Sensitivity Analysis and Forecasting in Network Epidemiology Models

Elaine O. Nsoesie

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Genetics, Bioinformatics, and Computational Biology

Richard J. Beckman, Co-Chair
Madhav V. Marathe, Co-Chair
Scott C. Leman
Josep Bassaganya-Riera
David R. Bevan
Ina Hoeschele

March 30, 2012
Blacksburg, Virginia

Keywords: Statistics, Epidemic Forecasting, Simulation, Individual-based Model,
Classification, Dirichlet Process, Optimization

Sensitivity Analysis and Forecasting in Network Epidemiology Models

Elaine O. Nsoesie

ABSTRACT

In recent years, several methods have been proposed for real-time modeling and forecasting of the epidemic curve. These methods range from simple compartmental models to complex agent-based models. In this dissertation, we present a model-based reasoning approach to forecasting the epidemic curve and estimating underlying disease model parameters. The method is based on building an epidemic library consisting of past and simulated influenza outbreaks. During an influenza epidemic, we use a combination of statistical, optimization and modeling techniques to either assign the epidemic to one of the cases in the library or propose parameters for modeling the epidemic. The method is presented in four steps. First, we discuss a sensitivity analysis study evaluating how minute changes in the disease model parameters influence the dynamics of simulated epidemics. Next, we present a supervised classification method for predicting the epidemic curve. The epidemic curve is forecasted by matching the partial surveillance curve to at least one of the epidemics in the library. We expand on the classification approach by presenting a method which identifies epidemics similar or different from those in the library. Lastly, we discuss a simulation optimization method for estimating model parameters to forecast the epidemic curve of an epidemic distinct from those in the library.

Dedication

To Mom, Dad, Cynthia and Ekole

Acknowledgments

This dissertation would not been completed without the help, prayers and advice from several individuals mentioned below and many more not listed here. I say "Thanks" to everyone who have helped me in one way or another throughout my life.

I thank my advisors Dr. Richard Beckman and Dr. Madhav Marathe for their guidance and support. The project discussed in this dissertation exists because of Dr. Beckman. He conceived the idea and has patiently guided me through my doctoral work. Together, Dr. Marathe and Beckman have helped me to develop better technical writing, problem solving and presentation skills. I am indebted to Dr. Marathe for inviting me to join the Network Dynamics and Simulation Science Lab (NDSSL). The experience I have gained while working with my advisors and members of NDSSL have taught me to appreciate collaborative research and excellence. I am grateful to have had advisors who cared about my progress and encouraged me to pursue a work-life balance.

I am also grateful to my committee members: Dr. Scotland Leman, Dr. Josep Bassaganya-Riera, Dr. David Bevan, and Dr. Ina Hoeschele. They have been supportive of my efforts and have provided constructive feedback on my work whenever needed.

I am indebted to all the members of NDSSL for making my work environment welcoming and pleasurable. I especially wish to thank Dr. Annette Feng for her help and timely response to my questions on Isis. In addition, I appreciate the friendship and support from staff, faculty, colleagues and fellow students: Kalyani Nagaraj, Samarth Swarup, Bryan Lewis, Chris Kuhlman, Sara Shashaani, Caitlin Rivers, Jose Jimenez, Nidhi Parikh, Ann Paul, Ashwin Aji, Suruchi Deodhar, Joyce Randall, Sharon Matchon, Sandra Wagener, Arron Dawson and Katy Bitely. I am also grateful to Dennie Munson for her kindness and help in navigating graduate school requirements.

I would also like to acknowledge the contribution of each of my co-authors to works described in chapters 2-5; technical reports and publications. The work in this dissertation was performed in collaboration with my advisors, Dr. Bryan Lewis, Dr. Scotland Leman, Kalyani Nagaraj and Sara Shashaani. Dr. Bryan Lewis' expert knowledge on public health epidemiology has been extremely useful in the selection of parameters for each of the studies in this dissertation. Dr. Scotland Leman has spent countless hours providing Bayesian expertise, guidance on research practice and has been extremely patient in responding to my

questions. Kalyani and Sara's knowledge on optimization methods and Perl programming were invaluable for the study presented in chapter 5.

