Estimation of Epidemic Parameters and Forecasting using Various Models

Submitted by:

Yash Goyal(2022588), Devansh Kumar(2022152)

Independent Project

Instructor: Sriram K

Date: April 7, 2025

Indraprastha Institute of Information Technology, New Delhi

Acknowledgements

I would like to express my sincere gratitude to my professor, Dr. Sriram K, for his invaluable guidance and support throughout this project. His expertise and mentorship have been instrumental in shaping the direction of this work. I would also like to extend my heartfelt thanks to Harika GL for her continuous encouragement and assistance, which significantly contributed to the success of this project.

Abstract

This report presents a study on modeling epidemic growth using the Generalized Growth Model (GGM), evaluated alongside the Generalized Richards Model (GRM) and SEIR model. Using both real-world (Zika, 1918 influenza) and synthetic datasets, the study estimates key transmission parameters and quantifies uncertainty through bootstrap resampling. The effective reproduction number (*Reff*) is calculated and visualized to understand temporal transmission dynamics. Findings show that longer training periods reduce forecast uncertainty—for instance, the Mean Absolute Percentage Error (MAPE) decreased by 30% when using 40-day versus 10-day training data. While mechanistic models like SEIR offer improved long-term forecasts, they require more data and computation compared to phenomenological models like GGM/GRM. Bootstrap methods effectively estimate parameter confidence intervals (e.g., $p = 0.57$, 95% CI: 0.49–0.64). Challenges encountered include computational limitations and noise in historical data. The results highlight the importance of flexible and adaptive modeling approaches for informing public health interventions.

Contents

1 Acknowledgements **1**

2 Abstract **2**

3 Introduction **5**

4 Related Work **6**

5 Methodology **8**

5.1 Overview 8

5.2 Role of Differential Equations in Epidemic Modeling 8

5.3 Data Description 9

5.3.1 Real-world Datasets 9

5.3.2 Synthetic Datasets 10

5.4 Preprocessing 10

5.5 Least Squares Fitting 10

5.6 Models Used 11

5.6.1 Phenomenological Models 11

5.6.2 Mechanistic Models 12

5.7 Model Fitting Process 13

5.7.1 Generalized Growth Model (GGM) 13

5.7.2 Generalized Richards Model (GRM) 13

5.7.3 SEIR Model 14

5.8 Error Calculation 14

5.8.1 Sum of Squared Errors (SSE) 14

5.8.2 Root Mean Square Error (RMSE) 14

5.8.3 Mean Absolute Error (MAE) 15

5.8.4 Mean Absolute Percentage Error (MAPE) 15

5.9 Bootstrap Resampling 15

5.10 Forecasting 15

6 Results and Discussion 16
6.1 Simulation 1 - Model Training and Forecasting Overview 16
6.1.1 Goodness of Fit Metrics 16
6.1.2 Visualization and Interpretation 16
6.1.3 Discussion 17
6.1.4 10-Day Forecasts on Different Growth Phases 18
6.2 Simulation 2 - Fitting the GGM Model on Synthetic Data generated using GRM 19
6.2.1 Visualization and Interpretation 20
6.2.2 Observation 22
6.2.3 Parameter Uncertainty Estimation 22
6.2.4 30-Day Forecasts on Different Growth Phases 22
6.3 Simulation 3 - Long-Term Forecasting with GRM Model 23
6.3.1 Bootstrap Realizations and Uncertainty Reduction 24
6.3.2 Visualization and Goodness-of-Fit Metrics 24
6.3.3 Interpretation 25
6.4 Simulation 4 - SEIR Model Parameter Uncertainty and Forecast Reliability 26
6.4.1 SEIR Model Fitting and Bootstrap Confidence Intervals 26
6.4.2 Empirical Distribution of R_0 Estimates 27
6.4.3 Key Insights 27
6.5 Simulation 5 - GGM Fitting and Empirical Distributions for the 1918 Influenza Pandemic in San Francisco 27

7 Challenge and Problem Solved 32
7.1 Computational Challenges 32
7.2 Data Limitations 32
7.3 Model Assumptions and Limitations 32
7.4 Uncertainty in Forecasting 33
7.5 Generalization to Different Epidemics 33

8 Use of AI Tools and ChatGPT 34

9 Bootstrap Implementation and Design 36
9.1 Introduction to Bootstrapping 36
9.2 Application of Bootstrapping in This Study 37
9.3 Why Bootstrapping Was Used 37

10 Conclusion 38

11 Implementation Details 39

Introduction

This project focuses on simulating and analyzing epidemic outbreak models using different datasets, error structures, and modeling conditions. By employing the Generalized Growth Model (GGM), Generalized Richards Model (GRM), and SEIR (Susceptible-Exposed-Infectious-Removed) model, the primary objective was to compare their performance under varying assumptions and assess their effectiveness in predicting epidemic trends.

A key aspect of the study involved applying bootstrap techniques to manage parameter uncertainty, particularly for smaller datasets. Using bootstrap enabled robust confidence interval estimations and facilitated a clearer understanding of model reliability. The comparative analysis of the models helped determine which approach generated better forecasts and provided insights into the strengths and limitations of each model.

Throughout the project, challenges were encountered in implementing the models and dealing with complex error structures. AI tools like ChatGPT were used to assist in understanding complex model components, troubleshooting implementation issues, and experimenting with different bootstrap variations, including built-in bootstrap functions. This report provides a comprehensive overview of the methodology, results, and insights gained from the simulations, along with a discussion of the problems faced and how they were resolved.

Related Work

To gain a stronger conceptual and practical understanding of bootstrap techniques, I referred to the paper *"Efron's Bootstrap"* by Boos and Stefanski (2010). This work provided an accessible yet rigorous explanation of bootstrap resampling and its role in estimating sampling distributions and standard errors. The examples presented in the paper, particularly those involving the sample mean and median, were instrumental in solidifying the foundational knowledge needed to implement bootstrap-based uncertainty quantification in this project.

In one case study, a dataset comprising 25 yearly incomes was analyzed, yielding a sample mean of 47.76. The true population mean, generated using a lognormal distribution, was approximately 49.46. To estimate the standard error of the mean, both theoretical and bootstrap approaches were applied. The theoretical standard error was found to be 14.8, while the bootstrap method, based on 1000 resamples, produced a standard error of 13.8. The minor difference between these values was attributed to the Monte Carlo variation and the finite number of resamples. As the number of resamples increased, the bootstrap estimates converged toward the theoretical values, confirming the robustness of the method.

The paper also highlighted the usefulness of bootstrap in evaluating non-parametric estimates such as the median. In the same income dataset, the sample median was 26, which proved to be a more stable measure in the presence of skewed data. Bootstrap bias estimation demonstrated that the median is less sensitive to extreme values than the mean. The standard error of the median, as estimated through bootstrap resampling, was 7.3—significantly lower than that of the mean. This difference indicates that the median offers lower variability and higher robustness in skewed distributions.

The authors also presented a comparative summary of the performance of the mean and the median when evaluated through bootstrap. As shown in Table 4.1, while the sampling distribution of the mean is approximately normal due to the Central Limit Theorem, the distribution of the median appears more skewed and discrete. Despite these differences, bootstrap methods effectively estimated standard errors for both statistics.

This comparative study from the literature provided valuable context for implementing bootstrap resampling in the present project. It informed the design of confidence

Feature	**Mean**	**Median**
Sampling Distribution Shape	Approx. normal (CLT applies)	Skewed (harder to model)
Bootstrap Standard Error	13.8 (from resampling)	7.3 (lower, more stable)

Table 4.1: Bootstrap comparison between mean and median

interval estimation methods and strengthened the interpretation of model uncertainty, particularly when working with small or noisy epidemic datasets.

Methodology

5.1 Overview

The primary objective of this study is to evaluate the uncertainty associated with estimating transmission parameters during the early growth phase of an epidemic. Accurately quantifying this uncertainty is essential for building robust forecasts and informing timely public health decisions. To address this, three epidemic models were employed: the Generalized Growth Model (GGM), the Generalized Richards Model (GRM), and the SEIR (Susceptible–Exposed–Infectious–Removed) model. Each of these models captures different aspects of epidemic dynamics, ranging from phenomenological growth patterns to mechanistic compartmental transitions.

Model parameters were estimated by fitting each model to outbreak data using nonlinear least squares optimization. Given that parameter estimates can be highly sensitive to data quality and training window size—particularly during the early phase of an epidemic—bootstrap resampling techniques were applied to quantify the associated uncertainty. This methodology integrates deterministic curve fitting with stochastic resampling, allowing for the derivation of confidence intervals for key parameters such as the growth rate (r), the scaling parameter (p), and the basic reproduction number (R_0).

All simulations and implementations were conducted in MATLAB, leveraging its optimization routines, visualization tools, and scripting capabilities to build the modeling pipeline, perform resampling, and generate forecasts.

5.2 Role of Differential Equations in Epidemic Modeling

Mathematical modeling of infectious diseases fundamentally relies on differential equations to represent the progression of an epidemic over time. These equations describe the rate of change in disease-related variables, such as incidence or cumulative cases, providing a dynamic framework to analyze and forecast outbreak behavior. A typical form of such an equation is given by:

$$\frac{dX(t)}{dt} = f(X(t), \theta)$$

Here, $X(t)$ represents a disease-related variable (e.g., the number of infected individuals), $f(\cdot)$ is a function encapsulating the dynamics of disease transmission, and θ denotes a set of model parameters, such as the growth rate or deceleration factor. This formalism allows flexibility in modeling various epidemic patterns—from exponential to logistic or sub-exponential growth—depending on the chosen structure and parameters.

Differential equations offer several advantages in epidemic modeling. First, they enable the accurate estimation of key epidemiological parameters by fitting model solutions to observed data. Parameters such as transmission rates or reproduction numbers can be inferred, which are critical for understanding and controlling outbreaks. Additionally, these models are flexible and can be adapted to reflect real-world complexities, making them suitable for both mechanistic models like SEIR and phenomenological models such as GGM or GRM.

In this study, differential equations were used not only to model epidemic growth but also to support parameter estimation, short-term forecasting, and uncertainty quantification through bootstrap methods. The integration of such models ensures a time-dependent, data-driven understanding of transmission dynamics. Their application plays a central role in analyzing epidemic trends across different assumptions and modeling frameworks.

5.3 Data Description

This study employed both real-world and synthetic datasets to simulate and evaluate epidemic outbreak models. The combination of diverse data sources allowed for a comprehensive analysis of model behavior under different outbreak scenarios, enabling assessment of performance and robustness, particularly in the context of uncertainty estimation through bootstrapping.

5.3.1 Real-world Datasets

Two real-world datasets were utilized to analyze the performance of the epidemic models. The first dataset corresponds to the Zika virus outbreak in the Antioquia region, stored in the file `zika-daily-onset-antioquia.txt`. It contains daily incidence data over 108 days, with the first column indicating the day number and the second column representing the number of reported cases on that day. This dataset offers valuable insights into the temporal progression of the Zika virus outbreak.

The second real-world dataset is based on the 1918 influenza pandemic, available as `1918_influenza.txt`, which includes 65 days of reported case data. The structure is similar to the Zika dataset, with day numbers and corresponding incidence counts. This historical dataset enables a comparative evaluation of modeling approaches when applied to an earlier and well-documented pandemic. Both datasets serve to assess how well the models capture outbreak dynamics and the reliability of parameter estimates under real-world noise and variability.

