Imperial College London

Department of Computing

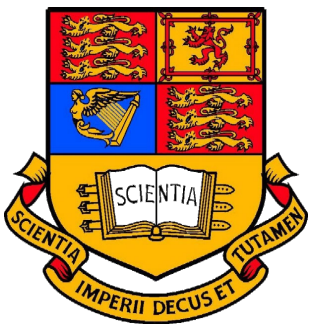
MEng Individual Project

Modelling and Prediction of Multiple Epidemic Phenomena

Author:
Thomas Wilding

Supervisors:
Dr. William Knottenbelt
Marily Nika

June 21, 2014



Abstract

Epidemiology is the study of the causes and transmission of infectious diseases. Since the 18th century, epidemiology has relied on mathematical models to explain and predict epidemic outbreaks. Today, epidemic modelling is a well established field, encompassing a large range of approaches and models. One of the most important uses of epidemic models is predicting the future development of an outbreak, as the predictions are used to inform critical outbreak control decisions.

Alongside technological advancements, epidemic modelling has expanded into unconventional applications and complex outbreak phenomenon have emerged. It has been observed that a number of determinants give rise to multiple peaks of epidemic activity. An interesting hypothesis is that the underlying phenomena manifest as sub-epidemics and the overall outbreak behaviour is explained as the superposition of sub epidemic parts. This concept is inspired by Fourier analysis, with basis functions replaced by sub epidemic models.

Existing epidemic modelling approaches are becoming increasingly inadequate due to the rise of multiple epidemic phenomenon. Current approaches are incapable if explaining the observed interactions due to their inherent inability to capture latent outbreaks. This report addresses the exigency for a dynamic epidemic model which is capable of explaining and predicting multiple concurrent epidemic outbreaks.

The main contribution of this report is the development of a multiple epidemic modelling approach, entitled Synthedemic modelling, to elucidate the observed outbreak behaviour. The Synthedemic model is formulated as the superposition of single sub epidemic models. Outbreaks are incorporated dynamically as they emerge, and redundant outbreaks are removed to ensure parsimony. To enhance the future predictions of the Synthedemic model a residual refinement stage is employed using autoregressive techniques. The development of the Synthedemic model has been undertaken within a multiple epidemic framework that facilitates the fitting, prediction, analysis and output of the Synthedemic model.

The Synthedemic model has been applied to a range of epidemic outbreak data from various domains. Synthetically generated, Influenza, and online music downloads datasets have been analysed, amounting to a total of over one thousand Synthedemic fittings. A high standard of fitting ability is attained over all datasets, showing the robustness of the fitting procedure and integrity of the multiple epidemic concept. The fitting and predictive ability of the Synthedemic model have been benchmarked against a number of existing epidemic modelling techniques. The most significant result is that, without exception, the Synthedemic model surpasses the archetype single epidemic model in its ability to explain and predict epidemic outbreaks.

Acknowledgements

I would like to thank my supervisor Dr. William Knottenbelt and my co-supervisor Marily Nika for their guidance. Also thanks to Dieter Fiems and Koen de Turck for their contribution to the methodology.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> *"In 1345, at one hour after noon on 20 March, there was a major conjunction of three planets in Aquarius. This conjunction, along with other earlier conjunctions and eclipses, by causing deadly corruption of the air around us, signifies mortality and famine"*
>
> Bubonic plague explanation from the University of Paris to King Phillip VI [1]

The understanding of the spread of infectious disease has developed very gradually over time. Dating as far back as Hippocrates in 370 BC, disease was first related to human and environmental factors [2]. However during the outbreak of the bubonic plague in the Middle Ages the causes of disease reverted to superstitious and celestial explanations [1]. One of the most significant advancements in the understanding of the spread of infectious disease arose from the use of epidemic modelling. Since Bernoulli formulated a mathematical model of the spread of Smallpox in 1766 [3], Epidemiology has relied on epidemic modelling to quantitatively describe epidemic dynamics. As a result the understanding of the spread of infectious diseases, and epidemic models, have shown significant advancements. Furthermore the ability of epidemic models to predict the future development of epidemic outbreaks has enabled critical decisions for founding disease control strategies. Epidemic model predictions potentially have global impacts on public health, economy, human behaviours and ultimately saves the lives of many.

More recently epidemic models have propagated into new applications, for example they have been applied within technological applications to predict outbreaks of computer viruses [4] or in online social networks [5] to predict social media outbreaks. With the expansion of epidemic modelling into new applications, it has been increasingly observed that unconventional outbreak phenomenon arise. This is often characterised by rise and fall patterns of epidemic activity. This project aims to provide a new approach to modelling the observed outbreak behaviours.

1.1 Motivation

With the expansion of epidemic modelling into new applications, complex outbreak behaviours have been observed. However, current epidemic modelling approaches show inability to capture epidemic outbreaks that can emerge from multiple unpredictable underlying phenomenon. The increasing use of epidemic models in new applications and the current insufficiency of epidemic modelling to explain the observed behaviour, show the clear need for the development of a multiple epidemic model.

1.2 Synthedemic Overview

The main contribution of this project is the development of an approach to multiple epidemic modelling, entitled Synthedemic modelling. The Synthedemic model is formulated as the superposition of a number of single sub epidemic models. Each sub epidemic is used to represent an underlying driving mechanism which contributes to the overall epidemic phenomenon. The sub epidemics are optimised simultanously, potentially influencing eachother, and their combined contributions constitute the overall outbreak patterns.

Given a dataset representing an outbreak with multiple underlying epidemics, the Synthedemic model decomposes it into sub epidemic parts by detecting the sub epidemic outbreaks and selecting the most suitable single epidemic model. The sub epidemic models are then synthesised to formulate the Synthedemic model, enabling the evaluation of the model and generation of future predictions. Residual refinement is then undertaken to enhance the Synthedemic model predictions.

1.3 Main Objectives

This project aims to provide solutions to challenges that have arisen within the latest applications of epidemic models. The main contribution is the Synthedemic modelling approach and the following sections details the main aspects of its implementation.

On-the-fly Epidemic Fitting An important area of current research within epidemic modelling is *on-the-fly* fitting of an epidemic. This involves parameter fitting to a single trace, in real-time as an epidemic unfolds. As the latest information becomes available, the model is adjusted to enable up-to-date predictions of the future evolution of the epidemic. Real-time epidemic models can be applied to systems with fast evolution times to enable the epidemic to be modelled as it unfolds.

Multiple Epidemics The realisation of a multiple epidemic model poses many challenges. Challenges investigated are:

- Decomposition of multiple epidemic outbreaks into sub epidemics
- Finding the start times for each sub epidemic.
- Selecting the underlying type of the sub epidemic models.
- Determining the most stable way to evaluate and optimise the model.
- Enhancing the predictions of the model using residual refinement techniques.

The simultaneous fitting of multiple epidemics may require exploration of techniques such as residuals analysis, basis functions and extended optimisation procedures. Parallel computing implementations and search heuristics may also be required to speed up the multiple epidemic optimisation process.

Detecting Epidemic Outbreaks A significant problem within multiple epidemic modelling is detecting the start times of the sub epidemic outbreaks. This is non trivial as the underlying mechanisms are unpredictable and may derive from different underlying fundamental types of outbreak. Characterising the start time and type of epidemic outbreaks is a fundamental requirement to enable the successful optimisation of a multiple epidemic model.

Residual Refinement Time series techniques such as autoregressive modelling can be used to refine the residuals of the multiple epidemic model, this may enhance the future predictions of the model.

Developing an Epidemic Framework The development of the Synthedemic model is undertaken within an epidemic framework that facilitates the fitting, prediction, analysis and output of the model.

1.4 Report Structure

Existing epidemic modelling techniques are researched within chapters 2.1 to 2.6 and recent interesting applications of such models are detailed in section 2.7.

Initial modelling techniques and implementations are detailed throughout section 3 starting with design considerations in section 3.1. The majority of techniques presented in this section are existing approaches within epidemic modelling that are required for the development of the Synthedemic model.

Chapter 4 presents the main contributions of this project, the details of the Synthedemic model development are broken up into Model Composition in section 4.1, Model Selection in section 4.3 and Residual Refinement in section 4.4. The Synthedemic fitting procedure is detailed in 4.5.1 and the Synthedemic results are presented in section 4.5.2

Evaluations of the model are undertaken using both simulated and real epidemic data sets within section 5 which is concluded with a discussion of the limitations and future developments for the project.

Chapter 2

Background

"Various methods can be used to carry out epidemiological investigations: surveillance and descriptive studies can be used to study distribution; analytical studies are used to study determinants" [6]

2.1 Epidemic Modelling

There are many different approaches and techniques currently employed within epidemic modelling. The models are based upon well established mathematical concepts and as with all models, assumptions are made within the model to reflect the underlying system. Epidemic model assumptions are taken according to the underlying characteristics of the epidemic being studied. As such, each application will have different underlying properties and give rise to a different model. The following section provides an overview of some of the models currently used, partitioned according to fundamental characteristics of epidemic models.

2.1.1 Deterministic and Stochastic models

Epidemic models can be divided onto two broad categories depending on their deterministic or stochastic nature. Deterministic models are probabilistic but with reproducible results. Stochastic models include a random aspect within the interactions between individuals, causing the results of the model to vary. In an extensive report on stochastic methods by *Anderson* and *Britton* [7] they note that although stochastic models are generally more accurate (as they more naturally represent uncertain parameters in the way that epidemics spread), they are more complex than deterministic

alternatives which can be used when the population size is large or as "introductory models when studying new phenomena".

The following section details some existing deterministic and stochastic models, however, in the remainder of this paper the main focus is on the use of deterministic models.

2.1.2 Continuous and Discrete models

Another fundamental attribute of epidemic models is the use of continuous or discrete time. The majority of traditional epidemic model approaches are continuous time models defined by differential transition equations. Discrete epidemic models have also been developed using difference equations to determine the transition dynamics between the discrete time intervals. Although discrete models are more complex, they may model the data recorded during an outbreak more naturally because it is sampled over discrete time [8, 9].

2.1.3 Compartment models

The majority of epidemic models consider a segregation of the population. Compartmental models divide the population into separate compartments according to the characteristics of the infection. The compartments often reflects the stages that an individual may progress through during an infection, and an individual can only belong to one compartment at any given time. An individual is initially Susceptible to an infection, on contracting the infection, they may proceed through a state of latency before becoming Infectious and after a period of being Infectious they may transition to an immune or Removed/Recovered state [10].

The Susceptible, Infectious, Recovered (*SIR*) model is one of the most frequently used compartmental models, but there are many more compartmental models with different compartments and complexity. Alternative approaches, such as Individual Event History (IEH) models, relax compartmental and diffusion homogeneity assumptions by allowing each individual to be unique [11].

The dynamics of compartmental systems are defined by specifying how individuals transfer between the compartments. During an epidemic, the transition of individuals between different compartments is modelled using transition rates and the number of individuals within each compartment changes as individuals transfer between the compartments over time. The relationships between the number of Susceptible, Infectious and Recovered individuals throughout an epidemic can be visualised as shown in Figure 2.1.

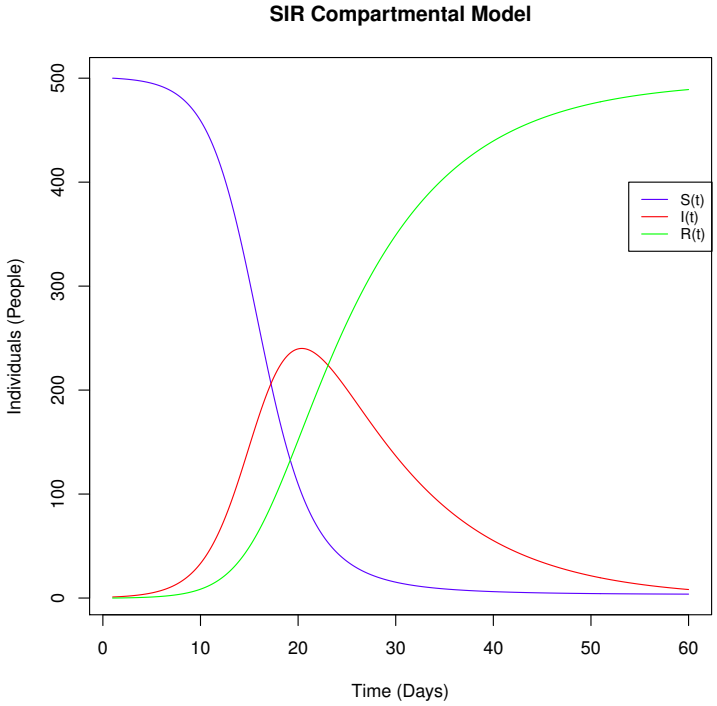


Figure 2.1: Relationship between the number of Susceptible $S(t)$, Infectious $I(t)$ and Recovered $R(t)$ individuals throughout an epidemic governed by an SIR model

2.2 Deterministic Models

2.2.1 Early Homogeneous Models

Epidemic model studies from the beginning of the twentieth century by Hammer and Ross postulated relations between compartments based on principles from previously established chemical reaction models. In particular they adopted the "Law of Mass Action" that states that the rate of a reaction is proportional to the product of the masses of its components. They assumed that individuals mix homogeneously. This enabled them to apply the "Law of Mass Action" into a compartmental approach whereby the number of Infectious individuals at time $t+1$ was proportional to the product of the number of Infectious and Susceptible individuals at time t [12].

2.2.2 The Kermack-McKendrick model

One of the most fundamental results in epidemic modelling was provided by Kermack and McKendrck in their papers (I [13], II [14] and III [15] published between 1927-1932) entitled "Contributions to the Mathematical Theory of Epidemics". They proposed key assumptions and a

fundamental approach to mathematical epidemic modelling that is still used in epidemic modelling.

Kermack and McKendrick defined the fundamental dynamics of compartmental models in which the flow between compartments is dependent on time. In particular they laid out some key assumptions and differential equations governing the rates of flow between different compartments and defined the archetype *SIR* epidemic model. Within their papers, they present a range of different models based on different assumptions, however one of the most significant models arises based on the assumptions that the population size is considered as constant (without birth and death rates) throughout the epidemic and the rate that individuals transfer between compartments is constant [13, (p.713, eq.29)]. These assumptions give rise to the following *SIR* model [16].

2.2.3 The Kermack-McKendrick SIR model

Considering an infection between three compartments Susceptible, Infected and Recovered, then Let:

- $S(t)$, $I(t)$, $R(t)$ represent the number of individuals in the Susceptible, Infectious and Recovered compartments at time t, respectively. And N be the population size $N = S(t)+I(t)+R(t)$
- β be the contact or infection rate, this represents the rate at which Infected individuals come into contact with other individuals.
- βN represents the number of other individuals that an Infectious individual comes into contact with. Of the βN individuals contacted, the fraction $\frac{S}{N}$ represents the proportion of Susceptible individuals. Therefore the rate that an Infectious individual contacts a Susceptible is $\beta N \times \frac{S}{N} = \beta S$ and so the transition rate at which Susceptible individuals transfer to the Infectious compartment is βIS.
- γ represents the rate at which individuals transfer from the Infected to the Recovered compartment, often called the recovery rate.

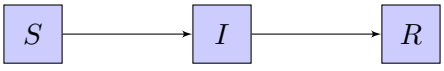


Figure 2.2: SIR Compartmental Diagram

Figure 2.2 shows the compartment and transitions of the SIR model. The system of ODEs (2.1) governs the population transition dynamics.

$$\frac{dS}{dt} = -\beta IS$$

$$\frac{dI}{dt} = \beta IS - \gamma I \tag{2.1}$$

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

In addition to defining the above equations to describe the epidemic dynamics, Kermack and McKendrick also used their mathematical model to show for the first time that the factor behind whether an epidemic will spread is determined not only by the number of Susceptible or Infected individuals remaining, but by a critical ratio of the population density. This is also known as the reproduction number R_0, representing the expected number of new infectious individuals introduced by a single Infected individual ($\beta S0$) over their infectious period ($\frac{1}{\gamma}$) in an otherwise uninfected population. Therefore $R_0 = \frac{\beta S0}{\gamma}$. If the basic reproduction number is $R_0 > 1$ then the epidemic spreads, otherwise it dies out. Towards the end of the epidemic, the chain of transmission breaks down due to a lack of number of infective people (not a lack of susceptible individuals) [17, 18].

2.2.4 Model Variations

Different models are conceived as a result of their application into different areas. This has led to the development of many variations of compartmental models that describe other characteristic that the infection may possess. For example *SEIR* models use an additional Exposed compartment to model a latent phases on contracting the infection. Compartment models can become very complex such as the *MSEIRS* model that incorporates passive immunity passed from parents along with latent periods and potential for recovered individuals to become Susceptible again. Other model variations arise from different means of transmission such as *Vector* or *Vertical* transmissions, considerations of Vital Dynamics (Births and Deaths in the population) and many other factors [19].

2.3 Stochastic Models

Stochastic models attempt to capture the randomness present in epidemic models, for example the rate at which an infection spreads will not necessarily be the same for each individual. Stochastic models achieve this by associating probabilities with transitions instead of rates (as in deterministic models).

2.3.1 The Reed-Frost Model

The Reed-Frost model was one of the first stochastic epidemic models. It is a stochastic, discrete time model that is based on a chain binomial model as it uses a binomial probability distribution to determine the compartment sizes at discrete time intervals as shown in Equation 2.2

$$\begin{aligned} &\Pr(I_{t+1} = i_{t+1} \mid S_0 = s_0, I0 = i_0, ..., S_t = s_t, I_t = i_t) \\ &= \Pr(I_{t+1} = i_{t+1} \mid S_t = s_t, I_t = i_t) \\ &= \binom{s_t}{i_{t+1}} (1 - q^{i_t})^{i_{t+1}} (q^{i_t})^{s_{t+1}} \end{aligned} \tag{2.2}$$

where q is the probability of not having contact with an Infectious individual and i_t is the number of Infected individuals at time t, (therefore $(1 - q^{i_t})$ is the probability of contact with at least one Infected individual assuming the Markov property holds). Although the Reed Frost model is simple and fast, it has relatively high error in predictions of real epidemic data [20, 12].

2.3.2 Markov Chain Models

Many stochastic epidemic models are based upon Markov Chains. Three types of stochastic modelling processes are Discrete Time Markov chains ($DTMC$), Continuous Time Markov Chains ($CTMC$) and Stochastic Differential Equation (SDE) models which have continuous state and time variables.

Stochastic simulations such as the Gillespie simulation provide numerical predictions of systems involving large interacting populations. They can be applied to epidemic models to simulate traces representing a "random walk" of an underlying model, given parameters and initial conditions. This can often be a useful source of synthetic data for evaluating the performance of epidemic models as the estimations of the parameters can be compared to the actual parameters used to generate the data.

2.3.3 Graph Models

Graph models are a stochastic based epidemic modelling technique that represents interactions as a graph and the spread across the network using stochastic agents. The Generalised Epidemic Mean-Field model ($GEMF$) [21] is a generalised graph model that can be applied to many different

situations, it models agents in different compartments using simple agent level descriptions of the stochastic process. The model assumes Markovian interacting agents and uses a first-order mean-field approach to approximate the exponentially increasing state occupancy probabilities.

2.4 Mathematical Techniques

The modelling of epidemics encompasses many different mathematical techniques and algorithms in order to construct useful and accurate models. Concepts from Statistics and Probability have also been used in order to give generality and degrees of certainty in the models created. Many different techniques are employed depending on the type of model created, however there are some common concepts that underpin most models are detailed in the following section.

2.4.1 Ordinary Differential Equations (ODEs)

Within compartmental models of epidemics, the dynamics of the transition between the different compartments of the model is described using Ordinary Differential Equations (*SIR*) of the transition rates to describe the rate of change of the compartment populations with time. For example the standard *SIR* model in (2.1) contains the set of three coupled (multiple dependent variables I(t), S(t), R(t) are dependent on the independent variable t) non-linear first order differential equations dS/dt, dI/dt and dR/dt to determine the dynamics of the Susceptible, Infectious and Recovered (respectively) compartment populations.

2.4.2 Optimisation

For compartmental epidemic models, the underlying form of the problem is non-linear multidimensional optimisation of the transitional parameters. In order to find the optimal parameters, it is required to find the solution of the differential equations determining the system. The equations are often coupled and non-linear systems of differential equations and exact analytical solutions are often difficult to obtain (depending on the model used) [22]. The standard *SIR* model has no generic analytical solution.

Numerical methods are therefore used in order to optimise the parameters using an objective function which determines how well the model fits the real data. The optimisation algorithm will then minimise / maximise this objective function and return the optimal parameters. Often an Ordinary Least Squares (OLS) or Maximum Likelihood Estimation (MLE) objective functions are used. OLS provides an optimal, minimum-variance, mean-unbiased estimation when the residuals have similar finite variance and are serially uncorrelated (Therefore autoregressive residual analysis may need to be undertaken). Furthermore OLS and MLE are equivalent when the errors are normally distributed [23].

There are many optimisation algorithms, one of the most robust is the Nelder-Mead unconstrained optimisation algorithm.

The Nelder-Mead optimisation method uses a simplex constructed from the data points $x_0, ..., x_n$ as the vertices. The function is then evaluated at the vertices of the simplex to enable the algorithm to determine whether to apply a Reflect, Expand, Contract or Shrink transformation to improve the objective function. Nelder-Mead is one of the most frequently used optimisation algorithms as it is one of the most robust; in many cases it provides an efficient reduction in function value over stiff and non stiff equations. Although the Nelder-Mead is a heuristic search method and can converge to non optimal solutions, in applications where the parameter estimations are subject to noise and a highly accurate solution is not required, the Nelder-Mead method provides efficient and reliable optimisation [24, 25].

There are many different control parameters that can be set for Nelder-Mead optimisation, such as tolerances and parameters for Contraction, Expansion and Reflection of the simplex. Different values of these settings may cause more accurate results in particular for multi-variable optimisation [26]. Furthermore, it is often required to run the optimisation repeatedly, check the convergence at each stage and increasing the maximum number of iterations as required.

Simulated Annealing [27] and Particle Swarm Optimisation [28] are stochastic optimisation techniques aimed at finding approximate values over a very rough optimisation surface. These methods incorporate a random variable to enable exploration of the rough surface space and can reduce the chance of finding local minima. Within advanced applications of multiple epidemic analysis, such approach may be applicable to get an initial estimate of the parameter values when the optimisation has a very rough optimisation surface, however they often take a long time to optimise and produce less accurate and robust results in comparison to the above methods.

General analytical solutions to epidemic models are typically considered hard to obtain, however attempts have been proposed for specific cases of *SIR* models by *Shabbir* et al. [22, 29] and the results have been shown to be consistent with numerical algorithm results for large population sizes. This could provide significant computational efficiency improvements, as optimisation is one of the most computationally intensive parts of epidemic modelling. If such analytical models can be adopted for epidemics of large parameters then further analysis to obtain gradients of the functions to guide optimisation would greatly enhance the current numerical optimisations approach commonly adopted such as Nelder-Mead and provide a significant enhancement in the computational complexity of the optimisation process.

2.4.3 Logistic Functions

Logistic functions may be a very useful in restricting the range of parameters within the optimisation procedure, the Logistic function transforms values in the range $(-\infty, +\infty)$ into the range $(0, 1)$ as

shown in 2.3

$$f(x) = \frac{1}{1 + e^{-x}}$$

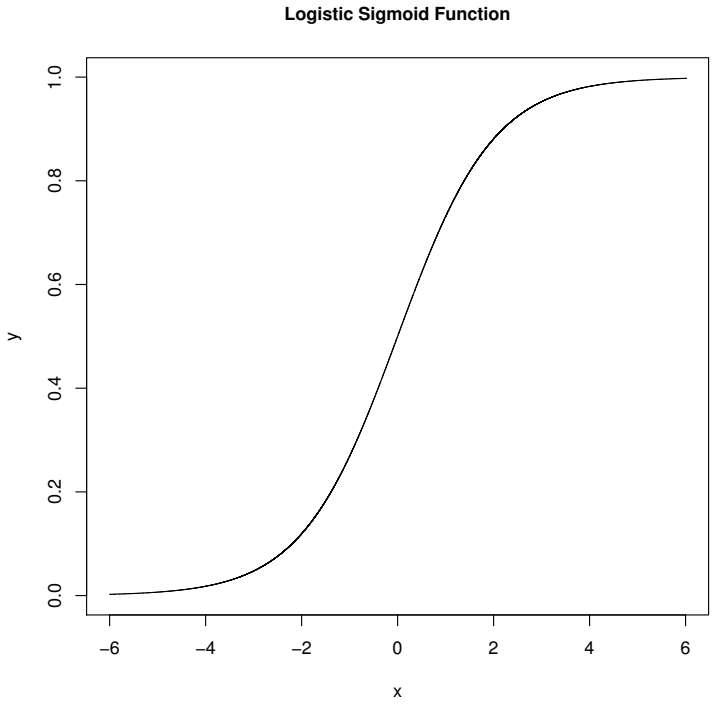


Figure 2.3: The Logistic Function

Adapting the Logistic function can enable lower and upper bounds to be specified on the optimisation parameters. The logistic function can enable an upper bound to be set on the start time, this can be achieved by using the following logistic function alternative,

$$f(x) = \frac{x_{max}}{1 + e^{-x}}$$

Where x_{max} is the upper bound of the parameter x. This function transforms values in the range $(-\infty, +\infty)$ into the range $(0, x_{max})$ because as $x \to -\infty, f(x) \to 0$ and as $x \to \infty, f(x) \to x_{max}$, ensuring that the output is within the required range and that the optimisation does not explore outside this range.