I thank Rose Nkiri Asong and Caitlin Rivers for proofreading parts of this dissertation.

Finally, I would like to thank my family: my parents, my sister Cynthia Nsoesie, my brother Nfoni Ekole and my extended family in the United States and in Cameroon for their love and encouragement. Thanks for always believing in my abilities, reading my research papers, and being an anchor to lean on. I thank friends at the Blacksburg Christian Fellowship and Graduate Christian Fellowship for spiritual guidance and encouragement. I thank Janet McCarthy and Evelyn McKoen for their motherly warmth and comfort.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A forecast can be defined as an attempt to quantitatively predict a future event. The term forecast is typically used in describing prediction of events such as the weather, which are not easily predictable. In this dissertation, we present a method which seeks to forecast the epidemic curve during an epidemic. The epidemic curve is a time series with observations representing the number of infected persons for the duration of an epidemic (see example in Figure 1.1). Forecasting the epidemic curve implies that given data $x_1, \ldots, x_t$, we seek to predict $x_{t+1}, \ldots, x_n$. Here x_i represents infected counts on day i and n is the expected duration of the epidemic. Accurate prediction of specific measures such as the peaking time, peak infected rate and number of infected individuals for the epidemic duration would be invaluable to public health officials. These measures would enable public health officials to make informed decisions regarding allocation of resources and introduction of interventions such as vaccinations and school closures [46, 109].

A good prediction of the epidemic curve would provide an estimate of disease severity during an epidemic. The epidemic curve can be predicted under the baseline scenario (no measures are introduced to control the epidemic spread) or under certain assumptions and hypotheses. These assumptions could include the presence of intervention measures and other "what if" scenarios patterning to changes in individual behavior [101]. The studies in this dissertation deal with the baseline scenario. However, the methods can be easily extended to include additional assumptions and hypotheses.

We also focus on predicting influenza epidemic curves although the methods can be applied to forecasting the epidemic curve for other infectious diseases. Influenza epidemics are typically observed during the winter season in most regions. Seasonal influenza outbreaks typically result in approximately 250,000 to 500,000 deaths globally per year [131]. On the contrary, influenza pandemics are rare and result from novel influenza viruses. The recent H1N1(A) influenza pandemic of 2009 resulted in an estimated 8,870-18,300 H1N1-related deaths from April 2009 to April 10, 2010 [33] in the United States. However, previous pandemics in the 20^{th} century were more severe [107, 118, 126]. The Spanish flu of 1918, Asian flu of 1957

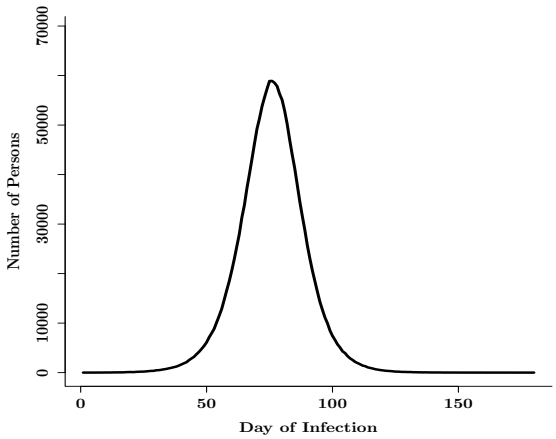


Figure 1.1: An Epidemic Curve

and Hong Kong flu of 1968 are estimated to have resulted in approximately 40-50 million, 1-4 million and 1-4 million deaths worldwide respectively [132]. Since the majority of the population would have no immunity to a novel human influenza virus, one would expect major effects on public health in addition to economic and social disruption [132]. It is therefore important to continue developing tools to aid in the planning and preparation for the next severe global pandemic.

