					
Spain	0.33	0.09	0.21	0.00				
France	0.19	0.22	0.08	0.11	0.00			
Germany	0.42	0.07	0.29	0.06	0.18	0.00		
Colombia	0.31	0.10	0.18	0.09	0.11	0.12	0.00	
Mexico	0.16	0.41	0.08	0.31	0.19	0.40	0.24	0.00

This process continues repetitively until we no longer have any country to cluster. We then obtain the linkages given by the dendrogram in Figure 16. UK, Canada and Peru being clustered together confirms the results we saw in our calculations. The dendrogram depicts the clusters' hierarchical relationship. Each leaf node represents a country, while the branches reflect the process of merging. The height of each branch shows the distance between clusters. The dendrogram depicts the clusters' hierarchical relationship.

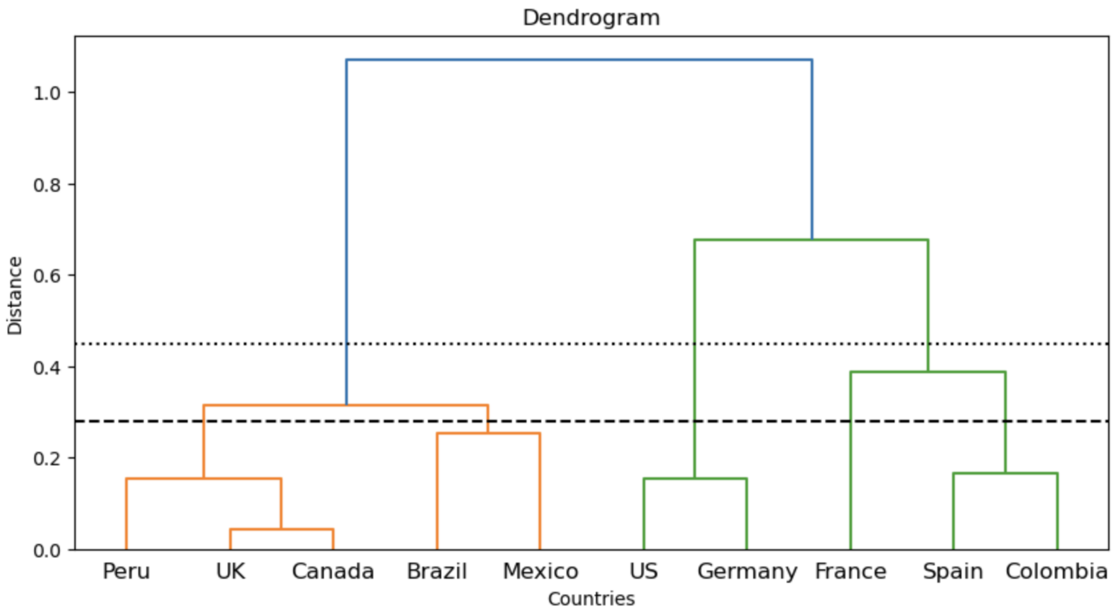

Figure 17: Dendrogram of Countries Based on Time Series Data

The graph takes an in-depth look at the entire linkage clustering approach for our time series data. Based on their time series patterns, the resulting dendrogram can provide insights into the similarities and hierarchical structure of the countries. From the dendrogram above, we count the number of vertical branches (cluster merges) intersecting this line at a height 0.45 and note the number of clusters generated at this height. We observe three clusters at this threshold. Figure 17 shows the clusters formed at the 0.45 threshold.