2.5 Model Analysis

The following section details methods in which the models fitting and prediction ability can be quantitatively assessed. Both quantitative and qualitative analysis are required to provide a comprehensive evaluation of the model.

2.5.1 Assessing the Model Fit

To fit the model to the data, we need to define a measure to determine how well the predicted model, given a set of parameters, fits the actual data. Optimisation of this criteria will then provide the best possible set of parameters for the model given the current known data. The criteria used to represent the fit is then used as the objective function of the optimisation.

One possible objective function is the sum of square error of the predicted data points from the actual observed data points. Considering the model predictions represented as a line through the observed values, minimising the squared error will intuitively enhance the fit of the model.

$$SSE = \sum_{t=1}^{N} (y_t - \hat{y}_t)^2$$

There are a number of different statistical approaches in order to determine a representative figure of how well the model fits the provided data.

The Coefficient of Determination, denoted R^2, can be considered as the "goodness of fit of a regression". Ranging between 0 (poor fit) and 1 (prefect fit), it describes how well a predicted function (f) fits the data (y). It is the inverse of the sum of squares of the distance of each data point to the mean in proportion to the sum of squares of each data point to the predicted value [30].

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

Where SS_{total} and $SS_{residual}$ are the sum of squared errors of the actual data from the mean and predicted values (respectively)

$$SS_{total} = \sum_{t} (y_t - \bar{y})^2$$

$$SS_{residual} = \sum_t (y_t - \hat{y}_t)^2$$

The Coefficient of Determination represents the proportion of variability in a data set that can be explained by the model [31]. and is a standard statistic for determining the *goodness* of fit of future predictions.

Another measure of how well the model fits the data is the Root Mean Square Error (RMSE). The RMSE also uses the sum of squared error, dividing SS_{total} by the sample size and taking the root gives the RMSE.

$$RMSE = \sqrt{\frac{SS_{total}}{n}}$$

The RMSE represents the sample standard deviation of the differences between predicted and observed values. Alternative techniques for analysing the model fit can be considered [32, 33] including using percentage errors instead of absolute errors. These metrics provide key statistics of the quality of fit and prediction of the model, however caution must be taken to not over-fit the model when using these approaches to determine outbreaks.

Although there are many different statistics to quantify the fit of the model, consideration of how to apply these techniques within iterative model fitting is required to provide the most representative measure. For example, to determine how well the model predicts future values it may be more appropriate to only apply the statistic over data in the future of the current time. Many other similar approaches need to be assessed in order to determine the most representative.

2.5.2 Uncertainty of the Model

Repeat runs of stochastic simulations can be used to provide upper and lower bounds of the predicted values of epidemic models. This can be predicted from the start of the dataset or from the current time interval within iterative fitting to show the expected bounds of future predictions. Repeating the simulation many times provides the expected bounds of the model, however it does not provide a quantitative measure of the uncertainty of the model.

Bounding predictions could also originate from the current time during the iterative fitting. This approach would provide bounds branching into the future from the current time value. For example Figure 2.4 shows the uncertainty of predictions of a time series into the future.

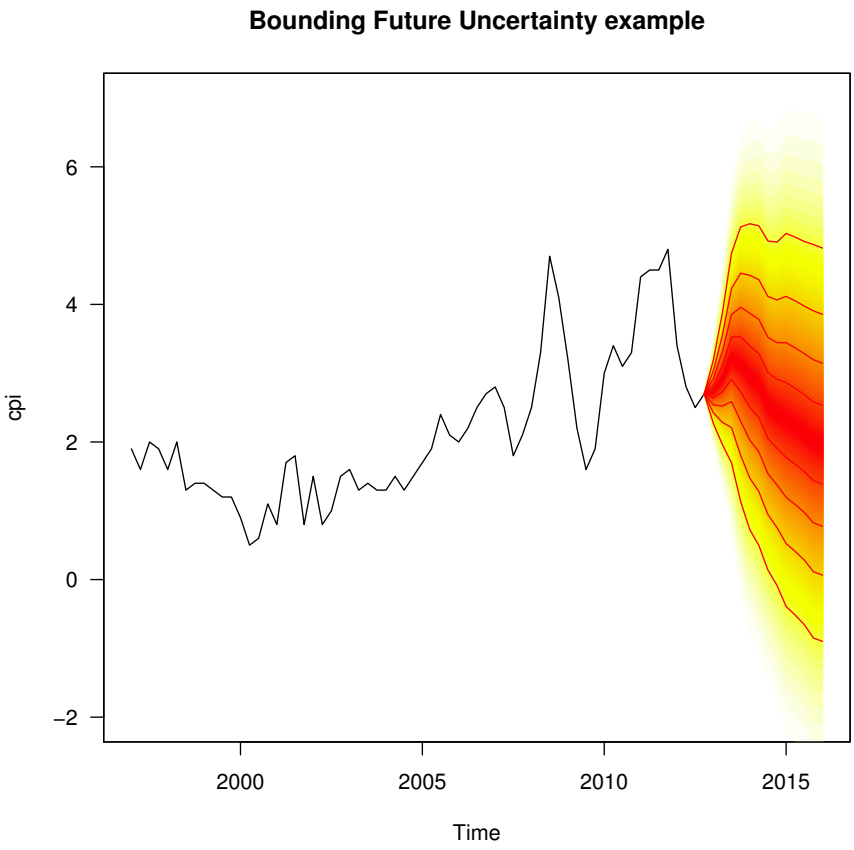


Figure 2.4: An example of future time series bounding predictions

Applying this analysis to epidemic models to enable similar future bounding uncertainty would be an ideal way to represent the uncertainty in future predictions of the model.

Likelihood based approaches use the likelihood of the observed data given the current parameters to formulate an objective function and may provide further insight into the uncertainty of the parameters. The likelihood can be represented as the product of the probability of observing the actual data assuming the error is distributed according to some Probability Density Function (PDF) f,

$$\mathcal{L}(\theta \,|\, x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta)$$

And the Log-Likelihood

$$\ln \mathcal{L}(\theta \,|\, x_1, \ldots, x_n) = \sum_{i=1}^{n} \ln f(x_i|\theta)$$

The Log-Likelihood can be maximised by the optimisation to find the optimal parameters of the model and assuming a Gaussian distribution of f is equivalent to using a Least Squares objective function. The use of Maximum Likelihood for optimisation enables a likelihood to be associated with the model parameters. Likelihood profiles or Likelihood contours can then be constructed representing the surface of the optimisation. Maximum Likelihood techniques can also be combined with statistical tests (such as the Chi-Squared test) to provide the percentile confidence levels associated with the optimal parameters of SIR epidemic models.

Other approaches to determining uncertainty in deterministic epidemic models include sensitivity analysis where parameters of the model are changed by a small amount to analyse the effect this has on the solution, but this does not provide any probabilistic techniques and so uncertainty cannot be associated to the parameters using sensitivity analysis [34].

The use of Bayesian techniques may enable uncertainty in deterministic models to be quantified. One such technique based on Bayesian Monte Carlo methods, assigns a probability distribution to each parameter and then uses sensitivity analysis, repeating the model fitting by sampling parameter values from their distributions. This approach enables probabilities to be associated to predictions, however, may underestimate prediction intervals. Other Bayesian approaches include Bayesian Synthesis and Approximate Bayesian Computation, these approaches construct a pre-model distribution over input and output variables. It then samples from the pre-model distribution and uses the model to run simulations for each sample. The simulations are then weighted and re-sampled with probabilities according to the weights [35, 17].

2.5.3 Cross Validation

To determine how well the model predicts future data is to repeat the fit of the model several times over different datasets, this is often referred to as cross validation. Cross validation splits the given dataset into training and testing data. The training data is used to optimise the model parameters, whereas the testing data is used to compare the model predictions against the actual data values. Often the data is compared in folds, using different randomly selected portions of the data for testing and training. This is often infeasible for epidemic modelling in practice, because the data is often only collected for a single outbreak of the epidemic. However in sequential fitting, past data can be considered as the training data and future data used as testing data to determine the uncertainty of predictions (fitting the model in hindsight of the epidemic).

2.5.4 Residual Analysis and Autoregressive Modelling

Analysis of the residuals of the model can provide insight into many different aspects of the model. Both quantitative tests and visualisation of residuals can be used to highlight many complex aspects

of the fit of the model [36]. Analysis of the epidemic model residuals using the Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) can determine if any variance remains in the residuals. The ACF shows the correlation between a time series and a lag of itself. The PACF shows the correlation between a time series and a lag of itself that does not include the correlation propagated from intermediate lags. The ACF and PACF functions can also be used to determine the order of autoregressive models.

An autoregressive (AR) model uses the previous values of a time series in order to predict future data points assuming that the observations are not independent and the data is stationary. An AR model defines the next data point using the previous data points,

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t \tag{2.3}$$

AR models such as ARIMA could be applied to the actual data of an epidemic in order to predict the next data points. Furthermore AR models could be applied to the residuals remaining after fitting an epidemic model in order to explain the remaining variability in the data and refine the epidemic residuals to enhance the model prediction [37, 38].

2.6 Related Decomposition Techniques

An emerging realisation within the application of epidemic models to complex domains is that there may be multiple underlying epidemics driving an overall epidemic process. In particular many applications have shown multiple peaks of epidemic activity. There is therefore a clear requirement for an epidemic model which can account for multiple epidemic outbreaks.

2.6.1 Fourier Series

One approach to building a multiple epidemic model is inspired from Fourier series decomposition of a signal. Fourier analysis can be used to decompose a periodic function into many sub functions that can be solved individually and then recombined to obtain an approximation of the original function. As a result of the superposition principle, each sub function can be treated separately to determine their individual contributions, and the original function can be represented as the superposition of these sub parts. The Fourier series is derived using equation 2.4. where a_0, a_n and b_n are coefficients determined by integrals of the function over l, $-l \leq x \leq l$ and n is the number of terms.

$$f(x) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi x}{l}\right) + \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{l}\right) \tag{2.4}$$

Decomposition of a square periodic function into sub periodic functions using Fourier series is shown in Figure 2.5.

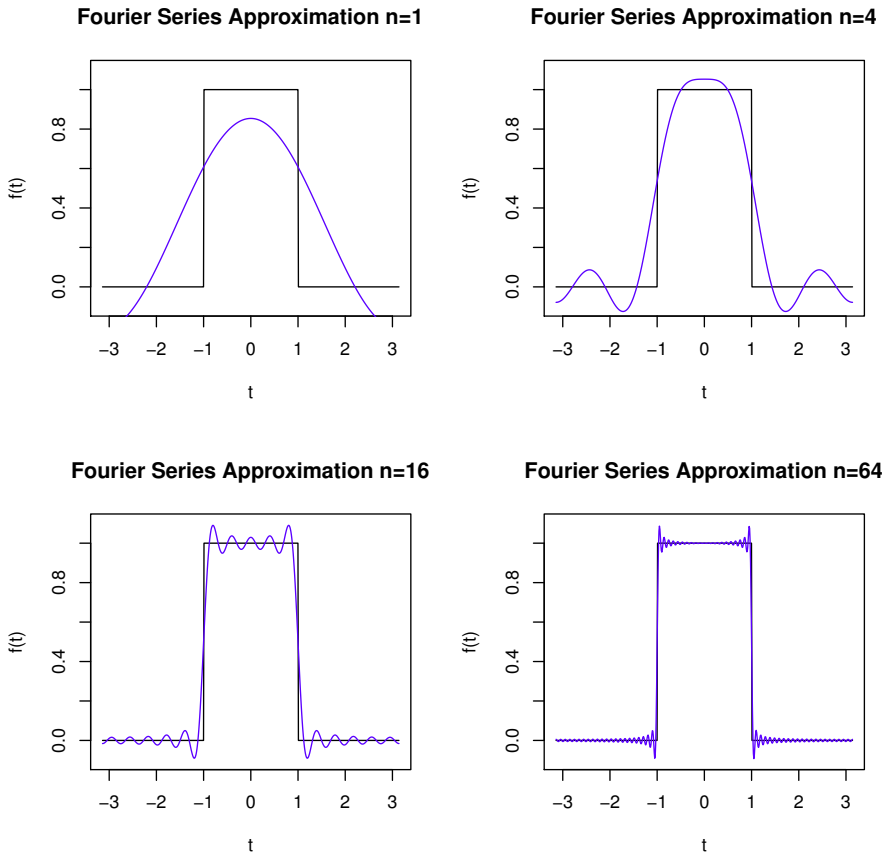


Figure 2.5: Fourier series of a square function using different numbers n of sub periodic functions

2.6.2 Model Composition Considerations

A significant complication that needs to be considered is how to determine the initial parameterisation of each sub epidemic. In particular the start times of the epidemics and both discrete and continuous approaches need to be considered.

Alternative approaches to fitting multiple epidemics could fit sub epidemic piece-wise or even adopt a basis function approach, using many sub epidemic models or other models such as Gaussian distributions to build up the overall epidemic. An Expectation Maximisation approach could also be investigated as a potential method to find the optimal fit.

Finally the idea of both positive and negative epidemics is also interesting, negative epidemics would enable fitting to scenarios such as immunisations that are provided throughout the population. Furthermore the algorithm could also drop epidemics if they become redundant.

In addition to different approaches to fitting the sub-epidemics, techniques for determining outbreaks and the type of the outbreaks are required. For example instead of using the R^2 as a guide of the fitting procedure, an alternative could analyse the residuals of the fit against the actual data points. This could be implemented though established residual analysis techniques, or a simple approach could use the final n residuals to determine if new epidemics should be introduced. Many different combinations of approaches to the model fit can be considered.

2.7 Recent Applications

The main application of epidemic models over the last century has been within biological epidemics. However in recent studies applications of epidemic models have proliferated into many new areas.

2.7.1 Current Disease Models

Epidemic and Pandemic Modelling Frameworks Current epidemic models for disease outbreaks include many different factors and assumptions specific to the disease being modelled. This gives rise to very complex epidemic / pandemic models. One of the most significant challenges in such models is to simultaneously consider the co-evolutionary dynamics of an epidemic - assessing the dependent effects on the Community, Economy, Behaviours and many other factors. Such models often incorporate spatio-temporal factors as a key aspect and simulate transportation globally. Vertical transmission (inherited transmission), Vector transmission (transmission through an indirect medium) and Demographic considerations are also considered. For example the impact of the airline transportation network, school closures and even meteorological patterns [39, 40, 41].

A general framework that simultaneously encompasses the co-evolutionary dynamics of an epidemic is the holy grail of current epidemiology disease modelling research, however to date the impressive yet limited frameworks such as the Eclipse project *STEM* - Spatio Temporal Epidemiological Modeller[1] and The *GLEAMviz*[2] project are the closest attempts. They enable fine grain stochastic epidemic simulations to be created by specifying the characteristics of the outbreak such as the outbreak type and transition rates. These models provide a good insight into some factors, in particular transportation and geographical aspects, however they are limited in the simultaneous impacts of other aspects such as Economy and Behavioural factors which are essential when using the model for determining the best course of future action to prevent a pandemic.

[1] http://www.eclipse.org/stem/
[2] http://www.gleamviz.org/

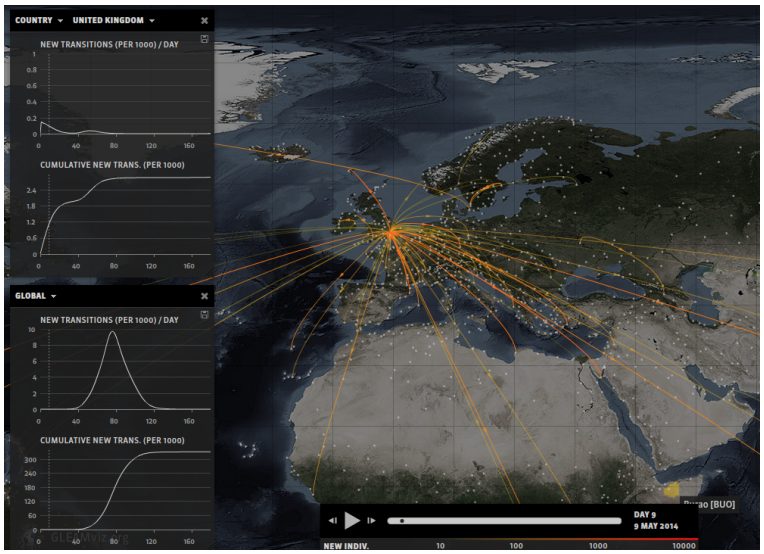


Figure 2.6: GLEAMviz pandemic modelling framework

Multiple Epidemic Models Within the spread of infectious diseases, recent work has looked at the changes caused via variation in social mixing within a heterogeneous population. In particular they found that the number of people infected with Influenza dropped when schools closed. This causes multiple peaks in the epidemic outbreak, they note that the "SIR epidemic clearly does not reproduce the dynamics" as the "SIR curve has one turning point, whereas the real epidemic curve has three". From this study the concept of independence in the multiple outbreaks, and in particular fitting the second epidemic using the final conditions of the first epidemic, has shown to be problematic [42].

2.7.2 Social Interactions

An interesting application of epidemic modelling is within the context of social interactions - applying epidemic models to the dissemination of social media events across the internet. The spread of such internet based phenomena is analogous to the spread of a biological infection. A certain number of initial Infected people exposed to the initial event spread the event (for example via posting, tweeting or linking) to Susceptible individuals that then become Infectious and may further propagate the event. The degree of similarity between the spreading processes remains an area of research within sociology with many complex factors such as electronic *"Word of Mouth"* effects, underlying network connectivity and increased spreading rates [5, 43]. Recent research papers have succeeded in applying epidemic models to predict the evolution of a single outbreak to social data. For example the spread of events on *Google Trends*[3] modelled by epidemic outbreaks [44] and celebrity events within On-line Social Networks (OSNs) [5]. There have also been a number of

[3] www.google.co.uk/trends

alternate approaches to study the spread of information within OSNs [45].

Bieber Fever A recent study on the spreading of events on *Google Trends* by Tweedle and Smith entitled *A Methematical model of Bieber Fever* [44], studies the highly infectious "Bieber Fever Pandemic". An interesting aspect of their model considered the positive and negative effects of media as the linear combination $M = \epsilon Pos + (1 - \epsilon)Neg$ enabling Susceptables to succumb to *"Bieber Fever"* though Positive media in addition to contact with existing *"Bieber-infected"* individuals. The customised form of SIR model adopted incorporates many other factors within the model, such as Maturation and Media rates. This leads to many possible outcomes of the analysis due to the different values that the additional factors can take. Furthermore these factors can only be estimated and may not even present real association to the underlying factors driving the epidemic, for example negative media may even result in more *"Bieber-infected"* individuals. As a result the model is overcomplicated and only enables hypothesised situations to be analysed (such as the *"Lindsay Lohan"* effect caused by excess negative media!).

Celebrity events A more general approach by *Nika* et al. [5] investigated the spread of celebrity event propagation throughout OSNs by analysing real-time epidemic outbreaks. Several case studies are provided in which SIR and SEIR models are fitted to simulated data sets and also real download. The model showed ability to predict with reasonable accuracy the future evolution of an outbreak within the early stages of the iterative fitting. A significant realisation from their results is that epidemic dynamics can be applied within social spreading and furthermore that more than one epidemic process may be responsible for the overall spreading mechanism (for example a recent event in a celebrity profile such as releasing a new song). This result directly influenced the conception of this project.

Social Network Graph based Approaches Recently the use of graph based approaches to model the outbreak of information over social networks have been proposed. An interesting recent application developed by *Fiems* [45] uses stochastic diffusion processes on directed graph based models to predict the spread of epidemics within networks such as Twitter[4]. In particular they argue that a directed graph model naturally models the asymmetric relations present in social networks. They adopt a compartmental approach and use the mean-field approach to reduce the state space of the stochastic model to derive the dynamics of the compartmental flow as differential equations in terms of the number of Infected and Recovered nodes of a certain in and out degree. They also present simulations of the mean-field model developed by first setting up the graph and then running the information diffusion process on the graph. The simulation results correspond closely to the proposed differential equation solutions.

[4]https://twitter.com/

Interactions in Overlapping Epidemics The effect of different epidemic types have been modelled by *Fiems, Nika and Knottenbelt* [46] by using a Markov model whose fluid limit reduces to a set of coupled SIR-type ODEs. In particular they looked at modelling the effects of Syndemics and Counter-Syndemics within different degrees of overlap between multiple populations. They present a novel and interesting model of the dynamics overlapping epidemics which could be used for future case studies by being applied to real epidemic data sets.

Computer Worm Viruses Another novel application of the use of epidemic models is within the prediction of the spread of computer virsues such as computer worms like CodeRed through a network. A paper by *Bradley* [4] analyses the spread of such worm attacks by simulating SIR models via using the modelling language PEPA to develop the underlying differential equations of the system. This is a more automated approach than manually developing customised epidemic models and they note the impact of throttling the bandwidth. A multiple epidemic approach may be applied within such a scenario to analyse the real-time spread and impact of throttling the bandwidth using multiple standard SIR epidemics.

Patterns of Information Diffusion An interesting study on temporal patterns in online content and how content grows and fades over time highlighted how temporal variation can be characterised by time series shapes. They observe the presence of six main temporal "shapes" including "bursty" and "spiky" peaks within social online data and develop a K-Spectral Centroid (K-SC) clustering to identify patterns of temporal variation in online media [47].

SpikeM An analytical model called *spikeM* [48] embodied the main ideas of the above paper. The model has an initial "shock", an infectivity rate β and a power law decay function which unifies the different patterns observed in social media. They show its ability to fit to many datasets, forecast, "reverse engineer" parameters, predict scenarios and also apply it to outlier detection. The model has four main advantages: unification, practicality, parsimony and usefulness. Overall they develop a completely novel model to characterise the "rise and fall patterns" observed in on-line media and compare it to many existing models. The evaluation of the model is qualitatively based (with the exception of the RMSE), and more extensive quantitative comparisons to other existing methods would enable better model evaluations. The main disadvantage of this model is that the periodicity frequency needs to be manually provided to the model and the periodicity is cyclic which is what enables the parsimony of the model. A more generic approach would enable epidemics to arise at irregular time intervals (but may require more parameters as a consequence). The *spikeM* project is highly relevant to this project as it attempts to explain the interaction of outbreak patterns in outbreak data.

There are many other areas in which epidemic models can be applied and the underlying techniques discussed in this paper can be used in a range of different fields.