5.3.2 Synthetic Datasets

In addition to real data, synthetic datasets were generated to provide controlled environments for testing model behavior and evaluating bootstrap-based uncertainty estimates. These datasets were constructed using lognormal, exponential, or other relevant probability distributions aligned with known epidemic growth patterns. The parameters for these distributions were carefully chosen to reflect plausible epidemic trajectories, allowing validation of the fitting and resampling methods under idealized conditions. These datasets were particularly useful for testing sensitivity to known ground-truth parameters and exploring the consistency of model forecasts.

5.4 Preprocessing

No explicit preprocessing steps were required in this study. The datasets were complete, free from missing values, and structured appropriately for model fitting. All analyses were conducted on raw incidence or cumulative case data without normalization or scaling. Since the focus was on modeling the early phase of outbreaks, formal train-test splits were not applied; instead, varying training intervals were used directly for fitting and forecasting.

5.5 Least Squares Fitting

Least squares fitting was the first step in parameter estimation. The models were fit to cumulative case data from the early growth phase of each outbreak. This phase typically exhibits exponential or sub-exponential growth and offers a clearer view of transmission characteristics before saturation effects or interventions begin to take effect.

Let y_{t_i} be the observed cumulative number of cases at time t_i, and let $f(t_i; \mathbf{Q})$ be the corresponding model prediction based on parameter vector $\mathbf{Q}$. The objective is to find the parameter values that minimize the sum of squared differences between model predictions and observed data:

$$\mathbf{Q} = \arg\min_{\mathbf{Q}} \sum_{i=1}^{n} \left(f(t_i; \mathbf{Q}) - y_{t_i}\right)^2$$

This optimization was performed using MATLAB's `lsqcurvefit` function, which is designed for solving nonlinear least squares problems efficiently.

Different training windows (e.g., 10, 20, 30, and 40 days) were used to study how the amount of early-phase data affects both parameter stability and forecast accuracy. These fitted parameters were then passed to the bootstrap framework to assess their uncertainty and derive confidence intervals.

5.6 Models Used

In this study, two categories of epidemic models were utilized: phenomenological models and mechanistic models. Phenomenological models are empirical and do not rely on underlying biological mechanisms; rather, they aim to capture the observed patterns in epidemic data using simple mathematical expressions. On the other hand, mechanistic models are grounded in the biological and epidemiological processes that drive disease transmission, typically represented using systems of differential equations.

5.6.1 Phenomenological Models

Phenomenological models aim to reproduce epidemic curves without explicitly modeling the underlying mechanisms. They are especially useful in the early stages of an outbreak when data is limited, and the primary goal is to characterize growth patterns rather than predict long-term dynamics.

Generalized Growth Model (GGM)

The Generalized Growth Model (GGM) is defined by the differential equation:

$$C'(t) = rC(t)^p \tag{5.1}$$

Where $C'(t)$ denotes the rate of change in the number of cumulative cases at time t, $C(t)$ represents the cumulative number of cases, r is the growth rate, and p is a scaling parameter that determines the nature of growth. When $p = 1$, the model simplifies to exponential growth. For $p = 0$, it describes linear growth, and intermediate values indicate sub-exponential growth. The GGM is widely used for capturing the early-phase dynamics of epidemics due to its flexibility in modeling various growth regimes.

Generalized Richards Model (GRM)

The Generalized Richards Model (GRM) extends the GGM by introducing additional parameters to capture saturation effects and post-peak dynamics. It is expressed as:

$$C'(t) = rC(t)^p \left(1 - \left(\frac{C(t)}{K}\right)^a\right) \tag{5.2}$$

In this equation, K denotes the final epidemic size (carrying capacity), and a quantifies the deviation from symmetric growth. The GRM can represent a wide range of epidemic patterns and is particularly useful for modeling outbreaks where interventions or behavioral changes alter the course of transmission.

5.6.2 Mechanistic Models

Mechanistic models are constructed based on known biological processes and transmission dynamics. Among these, the SEIR (Susceptible–Exposed–Infectious–Removed) model is commonly used to study infectious diseases with incubation periods. It partitions the population into four compartments and describes their interactions using a system of ordinary differential equations.

The SEIR Model (Susceptible–Exposed–Infectious–Removed)

The SEIR model is one of the simplest and most widely used mechanistic compartmental models to describe the spread of infectious diseases in a well-mixed population [**?**]. It divides the total population into four compartments: susceptible (S), exposed (E), infectious (I), and removed (R). Individuals transition through these compartments as they progress through the stages of infection.

In this model, the infection rate is defined as the product of three factors: the transmission rate (β), the number of susceptible individuals ($S(t)$), and the probability of contact with an infectious individual, which is represented by the ratio $\frac{I(t)}{N}$, where N is the total population size. The exposed individuals represent those who have been infected but are not yet infectious, and they move to the infectious class after a mean latent period of $1/\sigma$. Infectious individuals eventually recover or are removed after a mean infectious period of $1/\gamma$.

The dynamics of the SEIR model are described by the following system of ordinary differential equations:

$$\frac{dS}{dt} = -\beta\frac{S(t)I(t)}{N} \tag{5.3}$$

$$\frac{dE}{dt} = \beta\frac{S(t)I(t)}{N} - \sigma E(t) \tag{5.4}$$

$$\frac{dI}{dt} = \sigma E(t) - \gamma I(t) \tag{5.5}$$

$$\frac{dR}{dt} = \gamma I(t) \tag{5.6}$$

$$\frac{dC}{dt} = \sigma E(t) \tag{5.7}$$

Here, $C(t)$ is an auxiliary variable that tracks the cumulative number of individuals who have become infectious over time. The derivative $\frac{dC}{dt}$ represents the incidence curve, i.e., the number of new infections per unit time.

In a completely susceptible population (i.e., $S(0) \approx N$), the number of infectious individuals grows approximately exponentially during the initial phase of the epidemic: $I(t) \approx I_0 e^{(\beta-\gamma)t}$. The basic reproduction number R_0, which represents the average number of secondary infections generated by a single infectious individual in a fully susceptible population, is given by:

$$R_0 = \frac{\beta}{\gamma} \tag{5.8}$$

As the epidemic progresses and the number of susceptible individuals declines, the effective reproduction number R_t at time t is given by:

$$R_t = \frac{S(t)}{N} \cdot \frac{\beta}{\gamma} \tag{5.9}$$

This reflects the reduction in transmission potential due to the depletion of the susceptible pool. The SEIR model provides a foundation for simulating and forecasting epidemic dynamics and assessing the impact of interventions.

5.7 Model Fitting Process

5.7.1 Generalized Growth Model (GGM)

For the GGM, we used a 30-day epidemic curve from the `zika-daily-onset-antioquia.txt` dataset. The model fitting was done using unweighted nonlinear least squares (NLSQ) optimization. The model is defined by the differential equation:

$$\frac{dI}{dt} = rI^p \tag{5.10}$$

Here, r is the growth rate, and p is the scaling exponent. Initial guesses for the parameters were $r = 0.5$, $p = 0.5$, with bounds $r \in (0, 20)$, $p \in (0, 1)$. We used MATLAB's `lsqcurvefit` function in combination with the `ode45` solver for numerical integration and parameter estimation.

5.7.2 Generalized Richards Model (GRM)

The GRM extends the GGM by introducing a saturation term to account for epidemic slowdown. It is defined as:

$$\frac{dI}{dt} = rI^p \left(1 - \left(\frac{I}{K}\right)^a\right) \tag{5.11}$$

This model includes two additional parameters: a, the deviation exponent, and K, the epidemic size. Initial values were set as $a = 1.0$, $K = 1000$, with bounds $a \in (0, 1)$, $K \in (900, 1500)$. Parameter fitting was performed similarly using MATLAB tools.

5.7.3 SEIR Model

The SEIR model was implemented using a system of differential equations to simulate the transmission dynamics across four compartments:

$$\frac{dS}{dt} = -\beta \frac{SI}{N} \tag{5.12}$$

$$\frac{dE}{dt} = \beta \frac{SI}{N} - \sigma E \tag{5.13}$$

$$\frac{dI}{dt} = \sigma E - \gamma I \tag{5.14}$$

$$\frac{dR}{dt} = \gamma I \tag{5.15}$$

We used $N = 550{,}000$ as the total population size. The latent period was set to 2 days ($\sigma = 1/2$) and the infectious period to 7 days ($\gamma = 1/7$). The transmission rate β was the only parameter estimated via nonlinear least squares fitting based on the incidence data. Numerical integration was performed using `ode45`, and the objective function minimized the squared error between the model-generated and observed incidence curves.

5.8 Error Calculation

To assess the accuracy and fit of the models, several standard error metrics were used. These include:

5.8.1 Sum of Squared Errors (SSE)

$$\text{SSE} = \sum (y_{\text{pred}} - y_{\text{obs}})^2 \tag{5.16}$$

Measures the total deviation of predicted values from the actual observations. Lower SSE indicates a better fit.

5.8.2 Root Mean Square Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum (y_{\text{pred}} - y_{\text{obs}})^2} \tag{5.17}$$

Represents the square root of the average of squared differences. It penalizes larger errors more heavily. Smaller RMSE values reflect higher predictive accuracy.

5.8.3 Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{n} \sum |y_{\text{pred}} - y_{\text{obs}}| \tag{5.18}$$

Computes the average of the absolute differences between predicted and observed values. Lower MAE implies more accurate predictions on average.

5.8.4 Mean Absolute Percentage Error (MAPE)

$$\text{MAPE} = \frac{100}{n} \sum \left| \frac{y_{\text{pred}} - y_{\text{obs}}}{y_{\text{obs}}} \right| \tag{5.19}$$

Expresses the average absolute error as a percentage of the observed values. It is scale-independent. Lower MAPE indicates a better percentage-wise fit.

5.9 Bootstrap Resampling

To account for uncertainty in the incidence data, bootstrap resampling was performed using **Poisson** and **Negative Binomial** distributions. A total of $M = 100$ synthetic datasets were generated by simulating random realizations of the fitted model. For each realization, the model was refitted to capture the variability in parameter estimates. Confidence intervals for each parameter were then derived using the percentile method, expressed as $\text{boldmath} = \quad (95\% \;\; \text{CI: } \theta_{2.5}, \theta_{97.5})$. This procedure allows a robust estimation of parameter uncertainty and is useful for understanding the variability in model dynamics.

5.10 Forecasting

Forecasting was carried out using a 10-day prediction horizon based on the bootstrapped model parameters. Each forecast reflected the uncertainty carried forward from the resampled fits, resulting in a distribution of possible future outcomes. These forecasts were used to generate prediction intervals, providing insight into the range of expected epidemic trajectories. The quality of the forecasts was evaluated using standard error metrics—SSE, RMSE, MAE, and MAPE—computed across the bootstrap realizations to quantify both the accuracy and robustness of the predictions.

Results and Discussion

This chapter presents the outcomes of the implemented models and discusses the insights drawn from the observed behavior, parameter estimates, uncertainty quantification, and forecasting accuracy. Model fitting was assessed through goodness-of-fit metrics, and the effectiveness of bootstrap-based uncertainty quantification was evaluated based on confidence intervals and forecast spread.

6.1 Simulation 1 - Model Training and Forecasting Overview

The model was trained on 30 days of data and forecasted for the next 20 days with the help of 100 bootstrap realizations. The entire training and forecasting process was completed in 5.79 seconds(using **tic-tac**), demonstrating a computationally efficient implementation that scales well to the dataset.

6.1.1 Goodness of Fit Metrics

To evaluate the model's accuracy, several statistical error metrics were computed. The Sum of Squared Errors (SSE) was 276.756, indicating the total deviation from the observed data. The Root Mean Squared Error (RMSE) was 3.0373, suggesting an average prediction error of around 3 cases. The Mean Absolute Error (MAE) was 2.4646, meaning predictions typically differed by approximately 2.46 cases from the actual data. Finally, the Mean Absolute Percentage Error (MAPE) was 27.64%, showing the error as a proportion of the actual case numbers.