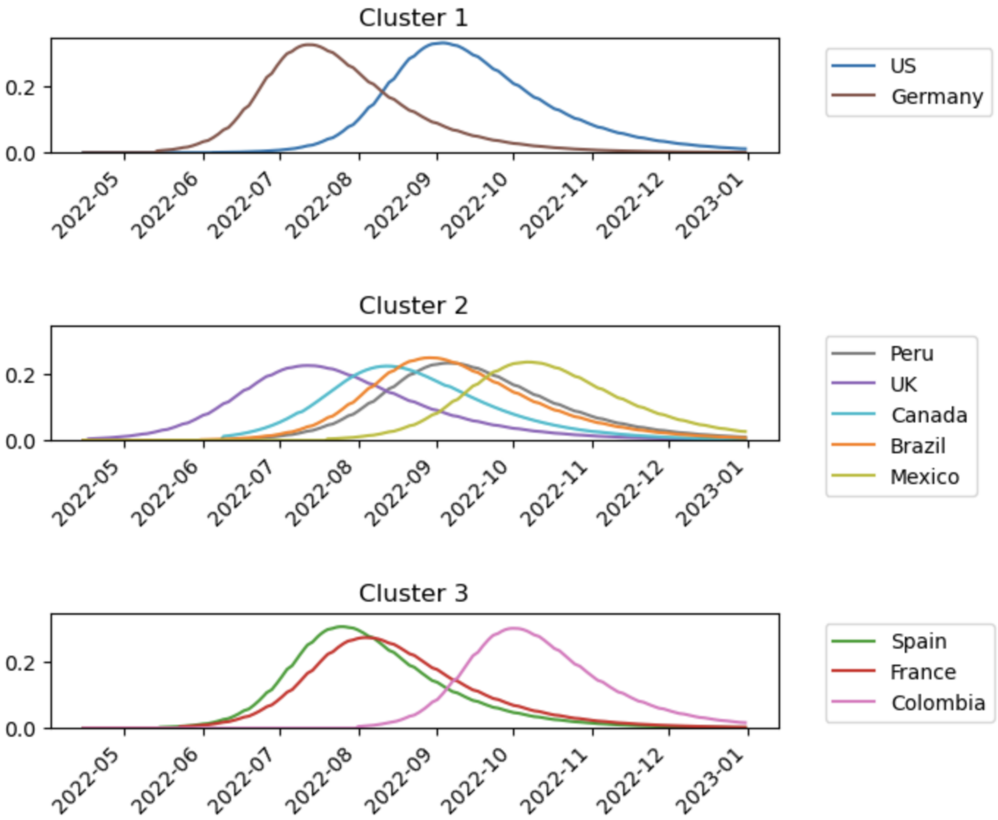

Figure 18: Clustering Results on Scaled Mpox Data

The examination of time series data from the top 10 countries (including the United States, Germany, Peru, the United Kingdom, Canada, Brazil, Mexico, Spain, France, and Colombia) found three unique clusters, each of which is important in understanding the dynamics of the outbreak across different nations. Cluster 1 consists of the United States and Germany, indicating that these countries share underlying characteristics that influence mpox cases. Cluster 2 contains Peru, the United Kingdom, Canada, Brazil, and Mexico, with similar tendencies but significant differences across time. This cluster aids in the identification of shared dynamics and potential outbreak contributors. Cluster 3 includes Spain, France, and Colombia, demonstrating mpox pattern similarities among these countries. Analyzing the similarities between these clusters provides useful insights for understanding the dynamics of the mpox outbreak and informing country-specific decision-making processes. Overall, the clustering findings give light on the links and patterns that exist between nations, assisting in the worldwide understanding of the mpox outbreak.

Also, at the threshold of 0.28, we count the number of vertical branches that intersect this line and we count the number of clusters. We identified five clusters at this threshold and we show these clusters in Figure 18. This thesis' clustering research demonstrates the significance of categorizing countries into various clusters based on their mpox epidemic dynamics. Cluster 1 includes the United States and Germany, indicating that these countries share underlying characteristics that influence mpox cases. Cluster 2 includes Brazil and Mexico, which represent countries with similar tendencies but with significant differences across time. Cluster 3 comprises of the United Kingdom, Peru, and Canada, showing that their mpox epidemics share common patterns and behaviors. Spain and Colombia are included in Cluster 4, indicating common traits and similarities in their mpox data. Finally, Cluster 5 represents France, which has a distinct pattern in comparison to the other countries. The clusters' visual representations provide an intuitive grasp of the temporal patterns in each group, aiding the discovery of potential common elements impacting the virus's transmission. Policymakers and researchers can gather useful insights into the mpox outbreak by evaluating the cluster relevance across different countries, identifying parallels, differences, and potential influencing factors, and making informed decisions to limit the impact of the virus.