Chapter 3

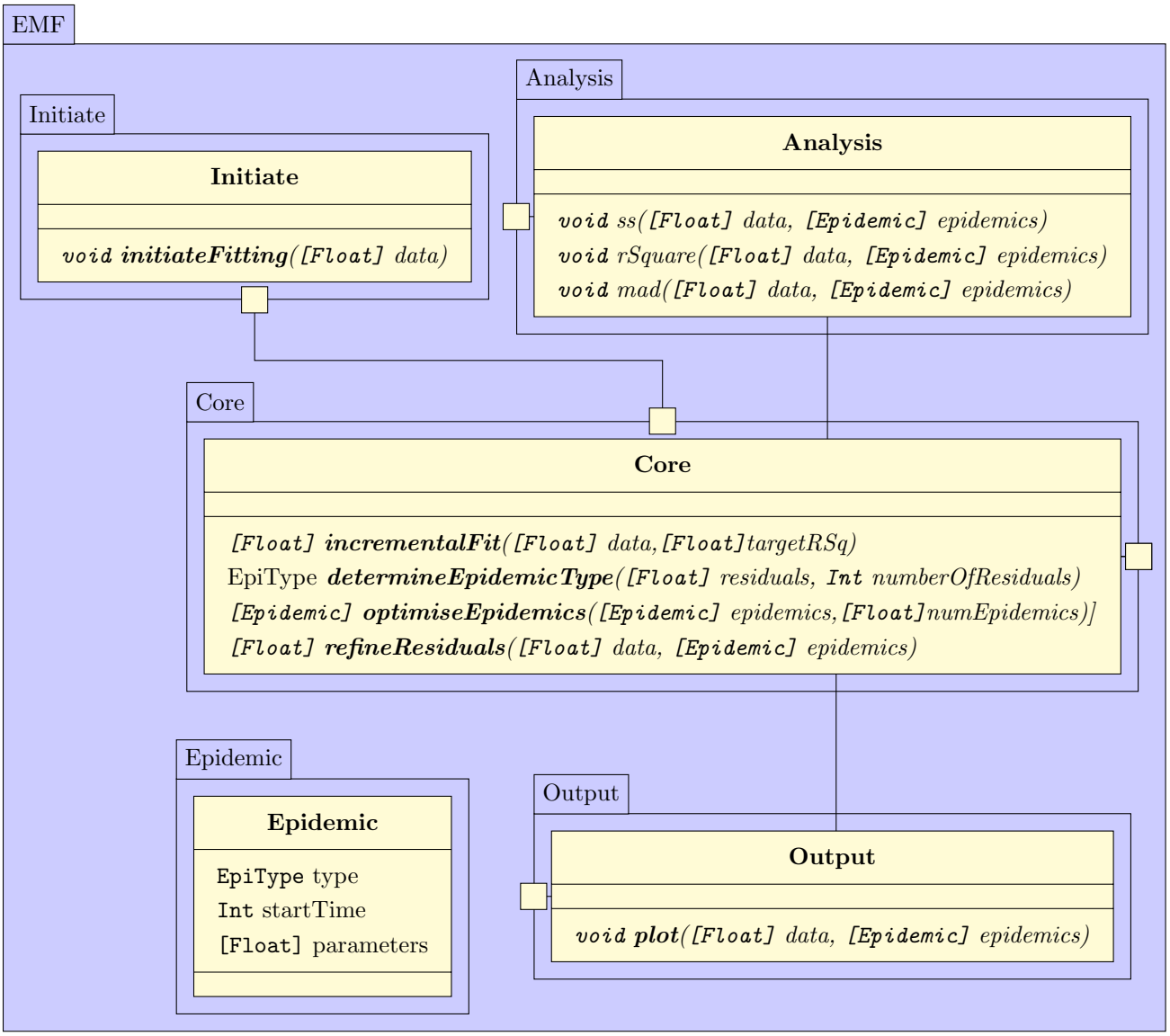
Single Epidemic Modelling

Existing epidemic modelling implementations are detailed throughout this section with the exception of the iterative, real-time fitting which is one of the main contributions of the project. At every new observation within the epidemic outbreak the model is re-optimised to adapt to the latest information. The techniques of parameter optimisation and synthetic modelling at each stage have been previously explored within existing epidemic models. More advanced modelling techniques that employ the iterative fitting method are described in section 4. The chapter starts with considerations of the implementation design.

3.1 Software Engineering Design Considerations

Before commencing development of the project, is was important to consider some design decisions to ensure the maintainability and extensibility of the epidemic modelling implementations. The contributions of this project have been developed within an epidemic modelling framework that is divided into four main modules; Model Initiation, Core fitting procedures, Model Output and Analysis. This enables for the extensive fitting procedure to be separated from the output and analysis of the model. Employing the *separation of concerns* design principle decouples the procedures and making them easier to maintain and extend. The key contributions of the project permeate throughout these main modules, with the main challenges implemented though additions or adjustments to one or more of the core modules. The following diagram outlines the main components of the framework

3.1.1 Epidemic Modelling Framework Design



3.1.2 Programming Language Considerations

An important decision in developing the epidemic modelling framework was determining the best language to implement the project. Due to the mathematical and data oriented nature of epidemic modelling, a numerical computing package such as Matlab, R or equivalent was required. R provides a vast range of different packages via the Comprehensive R Archive Network "CRAN"[1]. "CRAN" which provides a wide range of statistical and mathematical algorithms for data processing, analysis and formatting. Combined with the comparative ease of handling data in R, "CRAN" makes R an

[1] http://cran.r-project.org/web/packages/

ideal language for the framework.

Matlab was also a big contender as it was familiar and is overall well suited to the nature of the task due to its utilities for handling data. Matlab also offers some extended features that R does not provide (such as interactive graphs).

C++ and Java are not as suited to the handling data as Matlab and R - even though libraries such as GSL (GNU Scientific Library)[2] do provide data analysis functionality, the overall handling of data is more difficult. Less well known languages such as Sage and Maple were also considered, however they can't contend with the packages and support available from the more mainstream alternatives. R was chosen for the implementation due to the extensive range of extremely well suited and useful packages that R offered for the project, its comprehensive network of support and its suitability for handling data.

3.2 Core Components of the Epidemic Modelling Approach

To provide the foundations of the epidemic modelling framework, some core components need to be implemented, for example parameter optimisation. When developing the basic components of the framework some key considerations arise such as setting the initial conditions and deciding on an optimisation strategy.

3.2.1 Unknown initial conditions

In order to evaluate and optimise the epidemic model, the initial conditions and initial parameters of the epidemic need to be set, however they are often unknown. In particular the initial number of Susceptible individuals is often initially unknown and the start time of the epidemic may also be unknown. Traditionally, the initial conditions are estimated from observations of the first few cases of the epidemic [42], however in many applications this information is not available. Given the start time, the initial number of Infected individuals can be determined by the data point at the start time for a single epidemic. A reasonable estimation of the initial Susceptible population can be assumed to be significantly higher than the initial number of Infected individuals and throughout the fitting the number of initial Susceptible must be greater than the number of initial Infectious.

Setting the initial parameters to be arbitrary values is the simplest approach. For example setting $I0 = 1$ and $S0 = 10$, however this may not be general enough to deal with outbreaks with a large population. An alternative method is to set the initial number of Infected individuals to be the first data point and the initial number of Susceptibles to be an order of magnitude greater. This method sets the initial parameters within the range of the observed data however makes the initial conditions highly dependent on the first data point within the data set. Transforms and

[2] http://www.gnu.org/software/gsl/

bounds may also be applied to the initial conditions, this ensures that the optimisation does not proceed into infeasible values. Bounding parameter approaches are investigated in more detail within section 4.1.3.

3.2.2 Initialisation and Iterative fitting

Real-time fitting of an epidemic is simulated by adding a single observation at a time. at each stage within the fitting the model is re-optimised on the current truncated data set. The iterative fitting loop is outlined by the pseudo-code algorithm 1.

Algorithm 1 Iterative Fitting

```
1: function iterativeFitting(data, target)
2:     times = [1:length(data)]
3:     // Set minimum number of data points to optimise over
4:     minTruncation = 4
5:     initParams = []
6:     for i in (minTruncation :  length(data)) do
7:         optimParams = optimEpidemics(times[1:i], data[1:i], initParams)
8:         predictedInfectious = evalEpidemic(optimParams)
9:     end for
10: end function
```

3.2.3 Model Evaluation

In order to optimise the parameters of an epidemic model, it is required to be able to evaluate the model at given parameters. The model evaluation is obtained by solving a set of differential equations at the current parameters. The differential equations define the changes in the model compartments over time. To implement the model evaluation stage, firstly the model differential equations need to be implemented and then the `deSolve` package in R is used to obtain the evaluations at each time point as shown in Figure 2.

Algorithm 2 Evaluate Epidemic

```
1: function evalEpidemic(params, type)
2:     // Evaluate epidemic model at current parameters
3:     predictedInfectious = lsoda(params, type)
4: end function
```

Depending on the epidemic type, the associated set of differential equations are used by the ODE (Ordinary Differential Equation) solver to make the predictions given the initial parameters passed in `epidemic.parameters`. Different ODE solvers are available, the ODE solver used throughout this project is the R `lsoda` solver within the `deSolve` package.

The evaluations for each time point are then used as the predictions of the current model parameterisation, and the sum of squared error of the actual data points from these predictions can be

calculated.

3.2.4 Sum of Squared Error

Using the evaluation of the model at the current parameters, the sum of squared error of the current model to the actual data points is used as the objective function to guide the Least Mean Square optimisation.

3.2.5 Parameter Optimisation using Least Mean Squares

At each step within the optimisation, `optim` calls `sseEpidemic` to calculate the sum of squared error by evaluating the epidemic at the current parameters using `evalEpidemic`. The sum of squared error can be used as the objective function in order to guide the optimisation process and optimisation then progresses using Nelder-Mead optimisation to minimise this objective function. Optimisation is implemented by the `optim` procedure in R as shown in the following pseudocode 4.

Algorithm 3 Optimise epidemics

```
1: function optimEpidemics(data, params, type)
2:     // Optimise the model parameters
3:     optimParams = optim(params, sseEpidemic, times, data, type)
4:     return optimParams
5: end function
```

Algorithm 4 Sum of Square Error Objective Function

```
1: function sseEpidemics(data, params, type)
2:     // Optimise the model parameters
3:     epidemicPredictions = evalEpidemics(params, type)
4:     return pow(data - epidemicPredictions, 2)
5: end function
```

Although it appears that the optimisation and evaluation procedures (`optim` and `lsoda` respectively) could be called directly from within the `iterativeFitting` algorithm, they have been deliberately separated into two modules: to enable them to be easily extended and follow the *separation of concerns* design principle. This enables more complex optimisation and evaluation procedures to be implemented within separate modules, increasing the maintainability and extensibility of the framework.

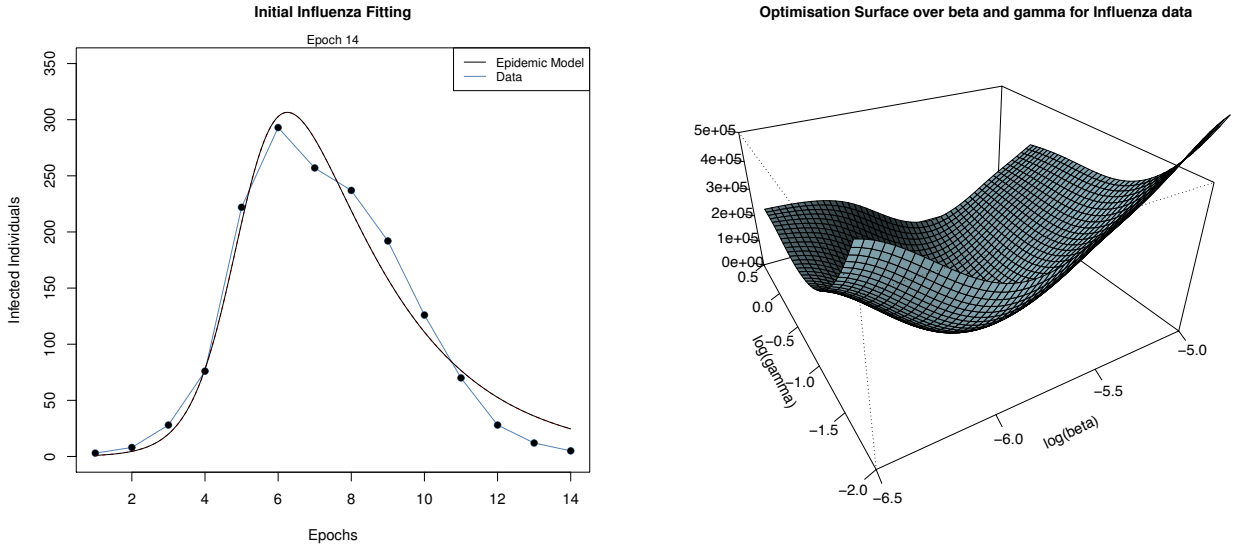
3.3 Basic Epidemic Model Implementation

3.3.1 Basic Epidemic Model

The compartmental SIR model is used throughout this entire project as the principal epidemic model. The SIR model has been chosen as it has a minimal number of parameters and is sufficient for evaluating new techniques [7]. The basic components outlined in chapter 3 are implemented in order to optimise the single epidemic model.

3.3.2 Parameter Optimisation

Initially just β (infection rate) and γ (recovery rate) are optimised, fixing the initial conditions $S0$ and $I0$. The change in SSE as a result of changing the parameters can then be visualised as an optimisation surface over the parameter space. The optimisation aims to find the minimum point of the optimisation surface. The log of the parameters are used in the optimisation to ensure that the values can not be negative. Figure 3.1 shows the change in the SSE over β and γ for a simple Influenza data set obtained from [49, 50] (Many of the optimisation techniques within this chapter are inspired by [49, 50]). A problem with only optimising over β and γ is that the initial $S0$ must be close to the actual value in order for the parameters to be determined. By including $S0$ in the optimisation, the initial $S0$, β and γ can be optimised simultaneously to produce the optimal parameters.



(a) Final Fitting

(b) Optimisation Surface over $\log(\beta)$ and $\log(\gamma)$

Figure 3.1: Optimisation Analysis for Influenza data

3.3.3 Synthetic SIR Iterative Fitting

Synthetic data sets can be generated using a Gillespie stochastic simulation as discussed in section 2.3.2. This enables the output parameters of the fitting procedure to be compared to ground truth values, enabling a quantitative comparison of how well the model reproduces the parameters. For real epidemic data comparing the final parameters to the actual parameters is not possible as the actual parameters of the epidemic are unknown. Although the parameters are expected to be within the correct range, there will be some variation as the synthetic data set is just one stochastic run starting from the current parameters. For the initial synthetic data set, an SIR epidemic was simulated with parameters $\beta = 0.001, \gamma = 0.1, I0 = 10, S0 = 500$. The final R^2 at the end of the fitting (day 58) is 0.997 showing that the epidemic fit is very close to the actual data. The reconstructed parameters from the optimisation of the final stage are $\beta = 0.000813, \gamma = 0.138, S0 = 667$.

The iterative, real-time, fitting process at three different times within the synthetic data fitting are shown in figure 3.2. The iterative fitting process involves fitting the model at every single data point within the graph. This is achieved by optimising the model at each truncated data set and graphs for every single time point are produced. At the start of the fitting process, the model fits the existing data well with a high R^2 of 0.998. However due to the increased uncertainty as a result of a small number of data points at this early stage within the fitting, future predictions of the number of infected individuals are over approximated. At time 24 of the outbreak, the peak of the outbreak and future number of infected individuals is predicted more accurately and the R^2 error over previous data remains similar. Towards the end of the fitting process, the future predictions are very close to the actual data points and the parameters are very near to the actual parameters used.

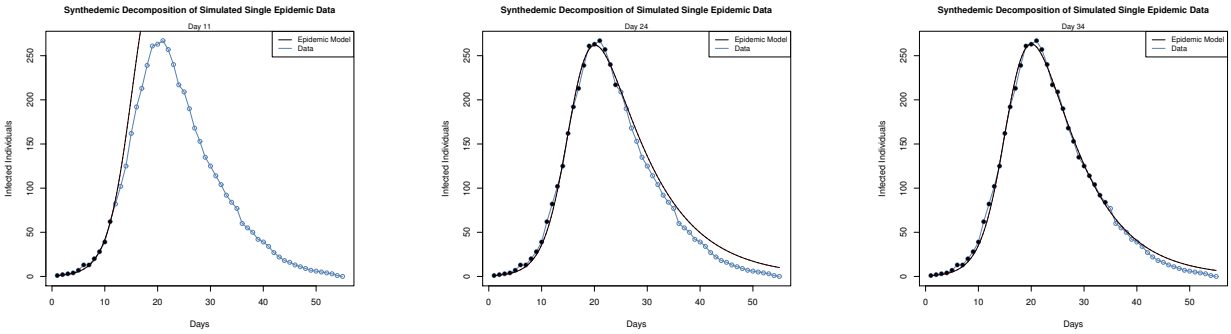


Figure 3.2: Single SIR Iterative fitting to synthetic data

It is important that the solid black dots show the data points that are being used for the optimisation (the black dots represent all previously observed data points), while the empty circles represent future data points (that have not yet been observed). The future empty circles enable qualitative future prediction comparisons to the model fit which is represented by the solid black line.

3.3.4 CDC Influenza Data

The Centres for Disease Control and Prevention (CDC) portal[3] is an excellent source of epidemic Influenza data providing detailed data on the latest Influenza outbreaks. The number of infected individuals sampled at each week throughout the 2012-2013 period is used as real data to fit the SIR model to actual data. 3.3.

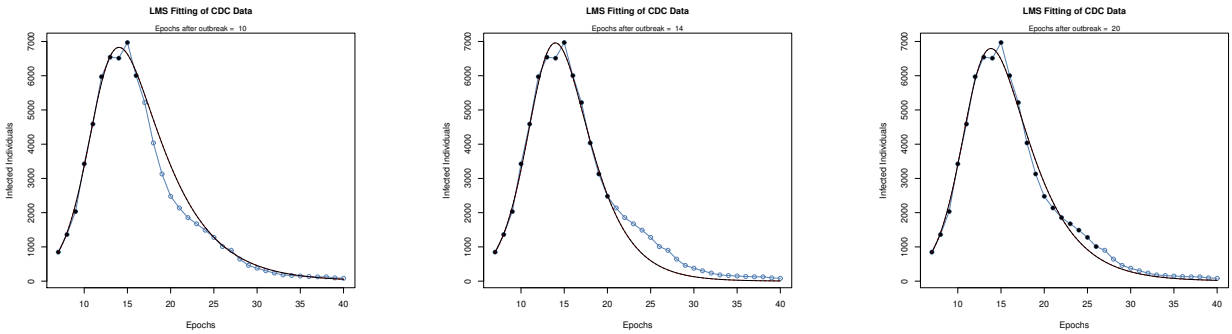


Figure 3.3: Single SIR Iterative fitting to CDC Influenza data using LMS

3.3.5 Parameter Optimisation using Maximum Likelihood

Maximum Likelihood Estimation (MLE) could be used as an alternative objective function to Least Mean Squares to optimise the epidemic parameters. Using an MLE approach would enable a likelihood to be associated to the parameters which can be used to produce confidence levels of the parameters.

The CDC results of Figure 3.3 have been replicated using an MLE based approach using the *mle2* function. The Real-time fitting with MLE optimisation has been undertaken over the entire CDC dataset and some of the interesting time points are shown in figure 3.4

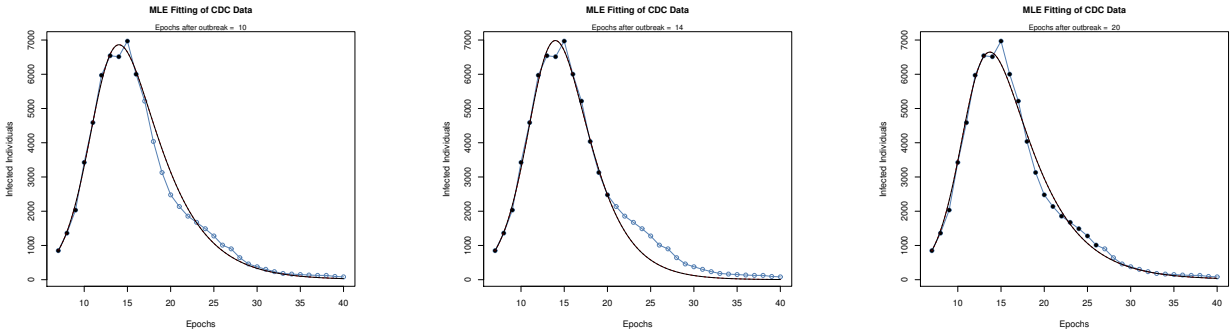


Figure 3.4: Single SIR Iterative fitting to CDC Influenza data using MLE

[3]http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html

Determining uncertainty is one of the most challenging areas of epidemic modelling. The level of inherent uncertainty in the model parameters and a confidence level in the model prediction needs to be provided in order for the results of the model to be evaluated properly. Many recent papers have been devoted to the challenge of characterising different sources of uncertainty and even showing that traditional approaches to determining uncertainty may not be as accurate as believed [51]. Traditionally, estimates of initial parameters are obtained from manual processes such as index case studies and contact tracing [42]. However in new applications of epidemic outbreaks, such methods are often infeasible and more comprehensive methods for determining the initial conditions and uncertainty of the model are required.

Table 3.1 shows the different values of β, γ, $S0$ and R^2 at the end of the CDC outbreak.

Table 3.1: Comparison of LMS and MLE fitting parameters for CDC data

	LMS	MLE
β	0.000040	0.000045
γ	0.28	0.24
$S0$	20700	16800
R^2	0.99	0.99

The main focus of this project is proposing new approaches modelling interacting epidemic behaviour. Therefore the use of MLE methods for the Synthedemic model remains an extension to this project for future work as discussed in section 6.3.

Chapter 4

Synthedemic Modelling

The underlying components of the Synthedemic model are presented throughout this chapter, constituting the main and novel contribution of the report. An overview of the Synthedemic components is provided in Figure 4

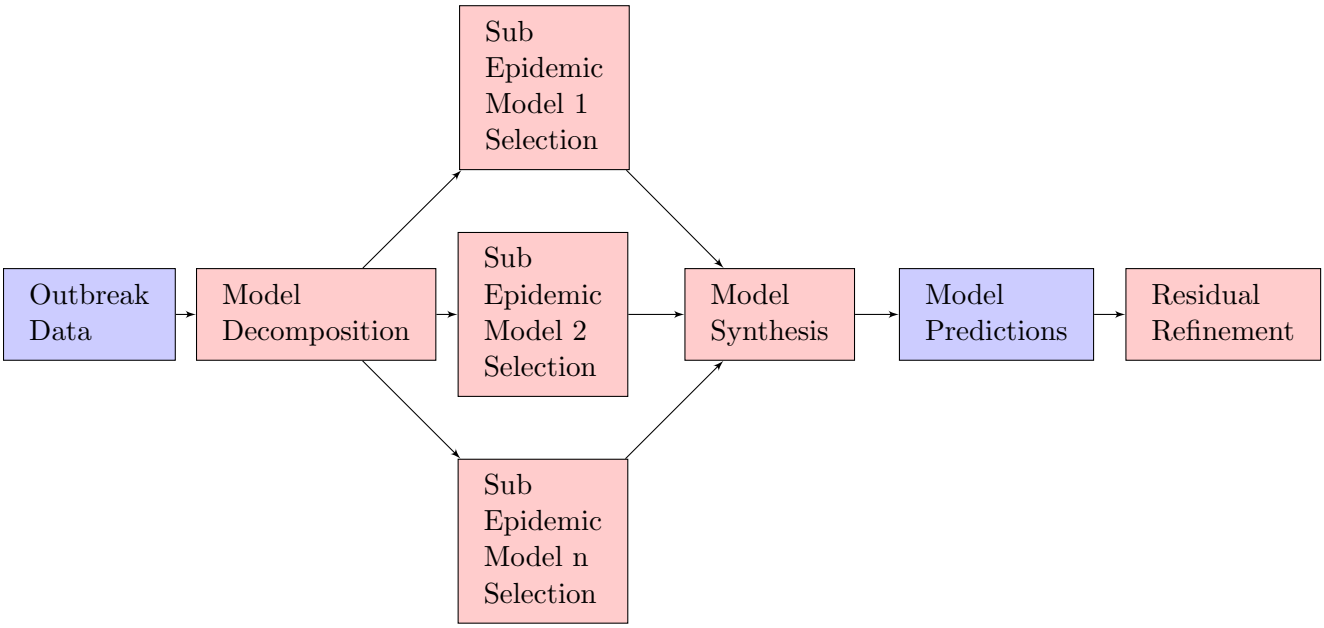


Figure 4.1: Synthedemic Modelling Overview (Red nodes correspond to the sections listed below).

Combined multiple oubreak data is input to the model which then dynamically detects the start times of each sub epidemic model within the Model Decomposition stage (detailed in Section 4.1). They type of each sub epidemic model is determined by the Model Selection stage (detailed in

Section 4.3). The sub epidemic components are then combined in the Synthesis phase which enables future predictions to be generated. Finally the Residual Refinement (detailed in Section 4.4) applies an autoregressive model to enhance the model predictions.

4.1 Synthedemic Model Decomposition

In recent applications of epidemic models (and even within traditional applications of epidemic models to the spread of disease) there are often multiple peaks of infection that emerge. Current research within this area has not attempted to model the simultaneous effects of multiple epidemics, and rather tries to establish initial conditions and model the outbreak as sequential individual epidemic models. This approach intuitively seems inadequate as the spread of an initial outbreak will have an effect on future outbreaks, furthermore multiple outbreaks may overlap due to different underlying causes of the epidemics. In fact fitting the second epidemic using the final conditions of the first, is shown to be problematic because the second epidemic can never take off due to $SR0 < 1$ and $\frac{dI}{dt} < 0$ [42]. This shows the need for a multiple epidemic model that optimises the parameters of the sub-epidemics simultaneously.

A potential solution to the observed inadequacy of the existing epidemic models could be a multiple epidemic model, composed of a number of sub-epidemics that are optimised simultaneously. Some of the key factors that need to be investigated are the initial conditions of each of the epidemic models, and in particular determining the start time of each sub epidemic is non trivial.

The main objective of the Synthedemic approach is to be able to fit a multiple epidemic model to a data sets containing many outbreaks. Ideally the algorithm should determine the number, and start times, of the sub-epidemics while parametrising the model. However when fitting outbreaks with many epidemics, this goal is often infeasible. Therefore the fitting is undertaken iteratively, enabling the start times of epidemics to be determined on the fly. Although the process is undertaken iteratively, given the start times of the outbreaks the model should be able to optimise the model starting at any arbitrary time during the epidemic outbreak as each optimisation should be independent of the previous. The following chapter details attempts at solving the challenges of sub epidemic Model Decomposition within the Synthedemic model.