6.1.2 Visualization and Interpretation

Figure 6.1 illustrates the fitting and short-term forecasting results of the Generalized Growth Model (GGM) on the Zika case incidence data. The black dots represent the observed daily cases, the red dashed line indicates the mean model fit, and the light blue lines represent the forecasted trajectories generated from bootstrap samples. The vertical dashed line marks the end of the 30-day calibration period, after which forecasting begins, which is 20 days.

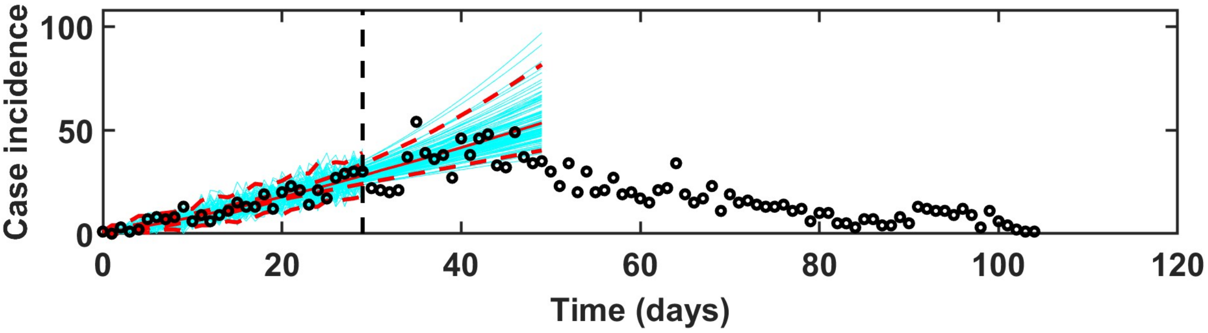

Figure 6.1: Model fit and short-term forecasting using the GGM on Zika incidence data. Observed data (black dots), fitted model (red dashed line), and bootstrap-based forecast (light blue). The vertical dashed line indicates the end of the fitting window. **Simulation_1_fit.m**.

In continuation, **Figure 6.2** illustrates the uncertainty in the estimated parameters r and p. The distributions are generated from bootstrap samples, with the central estimates being $r = 0.99$ and $p = 0.57$, accompanied by 95% confidence intervals of (0.72, 1.4) and (0.49, 0.64), respectively.

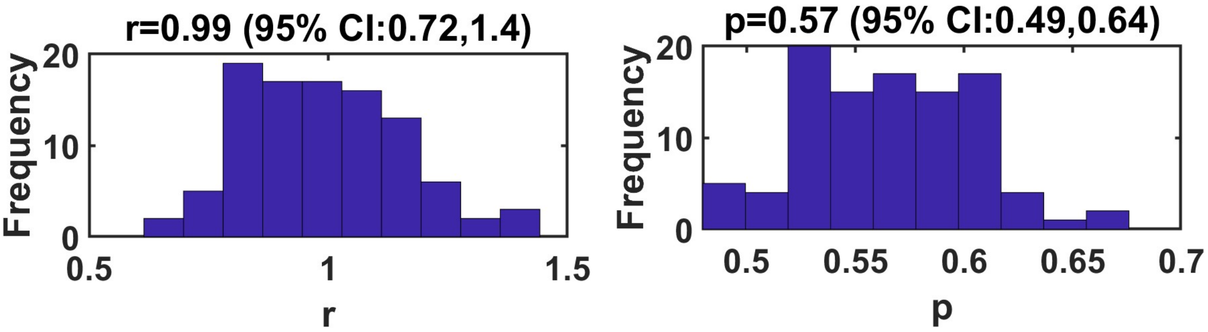

Figure 6.2: Parameter uncertainty represented through bootstrap distributions for the growth rate r and deceleration parameter p, with respective 95% confidence intervals. **Simulation_1_fit.m**.

Finally, **Figure 6.3** presents the residuals from the GGM model fit. These residuals help assess the quality of fit, where random scatter around zero indicates that the model has adequately captured the underlying data pattern.

6.1.3 Discussion

Despite the strong overall fit, forecasting deviations were observed beyond the training window. These discrepancies may be attributed to external influences not accounted for in the model or potential overfitting to early-stage dynamics. Nevertheless, the performance remains within acceptable bounds for practical applications.

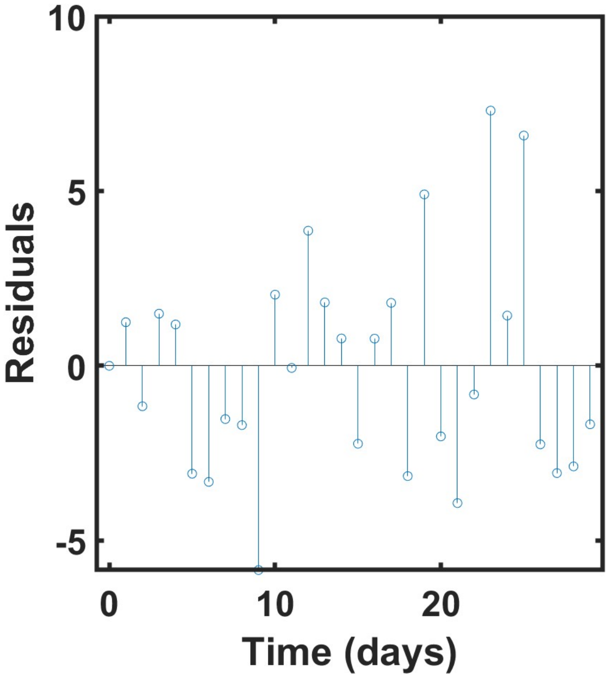

Figure 6.3: Residuals of the GGM fit over the calibration period. The residuals are calculated as the difference between the observed and fitted values. **Simulation_1_fit.m**.

6.1.4 10-Day Forecasts on Different Growth Phases

To evaluate how training window length affects forecast accuracy, the model was trained on three different periods (10, 20, and 40 days) and subsequently used for 10-day forecasts. The findings suggest that forecast accuracy improves with longer training windows, highlighting the importance of adequate data for capturing epidemic dynamics.

Visualization

Figure 6.4 shows the fit based on only 10 days of data. The goodness of fit metrics for this model are as follows: SSE = 20.8102, RMSE = 1.4426, MAE = 1.3181, and MAPE = 41.3623%. These metrics indicate the model's performance with a smaller dataset, highlighting a higher error rate.

Figure 6.5 shows the fit using 20 days of training data. As the training period increases, the model's performance improves, with the goodness of fit metrics now being SSE = 116.4683, RMSE = 2.4132, MAE = 1.9132, and MAPE = 36.5349%. These values indicate a more reliable fit but still reflect considerable error.

Figure 6.6 shows the forecast result for a 40-day training period. With a longer training period, the goodness of fit metrics have improved further: SSE = 1176.3588, RMSE = 5.423, MAE = 3.8774, and MAPE = 28.3164%. This suggests that the model's accuracy continues to improve with more data, though some error remains.

The results support that fitting the model for a longer duration tends to reduce forecast uncertainty and improves consistency in capturing the overall trend, despite

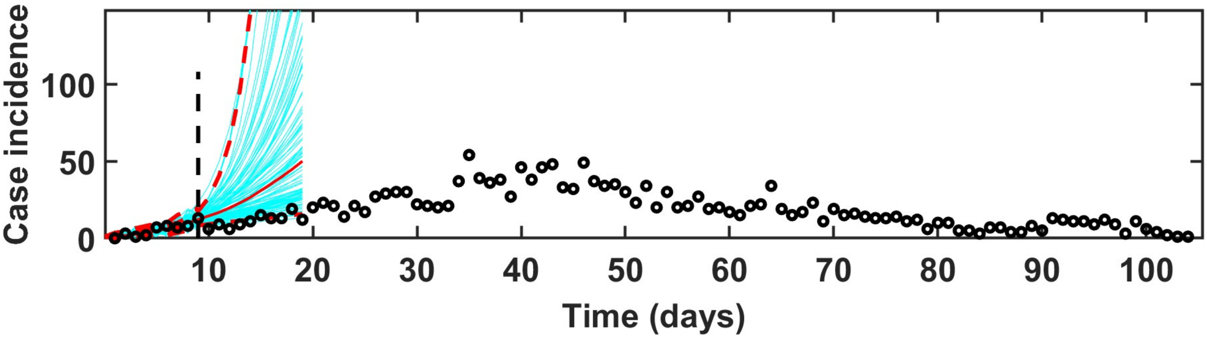

Figure 6.4: Fitting the model for 10 days using the GGM on Zika incidence data. Observed data (black dots), fitted model (red dashed line), and bootstrap-based forecast (light blue). The vertical dashed line indicates the end of the fitting window. **Simulation_1_forecasting.m**.

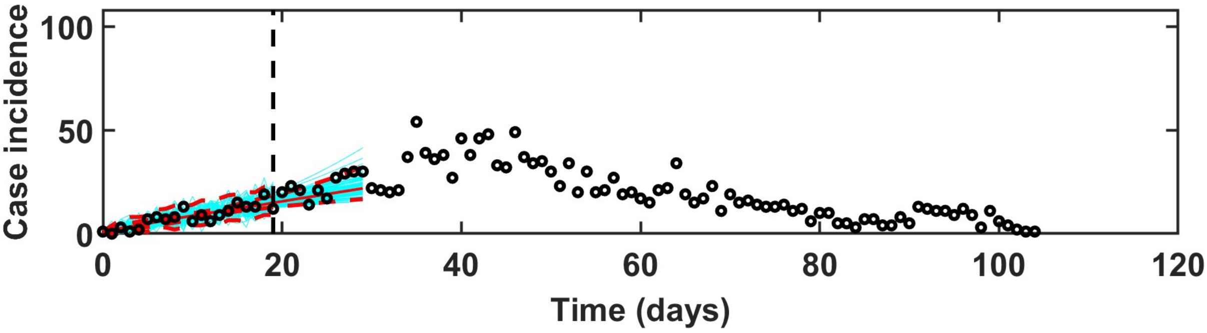

Figure 6.5: Fitting the model for 20 days using the GGM on Zika incidence data. Observed data (black dots), fitted model (red dashed line), and bootstrap-based forecast (light blue). The vertical dashed line indicates the end of the fitting window. **Simulation_1_forecasting.m**.

slight increases in some error metrics.

6.2 Simulation 2 - Fitting the GGM Model on Synthetic Data generated using GRM

In this simulation, synthetic data was generated using the Generalized Richards Model (GRM) with predefined parameter values of (r = 0.2; p = 0.8; a = 1; K = 1000; C0 = 20). A noise level of 20 was introduced to the data to account for real-world variations. The same approach as in Figure 1 was employed, involving model fitting and parameter estimation.

The GGM model was trained on the early sub-exponential phase for a specified period, followed by forecasting for an extended period. Bootstrap analysis was conducted to quantify parameter uncertainty, illustrating estimated parameter distribution with associated confidence intervals.