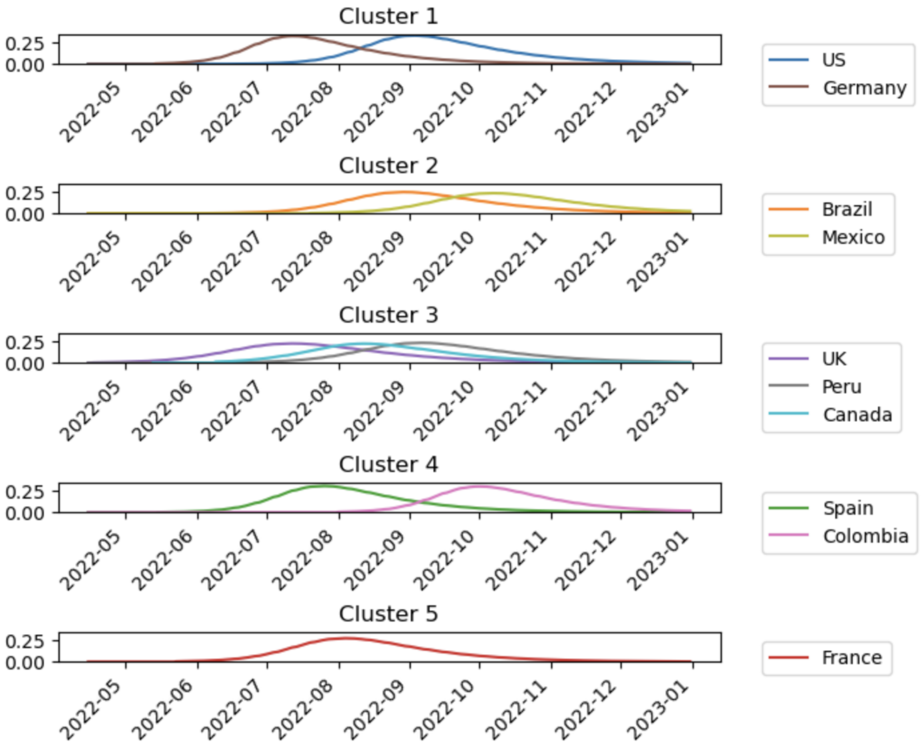

Figure 19: Clustering Results on Scaled Mpox Data

The clustering technique effectively identified five unique groups, providing important insights into the dynamics and patterns in the time series data of the top ten mpox countries. This clustering approach gives different results from simply looking at the total number of the infected cases (from Figure 1) as the total number of infections does not necessarily indicate similar trends. Clustering groups countries based on the similarity of their epidemic time series. When we compare the clustering findings to the R_0 values, we can see both similarities and differences in the groupings. Similarities emerge when countries with similar R_0 values are clustered together, indicating that their epidemic patterns are comparable. For example, the United States and Germany, which have greater R_0 values, are placed together in a cluster. Similarly, Brazil and Mexico, which have lower R_0 values, are grouped together. However, there are some distinctions. For example, France and the United Kingdom have similar R_0 values but are put in different clusters, indicating variances in their epidemic patterns despite comparable transmission rates. As a result, while R_0 gives essential information on disease transmissibility, clustering based on epidemic time series captures additional nuances and similarities in mpox dynamics, providing a more thorough understanding of the global epidemiological landscape. These findings add to a better understanding of country links and trends, which can aid in making informed decisions and developing suitable policies in a variety of fields.

The Dynamic Time Warping (DTW) method used in this section offers a significant improvement over the Ordinary Least Squares (OLS) method explained in Chapter 2.

First of all, due to its limitations, the OLS approach in Chapter 2 may have had difficulty defending cross-country comparisons. However, DTW allows for a more rigorous and reliable comparison of mpox trends across countries, using the entire time series for the similarity measurement. DTW, thus allows for a more comprehensive and relevant examination of similarities and dissimilarities between countries by generating a distance matrix and determining the DTW distance between pairs of series.