4.1.1 Determining Sub Epidemic Start Times

Searching the possible start times of multiple epidemics is a significant computational challenge. The times of the outbreak are an initial estimate of the epidemic start times however are often not the actual epidemic start times. The direct approach of searching all feasible discrete start times to find the optimal start time is too computationally expensive and the complexity increases linearly with the number of data points when fitting a single epidemic. Furthermore, when fitting multiple

epidemics, searching each combination of start times quickly becomes computationally infeasible and heuristics to refine the search need to be used such as only searching feasible start times of the most recently fitted epidemic. As there appears to be no ordering of the start time approaches, such as binary search algorithms, are inapplicable however searching using gradient descent may be possible.

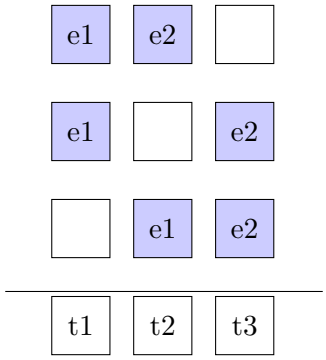


Figure 4.2: Discrete Feasible Start Times

As shown in Figure 4.2, there are many combinations of possible start times for the different epidemics. The number of combinations increases as the number of time points increases and the number of epidemics increases. For t time points and n epidemics (assuming that the order of the epidemics is not significant) then the number of combinations of start times to consider is tCn. To reduce the number of combinations, a range around an estimate of the start time of each epidemic needs to be searched.

To prevent the combinatorial explosion of start times, one approach is to only explore the start time of the most recently added epidemic. Consequently the number of combinations depends only on the remaining time after the start time of the previous epidemic. This simplification may lead to setting incorrect start times due to the previous epidemic start times not being updated in light of new data, however was initially implemented because even searching feasible start times for one epidemic has to be undertaken in parallel to complete in a reasonable time frame. The results for simulated data are shown in Figure 4.4.

Another challenge when fitting multiple epidemics with discrete start times is the problem of reducing the number of epidemics. If each epidemic is reconsidered and the start time of each epidemic needs to be searched again, and the range is uncertain, in the worst case each epidemic start time will have to be re-searched from the very start. The optimisation of start times provides a solution to this problem as the optimisation techniques aims to find the optimal start times for each epidemic simultaneously.

4.1.2 Initial Decomposition Approach

An initial approach at fitting multiple epidemics starts by fitting one SIR model and adding additional epidemics during the fitting process. At every time point within the fitting, both k and $k+1$ epidemics are fitted to the current data set. If the $k+1$ fit (represented by the RSq statistic $RSq^{(k+1)}$) is "significantly" enhanced (determined by the constant parameter *diff*) then another epidemic is introduced.

The combined set of parameters for each epidemic form the parameters of the overall model, let ($\beta^{(k)}$, $\gamma^{(k)}$) denote the transition rates and ($S_0^{(k)}$, $I_0^{(k)}$ and $R_0^{(k)}$) denote the initial conditions for the k^{th} SIR epidemic model. When fitting multiple epidemics each set of epidemic parameters ($\beta^{(k)}$, $\gamma^{(k)}$and $S_0^{(k)}$) are optimised to produce the optimal fit of the combined epidemic.

Algorithm 5 Epidemic Decomposition Initial Algorithm

```
1:  function incrementalSynthedemicFit(y, diff)
2:      // Set the number of epidemics
3:      k = 1
4:      // Start with one initial epidemic
5:      epidemics = [newEpidemic()]
6:      for (i in 1 :  length(y)) do
7:          // Fit k epidemics
8:          (epidemics, rSq) = optimInRange(sseEpidemics, y[1:i], epidemic, i)
9:          // Fit k+1 epidemics
10:         moreEpidemics = c(epidemics, newEpidemic())
11:         (epidemics', rSq') = optimInRange(sseEpidemics, y[1:i], moreEpidemics, i)
12:         // If the fit is significantly better set k+1 epidemics
13:         if (RSq' - RSq) ≥ diff then
14:             k = k + 1
15:         end if
16:     end for
17: end function
```

After implementing algorithm 5, it was realised that in most cases fitting an additional epidemic will increase the R^2 value as it introduces more parameters and more variability to explain the data. To prevent over-fitting, `k + 1` epidemics are only considered when `RSq` $\leq$ `limit`. This target value can be thought of as the limiting `RSq` value; if the `RSq` deteriorates below `target` then we try and improve the fit by adding another epidemic.

`newEpidemic()` instantiates a new epidemic object that contains the epidemic parameters, initialConditions, type and start time of the epidemic. The initial conditions and parameters are set to values approximately in the middle of the range of expected values. The choice of initial parameters will have an effect on the outcome of the model and even though successful optimisation should result in similar values, sensitivity analysis should be undertaken to assess the sensitivity of the algorithm to the initial conditions, chapter 3.2.1 discusses approaches to determining the initial parameters.

In order to test the discrete time iterative fitting algorithm, a synthetic data set with multiple un-

Algorithm 6 Epidemic Decomposition Second Algorithm

```
1: function incrementalSynthedemicFit(y, diff, limit)
2:     // Set the number of epidemics
3:     k = 1
4:     // Start with one initial epidemic
5:     epidemics = [newEpidemic()]
6:     for (i in 1 :  length(y)) do
7:         // Fit k epidemics
8:         (epidemics, rSq) = optimInRange(sseEpidemics, y[1:i], epidemics, i)
9:         // If the fit has deteriorated try k + 1
10:        if (rSq < limit) then
11:            // Fit k+1 epidemics
12:            moreEpidemics = c(epidemics, newEpidemic())
13:            (epidemics', rSq') = optimInRange(sseEpidemics, y[1:i], moreEpidemics, i)
14:            // If the fit is significantly better set k + 1 epidemics
15:            if (RSq' - RSq) ≥ diff then
16:                k = k + 1
17:            end if
18:        end if
19:    end for
20: end function
```

Algorithm 7 Start Time Search Range with Parallel Implementation

```
1: function optimInRange(sseEpideimcs, y[1:i], epidemics, i)
2:     // Fit in range in parallel
3:     for (i in (i - (range/2)) :  (i + (range/2))) doparallel
4:         // Fit k epidemics
5:         (epidemics, rSq) = optimInRange(sseEpidemics, y[1:i], c(epidemics), i)
6:         // If the fit has deteriorated try k + 1
7:         if (rSq < limit) then
8:             // Fit k+1 epidemics
9:             moreEpidemics = c(epidemics, newEpidemic())
10:            (epidemics', rSq') = optimInRange(sseEpidemics, y[1:i], moreEpidemics, i)
11:            // If the fit is significantly better set k + 1 epidemics
12:            if (RSq' - RSq) ≥ diff then
13:                k = k + 1
14:            end if
15:        end if
16:    end for
17: end function
```

derlying epidemics can be produced using the Gillespie algorithm. Each epidemic is simulated and then their contributions are offset appropriately and added together. Using the Gillespie algorithm, two different SIR epidemics were generated, the second epidemic was then offset by 20 days and the contributions of both epidemics summed. Note that in the initial approach the number of data points was reduced by only taking every other data point from the simulated data. The parameters of the SIR models are as follows,

$$\beta^{(1)} = 0.001,\ \gamma^{(1)} = 0.1,\ S_0^{(1)} = 500,\ I_0^{(1)} = 10$$
$$\beta^{(2)} = 0.002,\ \gamma^{(2)} = 0.2,\ S_0^{(2)} = 600,\ I_0^{(2)} = 10$$

Before fitting the Synthedemic model to the synthetic data set with two SIR epidemics, a single SIR model was applied to the data as shown in Figure 4.3.

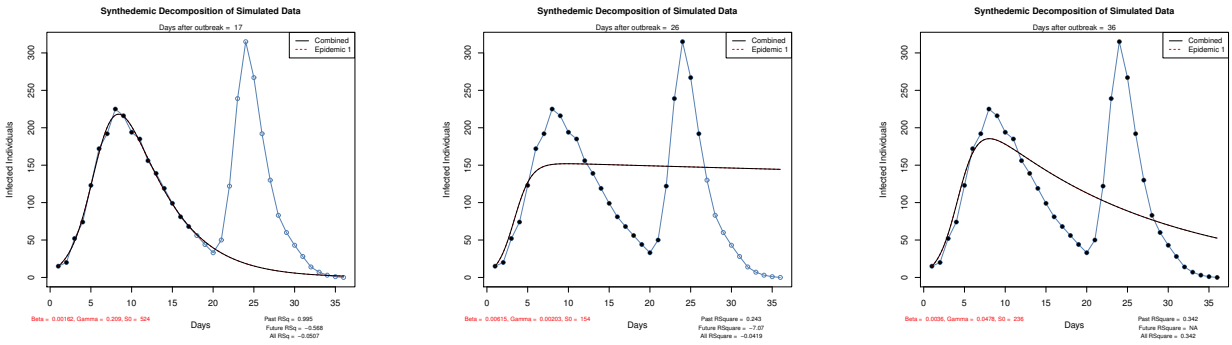


Figure 4.3: Single SIR iterative fitting to synthetic data

The single model is capable of fitting to the start of the synthetic data, as shown at time 17 model is fitting to the first epidemic. However as the contribution of the second epidemic starts at time 26 the single SIR model is clearly incapable of fitting to both underlying epidemics.

Figure 4.4 shows the initial multiple epidemic algorithm results on the data, the epidemic start times are discretely searched within a range around the detected outbreak time. It can be seen that at day 17 the single epidemic model is fitting the first epidemic well with a past R^2 value of 0.995. The algorithm then incorporates a second epidemic on day 22 when the R^2 falls below the limiting value. By day 26 the future prediction of the second epidemic is very accurate with a future R^2 of 0.992 and at the end of the outbreak the fit to the two epidemics has a high R^2 of 0.997 compared to an R^2 of 0.320 for fitting a single epidemic.

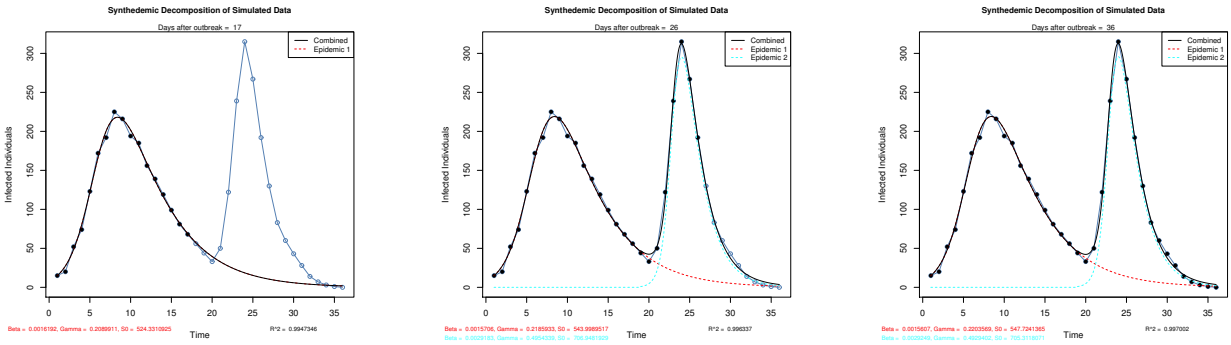


Figure 4.4: Multiple SIR Iterative fitting to Synthetic data

4.1.3 Detecting an Initial Outbreak

To make the above algorithm more generic and enable it to run directly on any dataset, the process needs to be initiated with no epidemics. This is accomplished by fitting a straight line through the initial data points. A buffer of data values with random variation can be appended to the start of the synthetic data to represent the data before any outbreaks. The model then needs to locate the start of the outbreak. Figure 4.5 shows the algorithm detecting the initial start time of the outbreak using residual analysis and a similar approach can be used to detect further epidemic outbreaks. An outbreak is detected using a Z test to determine when the last n residuals become greater than a number of standard deviations away from the mean of the previous residuals [52].

$$Z = abs(\frac{(mean - value)}{StandardDeviation})$$

There are many possible approaches to determining epidemic outbreaks and further considerations to outbreak detection are discussed in Section 4.3.3.

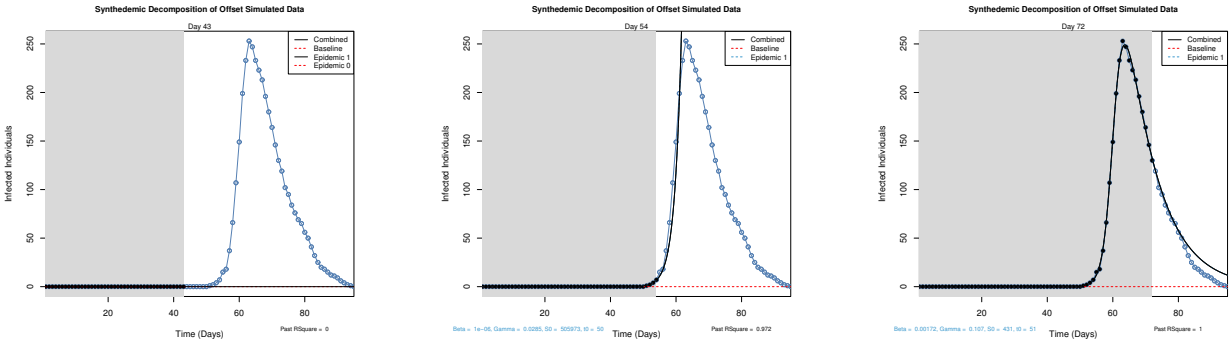


Figure 4.5: Offset SIR Iterative Fitting to a Synthetic data

Figure 4.5 shows a single outbreak detection using the above method. Note that the graphs have

a shaded grey region, this further highlights the current fitting time. Only the black filled data points within the grey region are used within the optimisation.

4.1.4 Synthedemic fitting with Parallel Start Time Search

When searching feasible epidemic start times, each optimisation for each start time is independent. This enables the search to be undertaken in parallel. A parallel implementation enables the full initial synthetic data set to be processed in a reasonable time. The full synthetic data can also be extended by offsetting the initial outbreak. Figure 4.8 shows the iterative algorithm applied to a multiple SIR synthetic data set starting with no epidemics, each sub epidemic has been generated with the initial parameters,

Synthetic 1:
$\beta^{(1)} = 0.001$, $\gamma^{(1)} = 0.05$, $S_0^{(1)} = 400$, $I_0^{(1)} = 1$
$\beta^{(2)} = 0.001$, $\gamma^{(2)} = 0.1$, $S_0^{(2)} = 400$, $I_0^{(2)} = 1$
The first epidemic is offset by 30 days and the second by 50 days from the start of the data.

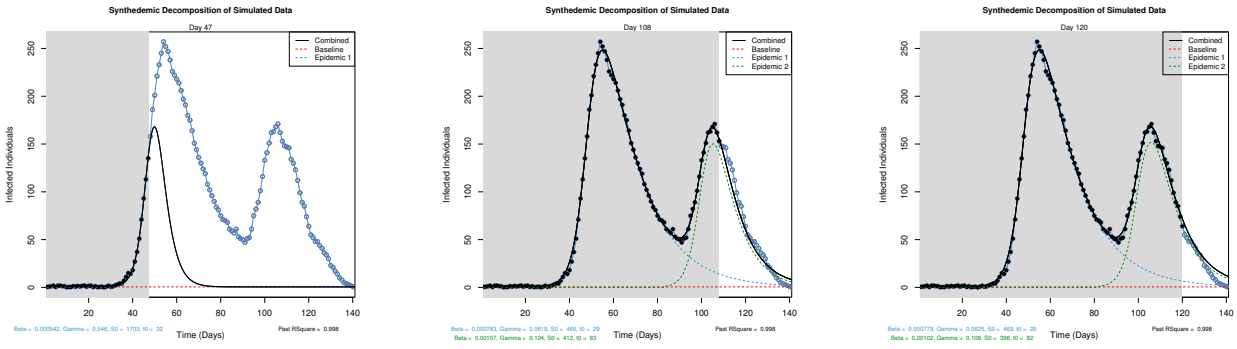


Figure 4.6: Synthedemic Fitting of Synthetic data with Parallel Start Time Search

In order to test the optimisation process further, a synthetic data set with a much higher $S0$ value was also fitted. This showed that the initial value of $S0$ is critical in optimising the epidemic. In particular it is important to ensure that the values of $S0$ is greater than I. This can be achieved by keeping a maximum value of observed data for the current epidemic and bounding the value of $S0$ to be above this value within the optimisation. It is also important to bound other parameters within the optimisation such as upper and lower limits for β and γ and ensuring that $\beta < \gamma$. In order for the epidemic to grow and not die out, it may also important to ensure that $SR0 > 1$ where $R0 = \beta/\gamma$ in a population without vital dynamics [42].

Figure 4.7 shows the synthetic data set with the initial parameters,

Synthetic 2:
$\beta^{(1)} = 0.0001$, $\gamma^{(1)} = 0.05$, $S_0^{(1)} = 4000$
$\beta^{(2)} = 0.0001$, $\gamma^{(2)} = 0.1$, $S_0^{(2)} = 4000$

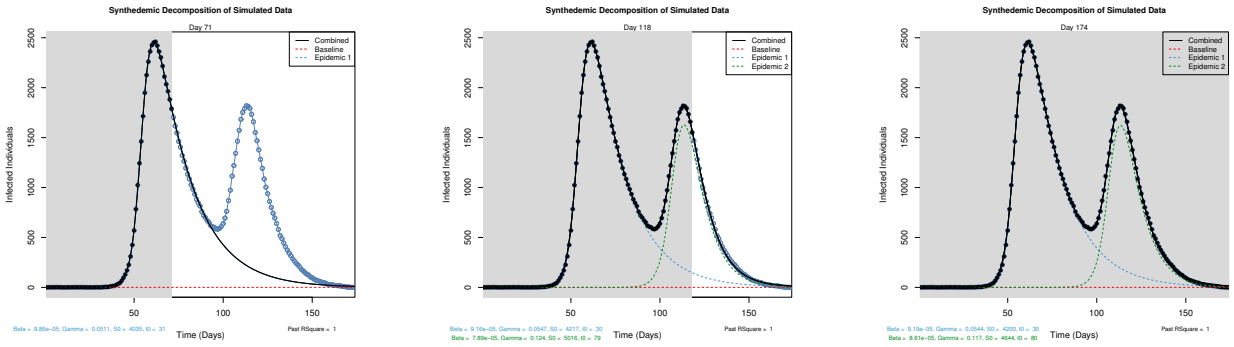


Figure 4.7: Multiple Epidemic Synthetic data with Large $S0$ – Parallel Start Time Search

4.1.5 Optimising Epidemic Start Times

An alternative approach to determine the start times of each epidemic is to pass the start time as a parameter to the optimisation. This approach reduces the overall number of optimisations because all combinations of start times do not need to be optimised. Optimising over the start time also does not require fixing previous epidemic start times as they are all simultaneously optimised. Optimising over the start time also enables the start time to be determined continuously, which is more realistic as it is not confined by the sample period.

A limitation of optimising the start time is that as the number of epidemics increases, the chances of the complexity of the optimisation increases due to the large uncertainty in the range of start times and epidemic parameters. In order to reduce the number of epidemics, a minimum time frame is set between detection of outbreaks. Furthermore, as more epidemics are added the optimisation stage becomes more complex and more likely to produce inaccurate solutions due to non convergence or getting stuck in local minima.

When reducing the number of epidemics, the problem of selecting which epidemic to remove remains, however the ranges of start values can be re-optimised more efficiently than searching all possible start times discretely.

4.2 Logistic Optimisation

Optimising the start time of an epidemic with unknown parameters is non trivial due to the large number of parameters. One technique that may provide a significant advancement in the start start time optimisation is using the Logistic Transform. It is common to use a log transform of parameters such as β and γ to prevent the optimisation making them negative and the logistic function can extend on this to bound the parameters [50].

4.2.1 Synthedemic Fitting with Logistic Time Optimisation

Further to specifying upper bounds, lower bounds can also be specified by using the logistic function,

$$f(x) = \frac{x_{max} - x_{min}}{1 + e^{-x}} + t_{min}$$

Where x_{min} is the lower bound and x_{max} is the upper bound of the parameter x. This function transforms values in the range $(-\infty, +\infty)$ into the range (x_{min}, x_{max}) because as $x \to -\infty, f(x) \to xmin$ and as $x \to \infty, f(x) \to x_{max}$. This function can be used to ensure that parameter values are optimised within a specific range which can be applied in the optimisation for the epidemic parameters and the start time of the epidemic. This technique is especially useful in restricting the range of optimisation of the epidemic start time to prevent optimisation over time from getting stuck in local minima, enabling the time and parameters to be optimised simultaneously.

In order to apply the logistic function to enhance the optimisation, upper and lower bounds on the parameters need to be obtained. For obtaining bounds on the start time of the epidemic, it is possible to apply upper bounds at the time when the epidemic is first detected and a lower bound in a region before this point. This is similar to the method adopted for discrete start time optimisation and produces a window of possible time values for the optimisation and narrows down the search. The main difference is that the optimisation process can now simultaneously optimise over all epidemic start times within their respective ranges, algorithm 8 shows the iterative outbreak detection with start start time optimisation. In addition to the optimisation over epidemic start times, this algorithm also commences the fitting process with no epidemics and detects outbreaks using the residual approach, furthermore it can also reduce the number of epidemics if the R^2 remains sufficient to ensure the model is as parsimonious as possible.

4.2.2 Synthedemic Fitting with Logistic Time Optimisation using Synthetic Data

The logistic start time optimisation synthedemic algorithm can be applied to the previous synthetic dataset with two SIR epidemics with parameters,

Algorithm 8 Synthedemic Fitting with Logistic Time Optimisation

```
1: function synthedemicLogisticFit(y, diff, limit)
2:     // Set the number of epidemics
3:     k = 0
4:     // Start with no epidemics
5:     epidemics = []
6:     for (i in 1 :  length(y)) do
7:         // Fit k epidemics
8:         (epidemics, rSq) = optim(sseEpidemics, y[1:i], epidemics)
9:         // If the fit is sufficient try k − 1
10:        if (rSq > limit) then
11:            lessEpidemics = epidmeics[1:(k-1)]
12:            (epidemicsLess, rSqLess) = optim(sseEpidemics, y[1:i], lessEpidemics)
13:            // If the fit is sufficient, set k − 1 epidemics
14:            if RSqLess ≥ limit then
15:                k = k - 1
16:                epidemics = epidemicsLess
17:            end if
18:        end if
19:        // If the fit has deteriorated try k + 1
20:        if (rSq < limit) then
21:            moreEpidemics = c(epidemics, newEpidemic())
22:            (epidemicsMore, rSqMore) = optim(sseEpidemics, y[1:i], moreEpidemics)
23:            // If the fit is significantly better, set k + 1 epidemics
24:            if (RSqMore - RSq) ≥ diff then
25:                k = k + 1
26:                epidemics = epidemicsMore
27:            end if
28:        end if
29:    end for
30: end function
```

$\beta^{(1)} = 0.0001$, $\gamma^{(1)} = 0.05$, $S_0^{(1)} = 4000$
$\beta^{(2)} = 0.0001$, $\gamma^{(2)} = 0.1$, $S_0^{(2)} = 4000$

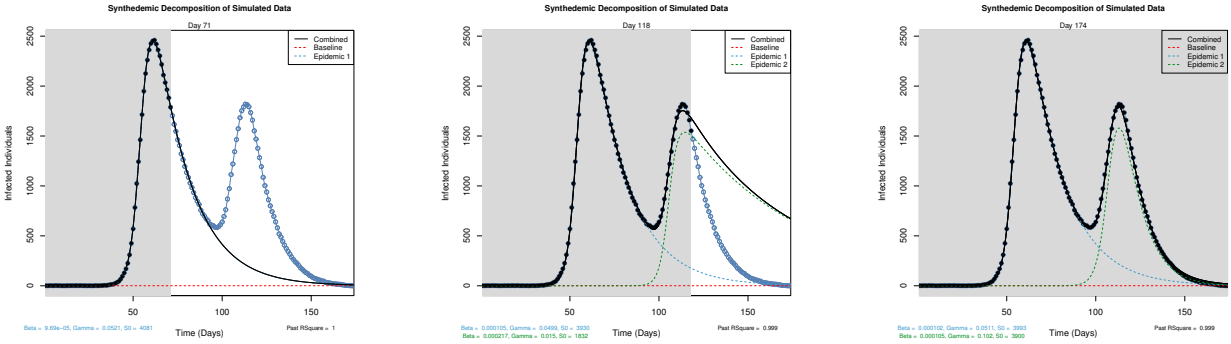


Figure 4.8: Synthedemic Fitting with Logistic Time Optimisation for synthetic data

The logistic start time optimisation enables the optimisation of time more accurately as it uses continuous time, the final time fitting shows a very high R^2 as a result. However the optimisation of the start time at the start of the epidemic when uncertainty in the other parameters are high means that the approach is not as stable throughout the fitting as the iterative fitting.