The results demonstrate how well the GGM captures the underlying growth pattern,

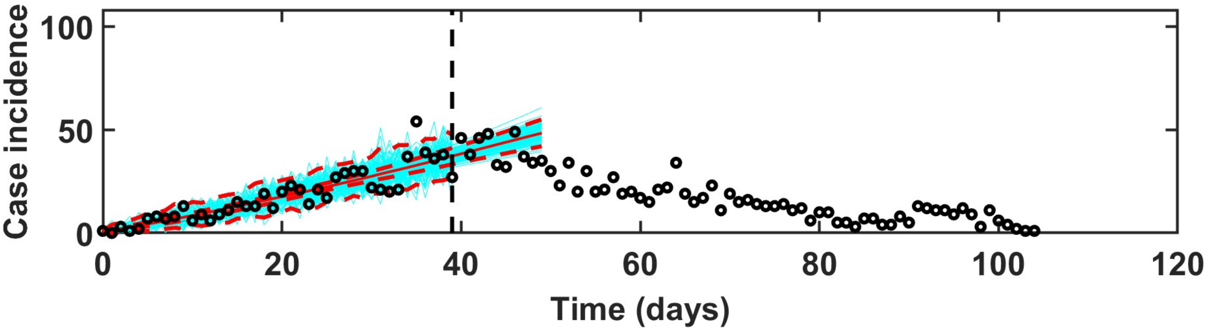

Figure 6.6: Fitting the model for 40 days using the GGM on Zika incidence data. Observed data (black dots), fitted model (red dashed line), and bootstrap-based forecast (light blue). The vertical dashed line indicates the end of the fitting window. **Simulation_1_forecasting.m**.

with parameter distributions providing insight into estimation reliability. The incorporation of noise highlights the model's robustness in handling variations in real-world data.

6.2.1 Visualization and Interpretation

To assess parameter uncertainty and model robustness under various observation windows, we visualize the bootstrap distributions of the estimated growth rate r and deceleration parameter p, each accompanied by 95% confidence intervals. As the duration of fitting increases, the range of values shows that the certainty in a parameter estimation increases.

With just 10 days of data, the bootstrap distributions for r and p (Figure 6.7) are relatively narrow, indicating high certainty in parameter estimation during the early growth phase. The model achieves strong fit quality, reflected in low error metrics: SSE of 0.12859, RMSE of 0.1134, MAE of 0.0888, and MAPE of 4.73%

Extending the fitting window to 20 days (Figure 6.8) introduces slightly more variability in the parameter estimates. While the model still maintains decent performance (SSE: 1.5163, RMSE: 0.2754, MAE: 0.2269, MAPE: 8.90%), the increase in uncertainty reflects the model's attempt to adapt to transitioning dynamics as the curve moves beyond the initial exponential-like regime.

At 40 days (Figure 6.9), a notable spread in the distributions is observed, particularly in p, indicating the onset of saturation and a more complex growth structure. The model performance also degrades, with higher errors (SSE: 12.8541, RMSE: 0.5669, MAE: 0.4917, MAPE: 10.93%), suggesting reduced predictive precision.

Finally, fitting over 50 days (Figure 6.10) shows a continuation of this trend. Although the MAPE slightly improves to 9.60%, the overall error (SSE: 16.2782) remains high. The distributions continue to highlight estimation variability yet remain bounded, showing that bootstrap-based uncertainty quantification still provides meaningful insight.

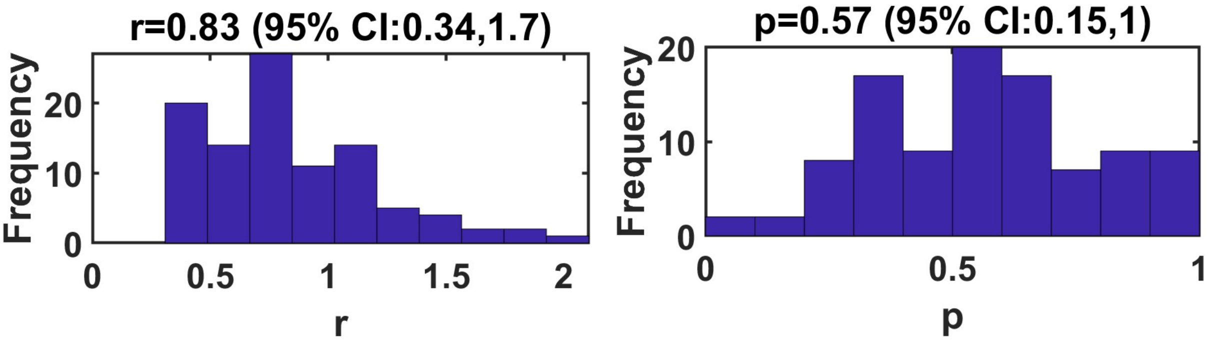

Figure 6.7: Parameter uncertainty (After fitting the model for 10 days): Bootstrap distributions for r and p with 95% confidence intervals. **Simulation_2.m**.

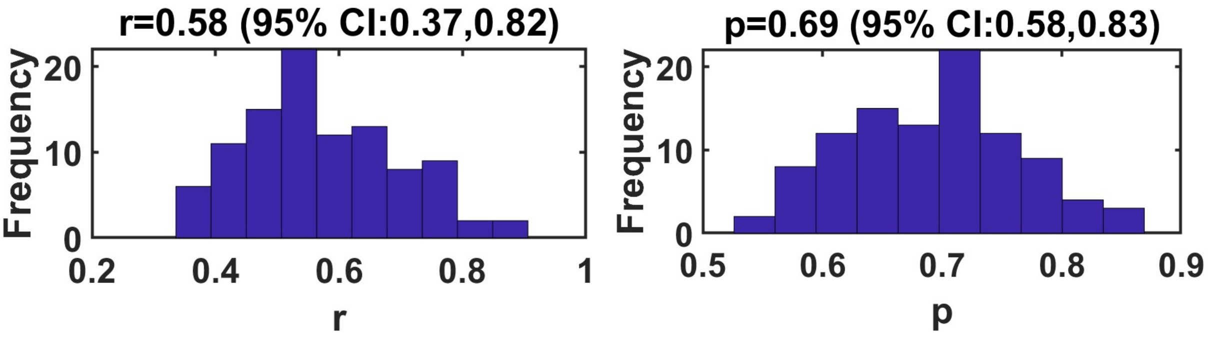

Figure 6.8: Parameter uncertainty (After fitting the model for 20 days): Bootstrap distributions for r and p with 95% confidence intervals. **Simulation_2.m**.

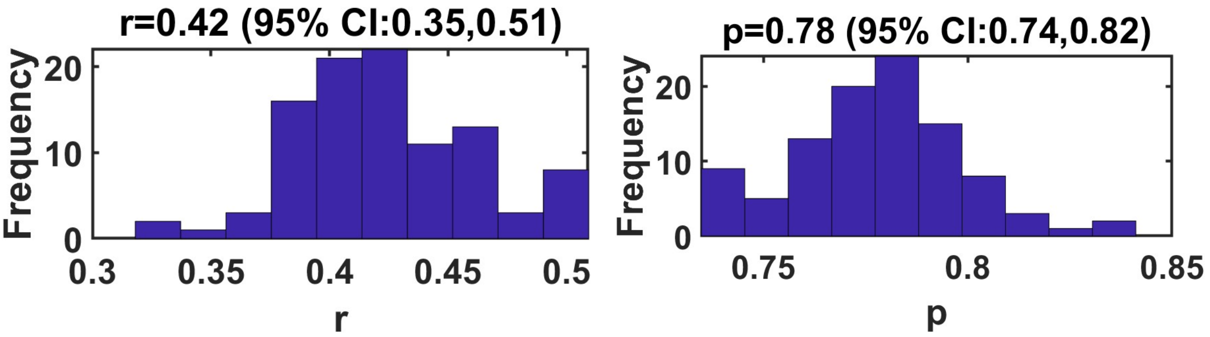

Figure 6.9: Parameter uncertainty (After fitting the model for 40 days): Bootstrap distributions for r and p with 95% confidence intervals. **Simulation_2.m**.

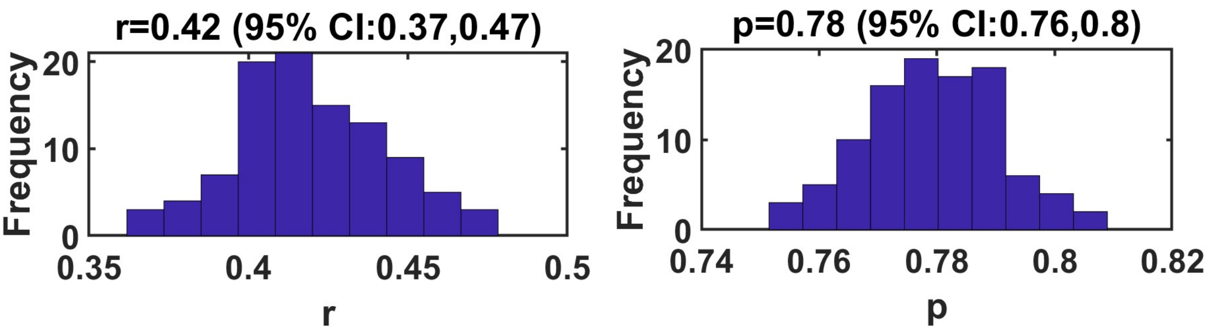

Figure 6.10: Parameter uncertainty (After fitting the model for 50 days): Bootstrap distributions for r and p with 95% confidence intervals. **Simulation_2.m**.

6.2.2 Observation

With only 10 days of data, the deceleration parameter p cannot be reliably estimated, as its confidence interval spans a wide range (0.5 to 1.0), making it difficult to distinguish between sub-exponential and exponential growth. As more early-phase data is included, the uncertainty in r and p reduces, and estimates become more stable and closer to the true values.

However, despite lower uncertainty, the goodness-of-fit metrics (SSE, RMSE, MAE, MAPE) worsen with longer fitting windows. This is because the GGM is designed for early growth phases, and its assumptions no longer align well as the epidemic evolves. Hence, while parameter estimates improve in confidence, the model fit deteriorates over time.

6.2.3 Parameter Uncertainty Estimation

Parameter uncertainty was assessed using the bootstrap method, generating multiple resampled datasets to estimate the variability in parameter values. The resulting frequency distributions highlight the confidence intervals and reflect how stable and reliable the estimates of r and p are across different realizations.

6.2.4 30-Day Forecasts on Different Growth Phases

We extend our analysis by performing 30-day forecasts after training the model on different durations of the growth phase. As shown in the visualizations, the precision of the forecasts improves with the length of the training data. This emphasizes the importance of sufficient early-phase data for reliable epidemic projections.

Visualization

Figure 6.11, Figure 6.12, Figure 6.13, and Figure 6.14 illustrate the GGM model's forecast after being fitted on 10, 20, 40, and 50 days of synthetic data, respectively. The shaded regions represent the forecast interval, which narrows as more data is used for fitting.

As evident, the forecast uncertainty reduces significantly as the fitting period increases, confirming that longer training phases yield more stable and accurate future projections.

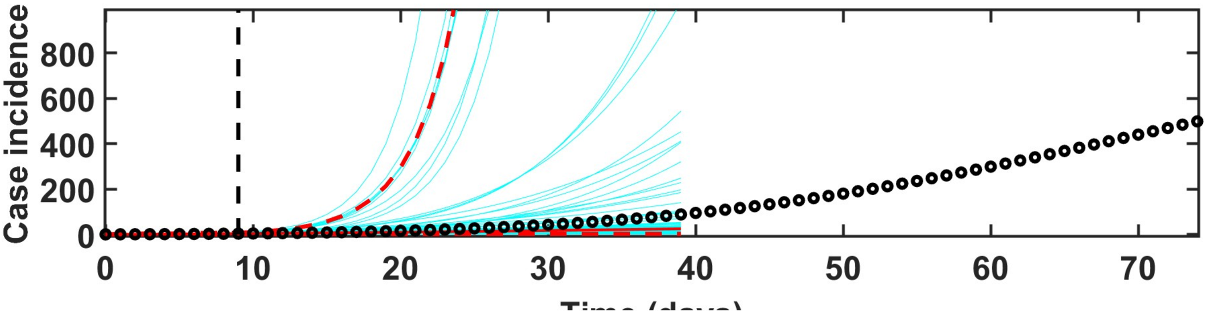

Figure 6.11: 30-day forecast after fitting the model for 10 days using the GGM on synthetic data. Observed data (black dots), fitted model (red dashed line), and bootstrap-based forecast (light blue). The vertical dashed line indicates the end of the fitting window. **Simulation_2.m**.