Also, OLS may not adequately align time series data, potentially resulting in curve fits that indicate maximum values much below the observed data. DTW, on the other hand, aligns time series by taking into account the local distances between matching items and selecting the optimum alignment path. This alignment ensures that cross-country comparisons are based on correct and meaningful data correspondences.

Moreover, in Chapter 4, we use hierarchical clustering based on the DTW distance matrix to uncover clusters and patterns in the mpox data. This approach extends beyond the restrictions of OLS, allowing for the examination of hierarchical linkages as well as the identification of similarities and differences between countries based on time series patterns. The clustering results provide useful insights into the dynamics and trends of the mpox across different countries.

Finally, by applying the DTW algorithm to the scaled trend data, the DTW method enables a more precise comparison of the mpox trends, enhanced alignment of time series data, and the finding of relevant clusters and patterns. These improvements help to gain a better understanding of the links and dynamics among countries in terms of mpox cases, which ultimately improves the analysis and outcomes of the study.

5 CONCLUSION AND FUTURE WORK

In conclusion, this thesis attempted to compare the 2022 mpox outbreak using mathematical modeling and time series clustering. The study sought to comprehend the patterns, dynamics, and potential control measures underlying this infectious disease. Several major conclusions have been clarified through the examination of data from the World Health Organization concerning the outbreak and the application of these quantitative methodologies.

To begin, the mathematical modeling approach used in this study provided useful insights into the transmission dynamics of the mpox virus. The choice of mathematical model for this research was heavily influenced by the AIC values for both the SEIR and SIR models. Our SIR model was able to mimic and forecast the disease's progression over time. We used Inverse Problems to estimate the effective transmission rate, the initial susceptible population, the basic reproduction number and the initial time of the first infected case for each country. Using the bootstrapping method to construct bootstrap intervals for the estimated parameters brought a critical aspect of statistical robustness to the modeling results. This resampling technique enabled the assessment of parameter uncertainties as well as the quantification of model reliability. The bootstrap intervals were useful for understanding the potential range of parameter values and the corresponding level of confidence in the model's predictions.

Second, the use of time series clustering with the scaled trend data from the mathematical model allowed the discovery of various temporal trends within the mpox outbreak, thus providing a better comparison between mpox outbreaks across countries. Clusters were generated to reflect different periods or stages of the epidemic by grouping similar time series data based on their characteristics and dynamics. This clustering approach aided in gaining a better understanding of the epidemic's progression by identifying important time points, changes in transmission patterns, and potential risk factors linked with certain clusters.

However, several limitations and opportunities for improvement in this research must be acknowledged. The mathematical model's accuracy and dependability are strongly reliant on the quality and availability of data. Limited data on mpox cases, particularly in the early phases of the outbreak, may have influenced the model's projections. Furthermore, the time series clustering analysis is very dependent on the clustering techniques and parameters used, which should be carefully studied and confirmed.

In the future, incorporating additional aspects and complexities, such as spatial dynamics, individual behavior, and environmental conditions, into the mathematical model could provide a more thorough understanding of mpox transmission. Future research should concentrate on refining and expanding the model in order to capture these nuances and evaluate the impact of targeted interventions. Also, investigating other clustering algorithms and approaches for time series analysis could provide additional insights into the outbreak's temporal patterns and heterogeneity. Comparative investigations of various clustering methodologies could improve clustering analysis and its applications in epidemic characterization.

We may enhance our understanding of mpox outbreaks and contribute to the development of more effective measures for outbreak preparedness, response, and control by exploring these prospective research directions.

BIBLIOGRAPHY

[1] Peter, O. J., Kumar, S., Kumari, N., Oguntolu, F. A., Oshinubi, K., & Musa, R. (2021). Transmission dynamics of Monkeypox virus: a mathematical modelling approach. *Modeling Earth Systems and Environment*, 1-12.

[2] Kumar, N., Acharya, A., Gendelman, H. E., & Byrareddy, S. N. (2022). The 2022 outbreak and the pathobiology of the monkeypox virus. *Journal of autoimmunity*, 102855.