The start of each epidemic has to be offset in order to evaluate the multiple epidemic model. This requires the start times to be discretised to enable an array of zeros to be constructed and appended to the start of each epidemic. At a coarse gain discretisation of for example 1 day, the epidemic start time is truncated to the nearest day. Figure 4.9 shows the difference in the SSE function over the start time of the first epidemic and it can be seen that when a coarse gain time is used, the SSE has structural breaks with values clustered at discrete levels creating a rugged optimisation surface. This could result in unexpected behaviour within the optimisation due to non smooth gradients, however by using a more fine grained start time discretisation the SSE function can be made smoother. Although this makes the optimisation more robust due to the smoother optimisation surface, it also means that the optimisation stage takes longer to compute.

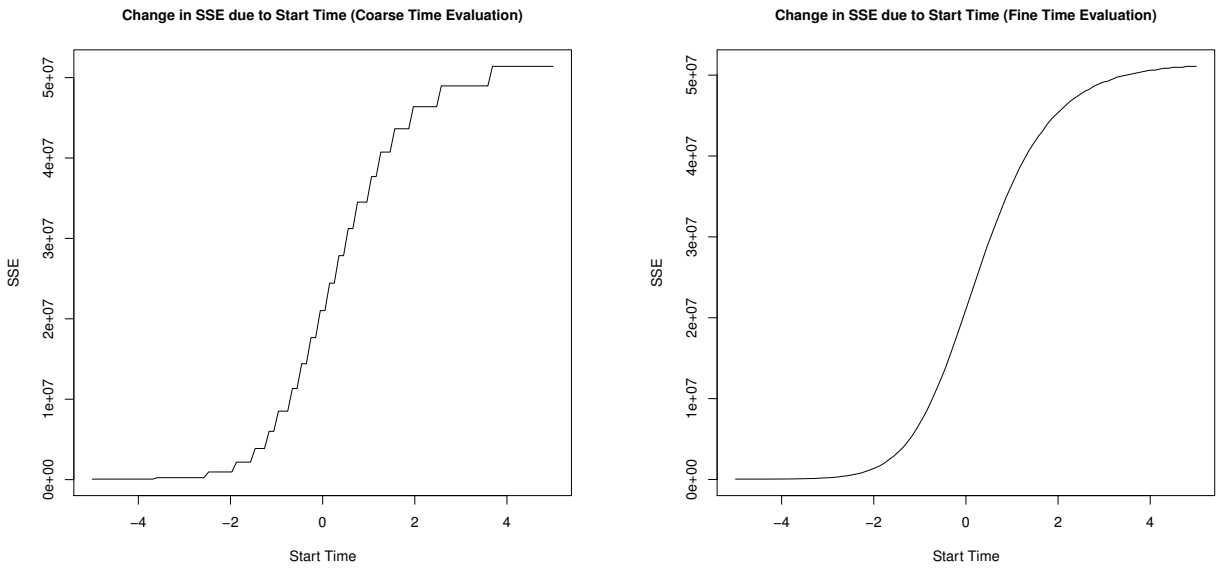


Figure 4.9: SSE optimisation over different granularity start times

Furthermore, after plotting the SSE graphs, it was realised that the logistic transform can have an unforeseen effect on the optimisation. If the minimum start time for the transformation is too large then regardless of how small the optimisation makes t the start time is bound by the minimum value. Therefore as shown in 4.9 with a minimum value of $t - 20$ the optimisation may not be able to reduce the epidemic start time past the minimum limit and so the SSE value does not start to increase as the start time decreases. This could cause the optimisation to make t very small as it tries to explore the shallow gradient and it is more stable to increase the range so that the SSE function forms a "well" for the optimisation to find the minimum. Figure 4.10 shows an updated SSE graph with a lower minimum start time.

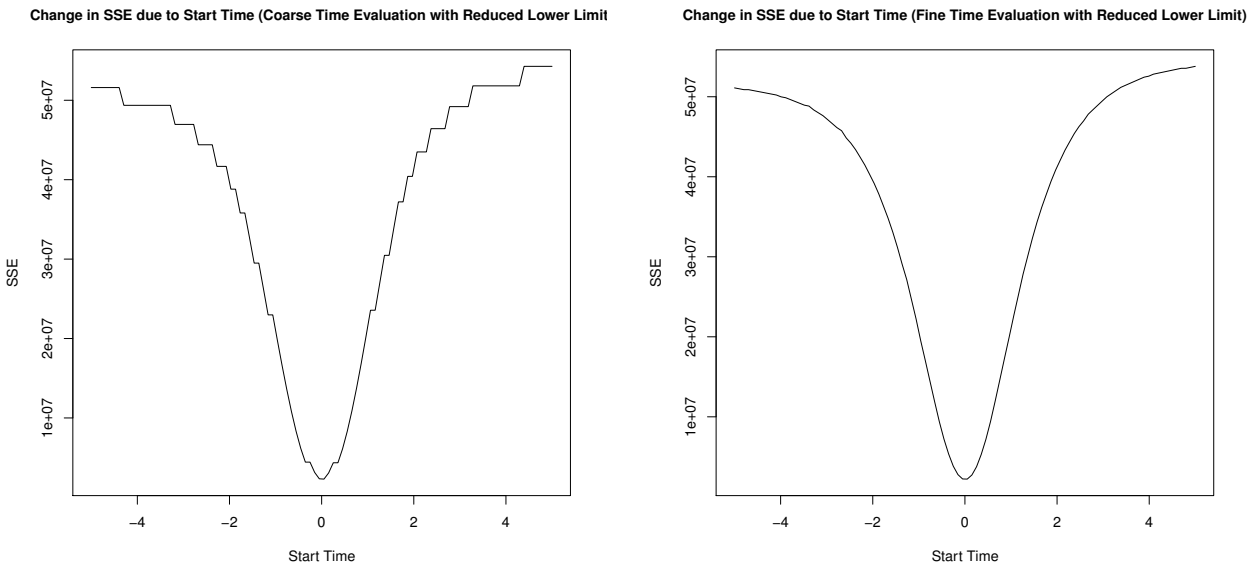


Figure 4.10: SSE optimisation over different granularity start times with a Lower Start Time Limit

(Note that the scale on the graphs is a logistic scale, so even though it appears that the start times are the same as the graph above, the minimum value is lower and the range of values is increased.)

4.2.3 Alternative approaches

An alternative approach is to initially have a high number of epidemic outbreaks within the model. At each discrete time value, an outbreaks is *seeded* to start at this time. During the fitting the epidemics with the optimal start times are then *grown* by the optimisation.

Another approach is to consider optimising over t0 and the epidemic parameters in different calls to `optim`, so that for each change in t0 values, the epidemic model parameters are optimised. This could have an impact on the way that the parameter space is searched by the optimisation and lead to different final parameter optimisations. This will also enable different t0 optimisation procedures to be used such as Simulated Annealing. Although this idea may be possible for a small number of epidemics, as the number of epidemics increases, the time for each optimisation becomes too large because in each stage of start time optimisation, the optimisation over the parameter space is undertaken. With more epidemics both the number stages in the start start time optimisation increases and the number of stages to optimise the increased number of epidemic parameters within each of these stages increases leading to an infeasible optimisation.

4.3 Synthedemic Model Selection

With the observation of multiple epidemic phenomena constituting an overall epidemic outbreak, it has also been noticed that the underlying processes may have different fundamental properties. This leads to the problem of classifying different types of sub epidemic types which is detailed in this section.

4.3.1 Different Types of Epidemic Outbreaks

Recently epidemic models have been applied within the context of social media outbreaks and a key observation is that different *types* of multiple epidemic outbreaks occur. It has been observed that there are gradual growth SIR type epidemics caused by social event media spreading throughout social networks at a relatively slow diffusion rate, and extremely rapid mass media events that cause spikes in media activity. For example when a tweet is posted by a popular Twitter user or an artist appears live on a show with a large audience the number of "Infectious" individuals within the social network increases significantly almost instantaneously. Previous research has characterised many different types of "temporal shapes" [47] and "rise and fall patterns" [48] within social media outbreaks. But the number of patterns used ranges from one to six and the actual number remains to be studied further.

More recently, an article has used the terms *Growth* and *Spike* to classify the gradual and rapid types of outbreak. The key characteristics of each types of outbreak as shown in Figure 4.11.

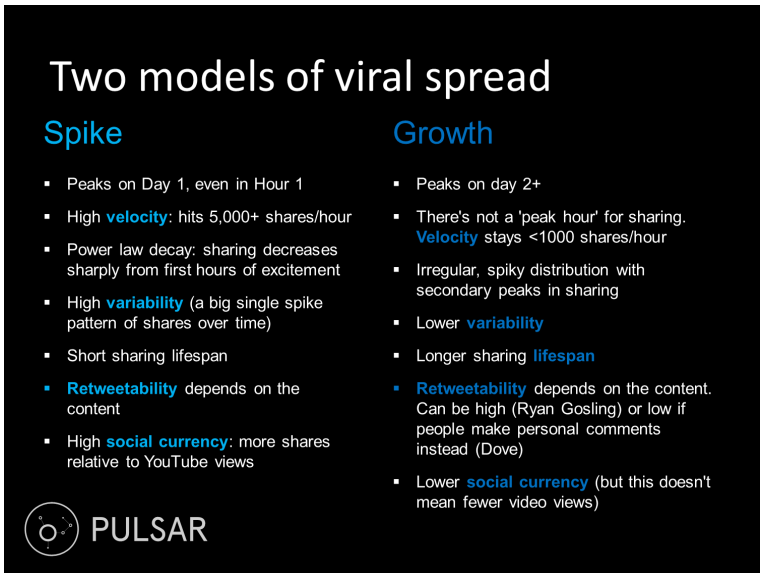


Figure 4.11: Classifying Outbreaks in Social Media Networks

Critically, the *Growth* and *Spike* epidemic classifications were developed at the same time as this

paper.

4.3.2 Epidemic Types

The decomposition of an overall outbreak into single epiemic models developed in the section 4.1 can be extended to enable different types of outbreaks inspired from the observed outbreak types. Theoretical analogues of these can be proposed using an SIR compartmental model to represent the gradual growth of event exposure and an exponential model consisting of an initial impulse followed by exponential decay to represent the rapid outbreak of mass media events.

Reflecting the natures of content diffusion discussed, the Synthedemic model is composed as a superposition of *spike* and *growth* type sub-epidemics.

Let M be the class of sub epidemic models that we are considering, in general this set can contain any type of epidemic models, in our methodology we consider two types; SIR and exponential decay models. At each time point we consider adding another sub epidemic model from the class M in order to improve the current model consisting of k epidemics. Let $M = \{\boldsymbol{f}_1^{(i)}, \boldsymbol{f}_2^{(i)}\}$ where $\boldsymbol{f}_1^{(i)}$ and $\boldsymbol{f}_2^{(i)}$ denote the following SIR and Exponential decay models for the i^{th} epidemic,

- SIR model $\boldsymbol{f}_1^{(i)}(\boldsymbol{\theta}^{(i)}, t)$ with parameter vector $\boldsymbol{\theta}^{(i)} = [I_0^{(i)}, S_0^{(i)}, \beta^{(i)}, \gamma^{(i)}]$

$$\begin{aligned} \frac{dS}{dt} &= -\beta IS, \\ \frac{dI}{dt} &= \beta IS - \gamma I \\ \frac{dR}{dt} &= \gamma I \end{aligned}$$

- Exponential decay model $\boldsymbol{f}_2^{(i)}(\boldsymbol{\theta}^{(i)}, t)$ with parameter vector $\boldsymbol{\theta}^{(i)} = [I_0^{(i)}, \gamma^{(i)}]$

$$\frac{dI}{dt} = -\gamma I$$

The parameter t enables the evaluation of the model at a specific time, giving the number of infected individuals at time t. $S_0^{(i)}$ and $I_0^{(i)}$ are the number of initial Susceptible and Infectious individuals and the parameters $\beta^{(i)}$ and $\gamma^{(i)}$ are the infection rate and recovery rate for the i^{th} epidemic.

Let the current number of sub-epidemics within the model be k, the combined epidemic model value at time t is then represented as the sum of the evaluated sub-epidemics,

$$\hat{y}(\boldsymbol{\theta}, t) = \sum_{i=1}^{k} \boldsymbol{f}^{(i)}(\boldsymbol{\theta}^{(i)}, t)$$

The general optimisation is the finding the optimal parameter vector $\boldsymbol{\theta}_t$ such that the Sum of Squared Error is minimised,

$$\boldsymbol{\theta}_t = \arg\min_{\theta} \sum_{i=1}^{k} (\hat{y}(\boldsymbol{\theta}, t) - y_{1:t})^2$$

Considering further the relationship between the SIR epidemic model and the exponential it can be seen that the SIR model reduces to an Exponential model when $\beta = 0$, and furthermore the Exponential collapses to a constant linear relationship when $\gamma = 0$, this hierarchy is particularly interesting because it shows that the models used are all sub types of the SIR model.

4.3.3 Outbreak Detection

In Section 4.1.3, an approach for detecting the initial outbreak was presented. The method detected an initial outbreak by fitting linearly at the start of the dataset. This can be considered as fitting an exponential model at the start of the fitting, with parameter γ=0. The residual analysis approach to outbreak detection can be extended in order to detect future outbreaks. One method to detect epidemic outbreaks is to track the standard deviation and mean of previous residuals and use them to formulate limits for future residuals. If a recent residual is more than 2 standard deviations away from the mean of the past residuals, then it indicates that the data has shifted away from the pattern observed in the past and may resemble the start of an outbreak. Furthermore, the magnitude of the distance of a residual from the mean may give insight into the type of the outbreak. For example if the residual exceeds 6 standard deviations from the mean then the epidemic may be a sudden "spike" epidemic.

When implementing this approach it is necessary to analyse a window of the most recent residuals (for example the two last residuals) to prevent epidemics being detected due to noise in the data. However this results in delayed detection, because for a window size n, all of the n residuals must be observed before an outbreak can be determined. Furthermore, this approach to outbreak detection may be dependent on the noise within the data, for example with synthetic data there is little noise in the data and the standard deviation is low, so even a small change in the residuals can be many standard deviations away from the mean and the wrong type of epidemic may be selected. Another approach to determining the type of the epidemic may be to detect a given outbreak by residual analysis and subsequently fit all possible epidemic types and take the type that best fits the data.

An alternative method of determining the epidemic start times and number of epidemics may be approached using gradient analysis of the actual data. A linear model can be fitted to a sliding window of data points and the point of maximum increase over the data set may give an indication of the time of the outbreak. This method may be particularly useful in conjunction with the optimisation of the start time to obtain estimates or upper bounds of each epidemic start time as this would enable all time points of a complete data set to be fitted simultaneously - without the need to iteratively determine the start times. As shown in Figure 4.12 a gradient based approach can determine the outbreak times. The method used to produce figure 4.12 uses a sliding window of 20 data points and locates the maximum gradient increase (from a positive gradient). Often there are multiple maximum increases around the same point and so it is required to take more than the required number of samples from the highest gradient increases and then only use samples with a time difference greater than 20 epochs. Alternatively clustering the highest increases of many samples may enable both the times of the outbreak and the number of outbreaks. The number of outbreaks was provided to this algorithm to determine their start times, however further complications arise when detecting outbreaks in different size data sets with unknown numbers of epidemics. The cumulative frequency plot may provide additional insight into the outbreak detection however this remains to be analysed.

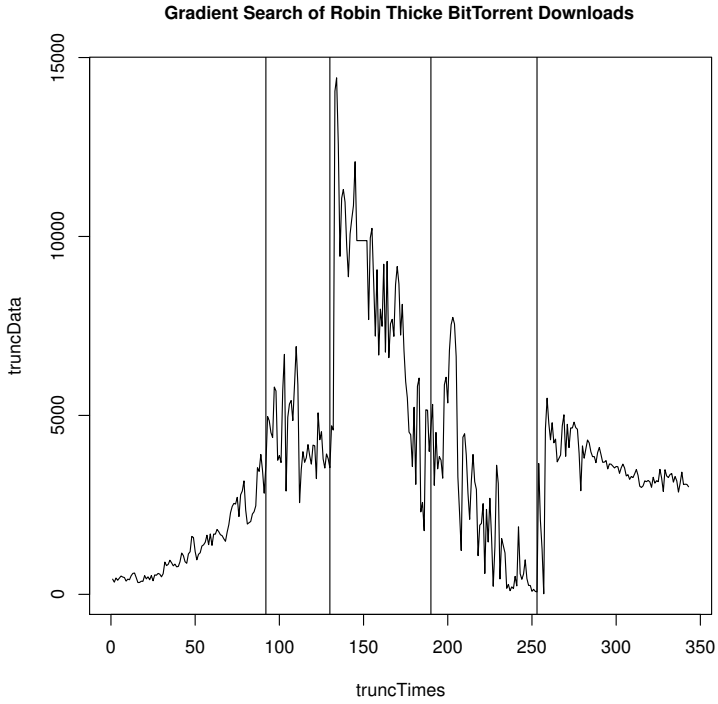


Figure 4.12: Gradient Search to determine epidemic start times

4.3.4 Synthetic Data with Multiple Epidemic Types

In addition to the synthetic generation of SIR epidemic data, an exponential *Spike* can be simulated using the Gillespie algorithm. Figure 4.13 shows the iterative algorithm with parallel time search applied to a multiple SIR synthetic data set, each sub epidemic has been generated with the initial parameters, Synthetic 3:
$\beta^{(1)} = 0.001$, $\gamma^{(1)} = 0.05$, $S_0^{(1)} = 400$
$\gamma^{(2)} = 0.2$, $S_0^{(2)} = 300$
The first epidemic is offset by 10 days and the second by 35 days from the start of the data.

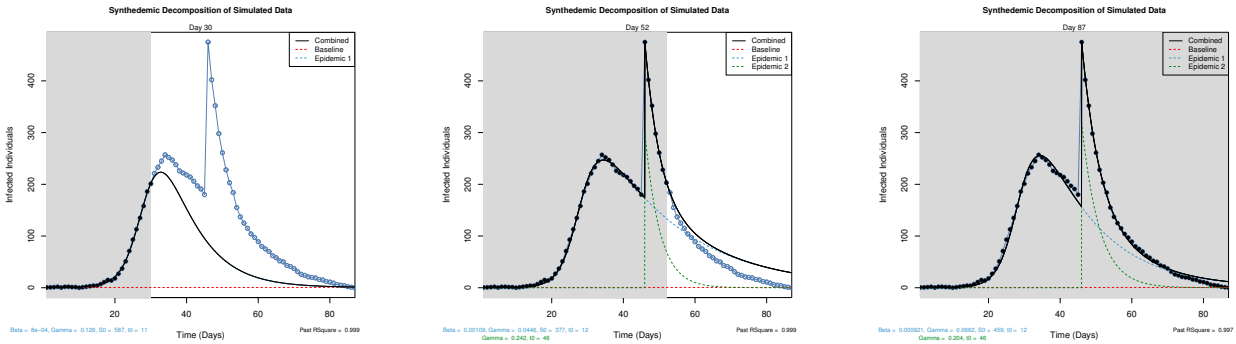


Figure 4.13: Single SIR iterative fitting to synthetic data

4.4 Residual Refinement

4.4.1 AR Model

The Synthedemic detects the underlying structure of outbreaks within the data. In order to capture any remaining variability in the data, an autoregressive model is fitted to the residuals of the multi epidemic model. The residuals from the epidemic model predictions are used to determine the order of the autoregressive model and then an autoregressive model is then used to refine the residuals.

4.4.2 Synthedemic Autoregressive Modelling

Figure 4.19 shows the results of applying an autoregressive model to the Synthetic dataset presented in section 4.1.4. The AR model order is determined using the ACF and PACF analysis shown in figure 4.15. The model uses the previous residuals from the epidemic model to refine the next prediction for $t+1$. It is clear that the AR model does not offer a significant improvement over the synthedemic fitting, this may be due to the low level of noise within the synthetic data set.

4.4.3 Residual Analysis

The residuals for the Synthetic data set are shown in Figure 4.14. The residuals are relatively small at the start of the fitting, up to approximately time 30 when the first epidemic outbreak starts. The residuals are not stationary, however it may still be worthwhile fitting an AR model to the data in hope of achieving a better fit. It may also be required to analyse relative residuals or use differencing to ensure the residuals are stationary. A more detailed comparison of the enhancement of the AR residual refinement is undertaken in the Evaluation chapter 5.1.3.

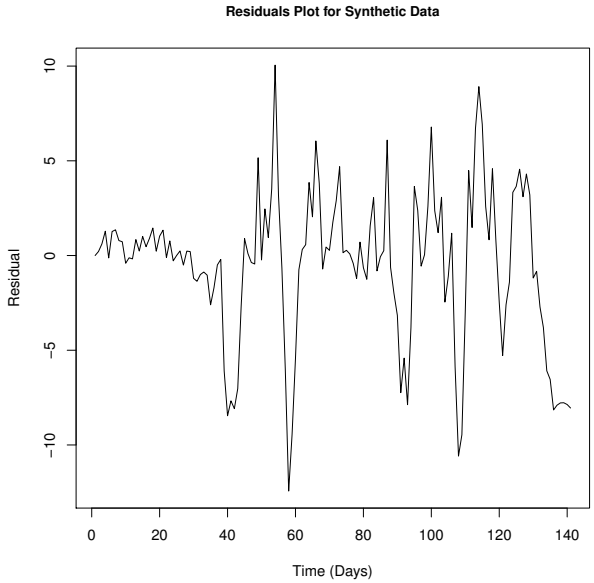


Figure 4.14: Residuals of the Synthedemic Fitting for Synthetic data

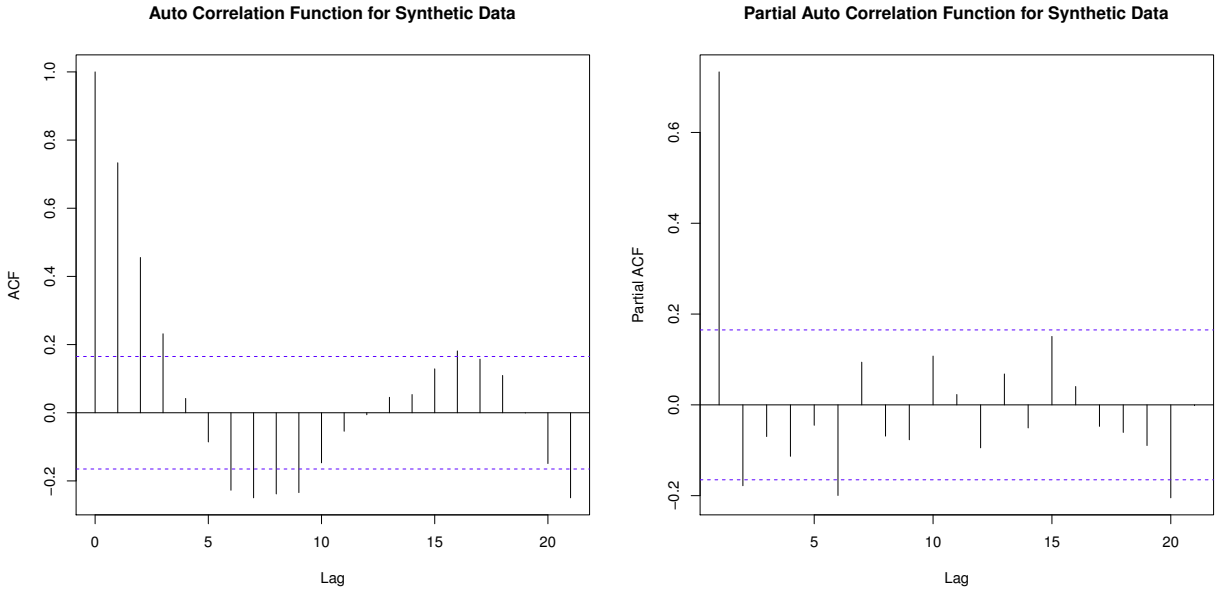


Figure 4.15: Synthedemic Residuals ACF and PACF Plots for Synthetic data

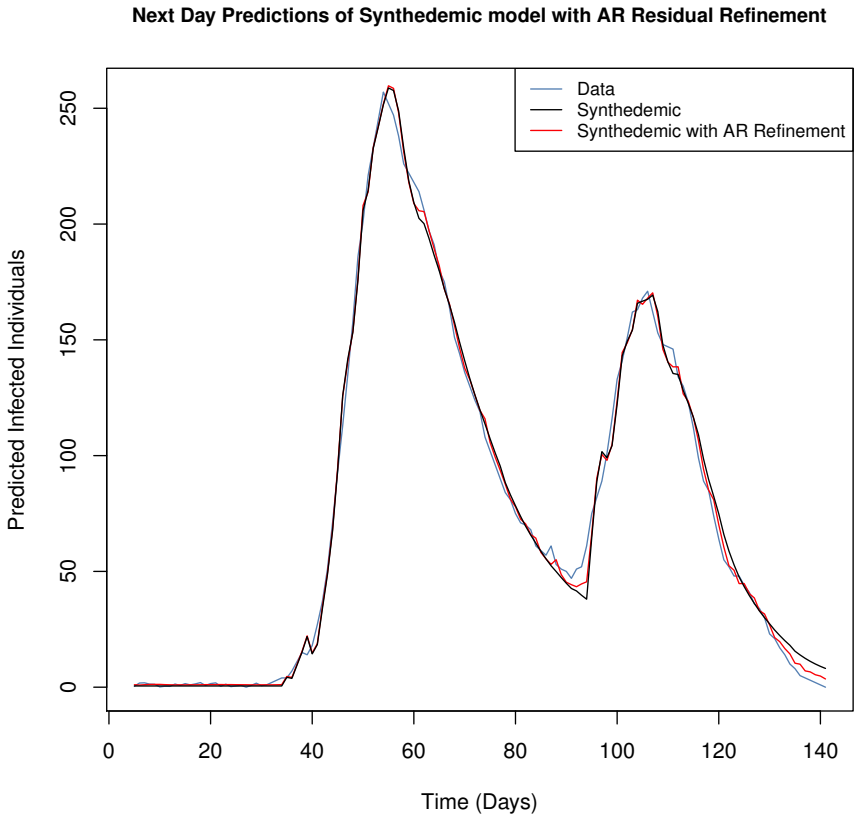


Figure 4.16: Synthedemic fitting with AR Residual Refinement for Synthetic data

4.5 Synthedemic Model

The Synthedemic model fitting procedure is summarised below.