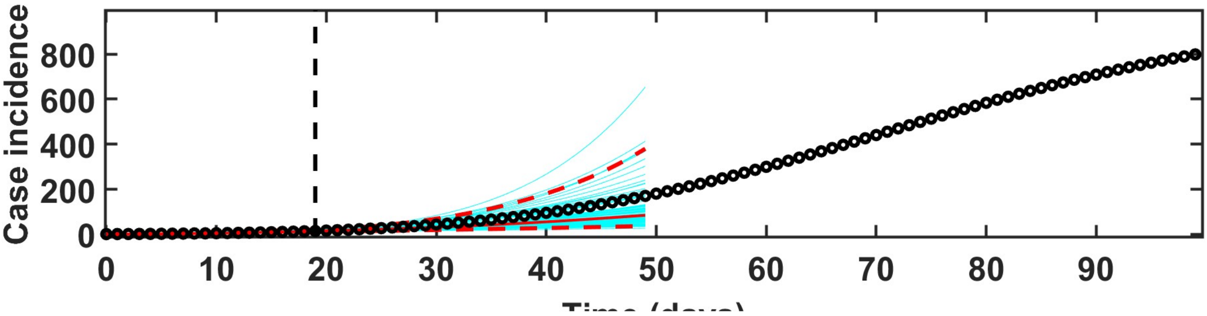

Figure 6.12: 30-day forecast after fitting the model for 10 days using the GGM on synthetic data. Observed data (black dots), fitted model (red dashed line), and bootstrap-based forecast (light blue). The vertical dashed line indicates the end of the fitting window. **Simulation_2.m**.

6.3 Simulation 3 - Long-Term Forecasting with GRM Model

In this simulation, we evaluate the long-term forecasting capability of the Generalized Richards Model (GRM) using real-world epidemic data. Specifically, we use the **zika-daily-onset-antioquia** dataaset to assess how the duration of the training period influences forecasting accuracy and parameter uncertainty. The model is fitted to varying lengths of the early growth phase and subjected to 100 bootstrap realizations to quantify variability in forecasts and parameter estimates.

As the training period increases, the uncertainty in forecast outcomes consistently decreases. This is expected, as more extensive data coverage provides a clearer view of the epidemic's growth pattern, leading to more stable and reliable parameter estimates.

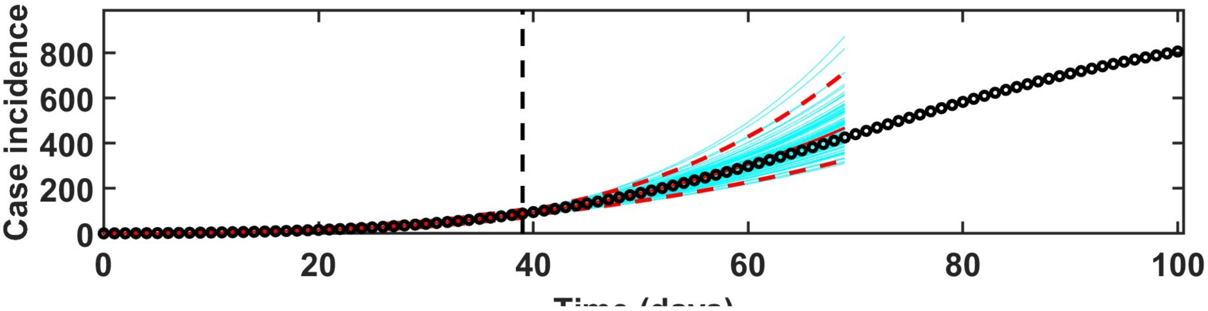

Figure 6.13: 30-day forecast after fitting the model for 10 days using the GGM on synthetic data. Observed data (black dots), fitted model (red dashed line), and bootstrap-based forecast (light blue). The vertical dashed line indicates the end of the fitting window. **Simulation_2.m**.

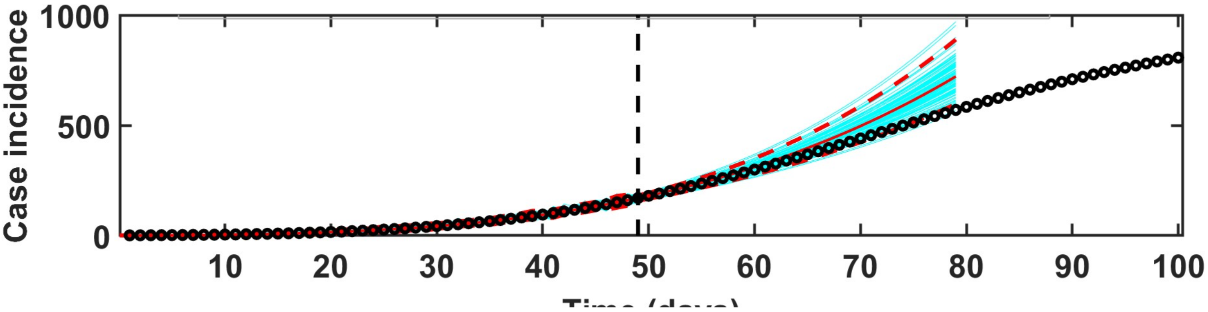

Figure 6.14: 30-day forecast after fitting the model for 10 days using the GGM on synthetic data. Observed data (black dots), fitted model (red dashed line), and bootstrap-based forecast (light blue). The vertical dashed line indicates the end of the fitting window. **Simulation_2.m**.

6.3.1 Bootstrap Realizations and Uncertainty Reduction

The bootstrap approach allows us to explore the distributions of parameter estimates and their influence on long-term forecasting. With shorter training durations, forecasts are highly variable and uncertain. In contrast, as the training window extends, the uncertainty reduces significantly, indicating stronger model confidence and parameter stability.

6.3.2 Visualization and Goodness-of-Fit Metrics

The following visualizations show the GRM model performance using the Zika dataset, with training durations of 10, 20, and 40 days. Each case includes the model forecast plot and bootstrap parameter distributions for r, p, a and K. The results demonstrate the impact of training length on prediction reliability and parameter estimation.

Training with 10 Days of Data

The GRM model fitted using 10 days of training data (Fig. 6.15) shows large forecast uncertainty, as seen in the wide confidence bands. The corresponding bootstrap distributions of parameters r, p, a and K (Fig. 6.16) highlight high parameter uncertainty.

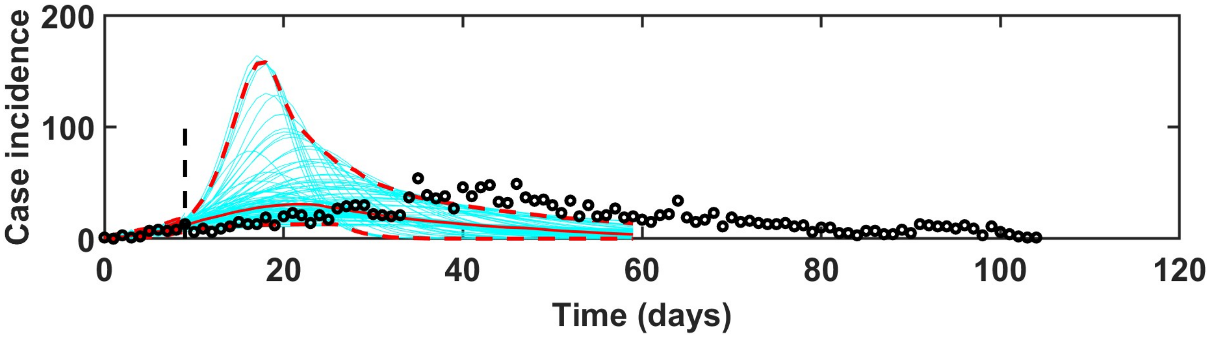

Figure 6.15: Fitting the model for 10 days using the GRM on Zika incidence data. Observed data (black dots), fitted model (red dashed line), and bootstrap-based forecast (light blue). The vertical dashed line indicates the end of the fitting window. **Simulation_3.m**.

Goodness-of-Fit (10 days): SSE = 20.7857, RMSE = 1.4417, MAE = 1.3263, MAPE = 40.94%

Training with 20 Days of Data

As shown in Fig. 6.17, the forecast uncertainty decreases compared to the 10-day case. The parameter distributions in Fig. 6.18 also become more concentrated, indicating improved parameter stability.

Goodness-of-Fit (20 days): SSE = 116.6549, RMSE = 2.4151, MAE = 1.9221, MAPE = 36.05%

Training with 40 Days of Data

In Fig. 6.19, the forecast is more constrained and closely follows the actual data. Bootstrap plots in Fig. 6.20 show even narrower distributions, indicating reduced parameter uncertainty.

Goodness-of-Fit (40 days): SSE = 1329.9875, RMSE = 5.7663, MAE = 4.1230, MAPE = 26.26%

6.3.3 Interpretation

The results (Fig. 6.15, 6.17, 6.19) confirm that longer training durations provide better forecasting performance, while parameter uncertainty (Fig. 6.16, 6.18, 6.20) is significantly reduced. Although forecasting error may not always monotonically decrease due to increased model complexity, the confidence in both the forecast and parameter estimates improves as more data becomes available.

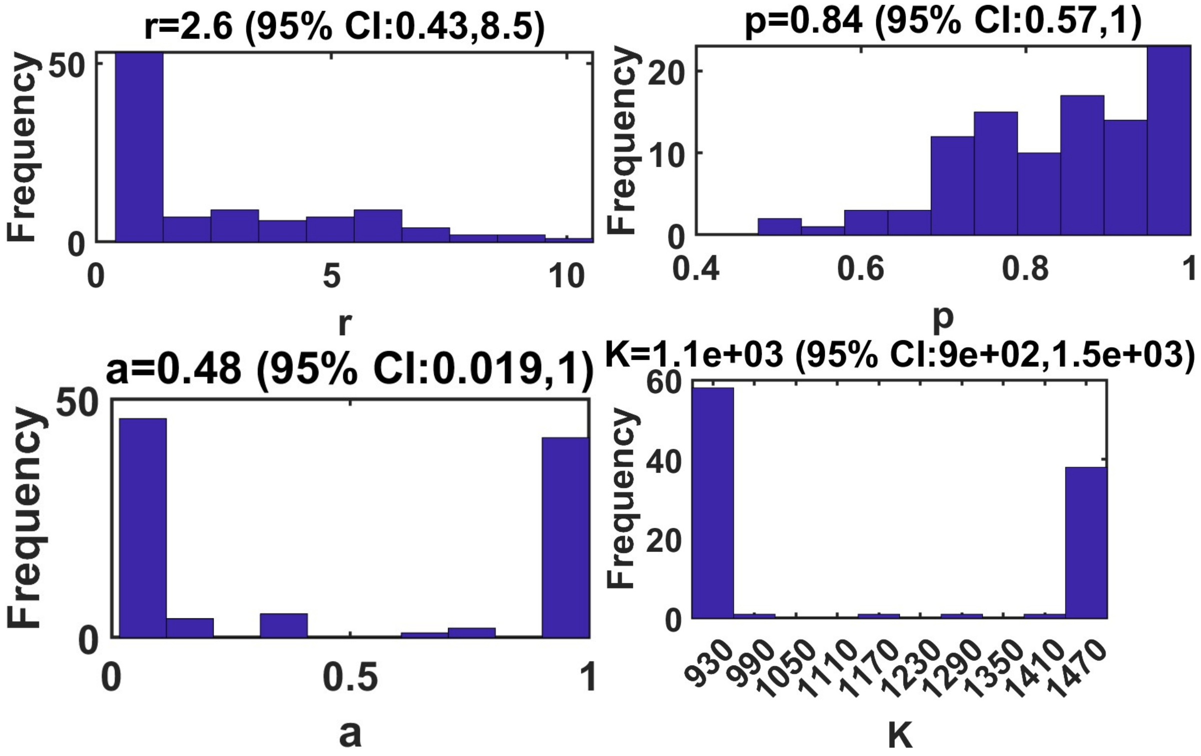

Figure 6.16: Parameter uncertainty (After fitting the model for 10 days): Bootstrap distributions for r, p, a and K with 95% confidence intervals. **Simulation_3.m**.