[3] Bunge, E. M., Hoet, B., Chen, L., Lienert, F., Weidenthaler, H., Baer, L. R., & Steffen, R. (2022). The changing epidemiology of human monkeypox—A potential threat? A systematic review. *PLoS neglected tropical diseases, 16(2)*, e0010141.

[4] Patel, V. M., & Patel, S. V. (2023). Epidemiological Review on Monkeypox. *Cureus*, 15(2).

[5] Thornhill, J. P., Barkati, S., Walmsley, S., Rockstroh, J., Antinori, A., Harrison, L. B., ... & Orkin, C. M. (2022). Monkeypox virus infection in humans across 16 countries—April–June 2022. *New England Journal of Medicine*, 387(8), 679-691.

[6] *2022 Outbreak Cases and Data* — MPOX — Poxvirus — CDC. (n.d.). https://www.cdc.gov/poxvirus/mpox/response/2022/index.html

[7] Khayyam, H., Jazar, R. N., Nunna, S., Golkarnarenji, G., Badii, K., Fakhrhoseini, S. M., ... & Naebe, M. (2020). PAN precursor fabrication, applications

and thermal stabilization process in carbon fiber production: Experimental and mathematical modelling. *Progress in Materials Science, 107*, 100575.

[8] Anuradha, M., & Rao, K. R. Overview on Monkey Pox an Emerging Viral Infection. A Review of Literature. *European Journal of Molecular* & *Clinical Medicine (EJMCM)*, 10(01), 2023.

[9] May, R. M. (1976). Simple mathematical models with very complicated dynamics. *Nature, 261*, 459-467.

[10] Renard, P., Alcolea, A., & Ginsbourger, D. (2013). Stochastic versus deterministic approaches. *Environmental modelling: Finding simplicity in complexity*, 133-149.

[11] Hall, J., & Lingefjärd, T. (2016). *Mathematical Modeling: Applications with GeoGebra*. John Wiley & Sons.

[12] Mbuvha, R., & Marwala, T. (2020). Bayesian inference of COVID-19 spreading rates in South Africa. *PloS one, 15(8), e0237126.*

[13] Adamu, H. A., Muhammad, M., Jingi, A., & Usman, M. (2019). Mathematical modelling using improved SIR model with more realistic assumptions. *Int. J. Eng. Appl. Sci, 6(1)*, 64-69.

[14] Rachah, A. (2018). Analysis, simulation and optimal control of a SEIR model for Ebola virus with demographic effects. *Communications Faculty of Sciences University of Ankara Series A1 Mathematics and Statistics*, 67(1), 179-197.

[15] Morshed, J., & Kaluarachchi, J. J. (1998). Parameter estimation using artificial neural network and genetic algorithm for free-product migration and recovery. *Water Resources Research, 34(5)*, 1101-1113.

[16] Khan, R., Hossain, M. J., Roy, A., & Islam, M. R. (2023). Decreasing trend of monkeypox cases in Europe and America shows hope for the world: Evidence from the latest epidemiological data. *Health Science Reports*, 6(1).

[17] Liu, X., & Stechlinski, P. (2017). Infectious disease modeling. *A Hybrid System Approach. Cham: Springer.*

[18] Banks, H. T., Hu, S., & Thompson, W. C. (2014). *Modeling and inverse problems in the presence of uncertainty*. CRC Press.

[19] Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The computer journal*, 7(4), 308-313.

[20] Meuleman, B., Loosveldt, G., & Emonds, V. (2015). Regression analysis: Assumptions and diagnostics. *The SAGE handbook of regression analysis and causal inference*, 83-110.

[21] Strang, G., & Aarikka, K. (1986). *Introduction to applied mathematics*. Taiwan: Wellesley-Cambridge Press.

[22] Cintrón-Arias, A., Castillo-Chávez, C., Betencourt, L., Lloyd, A. L., & Banks, H. T. (2008). *The estimation of the effective reproductive number from disease outbreak data*. North Carolina State University. Center for Research in Scientific Computation.