4.5.1 Details of the Synthedemic Algorithm

Key details of the synthedemic algorithm:

- A target proportion of explained variance, R^2_{target}, is passed as a parameter to the fitting.
- The procedure starts with no epidemics.
- At each time point, the multiple epidemic model is optimised.
- If the R^2 is above the R^2_{target}, then we consider dropping an epidemic.
- If the R^2 is below the R^2_{target}, then we consider adding an epidemic.

The optimisation over the multiple epidemic model uses the auxiliary function `evalEpidemics` to evaluate the epidemic using the current epidemic parameters. This enables the Sum of Squared Error to the actual data to be calculated via the function `sseEpidemics` which is used as the objective function of the optimisation.

At each time point, the start time of the most recent epidemic is determined by searching every feasible start time within a range. This involves re-optimising the epidemic for each feasible start time (in parallel). To prevent the algorithm from being infeasible to compute (in particular avoiding the large number of possible start time combinations) the following heuristics are used:

- Only the latest added epidemic start time is searched within a range.
- At most one epidemic is added or subtracted at any stage.
- If an epidemic is added, its type is determined by the previous residuals.
- SIR-type processes start with a single infected individual.
- Spike-type process start with an initial infected individual estimated from the difference between the predicted and observed data point at the time of the outbreak.

The following pseudo-code describes the algorithm in more detail.

Algorithm 9 Multiple Epidemic Optimisation

```
1: function incrementalSynthedemicFit(y, targetRSq)
2:     // Start fitting with no epidemics, k tracks the number of epidemics
3:     k = 0
4:     // Set the minimum number of data points for fitting an epidemic
5:     minFitSize = 4
6:     epidemics = []
7:
8:     for (i = minFitSize to length(y)) do
9:         [epidemics, rSq] = fitEpidemics(y[1:i], epidemics, i)
10:        if ((rSq > targetRSq) && (k > 0)) then
11:            [epidemicsLess, rSqLess] = fitEpidemics(y[1:i], epidemics[1:(k - 1)], i)
12:            if (rSqLess > targetRSq) then
13:                epidemics = epidemicsLess
14:                rSq = rSqLess
15:                k = k - 1
16:            end if
17:        end if
18:        epidemicType = determineEpidemicType(y, epidemics, startTime)
19:        if ((rSq < targetRSq) && (epidemicType != NONE)) then
20:            initialEpidemicsMore = updateParameters(epidemics, epidemicType)
21:            [epidemics, rSq] = fitEpidemics(y[1:i], initialEpidemicsMore, i)
22:            k = k + 1
23:        end if
24:        ŷ[i + 1] = makePrediction(y, epidemics, i)
25:    end for
26:    return ŷ
27:
28: end function
```

At each stage within the multiple epidemic fitting, the latest residuals are checked for signs of a new outbreak. The mean and standard deviation of the residuals from the current epidemic are used to determine outbreaks. Once an outbreak is triggered and the type is set by the `determineEpidemicType` function, and a new epidemic is added to the model, if this improves the current R^2 statistic then the new epidemic is included into the model.

A significant challenge when optimising multiple epidemics is the unknown start times $t0$ of each epidemic. For spike epidemics the $t0$ value is assumed to be at the point where the outbreak is detected due to the sudden impulse increase in exposure. Whereas for the growth epidemic the range of possible $t0$ values is broader, and as a result a window of feasible $t0$ values needs to be explored. This increases the computational complexity of the optimisation process because for each new data point, the multiple parameter model needs to be optimised repeatedly. It was found that exploring (in parallel) epidemic start times within a window around the time of the outbreak detection reduced the time taken and did not affect the results (in comparison to fitting over all possible start times).

In order to avoid the explosion in time taken for searching all possible combinations of all sub epidemic start times, only the start time of the most recent epidemic is explored in the current

methodology.

To prevent over fitting, at each stage $k-1$ epidemics are optimised. The current methodology only considered removing the final epidemic due to the large number of possible combinations of start times. A more comprehensive approach would provide a method of determining the start times for each epidemic more efficiently than linearly searching all feasible values, enabling enhanced re-fitting of the model when adding and removing epidemics. An alternative approach may include the start times of each epidemic into the optimisation procedure. Initial experiments show that this method of optimisation may be possible however is more unstable.

Algorithm 10 Fit Epidemics

```
1: function fitEpidemics(y, epidemics, i)
2:     allOptimEpidemics = []
3:     all$R^2$ = []
4:     // Fit Epidemics in parallel over time range
5:     for (t in [epidemics.times[length(epidemics) - 1] :  i]) do par
6:         epidemics' = epidemics
7:         epidemics'.times = c(epidemics.times, t)
8:         optimEpidemics = optimise(sseEpidemics, y, epidemics')
9:         // Store optimised epidemics at this start time
10:        allOptimEpidemics[t] = optimEpidemcis
11:        all$R^2$[t] = rSquare(optimEpidemics, y)
12:    end for
13:    // Find optimal start time
14:    optim$R^2$ = max(all$R^2$)
15:    optimEpidemics = allOptimEpidemics[maxIndex(all$R^2$)]
16:    return [optimEpidemics, optim$R^2$]
17: end function
```

`sseEpidemics` is the objective function called by `optim`. It uses `evalEpidemics` to get the predictions of the model and then computes the sum of squared error from the actual data `y`.

Algorithm 11 Determine epidemic type

```
1: function determineEpidemicType(y, epidemics, startTime)
2:     nResiduals = 2
3:     residuals = evalEpidemics(y, epidemics) - y
4:     epiType = NONE
5:
6:     if (resLength > (nResiduals + 1)) then
7:         inRangeResiduals = residuals[startTime :  (length(residuals) - nResiduals)]
8:         // Calculate Mean and Standard deviation of residuals
9:         meanOfResiduals = mean(inRangeResiduals)
10:        sdOfResiduals = sd(inRangeResiduals)
11:        // Limit for different epidemic types
12:        SIRLimit = meanOfResiduals + (2 * sdOfResiduals)
13:        EXPLimit = meanOfResiduals + (6 * sdOfResiduals)
14:        // Set types
15:        SIRResidual = residuals[length(residuals)]
16:        EXPResidual = min(residuals[(resLength - 2) :  length(residuals)])
17:        if (EXPResidual > EXPLimit) then
18:            epyType = EXP
19:        else if (SIRResidual > SIRLimit) then
20:            epiType = SIR
21:        end if
22:    end if
23:    return epiType
24: end function
```

The function, `evalEpidemics`, is called within the optimisation of the multiple epidemic model. It is required to evaluate the epidemic model at the current parameters within the optimisation. Using the parameters passed to the function, the number of infected individuals at each time point is predicted. These infected values are then used to compute the sum of squared error for the current parameters – which is used as the objective function for the optimisation. The optimised parameters for each sub epidemic are obtained and used to calculate their contributions, the contributions are then offset and added together to give the overall multiple epidemic model.

Algorithm 12 Evaluate epidemics

```
1: function evalEpidemics(y, epidemics)
2:     // Initialise multiple epidemic predictions
3:     multiplePredictions = []
4:     // Initialise I0
5:     I0 = epidemics[1].params
6:     // Evaluate epidemic model at current parameters
7:     for (i = 1 to k) do
8:         // Get parameters for current epidemic
9:         currentEpidemic = epidemics[k]
10:        currentParams = currentEpidemic.params
11:        // Set predicted I0 for epidemic
12:        currentParams[1] = I0
13:        predictedInfectious = ode(currentParams, currentEpidemic.type)
14:        // Offset predictions for epidemics
15:        startOffset = zeros(currentEpidemic.startTime)
16:        offsetPredictions = c(startOffset, predInfectios)
17:        // Include current offset epidmeic into multiple epidemic prediction
18:        multiplePredictions = multiplePredictions + offsetPredictions
19:        // Set I0 of next epidemic
20:        if (i < k) then
21:            nextStartTime = epidemics[i + 1].startTime
22:            nextStartInfectious = predInfectious[nextStartTime]
23:            // Set I0 as difference between data and prediction
24:            I0 = max(data[nextStartTime] - nextStartInfectious, 1)
25:        end if
26:        return multiplePredictions
27:    end for
28: end function
```

After the multiple epidemic model has been optimised to determine the outbreaks within the data, an autoregressive model is fitted to capture the remaining variability in the data. This process first requires the order of the AR model to be determined. The function `getAROrder()` uses autocorrelation to find the most significant order for the AR model. The offset `predOffset` is set as the offset for predictions of the model. The fitted epidemic model at `prevOffset` in the past is used to obtain the prediction for the current time and the AR model is used to augment this prediction. Evaluations of the models predictive ability are later undertaken using predictions `prevOffset` in the future.

Algorithm 13 AR Model Fitting

```
1: function makePrediction(y, epidemics, i)
2:     // Offset of future predictions
3:     epidemicPredictions = evalEpidemics(y, epidemics)
4:     residuals = y - epidemicPredictions
5:     // Get AR model order
6:     arOrder = determineAROrder(residuals)
7:     // Fit AR model to residuals
8:     arModel = ar(residuals, arOrder)
9:     // Use AR model to predict perturbation to prediction
10:    nextResidual = predict(arModel, 1)
11:    return (epidemicPredictions[i + 1] + nextResidual)
12: end function
```

The final stage of the modelling technique is to analyse the fitted multiple epidemic AR model. This enables comparison with benchmark fitting procedures such as a single epidemic model. In order to quantitatively assess the *goodness* of fit and predictive ability of the model, a range of statistical tests are undertaken including comparing the rSquare over future predictions at the given offset

4.5.2 Synthedemic Decomposition Results

In the previous chapters, results of the Synthedemic Algorithm applied to synthetic data sets have been presented. The synthedemic Algorithm has also been applied to a range of real world data sets including the 2009 H1N1 Influenza outbreak in England and online music downloads from Music Metrics[1].

2009 H1N1 Data The 2009 H1N1 viral outbreak, also known as *Swine flu*, hit the UK in May 2009. Infectious cases fell rapidly during the first week of August when the schools closed for summer holidays. From the start of September cases began to rise again, coinciding with the normal flu season. This created a second peak of infection at the end of October. Figure 4.17 shows the synthetic algorithm applied to the outbreak data obtained from the Health Protection Agency. The model is capable of capturing the second outbreak of the flu during October and predicts the second peak infection rates with a high R^2 of 0.945 as shown.

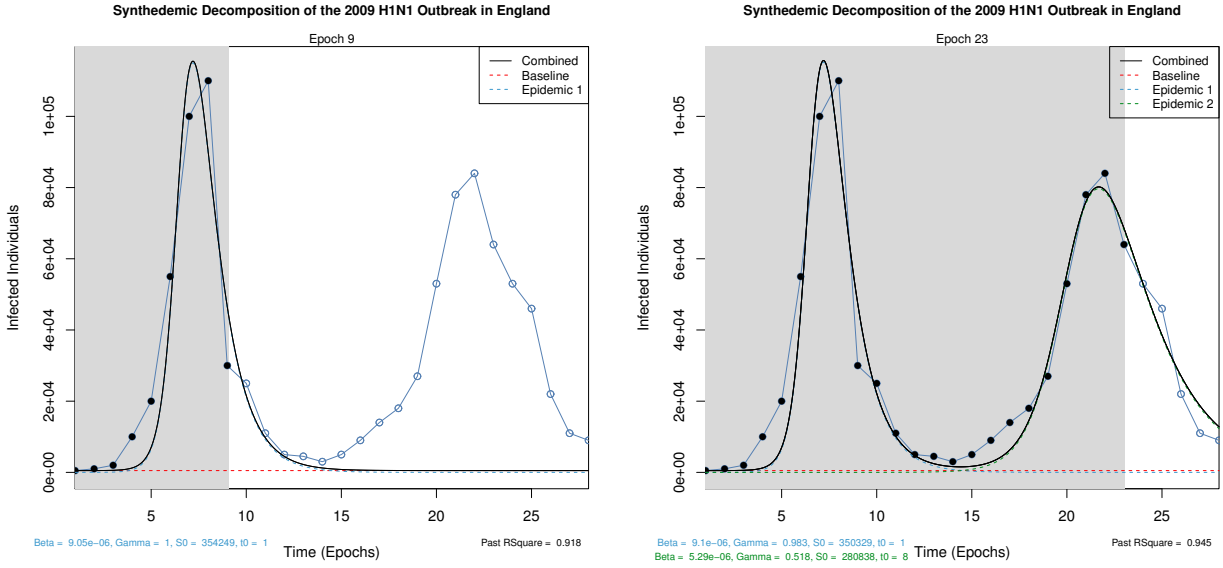


Figure 4.17: Synthetic fitting to 2009 H1N1 Outbreak

Robin Thicke Blurred Lines Music Downloads Robin Thickes Blurred Lines was released in March 2013 and sold over 10 million copies within 30 weeks, making it one of the fastest downloaded song in digital history. As seen in Figure 4.18, on day 101 the model has two sub-epidemics and has a high R^2 of 0.911, furthermore the future decay of the current epidemic is predicted accurately. On day 133 a sudden peak in the data is observed, which was caused by Robin Thickes performance of Blurred Lines on the US TV show *Jimmy Kimmel Live*. This is captured very quickly by the model

[1] www.musicmetric.com

by a *Spike* epidemic. On day 208 the model has already observed that a new epidemic has started, corresponding to the infamous live performance of Blurred Lines by Robin Thicke and Miley Cyrus at the 2013 *MTV Video Music Awards*. At the end of the fitting a final epidemic has been added, caused by Thicke's appearance on the TV show *X Factor live*, and on day 342 the Synthedemic model is fitting with a high past R^2 of 0.895.

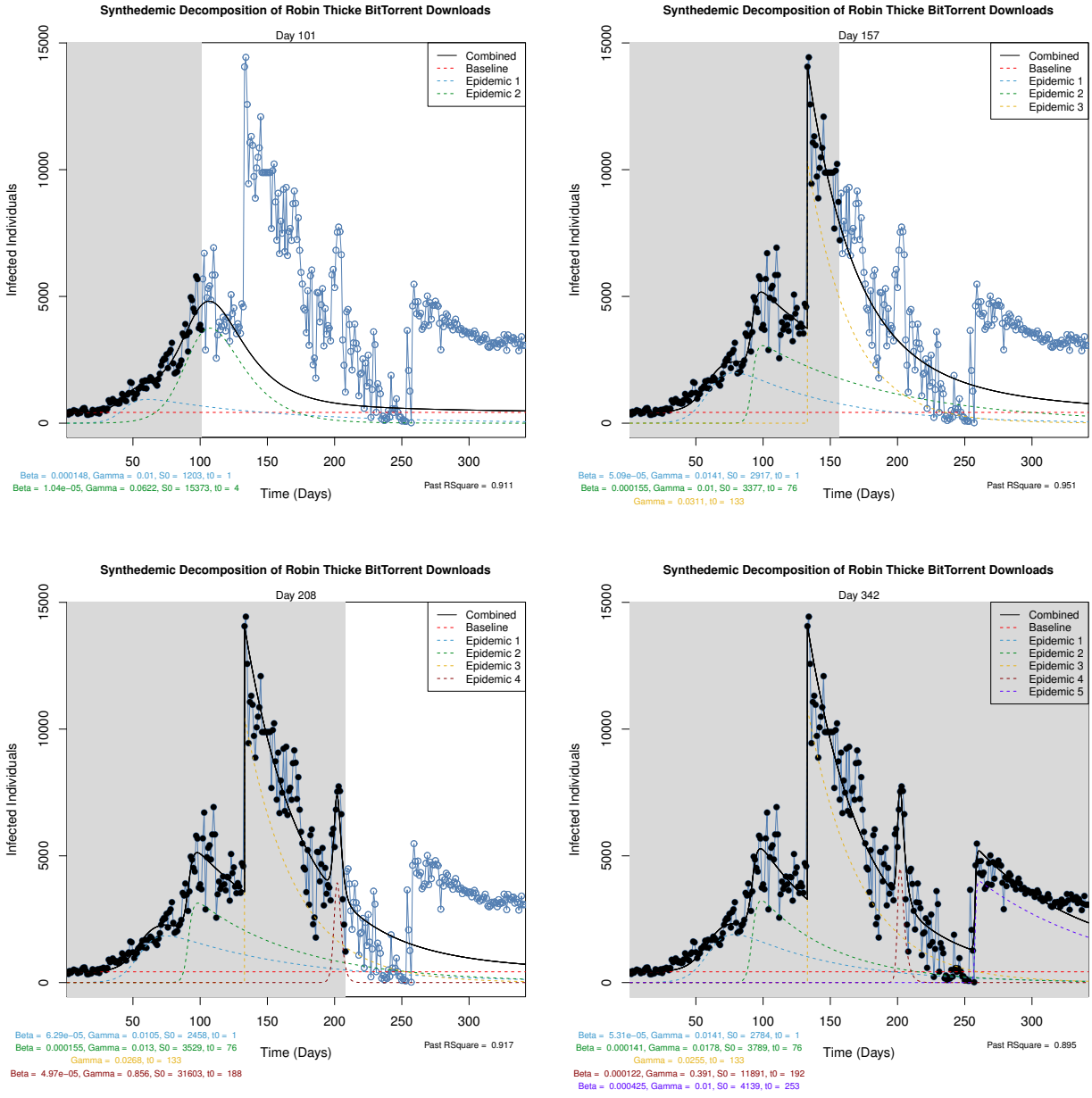


Figure 4.18: Synthedemic fitting to Robin Thicke BitTorrent Downloads

Figure 4.19 shows the AR residual refinement for next day predictions over the course of the Robin Thicke dataset and Figure 4.20 shows the residual ACF and PACF plots for the final fitting.

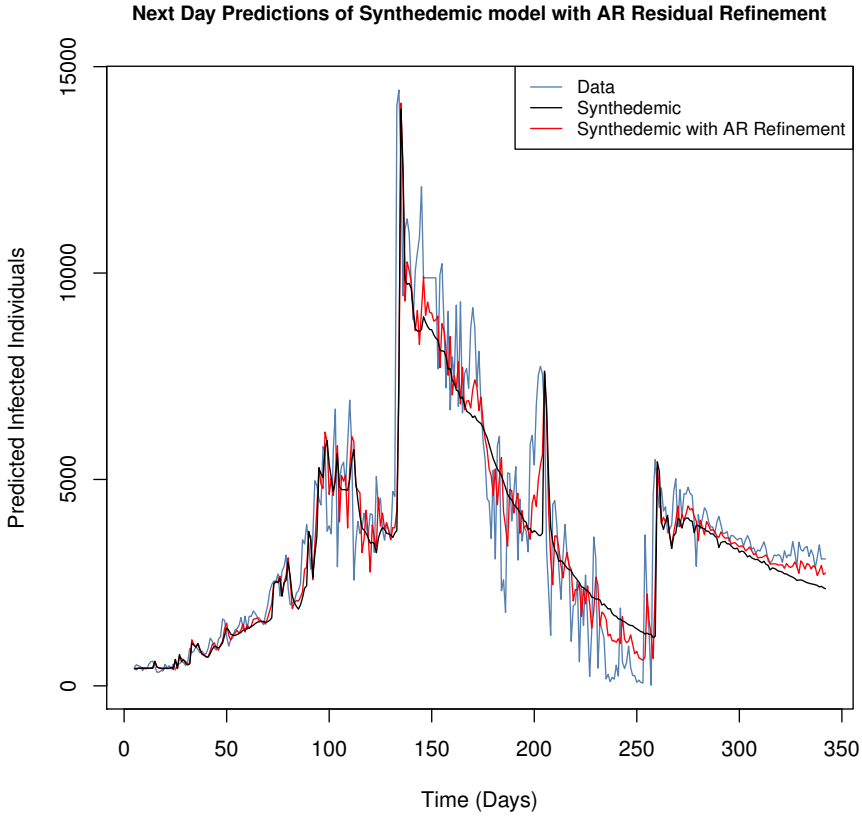


Figure 4.19: Synthedemic fitting with AR Residual Refinement for Robin Thicke downloads

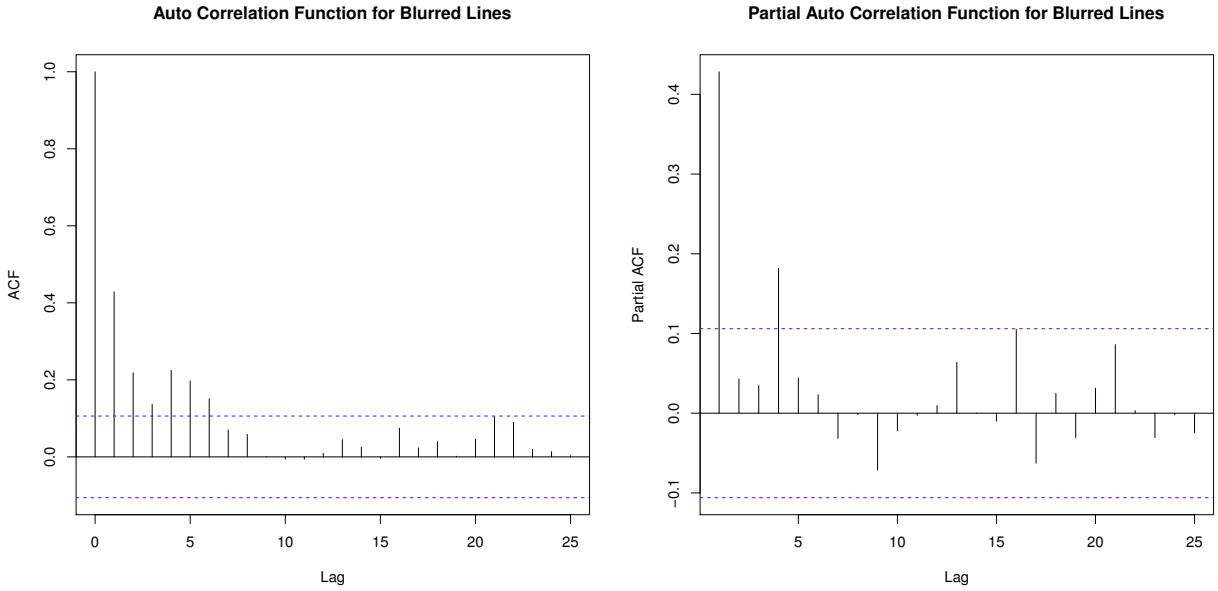


Figure 4.20: Synthedemic Residual ACF and PACF Plots for Robin Thicke downloads

Carly Rae Jepson Music Downloads Caly Rae Jepsens song Call Me Maybe reached worldwide sales of over 13 million copies as of May 2013. As seen in Figure 4.21, at day 116 the model exhibits high near-term accuracy and the quality of the fit to past data is high with an R^2 of 0.977 with three epidemics at this stage. At day 214, the model has encountered another SIR epidemic, fitting a total to four epidemics. At day 244 of the outbreak, the model successfully detects the decaying epidemic trend and has reduced the number of epidemics to three whilst maintaining a high past R^2 of 0.911. At the end of the fitting, the model has successfully fitted the entire dataset with a very good overall accuracy with an R^2 of 0.914.