6.4 Simulation 4 - SEIR Model Parameter Uncertainty and Forecast Reliability

In this section, we evaluate the performance of the SEIR (susceptible- exposed- infectious-recovered) model in capturing the dynamics of the epidemic using real data. The model is calibrated over different training durations, and the estimated values of the basic reproduction number (R_0) are examined using bootstrap-based empirical distributions. This simulation aims to explore how the availability of data influences model fitting and parameter uncertainty.

6.4.1 SEIR Model Fitting and Bootstrap Confidence Intervals

The SEIR model is first fitted to the epidemic data using a 20-day training period. Figure 6.21 shows the predicted infected population over time, with red open circles representing actual data, the solid black line denoting the model's prediction, and the blue shaded region indicating the 95% confidence interval obtained through bootstrap resampling.

The model fits the data well, particularly during the mid-phase of the time window. The increasing width of the confidence interval towards later days reflects growing uncertainty in forecasts. Notably, the actual values at the end (Days 18–20) fall within the

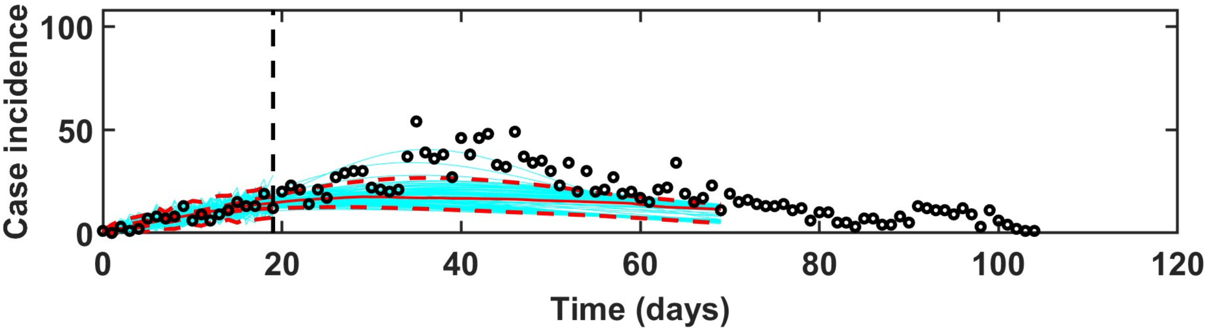

Figure 6.17: Fitting the model for 20 days using the GRM on Zika incidence data. Observed data (black dots), fitted model (red dashed line), and bootstrap-based forecast (light blue). The vertical dashed line indicates the end of the fitting window. **Simulation_3.m**.

confidence bounds, demonstrating the reliability of the fit.

6.4.2 Empirical Distribution of R_0 Estimates

To evaluate how training duration influences parameter estimation, we perform bootstrap-based estimation of the basic reproduction number (R_0) using training data for 16, 18, and 20 days. Figure 6.22 illustrates the empirical distributions of R_0 for each case.

The distributions exhibit a roughly symmetric, bell-shaped form in each case. As the training period increases, the distributions become taller and narrower, indicating reduced uncertainty and more stable estimation. The peak values of R_0 gradually increase with data availability—approximately 2.10 for 16 days, 2.14 for 18 days, and 2.18 for 20 days—suggesting slight refinement in the estimate with more information.

6.4.3 Key Insights

These results collectively indicate that longer training durations not only enhance the model's fit but also significantly reduce uncertainty in key parameter estimates. The narrowing of the R_0 distribution with more data highlights the benefit of extended calibration periods for stable and credible epidemic forecasting.

6.5 Simulation 5 - GGM Fitting and Empirical Distributions for the 1918 Influenza Pandemic in San Francisco

In this simulation, we apply the Generalized Growth Model (GGM) to the early phase of the 1918 Influenza Pandemic data from San Francisco. We aim to estimate and analyze key epidemiological parameters, including the growth rate (r), deceleration parameter

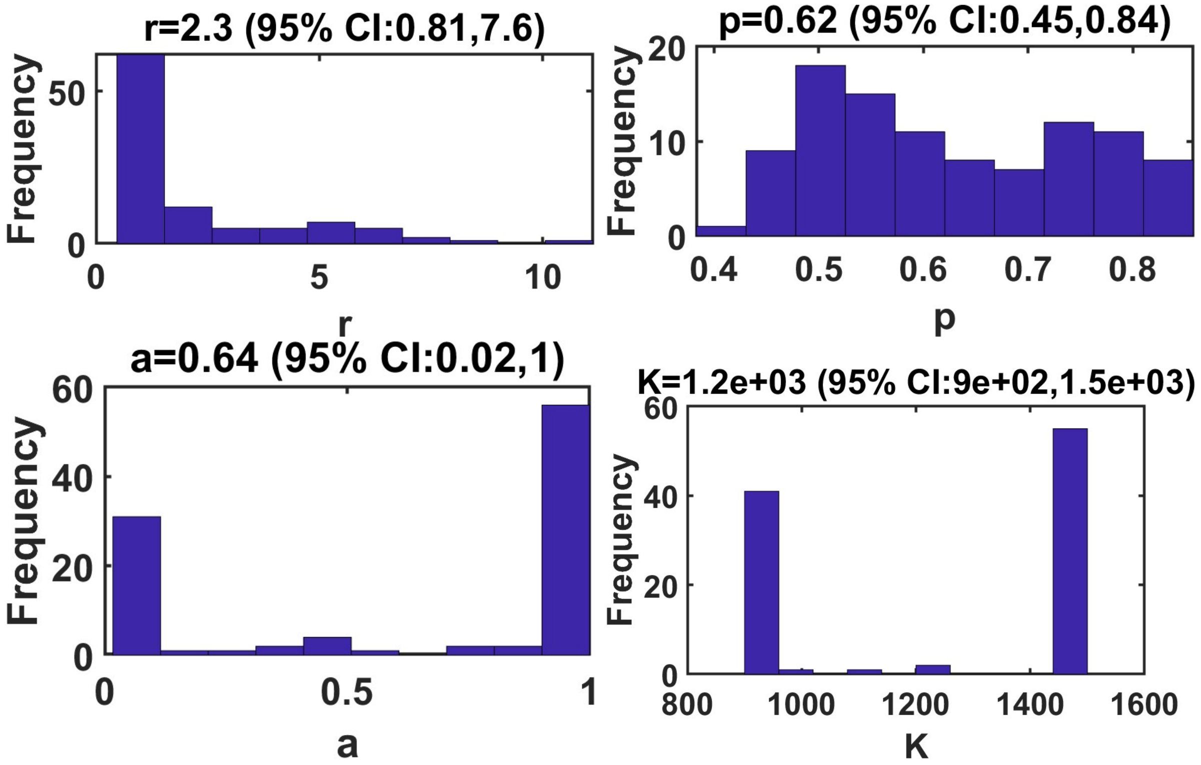

Figure 6.18: Parameter uncertainty (After fitting the model for 20 days): Bootstrap distributions for r, p, a and K with 95% confidence intervals. **Simulation_3.m**.

(p), and the time-varying effective reproduction number (R_t). We fitted the model for 20 days and performed 200 bootstrap realizations to quantify the uncertainty in these parameter estimates.

Figure 6.23 presents the empirical distributions of the growth rate (r), deceleration parameter (p), and the effective reproduction number (R_t). These histograms illustrate the variability and confidence associated with each parameter. The left panel displays the distribution of the growth rate r, which reflects the speed of the epidemic's early growth. The middle panel shows the distribution of the deceleration parameter p, capturing the deviation from exponential growth and accounting for sub-exponential dynamics often observed in real epidemics. The right panel depicts the distribution of the effective reproduction number R_t, which measures the average number of secondary infections generated by a single infectious individual over time. All three distributions appear approximately symmetric, indicating stable and well-converged bootstrap estimates.

The effective reproduction number R_t is estimated dynamically using the formula:

$$R_{tj} = \frac{I_{tj}}{\sum_{s=1}^{t_j} I_s \cdot w(t_j - s)},$$

where I_{tj} is the number of new infections at time t_j, I_s represents the number of infectious individuals at previous time points s, and $w(t_j - s)$ denotes the discretized generation interval distribution. We assume the generation interval follows an exponential

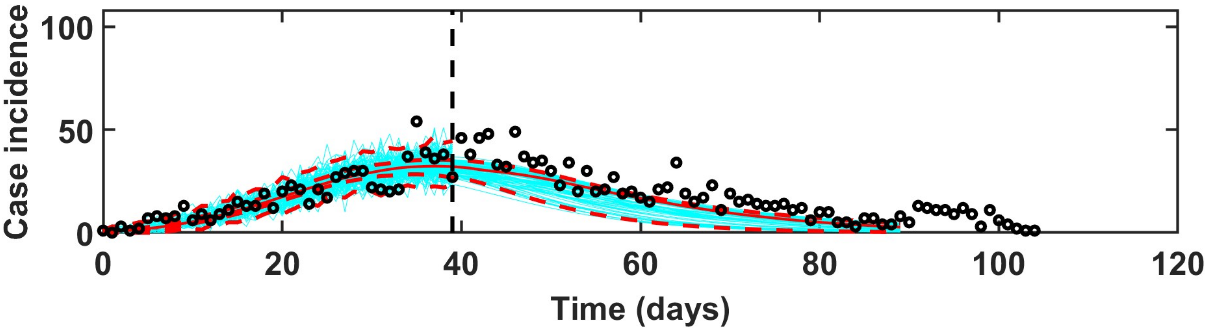

Figure 6.19: Fitting the model for 40 days using the GRM on Zika incidence data. Observed data (black dots), fitted model (red dashed line), and bootstrap-based forecast (light blue). The vertical dashed line indicates the end of the fitting window. **Simulation_3.m**.

distribution with a mean of 4 days, corresponding to a rate parameter $\lambda = 0.25$. The discretized generation interval is computed as:

$$w(s) = \Pr(s - 1 < X < s) = e^{-\lambda(s-1)} - e^{-\lambda s}, \quad s = 1, 2, \ldots, 20.$$

This formulation accounts for the contribution of prior infections to current incidence while incorporating temporal delays in transmission, thereby allowing for a realistic and time-sensitive estimation of R_t.