The aim of this chapter is to introduce the reader to the problem of forecasting the epidemic curve and our approach towards it. Section 1.1 provides background information on epidemic models, and section 1.2 briefly describes the problem and existing approaches. Section 1.3 explains the methods we have used and provides a chapter-to-chapter summary of the dissertation.

1.1 Epidemic Models

The method presented in this dissertation can serve as a tool for decision making during a pandemic. The method combines individual-based modeling, classification, statistical and optimization techniques to predict the epidemic curve and infer disease parameters. The individual-based model consists of a network model and a disease model. Network models provide a means for describing interactions between individuals and their environment [91]. Contact networks are typically defined using a weighted undirected graph [113]. Nodes in the network represent agents (individuals) and edges are used to represent contacts between individuals. Weights on the edges represent either the contact duration or the probability of contact. The structure of a contact network including the level of heterogeneity directly influences the spread of a disease through a network [113]. Therefore, assumptions made in constructing the network structure are extremely important.

Synthetic individuals in the social networks are assigned demographic details such as age and household income. Individuals are also assigned possible (disease) states based on the disease model. This implies that at every time point, each individual in the population can be either **s**usceptible, **i**nfectious or **r**ecovered if an SIR model is used. The movement between disease states is probabilistic depending on factors such as the disease characteristics, and the contact duration. For each simulation of a disease outbreak, we note contact between individuals, time of contact and whether disease transmission occurs [17]. Details about the individual-based model used in this study are presented in chapters 2-5. We therefore refrain from providing extensive details on this particular model.

The individual-based model allows the study of how individual behavior affects the outcomes observed in the system and can also be used to study how changes in the system influences individual behavior [64]. However, there are pros and cons to using the individual-based model for studying epidemics. One major difficulty is the limited understanding and ability to predict human behavior, which is variable [84]. Also during disease outbreaks, the focus is mainly at the population level; data is typically collected on the number of individuals reported to be infected by geographical region and not on infections resulting from individual-to-individual interactions. The lack of such data therefore makes it difficult to find appropriate parameters for individual-based models. Population-level models are typically used in most short term predictions, while individual-based models are used as tools for public policy [8, 17, 45, 67, 84].

Population-level models are typically defined using differential equations. Kermack and McKendrick developed the first known mathematical and population-level model applicable to the study of influenza outbreaks [86]. The model is divided into three compartments: **S**usceptible, **I**nfectious and **R**ecovered. Such models are known as compartmental models. The Kermack-McKendrick model describes the number of infected individuals over time in a closed population with homogeneous mixing. The version of the SIR model used in many publications is without social or spatial structure. Individuals in the population transmit influenza with probability β to susceptible individuals as shown in (eqn 1.1). The rate at which a susceptible individual gets infected is generally assumed to be proportional to the incidence of infection in the population. Infected individuals are also assumed to recover at a constant rate γ [63]. The rates of infection β and recovery γ are the two parameters estimated in the basic SIR model.

$$\begin{aligned} S'(t) &= \frac{\beta S(t) I(t)}{N} \\ I'(t) &= \frac{\beta S(t) I(t)}{N} - \gamma I(t) \\ R'(t) &= \gamma I(t) \end{aligned} \tag{1.1}$$

S(t), *I(t)* and *R(t)* in eqn 1.1 represent number of susceptible, infectious and recovered individuals at time t. There have been extensions to the SIR model to include an exposed state (SEIR), spatial dynamics, and interventions. See [63] for a detailed review of models for infectious disease transmission.

We use the individual-based model in this dissertation for several reasons. The model is available and efficient. Studies focusing on specific aspects of the population demographic (such as age groups) are easy to design and implement. In addition, we can study different populations by constructing a synthetic social network for each population. Compartmental models or other aggregated models can also be used to replicate the results observed in this dissertation.