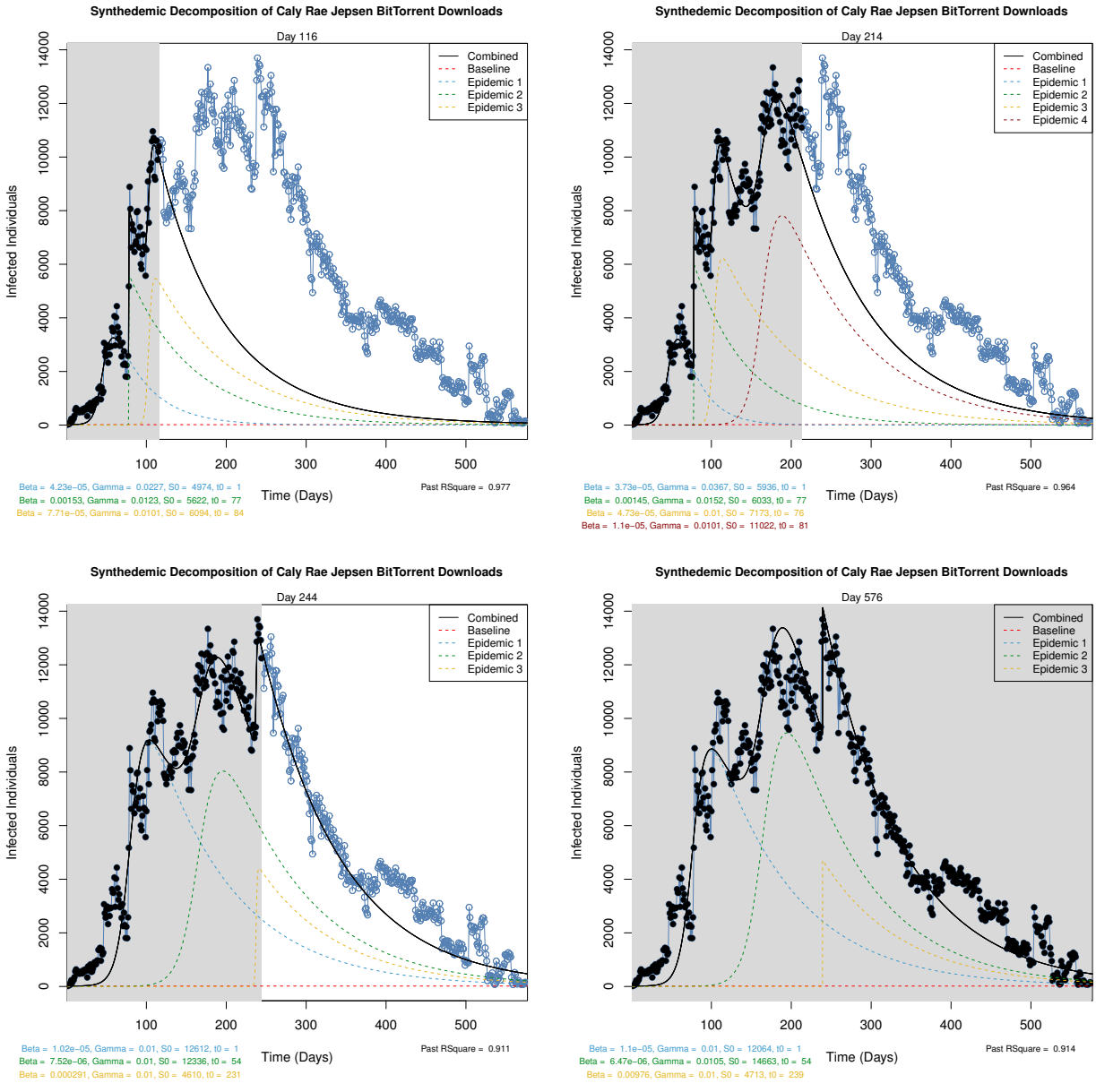


Figure 4.21: Synthedemic fitting to Carly Rae Jepson BitTorrent Downloads

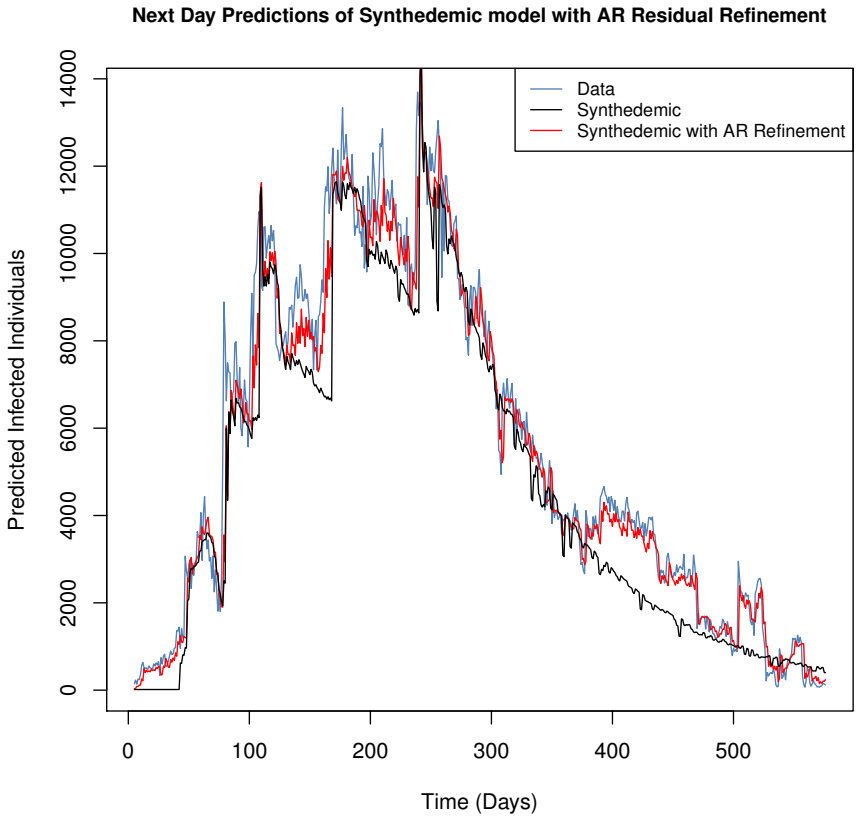


Figure 4.22: Synthedemic fitting with AR Residual Refinement for Carly Rae Jepson downloads

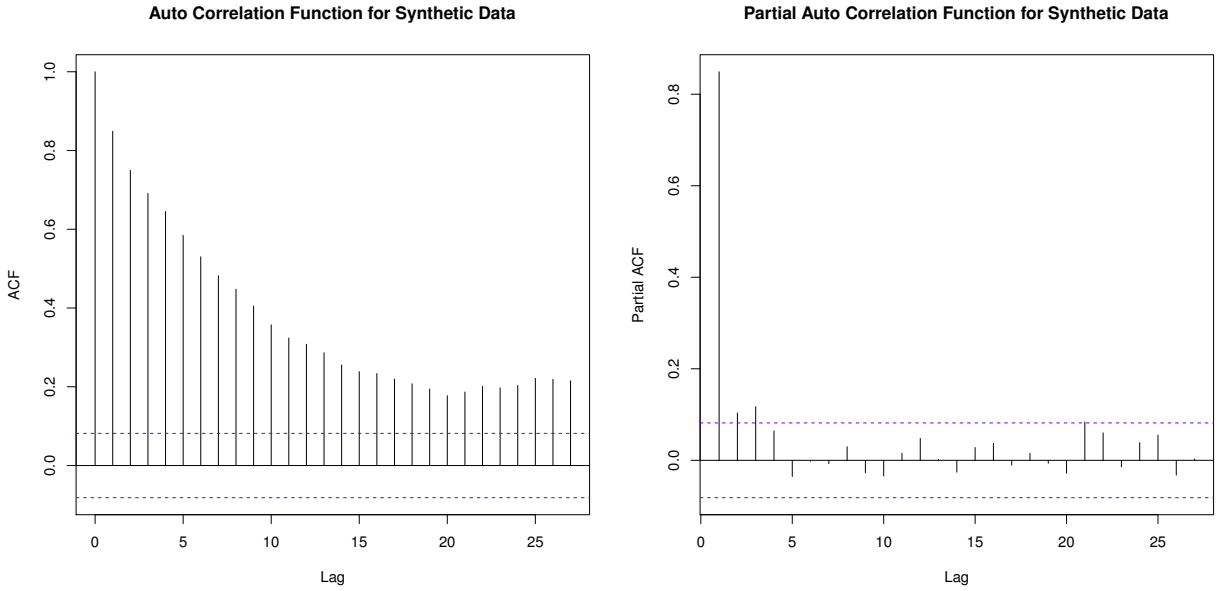


Figure 4.23: Synthedemic Residual ACF and PACF Plots for Carly Rae Jepson downloads

Chapter 5

Evaluation

The proposed model is evaluated both quantitatively and qualitatively over a range of synthetic and real world data sets. Section 5.1 evaluates the models fitting quality, and section 5.1.3 evaluates its forecasting ability. In both chapters, the model is benchmarked against a number of existing techniques. Finally the run time of the fitting process is considered and limitations of the algorithm are discussed.

5.1 Model Fitting Evaluation

Three key properties are used in order to evaluate the model fit:

- *accuracy* - How well the model explains the observed data.
- *robustness* - The ability of the model to fit to different types of data.
- *parsimony* - Ensuring that the model is as simple as possible.

5.1.1 Fitting Metrics

A range of different metrics are used to evaluate the *accuracy* of the model. Table 5.1 shows a summary of fitting statistics for each dataset at the final time.[1] The R^2 is one of the most insightful indicator used as it characterises the models fitting ability and can be compared between different data sets. The ability of the model to fit to a variety of independent data sources with high *accuracy* indicates the *robustness* of the fitting procedure.

[1] R^2=Coefficient of Determination, SSE=Sum of Square Error, MAD=Median Absolute Deviation, MAPE=Mean Absolute Percentage Error, RMSE=Root Mean Square Error, RAE=Relative Absolute Error

Table 5.1: Final Time Fitting Metrics (RT = Robin Thicke, CRJ = Carly Rae Jepson)

Dataset	R^2	SSE	MAD	MAPE	RMSE	RAE
Synthetic1	0.997	2.29×10^3	2.84	0.508	4.03	0.0446
Synthetic2	0.999	2.39×10^4	8.90	0.644	12.2	0.0143
Synthetic3	0.997	3.17×10^3	4.72	0.593	6.04	0.0521
H1N1	0.945	1.50×10^9	5000	0.273	7310	0.913
RT	0.895	2.73×10^8	629	0.980	893	0.311
CRJ	0.914	7.74×10^8	880	0.431	1160	0.251

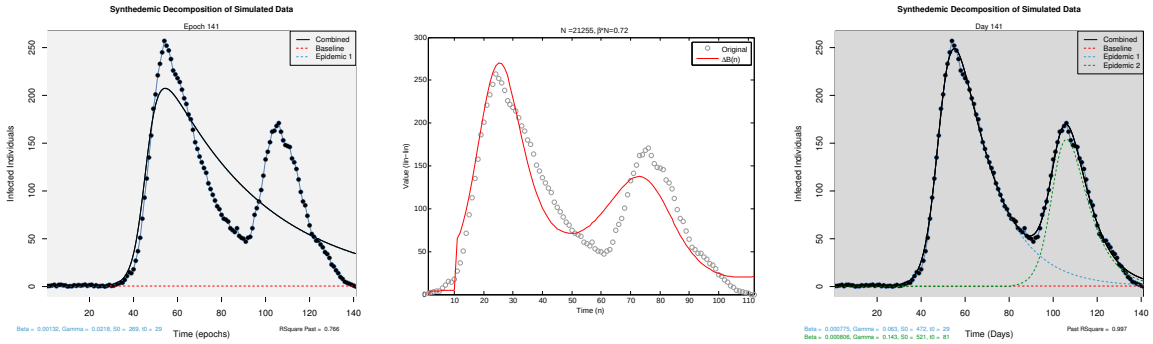
The ability of the model to fit to a variety of different data sources with high *accuracy* highlights the *robustness* of the fitting procedure.

5.1.2 Benchmarking the Fitting Ability

To evaluate the fitting ability of the Synthedemic model, it is benchmarked against existing epidemic modelling techniques. The models are compared quantitatively using error analysis, and qualitatively using the final fitting graph of the epidemic. The models are also compared in terms of the three key factors, *accuracy*, *robustness* and *parsimony*.

Figure 5.1 shows a direct comparison between the single epidemic model, *spikeM* and the Synthedemic model at the final time fitting for the synthetic data set.

The single epidemic model is implemented by limiting the Synthedemic model to a single epidemic outbreak. The *spikeM* model is detailed in the *spikeM* paper [48] and source code[2] was edited to enable the R^2 of the fit to be reported. The *spikeM* model has been fit to each data set using 40 iterations. The total outbreak time is set within the algorithm and the periodicity set accordingly for each dataset.



(a) Single Epidemic (b) SpikeM (c) Synthedemic

Figure 5.1: Direct comparison of *spikeM* and epidemic models on synthetic data

[2] `http://www.cs.kumamoto-u.ac.jp/~yasuko/software.html`

Table 5.2 shows a comparison of the final time R^2 of each data set for the Single Epidemic, *spikeM* and Synthedemic models.

Table 5.2: Final Time R^2 Model Comparison (RT = Robin Thicke, CRJ = Carly Rae Jepson)

Dataset	Single Epidemic	*spikeM*	Synthedemic
Synthetic1	0.766	0.928	0.997
Synthetic2	0.774	0.83	0.999
Synthetic3	0.807	0.789	0.997
H1N1	0.158	0.54	0.945
RT	0.487	0.571	0.895
CRJ	0.508	0.823	0.914

The single epidemic model is clearly incapable of capturing the multiple underlying epidemic outbreaks and produces the lowest R^2 value over all of the data sets. The *spikeM* model produces mid ranged R^2 values that in most cases show a significant improvement over the single epidemic model. In all cases the Synthedemic model shows the highest R^2 fitting statistic, this highlights the *robustness* and accuracy of the approach.

Single Epidemic and Synthedemic Model Comparison Comparing the Single Epidemic and Synthedemic model further, Figure 5.2 shows the change in R^2 over the entire fitting process for Robin Thicke. The R^2 is lower at the start due to higher uncertainty in the parameters as a result of less data. At around time 130, the single epidemic R^2 significantly drops as it is incapable of capturing the large *Spike* outbreak. Throughout the entire fitting the Synthedemic consistently has a higher R^2, representing a more *accurate* fit.

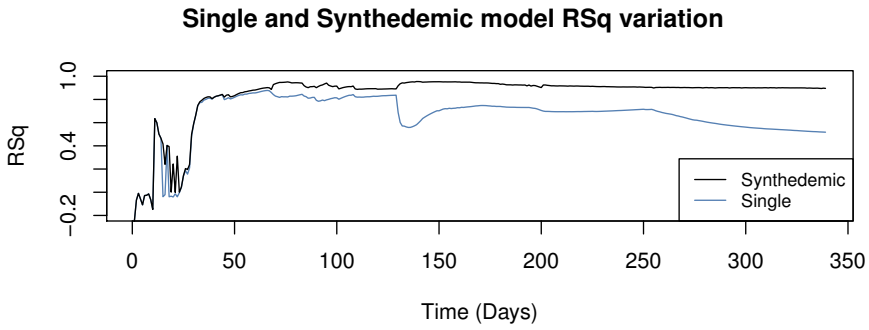
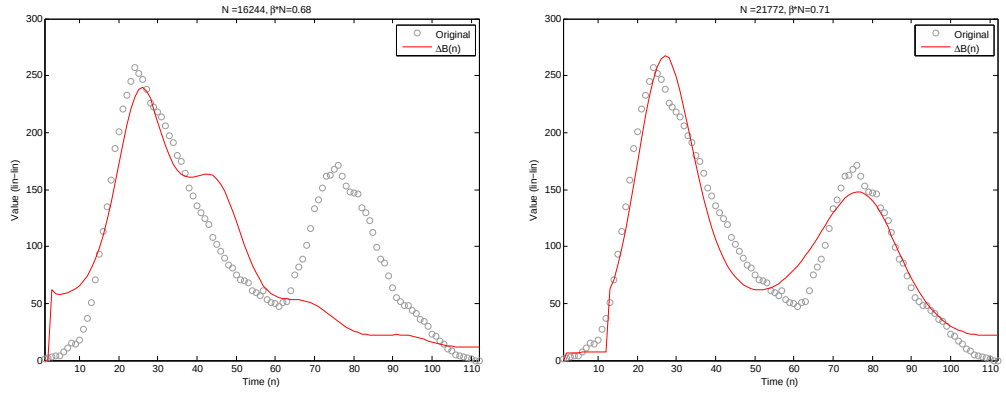


Figure 5.2: Change in R^2 over time of the Single and Synthedemic model for Robin Thicke fitting

***SpikeM* and Synthedemic Model Comparison** For the *spikeM* algorithm, the total outbreak duration (which is often unknown) and number of iterations need to be set at the start of the fitting. This reduces the generality of the method, however the most significant factor which affects the

spikeM fitting the periodicity. It is interesting that without setting the periodicity to match the data set, the *spikeM* model only achieves an R^2 of 0.436 for the Synthetic1 data set as it fails to produce a reasonable fit to the epidemic as shown in Figure 5.3. Furthermore, for the Synthetic3 data set, consisting of an SIR and an Exponential, the R^2 is better for the single epidemic model than the *spikeM* model, this is because with periodicity set to 60 and the *spikeM* model does not classify the *Spike* outbreak. A significant disadvantage of the *spikeM* method is therefore having to set the periodicity and duration of the outbreak when these quantities are often completely unknown at the start of the outbreak.

The assumption of epidemic outbreaks having regular periodicity allows the *spikeM* model to remain parsimonious, however this simultaneously compromises the agility of the model to fit to epidemics with irregular periodicity. This reduces the overall *robustness* of the model as it is unable to adapt to the irregular periodicity epidemic patters which are observed in real world data.



(a) Period Frequency set to 24 (default), 40 Iterations. RSq=0.436

(b) Period Frequency set to 60, 40 Iterations. RSq=0.928

Figure 5.3: SpikeM fitting on Synthetic data with different Periodicities

The parsimony of the *spikeM* approach is reduced by only having a small number of parameters. The number of parameters within the Synthedemic model depends on the number of sub-epidemics added throughout the fitting process. Each *SIR* epidemic has four parameters and each *Spike* epidemic has two parameters. Although the number of parameteres in the Synthedemic model can potentially become large, for the synthetic data set above the Synthedemic model with two *SIR* sub-epidemics has a total of eight parameters, which is the same as the *spikeM* model. Section 5.3 details the compromise between parsimony and accuracy in more detail.

The *spikeM* fittings for all data sets (and further experimentation with iterations and periodicity) are added in the appendix 6.3.

The Synthedemic model requires a target R^2 to be specified as a lower tolerance of the R^2 value. If the fitting procedure falls below the R^2 target then more epidemics are considered, if the fitting procedure is above the target then reducing the number of epidemics is considered. Table 5.3 shows the effect of the final time fitting R^2 and number of parameters within the model as a result of setting different target R^2 values. (The simulated data sets are excluded as the data does not have a significant level of noise and so the target R^2 can be set to be very high and lower R^2 values do not show any difference on the final R^2 value.)

Table 5.3: Final Time Synthedemic Fitting R^2 (RT = Robin Thicke, CRJ = Carly Rae Jepson)

	Target $R^2 = 0.8$		Target $R^2 = 0.85$		Target $R^2 = 0.9$	
Dataset	AttainedR^2	Params	AttainedR^2	Params	AttainedR^2	Params
H1N1	0.945	8	0.945	8	0.945	8
RT	0.671	6	0.861	10	0.895	18
CRJ	0.909	8	0.909	8	0.914	10

Table 5.3 shows that the R^2 attained by the Synthedemic fitting is often very close to the R^2 target specified. The total number of parameters within the Synthedemic model are also provided. It is interesting that the for the H1N1 and CRJ datasets some results remain the same, this is because the fitting converges to the same outcome with two *SIR* epidemics.

Logistic Time Optimisation Model Comparison A comparison between optimising over the start time and iteratively searching the start time enables the most suitable method to be selected. Figure 5.4 shows the change in R^2 throughout the Synthetic1 fitting for both the start time optimisation and parallel time search approaches.

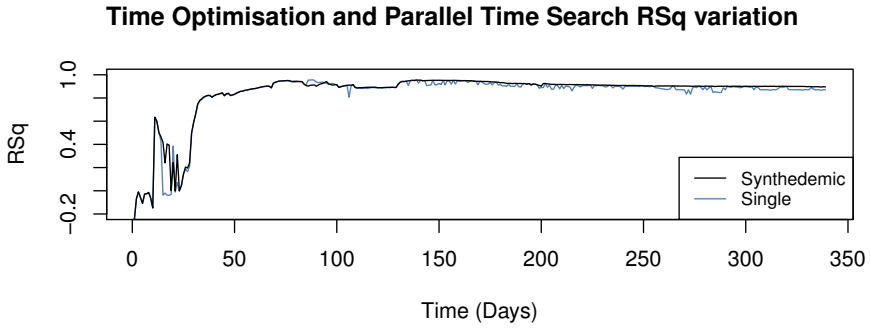


Figure 5.4: Change in R^2 for Time Optimisation and Parallel Time Search models for Robin Thicke fitting

An important result from Figure 5.4 is that the R^2 of the time optimisation approach is similar to the iterative approach. However at the start of new epidemic outbreaks, when the uncertainty

in the optimisation over the other parameters is high, the start time optimisation approach has a lower R^2 than the parallel search start time approach. The time optimisation approach is capable of producing good final time fittings, however due to the increased uncertainty, this method is less consistent.

The results presented in this chapter show that the Synthedemic procedure is both *robust* and *parsimonious* as it is capable of fitting with high *accuracy* to many different datasets and reduces the model dynamically. In the next section, the predictive quality of the model is benchmarked against alternative techniques.

5.1.3 Benchmarking the Predictive quality

One of the most important uses of epidemic models is in predicting the future evolution of an outbreak. The following analyses the predictive quality of the Synthedemic model and benchmarks it against other existing techniques.

5.1.4 Prediction Metrics

To evaluate the models ability to predict future data, a range of metrics are calculated over the next day predictions for the different datasets as shown inTable 5.4.

Table 5.4: Next Day Synthedemic Prediction Metrics (RT = Robin Thicke, CRJ = Carly Rae Jepson)

Dataset	R^2	SSE	MAD	MAPE	RMSE	RAE
Synthetic1	0.994	4.69×10^3	4.33	0.485	5.85	0.0680
Synthetic2	0.993	5.90×10^5	24.3	0.524	58.9	0.0388
Synthetic3	0.895	1.01×10^5	10.54	0.564	34.9	0.115
H1N1	0.420	1.37×10^{10}	14912	0.542	23919.8	0.556
RT	0.777	5.71×10^8	820	0.871	1300	0.407
CRJ	0.889	9.84×10^8	947	0.405	1311	0.271

5.2 Benchmarking the Predictive ability

The R^2 and SSE of next day predictions over the entire fitting are compared inTable 5.5.

Table 5.5: R^2 and SSE averages over all times

	Single Epidemic		Synthedemic		SynthedemicAR		TT	
Dataset	R^2	SSE	R^2	SSE	R^2	SSE	R^2	SSE
Synthetic1	0.749	1.93×10^5	0.994	4690	0.996	3040	0.988	8907
Synthetic2	0.761	2.22×10^7	0.993	590300	0.998	159000	0.991	793000
Synthetic3	0.762	2.30×10^5	0.895	101100	0.897	99030	0.887	108000
H1N1	- 0.712	4.05×10^{10}	0.420	1.37×10^{10}	0.425	1.36×10^{10}	0.44	1.32×10^{10}
RT	0.487	1.32×10^9	0.777	5.71×10^8	0.816	4.72×10^8	0.838	4.16×10^8
CRJ	0.486	4.59×10^9	0.889	9.84×10^8	0.959	3.65×10^8	0.977	2.00×10^8

As expected the R^2 and SSE values for the Synthedemic model future predictions are lower than the R^2 values of the past fitting inTable 5.1. The Synthedemic model shows the ability to predict with a high R^2 value for all of the datasets and consistently produces an improved prediction error in comparison to the single epidemic model. The AR residual refinement of the Synthedemic model shows a further enhancement to the prediction. This can be seen visually in the previously presented AR model graphs in section 4.5.2. The Tomorrow equals Today (TT) prediction is surprisingly accurate with a high average R^2 of 0.85 over all datasets, this makes it is hard to show a significant improvement on the TT prediction, however a slight improvement is made for the synthetic data. It is interesting that the Synthedemic model next day predictions are outperformed by the TT model for the later datasets, however the significance of the enhancement is relatively small and both models possess a high predictive quality. The most significant factor which reduces the Synthetic prediction is the residuals at the start of unpredictable *Spike* outbreaks. A more robust statistic such as the Median Average Deviation shows that the median error for the Synthedemic next day predictions (2780) is lower in comparison to the TT model (3100).

The results of evaluating the Synthedemic show that the Synthedemic model often provides enhanced fitting and prediction ability compared to existing approaches. The key outcomes of the benchmark results are summarised in section 6.2.

5.3 Limitations

Compromises within the existing approach arise in many aspects of the fitting procedure. The most important trade off is between the accuracy and parsimony of the Synthedemic approach. Although the number of parameters of the Synthedemic model can potentially become large, diminishing the parsimony of the model, at each time within the fitting a reduction stage is undertaken. This determines if the number of epidemics can be decreased while maintaining a sufficient R^2. The dynamic approach results in a model that is parsimonious while maintaining fitting accuracy. A key factor in the control of the trade off between parsimony and accuracy is therefore the target R^2 provided to the model. When the target is low then the model is more parsimonious however

as the R^2 is increased the model incorporates more epidemics as required to obtain the level of accuracy.