Figure 6.24 shows the GGM model fit to the early outbreak data. The red circles denote the actual case data, the solid black line represents the model's predicted number of cases, and the blue shaded region indicates the 95% bootstrap confidence interval. The model closely tracks the observed cases, particularly during the mid-range of the epidemic curve, and maintains consistency within the confidence bounds toward the end of the observed period.

To assess model accuracy, we compute several goodness-of-fit metrics. The Sum of Squared Errors (SSE) is 1531.1311, the Root Mean Squared Error (RMSE) is 8.7497, the Mean Absolute Error (MAE) is 6.0775, and the Mean Absolute Percentage Error (MAPE) is 30.4587%. These values indicate a reasonable level of model performance, given the stochastic nature of the epidemic data.

From the bootstrap analysis, we find the deceleration parameter p to be approximately 0.95, with a 95% confidence interval of (0.95, 1.0). This suggests that the early trajectory of the epidemic in San Francisco was nearly exponential. Overall, the GGM model offers a reliable approximation of early outbreak dynamics, and the bootstrap-based analysis provides robust measures of uncertainty for key epidemiological parameters.

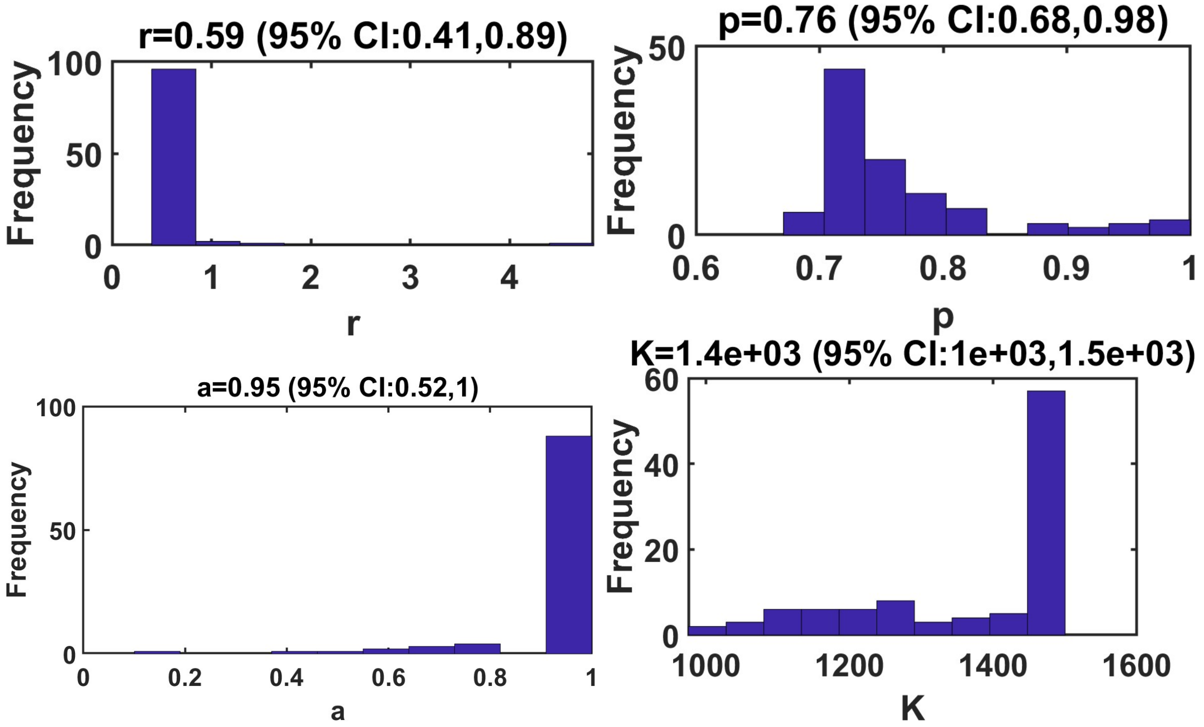

Figure 6.20: Parameter uncertainty (After fitting the model for 40 days): Bootstrap distributions for r, p, a and K with 95% confidence intervals. **Simulation_3.m**.

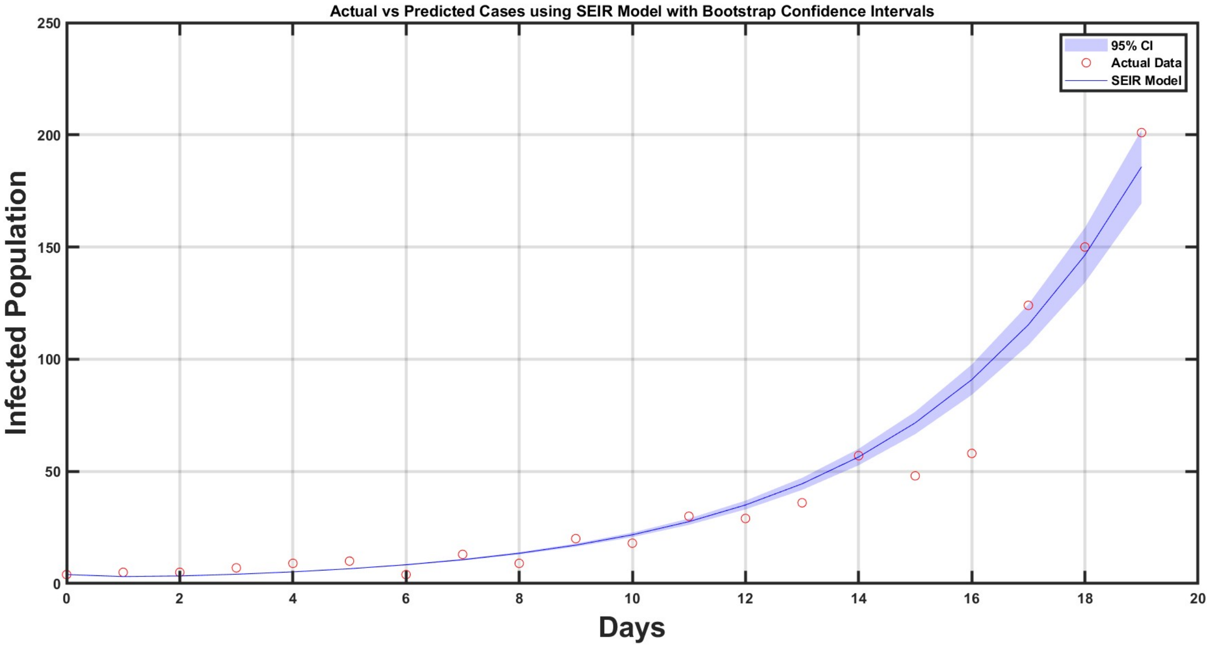

Figure 6.21: Actual vs predicted infected cases using the SEIR model. The blue shaded area represents the 95% confidence interval from bootstrap realizations. **Simulation_4.m**.

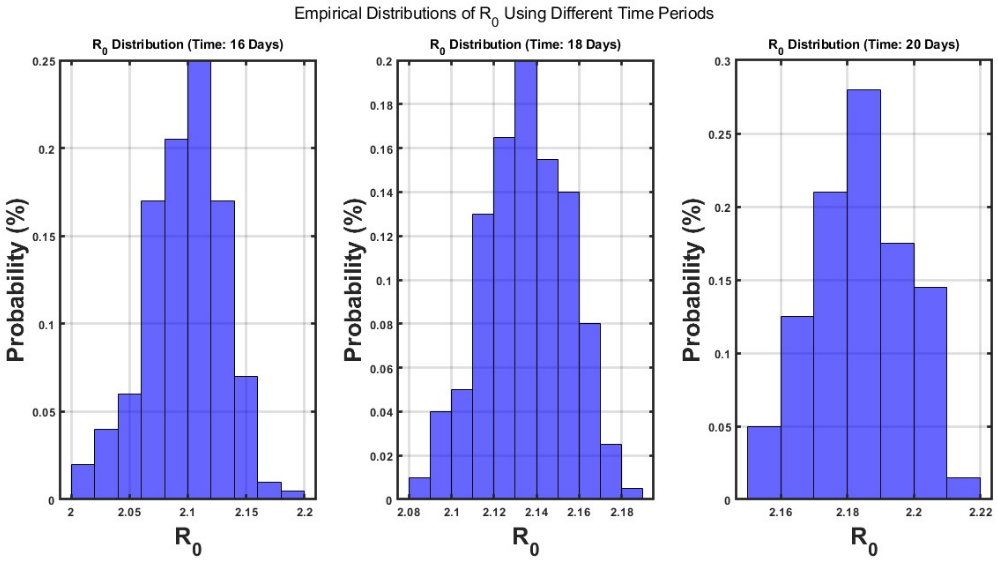

Figure 6.22: Empirical distributions of R_0 estimated from bootstrap realizations using 16, 18, and 20 days of training data. **Simulation_4.m**.

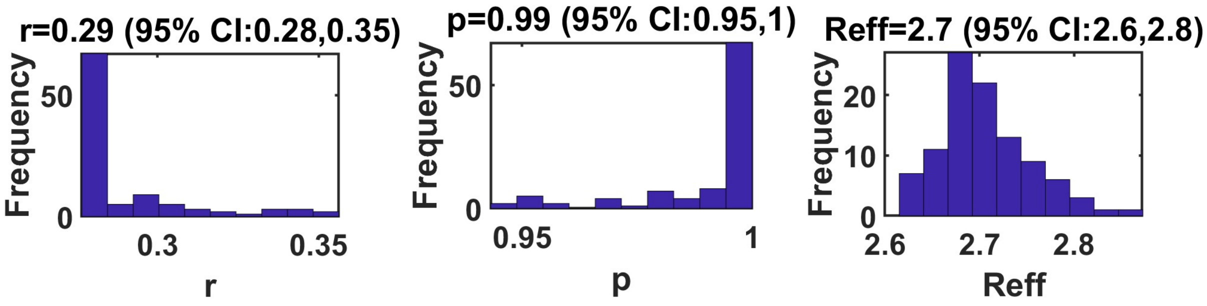

Figure 6.23: Empirical distributions of key model parameters based on 200 bootstrap realizations. **Simulation_5.m**.

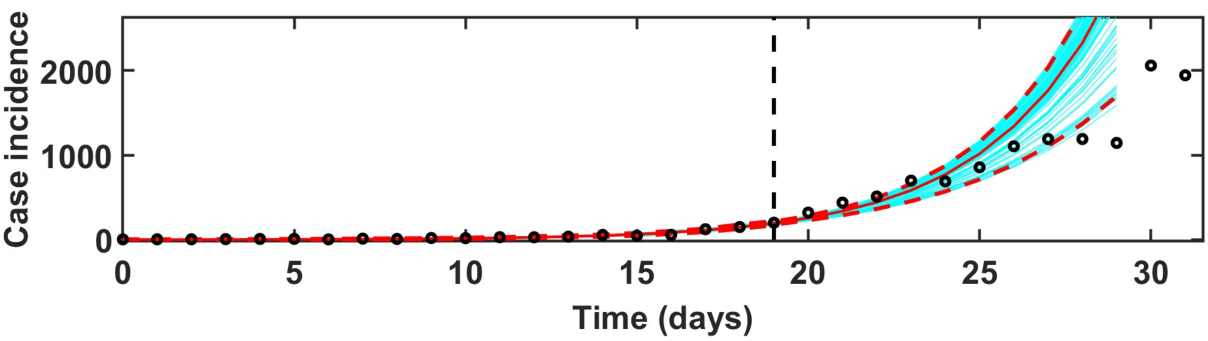

Figure 6.24: GGM model fit to the 1918 Influenza data in San Francisco. Red circles represent actual data, the black line is the model prediction, and the blue shaded region shows the 95% confidence interval from bootstrapping. **Simulation_5.m**.