[23] Hu, S. (2007). Akaike information criterion. *Center for Research in Scientific Computation*, 93, 42.

[24] Banks, H. T., & Joyner, M. L. (2017). AIC under the framework of least squares estimation. *Applied Mathematics Letters, 74*, 33-45.

[25] den Driessche, P. V. (2017, June 29). *Reproduction numbers of infectious disease models*. PubMed Central (PMC). https://doi.org/10.1016/j.idm.2017.06.002

[26] Dietz, K. (1993). The estimation of the basic reproduction number for infectious diseases. *Statistical methods in medical research, 2(1)*, 23-41.

[27] Delamater, P. L., Street, E. J., Leslie, T. F., Yang, Y., & Jacobsen, K. H. (2019). Complexity of the Basic Reproduction Number (R0). *Emerging Infectious Diseases, 25(1)*, 1-4. https://doi.org/10.3201/eid2501.171901.

[28] Gao, F., & Han, L. (2012). Implementing the Nelder-Mead simplex algorithm with adaptive parameters. *Computational Optimization and Applications, 51*(1), 259-277.

[29] Luersen, M. A., & Le Riche, R. (2004). Globalized Nelder–Mead method for engineering optimization. *Computers & Structures*, 82(23-26), 2251-2260.

[30] Ugrinowitsch, C., Fellingham, G. W., & Ricard, M. D. (2004). Limitations of ordinary least squares models in analyzing repeated measures data. *Medicine and Science in Sports and Exercise*, 2144–2148. https://doi.org/10.1249/01.mss.0000147580.40591.75

[31] Barber, J. A., & Thompson, S. G. (2000). Analysis of cost data in randomized trials: *an application of the non-parametric bootstrap. Statistics in medicine, 19(23)*, 3219-3236.

[32] Moffat, R. J. (1985). Using uncertainty analysis in the planning of an experiment.

[33] Meesrikamolkul, W., Niennattrakul, V., & Ratanamahatana, C. A. (2012). Shape-based clustering for time series data. *In Advances in Knowledge Discovery and Data Mining: 16th Pacific-Asia Conference, PAKDD 2012, Kuala Lumpur, Malaysia*, May 29-June 1, 2012, Proceedings, Part I 16 (pp. 530-541). Springer Berlin Heidelberg.

[34] Keogh, M. V. J. L. E., & Gunopulos, D. A Wavelet-Based Anytime Algorithm for K-Means Clustering of Time Series.

[35] Tormene, P., Giorgino, T., Quaglini, S., & Stefanelli, M. (2009). Matching incomplete time series with dynamic time warping: an algorithm and an application to post-stroke rehabilitation. *Artificial intelligence in medicine, 45(1)*, 11-34.

[36] Salvador, S., & Chan, P. (2007). Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis, 11(5)*, 561-580.

[37] Mensch, A., & Blondel, M. (2018, July). Differentiable dynamic programming for structured prediction and attention. *In International Conference on Machine Learning* (pp. 3462-3471). PMLR.

[38] Mondal, T., Ragot, N., Ramel, J. Y., & Pal, U. (2014, September). Flexible sequence matching technique: application to word spotting in degraded documents.

In 2014 14th International Conference on Frontiers in Handwriting Recognition (pp. 210-215). IEEE.

[39] Alizadeh, E. (2020, October 11). *An illustrative introduction to Dynamic Time Warping*. Essi Alizadeh. https://ealizadeh.com/blog/introduction-to-dynamic-time-warping/ (Accessed: June 22, 2023).

[40] Wan, Y., Chen, X. L., & Shi, Y. (2017). Adaptive cost dynamic time warping distance in time series analysis for classification. *Journal of Computational and Applied Mathematics, 319*, 514-520.

[41] Berkhin, P. (2006). A survey of clustering data mining techniques. *Grouping multidimensional data: Recent advances in clustering, 25-71.*

[42] Sharma, S., & Batra, N. (2019, February). Comparative study of single linkage, complete linkage, and ward method of agglomerative clustering. *In 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)* (pp. 568-573). IEEE.