1.2 Forecasting the Epidemic Curve

Similar to current approaches, the methods in this dissertation aim to predict the epidemic curve by predicting both the daily infected-counts and other aspects such as the peak infected rate, time to the peak and magnitude of the epidemic [38, 65, 101, 104, 105]. Both individual-based and population-level models have been used in forecasting the epidemic curve. However, the available literature on real-time forecasting of the epidemic curve is limited. Existing methods are discussed in Chapters 2 to 5.

Several of the existing approaches are based on compartmental models, which assume homogeneous mixing in the population [101, 104, 105]. On the contrary, individual-based models such as that used in this dissertation have been shown to more closely mimic realistic social networks [9, 47]. In addition, existing methods that rely heavily on the branching process and on maximum likelihood procedures for prediction have been shown to be unreliable. Limitations inherent to these approaches include: difficulty in estimating likelihood based parameters since the infection and removal times cannot be observed for each infected individual and calculation of the likelihood can be intractable for large populations [94].

The model driven approach presented in this dissertation is different from existing methods, which aim to predict the next point on the time series epidemic curve by using the previous points as a best fit solution. In this scenario, we try to learn the parameters of a causal model that might have given rise to the dynamical phenomenon observed during an epidemic. The novelty of this approach lies in: (i) the combination of different methods to predict and update assumptions about an epidemic in real-time, (ii) the use of an individual-based model with a network representation of different populations and (iii) the creation of a library with historical data to improve the efficiency of forecasting.

There are many challenges to forecasting the epidemic curve using surveillance data. These include difficulties in obtaining detailed real-time epidemic data and a meager understanding of the natural history of a novel disease. Real-time prediction of the epidemic curve also requires a combination of good surveillance systems and epidemic modeling assumptions

[38, 90, 105]. Model assumptions affect parameter estimates and subsequent predictions.

1.3 Methodology and Contribution of this Dissertation

This dissertation presents a model-based reasoning approach for forecasting the epidemic curve. The method used in this study has not been previously used in real-time prediction of the epidemic curve. The problem of forecasting the epidemic curve using a model-based reasoning approach can be divided into three parts, which constitute the specific aims for this dissertation:

1. Quantify the relationship between changes in the three major disease parameters (infectious period distribution, incubation period distribution and transmissibility) to the dynamics of simulated epidemics. This aids in developing a better understanding of how changes in these parameters would affect the dynamics of predicted epidemics.

2. Differentiate between epidemic curves from different disease models given that the data are sequentially updated on each day of the epidemic. This involves finding a method or combination of methods that would enable timely and accurate prediction of the epidemic curve given a library of past and simulated epidemics. The method(s) should correctly identify epidemics similar to those in the library and epidemics different from those in the library based on the partial epidemic curve.

3. Find new parameters to model an epidemic different from those in the library. The search method would enable early prediction of the disease model parameters, which accurately represent the ongoing epidemic. The efficiency of the method is extremely important since results are needed in real-time.

To fulfill these aims, in Chapter 2, we perform a sensitivity analysis on the disease parameters in the individual-based model. The goal of the sensitivity analysis is to gain a proper understanding of the mechanism controlling the dynamics of the simulated epidemics. The dynamics of a simulated epidemic can be influenced by the structure of the interaction network across which the outbreak spreads and details of the disease model [17, 46]. Eubank et al. [46] developed a method to show that the amount of detail available in the network model is sufficient implying that the introduction of additional detail does not reduce the variability in possible outcomes. Assuming the structure of the network is sufficient, we focus on the effects of changes in the disease model parameters to the dynamics of the simulated epidemics.