A further compromise is between the speed of each fitting stage and the accuracy of the solution. This is derived from the accuracy of the optimisation procedure and the time that is given for each optimisation routine. To classify as real-time epidemic modelling it is important that the time taken for each stage does not exceed the time between observations. This does not present an immediate problem as each stage takes approximately five minutes to run, depending on the size of the observed data and number of sub-epidemics. The run time of the algorithm is detailed further in section 5.3.1. However if the data is recorded in shorter time intervals then the number of optimisation iterations may need to be reduced.

The main limitations of the approach are derived from the dynamic fitting procedure. The following explains some of the cases where the algorithm may have a delayed or inaccurate response to outbreaks:

Negative epidemics In some cases it has been observed that negative epidemics may be present within the overall epidemic phenomena. Negative epidemics can arise from external factors such as Vaccinations. The current methodology does not handle negative epidemics and as such may fail to detect or adjust the model parameters effectively to handle the lack of infectives. This may be particularly significant if the number of infected individuals falls below the baseline of the fitting, due to the addition of sub epidemic parts it is currently not possible to model epidemic processes that fall below the baseline at future times.

Residual analysis The current algorithm dynamically determines outbreaks by analysing the standard deviation and mean of the current epidemic residuals and uses the last two residuals to determine if an outbreak has occurred. This heuristic can lead to problems when the rise of observed data points is on the limit of determining the outbreak. When each limiting residual is added, if an outbreak is on the limit of being detected then on the next iteration the standard deviation will increase significantly. This could cause the detection to become less sensitive for the next residual and potentially cause a chain undetected outbreak residuals. Independent of number of standard deviations used to determine an outbreak this problem will always exist and so alternative approaches of gradient analysis or clustering may be required in order to prevent this rare case.

Classification of epidemics A very significant problem is the potential incorrect classification of sub epidemic types, in particular if an SIR outbreak is determined to be a *Spike* then the future residuals of the SIR will not be characterised by the *Spike* decay. This can be prevented by "swapping" the current epidemic type during the fitting to ensure that the most suitable type

is currently being used. However this was not observed within the current fitting process due to ensuring that a *Spike* epidemic is only determined for very high residuals.

Unknown epidemic development Throughout the fitting process, it has been observed that at the start of the fitting the procedure is more sensitive to small epidemic outbreaks, which are eventually dropped when larger future epidemic observations are considered later in the fitting process. This occurs because at the time the algorithm is not aware of the significance of the current data, because it has no information about the scale of the future epidemic behaviour. The effect of this may be reduced by limiting the possible value of initial Susceptible $S0$ with an estimate for the specific application.

The compromises and limitation of the current procedure highlight that although the algorithm shows promising results on the analysed data sets, there may be certain situations where the fitting process does not capture the complete underlying epidemic behaviour and future work is required to produce an even more general and robust procedure.

5.3.1 Algorithm Run Time Analysis

This analysis is based on the number of evaluations of the `sseMulti` objective function used by the optimisation. In a single evaluation of the `sseMulti` the residual of each prediction from the data points is calculated, therefore the run time is $\prime(n)$ where n is the number of data points in the complete sample. In the worst case, the optimisation algorithm hits the limit of evaluations l each time the optimisation runs, and the optimisation is repeated o times. Therefore at each time point, the call to the `sseMulti` function is repeated $l * o$ times and the overall complexity for each time point is $\prime(n * l * o)$. For all data points, the start time of the final epidemic is searched within a window around its current start time. Let the size of the window be w, this means that the time taken time taken is $\prime(n^2 * l * o * w)$ For the current implementation, $l = 1000$, $o = 5$, $w \approx 20$ $n \approx 200$ therefore the number of calls to `sseMulti` is approximately 20 million. This means that the iterative optimisation fitting can take a very long time to complete all data points. In order to speed up the fitting process, the start times are searched in parallel reducing the w to a constant time and making the run time $\prime(n^2 * l * o)$. The run time of the algorithm could be further enhanced by dropping epidemics which have died out at the current time. Reducing the number of parameters makes optimisation more efficient and makes the model more parsimonious.

Chapter 6

Conclusions and Future Improvements

This project has proposed an approach to the real-time fitting of outbreaks which have multiple underlying epidemic phenomena. The Synthedemic fitting algorithm dynamically determines the start times of each epidemic and simultaneously optimises over the parameters. The ability of the model to explain and predict multiple epidemic phenomena highlights the acceptability of the original hypothesis that the underlying phenomena manifest as sub-epidemics and the overall outbreak behaviour is explained as the superposition of sub epidemic parts.

The following summarises the main contributions and results of the project:

6.1 Main Contributions

The main contribution of this project is the development of the Synthedemic model within an epidemic modelling framework. The following sub tasks have been achieved through the development of the Synthedemic model:

- **Iterative fitting procedure:** Fitting of the Synthedemic model is undertaken in an iterative manner, incorporating single data observations at a time to fit at each stage throughout entire development of the epidemic outbreak.
- **Outbreak Detection:** Outbreaks are integrated dynamically into the Synthedemic model using outbreak detection, this uses residual analysis to determine the start time and type of sub epidemic outbreaks.
- **Determining Sub Epidemic Start Times:** The start time of the current epidemic is searched within a feasible range, in parallel, to enable the start time of the epidemic to be

determined. This was identified as the most stable method of determining epidemic start time with feasible complexity.

- **Synthedemic Model with Autoregressive Residual Refinement:** Autoregressive residual refinement has been included into to the Synthedemic modelling procedure. This has been shown to enhance the predictive ability of the Synthedemic model.

6.2 Main Results

The main results of the Synthedemic model benchmark evaluations are summarised:

- The Synthedemic model has been shown to be a *robust* fitting procedure, capable of attaining a high R^2 (averaging 0.97) over a range of six datasets obtained from different sources (containing a total of over 1000 fittings). Moreover, the dynamic nature of the Synthedemic model enables the model to attain the required *accuracy* whilst maintaining *parsimony*.
- One of the most significant results of the project is that the Synthedemic approach consistently out performs the fitting ability of the single epidemic model over all six datasets
- The Synthedemic model is shown to have enhanced fitting ability over the datasets when compared to the *spikeM* model.
- The next day predictive ability of the Synthedemic model is also superior to the future predictions of the single epidemic model over all datasets
- Predictive quality of the Synthedemic model is further improved by the autoregressive residual refinement

Throughout the development, alternative approaches to the proposed Synthedemic methodology have been investigated, such as optimising the start times of epidemics. The most appropriate methods have been incorporated into an epidemic modelling framework which enhances the ease of fitting, prediction and analysis of the Synthedemic model.

6.3 Future Work

There are many future extensions that can enhance the current work:

Uncertainty Replicating the results of this project using Maximum Likelihood objective function would be a significant extension to this project. Within the epidemic modelling framework, the objective function is passed to the fitting process, and therefore by setting the fitting mode to `MLE` the fitting process will, in theory, be undertaken in the same way but using a MLE based objective

function. MLE fitting has been implemented for a single epidemic model as shown in section 3.3.5. Extending the Synthedemic model to incorporate uncertainty estimates would be a very significant improvement to the project as it would enable uncertainty and confidence in the parameters of the multi epidemic model to be quantified, providing a more comprehensive model. One of the main challenges is determining the initial conditions for MLE based fitting and also optimising the multiple epidemic parameters using the *mle2* fitting procedure.

Alternative Approaches The parallel sub epidemic start time search approach has been evaluated against an alternative technique of optimising the start time of each epidemic. Many other techniques for determining the epidemic start times have been proposed, however qualitative and quantitative analysis is required to determine the success of each. Furthermore alternative methods for outbreak detection and search heuristics, such as searching only the final epidemic start time, also need to be analysed in more detail in future.

Alternative Models The current Synthedemic algorithm incorporates two types of epidemic, a gradual growth *SIR* model and a rapid outbreak *Spike* exponential decay model. Analysis of the use of different types of model (such as the power law decay model used in *spikeM*) may enhance the prediction of future data to actual datasets if the underlying outbreak is characterised better by the model. Further evaluations with a range of different models, for example $SEIR$ or power law decay models, are required to select on the most suitable models.

Bibliography

[1] Diane Zahler. *The Black Death*. Twenty-First Century Books, 2009. p.48.

[2] Volker Langholf. *Medical Theories in Hippocrates: Early Texts and the "Epidemics"*, volume 34. Walter de Gruyter, 1992.

[3] Klaus Dietz and JAP Heesterbeek. Daniel Bernoullis epidemiological model revisited. *Mathematical biosciences*, 180(1):1–21, 2002.

[4] Jeremy T Bradley, Stephen T Gilmore, and Jane Hillston. Analysing distributed internet worm attacks using continuous state-space approximation of process algebra models. *Journal of Computer and System Sciences*, 74(6):1013–1032, 2008. `http://pubs.doc.ic.ac.uk/continuous-pepa-worms/continuous-pepa-worms.pdf`.

[5] Marily Nika, Gergana Ivanova, and William Knottenbelt. On celebrity, epidemiology and the internet. 2013. `http://www.doc.ic.ac.uk/~wjk/publications/nika-ivanova-knottenbelt-valuetools-2013.pdf`.

[6] Epidemiology definition. `http://www.who.int/topics/epidemiology/en`. World Health Organization.

[7] Håkan Andersson and Tom Britton. *Stochastic Epidemic Models and their Statistical Analysis*, volume 4. Springer New York, 2000.

[8] Fred Brauera, Zhilan Fenga, and Carlos Castillo-Chaveza. Discrete epidemic models. *Mathematical Biosciences*, 7:1, 2010. `http://math.la.asu.edu/~chavez/CCCPUB/Discrete%20epidemic%20models.pdf`.

[9] Jianquan Li, Zhien Ma, and Fred Brauer. Global analysis of discrete-time SI and SIS epidemic models. *Mathematical biosciences and engineering: MBE*, 4(4):699–710, 2007. `http://www.ncbi.nlm.nih.gov/pubmed/17924720`.

[10] O Diekmann, JAP Heesterbeek, and MG Roberts. The construction of next-generation matrices for compartmental epidemic models. *Journal of the Royal Society Interface*, 7(47):873–885, 2010. `http://rsif.royalsocietypublishing.org/content/early/2009/11/04/rsif.2009.0386.full#p-6`.

[11] James S Koopman, Geoffrey Jacquez, and Stephen E Chick. New data and tools for integrating discrete and continuous population modeling strategies. *Annals of the New York Academy of Sciences*, 954(1):268–294, 2001. http://onlinelibrary.wiley.com/doi/10.1111/j.1749-6632.2001.tb02756.x/abstract.

[12] Daryl J Daley, Joe Gani, and Joseph Mark Gani. *Epidemic modelling: an introduction*, volume 15. Cambridge University Press, 2001.

[13] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A*, 115(772):700–721, 1927. http://rspa.royalsocietypublishing.org/content/115/772/700.

[14] W. O. Kermack and A. G. McKendrick. Contributions to the mathematical theory of epidemics. II. the problem of endemicity. *Proceedings of the Royal Society of London. Series A*, 138(834):55–83, 1932. http://rspa.royalsocietypublishing.org/content/138/834/55.

[15] W. O. Kermack and A. G. McKendrick. Contributions to the mathematical theory of epidemics. III. further studies of the problem of endemicity. *Proceedings of the Royal Society of London. Series A*, 141(843):94–122, 1933. http://rspa.royalsocietypublishing.org/content/141/843/94.

[16] SIR Model Derivation. http://en.wikipedia.org/wiki/Epidemic_model#Deterministic_compartmental_models.

[17] An Introduction to Disease Dynamics. http://www.unc.edu/~rls/s940/samsidisdyntut.pdf.

[18] Compartmental models in epidemiology. http://en.wikipedia.org/wiki/Compartmental_models_in_epidemiology.

[19] Compartmental Models. http://en.wikipedia.org/wiki/Epidemic_model#Models_with_more_compartments.

[20] Helen Abbey. An examination of the Reed-Frost theory of epidemics. *Human biology*, 24(3):201–233, 1952.

[21] F.D. Sahneh, C. Scoglio, and P. Van Mieghem. Generalized epidemic mean-field model for spreading processes over multilayer complex networks. *Networking, IEEE/ACM Transactions on*, 21(5):1609–1620, Oct 2013. http://ieeexplore.ieee.org/xpls/icp.jsp?arnumber=6423227.

[22] Exact solution of SIR and SIS epidemic models. http://www.researchgate.net/publication/48173364_A_note_on_Exact_solution_of_SIR_and_SIS_epidemic_models.

[23] Ordinary Least Squares. http://en.wikipedia.org/wiki/Ordinary_least_squares.

[24] S. Singer and J. Nelder. Nelder-Mead algorithm. *Scholarpedia*, 4(7):2928, 2009. http://www.scholarpedia.org/article/Nelder-Mead_algorithm.

[25] J. A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965. http://comjnl.oxfordjournals.org/content/7/4/308.full.pdf.

[26] Fuchang Gao and Lixing Han. Implementing the nelder-mead simplex algorithm with adaptive parameters. *Computational Optimization and Applications*, 51(1):259–277, 2012. http://www.webpages.uidaho.edu/~fuchang/res/ANMS.pdf.

[27] B. F. Finkenstdt, A. Morton, and D. A. Rand. Modelling antigenic drift in weekly flu incidence. *Statistics in Medicine*, 24(22):3447–3461, 2005.

[28] Maurice Clerc. *Particle swarm optimization*, volume 93. John Wiley & Sons, 2010.

[29] Sadik Olaniyi Maliki. Analysis of numerical and exact solutions of certain sir and sis epidemic models. *Journal of Mathematical Modelling and Application*, 1(4):51–56, 2011. http://proxy.furb.br/ojs/index.php/modelling/article/view/2219/1713.

[30] Coefficient of Determination. http://www.coefficientofdetermination.com/#post-body-4789529864446302815.

[31] RSquared in Matlab. http://www.mathworks.co.uk/help/stats/coefficient-of-determination-r-squared.html.

[32] Model Evaluation. http://www.saedsayad.com/model_evaluation_r.htm.

[33] Forecast Accuracy. http://en.wikipedia.org/wiki/Forecasting#Forecasting_accuracy.

[34] Geof H Givens and James P Hughes. A method for determining uncertainty of predictions from deterministic epidemic models. In *Proceedings of the 1995 SCS Western Multiconference on Health Science*. Citeseer, 1995. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.68.5809&rep=rep1&type=pdf.

[35] Geof H Givens. *A Bayesian framework and importance sampling methods for synthesizing multiple sources of evidence and uncertainty linked by a complex mechanistic model*. PhD thesis, Citeseer, 1993. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.68.6247&rep=rep1&type=pdf.

[36] Residual Analysis. http://www.itl.nist.gov/div898/handbook/pmd/chapter4/pmd44.htm.

[37] Time Series Analysis, autocorrelation. http://users.ecs.soton.ac.uk/jn2/teaching/timeSeries.pdf.

[38] Identifying the numbers of AR or MA terms. http://people.duke.edu/~rnau/411arim3.htm.

[39] Alessandro Vespignani, Vittoria Colizza, Alain Barrat, and Marc Barthlemy. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences of the United States of America*, 103(7):2015–2020, 2006. http://www.pnas.org/content/103/7/2015.full.

[40] Thomas House, Marc Baguelin, Albert Jan Van Hoek, Peter J White, Zia Sadique, Ken Eames, Jonathan M Read, Niel Hens, Alessia Melegaro, W John Edmunds, et al. Modelling the impact of local reactive school closures on critical care provision during an influenza pandemic. *Proceedings of the Royal Society B: Biological Sciences*, 278(1719):2753–2760, 2011. http://www.ncbi.nlm.nih.gov/pubmed/21288945.

[41] TA Abeku, SJ De Vlas, GJJM Borsboom, A Tadege, Y Gebreyesus, H Gebreyohannes, D Alamirew, A Seifu, NJD Nagelkerke, and JDF Habbema. Effects of meteorological factors on epidemic malaria in ethiopia: a statistical modelling approach based on theoretical reasoning. *Parasitology*, 128(06):585–593, 2004. http://www.ncbi.nlm.nih.gov/pubmed/15206460.

[42] Ellen Brooks-Pollock and Ken TD Eames. Pigs didnt fly, but swine flu. *Math. Today (Southend-on-Sea)*, 47(1):36–40, 2011. http://www.ima.org.uk/_db/_documents/mt_feb11_pigs_didnt_fly.pdf.

[43] Benjamin Doerr, Mahmoud Fouz, and Tobias Friedrich. Why rumors spread so quickly in social networks. *Commun. ACM*, 55(6):70–75, 2012. http://dl.acm.org/citation.cfm?id=2184338.

[44] Valerie Tweedle and Robert J Smith. A mathematical model of bieber fever: The most infectious disease of our time? *Understanding the dynamics*, 2012. http://mysite.science.uottawa.ca/rsmith43/BieberFever.pdf.

[45] Konstantin Avrachenkov, Koen De Turck, and Balakrishna Fiems, Dieter a Prabhu. Information dissemination processes in directed social networks. *arXiv*, 2013. http://arxiv.org/pdf/1311.2023v1.pdf.

[46] Marily Nika, Dieter Fiems, Koen Turck, and William Knottenbelt. Modelling interacting epidemics in overlapping populations. 8499:33–45, 2014. http://dx.doi.org/10.1007/978-3-319-08219-6_3.

[47] Jaewon Yang and Jure Leskovec. Patterns of temporal variation in online media. pages 177–186, 2011. http://doi.acm.org/10.1145/1935826.1935863.

[48] Yasuko Matsubara, Yasushi Sakurai, B Aditya Prakash, Lei Li, and Christos Faloutsos. Rise and fall patterns of information diffusion: model and implications. 2012. http://dl.acm.org/citation.cfm?id=2339537.

[49] Aaron King Tutorial on epidemic parameter optimisation. http://kinglab.eeb.lsa.umich.edu/EEID/eeid/2011_eco/EEID2011_Fitting.pdf.

[50] Aaron King Parameter estimation tutorial. http://kinglab.eeb.lsa.umich.edu/EEID/eeid/2011_eco/mle_2011.pdf.

[51] Tom L Burr and Gerardo Chowell. Observation and model error effects on parameter estimates in susceptible-infected-recovered epidemiological models. *Far East Journal of Theoretical Statistics*, 19(2):163–183.

[52] Engineering statistics handbook. http://www.itl.nist.gov/div898/handbook/eda/chapter3/eda35h.htm.

Appendix: SpikeM Model

.1 *SpikeM* Alternative Periodicity and Iterations Fittings

.1.1 *SpikeM* Synthetic Fitting with different Iterations

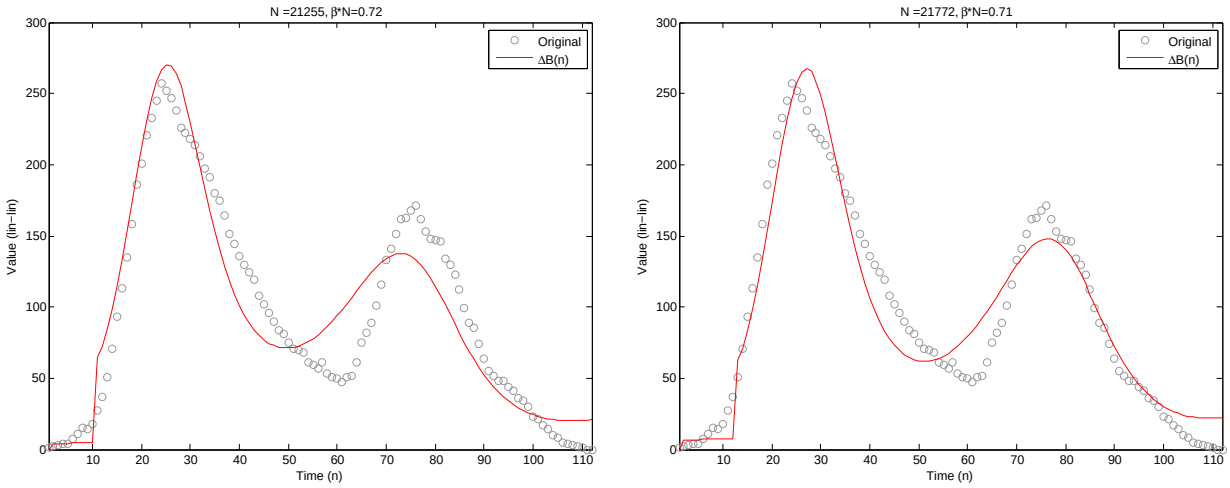


Figure 1: Synthetic Fitting, Iterations=20 (LHS) Vs. Iterations=40 (RHS)

.1.2 *SpikeM* Robin Thicke Fitting with different Periodicities

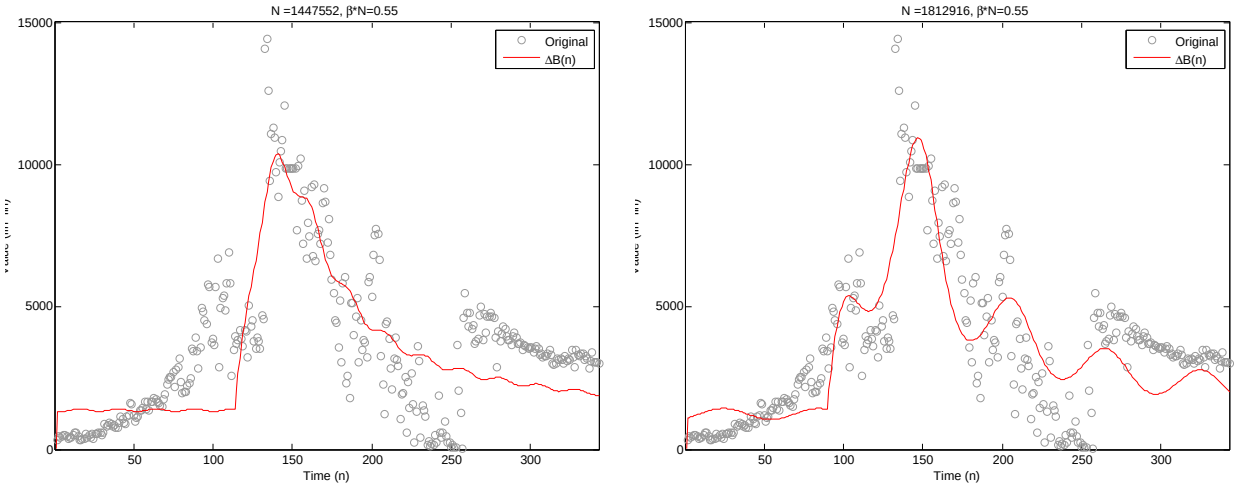


Figure 2: Robin Thicke *spikeM* Fitting, Periodicity=24 (LHS) Vs. Periodicity=60 (RHS)

.1.3 *SpikeM* H1N1 Fitting with different Iterations

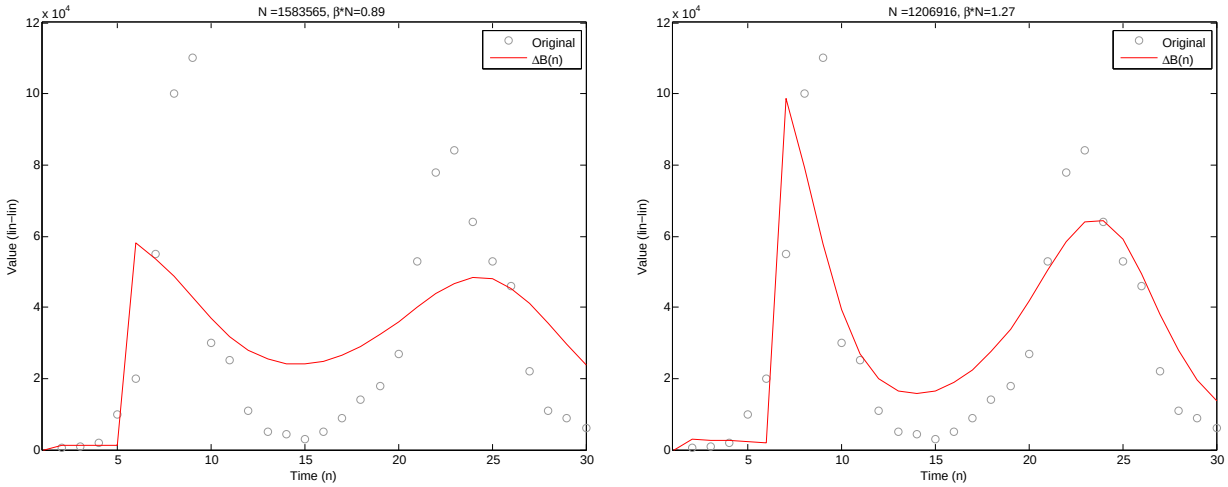


Figure 3: H1N1 *spikeM* Fitting, Iterations=20 (LHS) Vs. Iterations=40 (RHS)