[43] Mpox (Monkeypox) outbreak 2022 (no date) World Health Organization. World Health Organization. Available at: https://www.who.int/emergencies/situations/monkeypox-outbreak-2022 (Accessed: April 3, 2023).

APPENDICES

0.1 Mathematical Model Example Code Via MATLAB

```
%% load data

load('WHO_data_US.mat')

index_greater1 = find(new_cases_smoothed1 > 0.05*max(new_cases_smoothed1));

data = new_cases_smoothed1(index_greater1);

[data, ia, ic] = unique(data,'first');

tdata = datenum(date1(index_greater1));

tdata = tdata(ia);

[tdata,I] = sort(tdata,'ascend');

data = data(I);


%% Initial model

% model with an input of the optimal value of beta which was found above

t0 = datenum(date1(1))-1;

k0 = 2.7;

Sh0 = k0*max(data);

Ih0 = 1;

Rh0 = 0;

tplot = linspace(t0,datenum(date1(end)),500);

[tplot1,wplot] = ode45(@rhs_MP_cost_fixedg, tplot, w0, [], qopt);
```

0.2 Bootstrap Example Code Via MATLAB

```
%% Bootstrap

no_param = 3;

bootstrap_size = 500; (500−1000 for a good estimate)

paramests_matrix = zeros(bootstrap_size,length(q0));

rsdl = sqrt(length(data)/(length(data)−no_param))*(data − iapprox);

    figure(2) % starts a new figure for error bounds

    figure(3) % starts a new figure for bootstrap data sets

    for m=1:bootstrap_size

            disp(sprintf('bootstrap␣iteration␣=␣%d',m))


            srs_residuals = randSample(rsdl,length(rsdl),true);


            i_srs = iapprox + srs_residuals;


            index_greater_srs = find(i_srs > 0.05*max(i_srs));

            i_srs = i_srs(index_greater_srs);

            tdata_boot = tdata(index_greater_srs);

    end
```

0.3 Distance Matrix Code Via Python

```
#Importing Libraries
import pandas as pd
import numpy as np
from tslearn.utils import to_time_series_dataset
from tslearn.metrics import dtw
import scipy.cluster.hierarchy as sch
from scipy.cluster.hierarchy import linkage, dendrogram
n_series = formatted_dataset.shape[0]
distance_matrix = np.zeros(shape=(n_series, n_series))
# Build distance matrix
for i in range(n_series):
    for j in range(n_series):
        x = formatted_dataset[i]
        y = formatted_dataset[j]
        if i != j:
            dist = dtw(x, y)
            distance_matrix[i, j] = dist

print(distance_matrix)
```

0.4 Hierarchical Clustering Code Via Python

```
n_series = formatted_dataset.shape[0]
distance_matrix = np.zeros(shape=(n_series, n_series))
for i in range(n_series):
    for j in range(n_series):
        x = formatted_dataset[i]
        y = formatted_dataset[j]
        if i != j:
            dist = dtw(x, y)
            distance_matrix[i, j] = dist
Z = sch.linkage(distance_matrix, method='complete')
plt.figure(figsize=(10, 5))
dendrogram = sch.dendrogram(Z,labels=labelList)
plt.title('Dendrogram')
plt.xlabel('Countries')
plt.ylabel('Distance')
plt.axhline(y=0.28, color='k', linestyle='—')
plt.axhline(y=0.45, color='k', linestyle=':')
plt.show()
```

VITA

MARK-DANIELS TAMAKLOE

Education:	M.S. Mathematical Sciences, East Tennessee State University (ETSU), Johnson City, Tennessee, 2023 B.Sc. Mathematics with Economics, of Cape Coast, Cape Coast, Ghana, 2019
Professional Experience:	Graduate Assistant and Tutor, (ETSU), Johnson City, Tennessee, 2021–2023 High School Mathematics Teacher, Hope College, Gomoa-Fetteh, Ghana, 2020–2021
Professional Development:	Statistical and Mathematical R, Matlab, Python, SQL Microsoft Office Suite: MS Access , Word, Excel, PowerPoint, Publisher