Challenge and Problem Solved

7.1 Computational Challenges

The computational challenges faced during the study were significant. The execution time for bootstrapping and parameter estimation increased as the dataset grew, especially during long-term forecasting with multiple realizations. Optimization techniques required careful tuning to ensure convergence, particularly in models that depended on iterative procedures for parameter estimation. High memory usage was observed when performing simulations with a large number of bootstrap realizations, demanding efficient memory management strategies.

7.2 Data Limitations

Data limitations also posed obstacles. Real-world data availability was occasionally restricted, necessitating the use of synthetic data for validation, which, while useful, may not perfectly replicate real-world epidemic conditions. Noise in the data impacted the accuracy of parameter estimation, highlighting the need for careful preprocessing and smoothing techniques. Furthermore, uncertainties in fitting models to past outbreaks, such as the 1918 influenza pandemic, arose from the assumptions made about initial conditions and historical datasets.

7.3 Model Assumptions and Limitations

The assumptions underlying the Generalized Growth Model (GGM) and Generalized Reproduction Model (GRM) also posed limitations. For instance, the assumption of constant growth parameters may not fully capture the complexities of real epidemic dynamics. The SEIR model's reliance on assumptions about latent and infectious periods, which can vary between diseases, further limited its application. Additionally, sensitivity to initial parameter selection introduced variations in forecast accuracy, emphasizing the need for robust initialization techniques.

7.4 Uncertainty in Forecasting

Uncertainty in forecasting was a key challenge, with forecast accuracy heavily dependent on the duration of the training data. Shorter training periods resulted in higher uncertainty in predictions, and confidence intervals around the forecasts were sometimes wide, reflecting the inherent difficulty in long-term predictions. The impact of real-world interventions, such as social distancing and vaccination, was not always explicitly included in certain models, which could have led to discrepancies between forecasts and observed epidemic trends.

7.5 Generalization to Different Epidemics

The generalization of the models to different epidemics presented its own set of challenges. The transferability of the models between epidemics required adjustments in parameter assumptions, limiting their direct applicability. Differences in population dynamics, reporting accuracy, and healthcare interventions further complicated the cross-epidemic application of the models.

Despite these challenges, the study provides valuable insights into epidemic modeling, demonstrating both the strengths and limitations of different approaches in forecasting and parameter estimation. Addressing these issues in future work could enhance the reliability and applicability of epidemiological models.

Use of AI Tools and ChatGPT

AI tools, including GPT, played a crucial role in various aspects of this project, significantly enhancing efficiency and streamlining tasks. Key areas where GPT contributed include code generation and debugging, research and literature review, content structuring, and data analysis.

GPT assisted in writing and optimizing code for data preprocessing, model fitting, and visualization. It also helped identify and debug syntax errors, suggest corrections, and improve computational efficiency. Additionally, GPT supported the implementation of mathematical formulations for the epidemiological models used in the study, such as the Generalized Growth Model (GGM), Generalized Reproduction Model (GRM), and SEIR.

When conducting the literature review, AI tools helped gather relevant research papers, summarize key concepts, and compare different epidemiological models. GPT also provided valuable insights into their strengths and weaknesses, aiding the understanding of existing methodologies and approaches. In structuring the report, GPT made recommendations to ensure clarity and coherence between sections. It helped summarize findings, articulate results, and ensure a logical flow throughout the document, refining explanations of complex concepts for better readability.

In terms of data analysis, GPT was useful for interpreting statistical results, providing insights into metrics like RMSE, MAPE, and confidence intervals. It also suggested effective ways to visualize model performance and uncertainty, improving the presentation of results.

However, despite the significant contributions of AI tools, several challenges were encountered. GPT occasionally misinterpreted the problem statement, leading to suggestions that were not fully aligned with the project's needs. This required verification and careful adjustments to ensure relevance. In coding, although GPT provided useful assistance, some generated code snippets contained inefficiencies or errors, requiring additional manual intervention to optimize performance.

Another challenge was GPT's tendency to provide generic explanations that lacked domain-specific insights, especially when dealing with complex aspects of epidemiological modeling. This was particularly evident in areas such as parameter estimation, where the AI struggled with more nuanced modeling techniques. Moreover, there was a risk of over-relying on AI suggestions, which could lead to inaccuracies if not critically evaluated. Ensuring academic rigor required cross-referencing AI-generated insights with established literature.

Additionally, GPT's capabilities in handling large datasets were limited, necessitating the use of specialized computational tools for complex simulations. Suggestions for memory optimization also needed refinement to be fully effective.

Despite these challenges, the integration of AI tools, including GPT, greatly improved the efficiency of the project. The assistance provided across various stages—from coding and content structuring to research and data interpretation—helped streamline the process. Proper validation and manual review of AI-generated content helped mitigate potential issues, ultimately enhancing the overall quality of the work.

Bootstrap Implementation and Design

9.1 Introduction to Bootstrapping

Bootstrapping is a resampling technique used to estimate the distribution of a statistic by repeatedly sampling with replacement from the observed data. This method allows for robust uncertainty estimation and confidence interval calculation, making it particularly valuable in statistical modeling and forecasting, as illustrated in **Figure-9.1**. The key advantage of bootstrapping lies in its ability to provide inference without relying on strong parametric assumptions, making it a flexible tool for a variety of data scenarios.

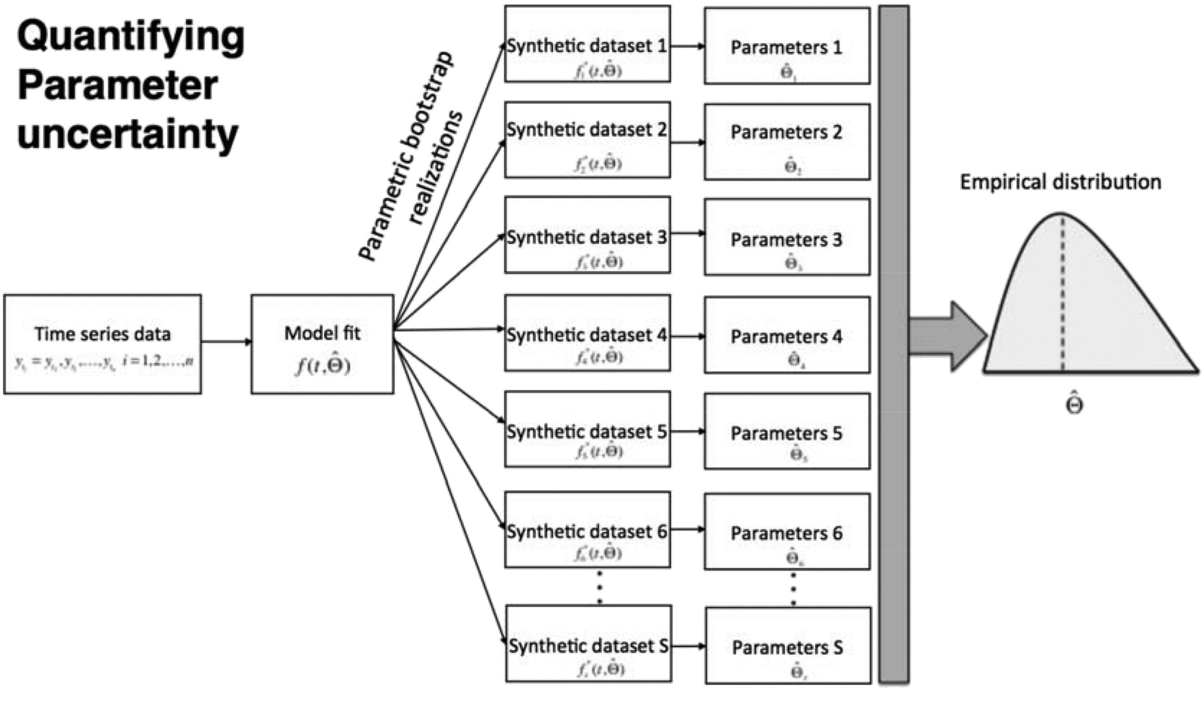

Figure 9.1: A graphical representation of the bootstrapping process. The figure shows the resampling of the original dataset to generate multiple bootstrap samples, each of which is used to estimate model parameters and quantify uncertainty. This iterative process allows for the construction of confidence intervals and the generation of multiple realizations, aiding in the robustness and reliability of the model.

9.2 Application of Bootstrapping in This Study

In this study, bootstrapping was employed to estimate model parameters and quantify uncertainty in forecasting epidemic growth. The primary applications of bootstrapping included parameter estimation, uncertainty quantification, and evaluating long-term forecasting stability.

Bootstrapping was used to generate multiple realizations of parameter values by resampling the original dataset. This process enabled the assessment of the variability in key epidemic parameters, such as the growth rate, the deceleration of growth, and the reproduction number. By performing multiple bootstrap simulations, we were able to obtain confidence intervals for these estimated parameters, providing insights into the reliability and robustness of the fitted models. Additionally, the stability of long-term forecasting was evaluated by using bootstrap realizations to assess the impact of the training duration on forecast uncertainty. The results indicated that longer training periods helped reduce forecasting uncertainty, highlighting the importance of adequate data coverage for improved forecasting accuracy.

9.3 Why Bootstrapping Was Used

Bootstrapping was selected for this study due to its unique advantages. It effectively handles small sample sizes by generating multiple synthetic samples, which is particularly useful when data is limited. The method does not require strict distributional assumptions, making it well-suited for complex epidemic data, which often exhibits non-standard distributions. Furthermore, bootstrapping provides confidence intervals that better reflect real-world variability and parameter uncertainty. Finally, by averaging over multiple realizations, bootstrapping enhances model robustness, reducing the impact of noise and outliers on the final estimates.

Conclusion

This study explored various epidemiological models, including the Generalized Growth Model (GGM), Generalized Richards Model (GRM), and the SEIR model, to analyze disease spread and forecasting. Through extensive simulations and parameter estimation techniques, several key insights were gained into the strengths and limitations of these models in epidemic prediction.

One of the primary observations was the impact of training duration on forecasting accuracy. Longer training periods led to reduced uncertainty and more precise predictions, particularly in the case of GRM-based long-term forecasting. Additionally, the use of bootstrapping helped quantify parameter uncertainty, reinforcing the importance of robust statistical techniques in epidemiological modeling.

Challenges such as computational constraints, data limitations, and sensitivity to model assumptions highlighted the need for adaptive modeling approaches. The variability in real-world epidemic conditions further emphasized that no single model can be universally relied upon for disease forecasting. Instead, hybrid methodologies and ensemble modeling approaches may offer more comprehensive solutions.

The study also demonstrated the role of AI-assisted tools in content structuring and research facilitation. However, AI tools had limitations in handling domain-specific complexities, requiring manual validation and refinement to ensure accuracy.

Future research should focus on improving model adaptability by integrating real-time data streams and refining uncertainty quantification techniques. Incorporating intervention strategies, such as vaccination and public health measures, into forecasting models could further enhance predictive reliability. Despite existing challenges, this research contributes valuable insights into the evolving field of epidemic modeling and its applications in public health decision-making.

Implementation Details

The forecasting model was implemented in MATLAB, utilizing built-in statistical and computational functions for efficient data analysis and prediction.

References

GitHub Repository: https://github.com/ygyashgoyal

Paper related to this study: https://www.sciencedirect.com/science/article/pii/S2468042717300234?via%3Dihub

MatLab product: https://www.mathworks.com/products/matlab.html