With the exclusion of the sensitivity analysis study, the steps of the proposed approach are summarized in Figure 1.2. In Chapter 3, we compare seven supervised classification methods (random forest, support vector machines, nearest neighbor with three decision rules, linear

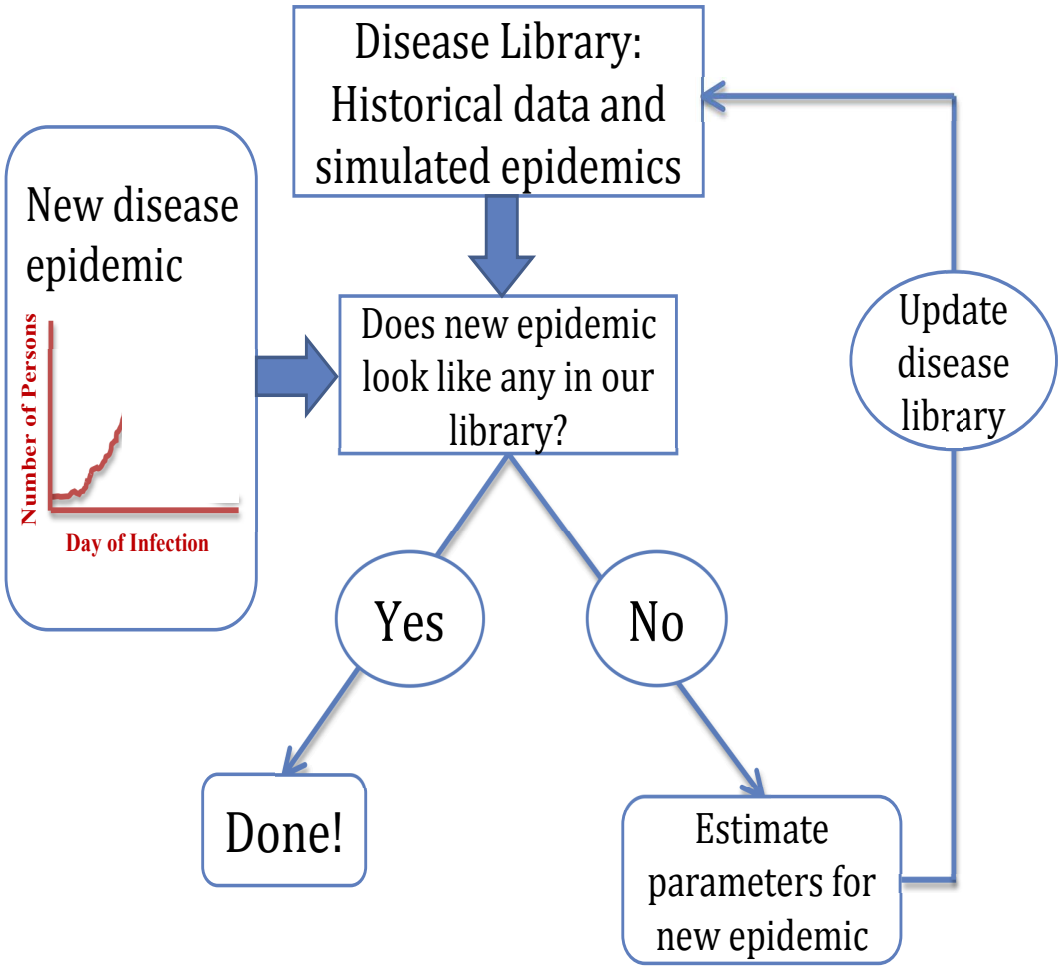


Figure 1.2: Summary of methodology. We develop a library of past and simulated epidemics. Given surveillance data for a current epidemic, we compare the partial surveillance epidemic curve to those in the library. The novel epidemic is either assigned to a case in the library or identified as being different from those in the library. If the epidemic is different from those in the library, we estimate the model parameters, forecast the epidemic curve and update the library.

and flexible discriminant analysis) with the goal of finding a method for identifying epidemic curves similar to those in the library. The accuracy of the methods is compared using a performance metric based on the McNemar test.

The methods in Chapter 3 have the limitation of being fully supervised implying a new outbreak will always be assigned to the library. We therefore present a method to identify epidemics different from those in the library in Chapter 4. Given surveillance data, the Dirichlet process model is used to either classify the novel epidemic into the library or identify the epidemic as being different from those in the library. If the epidemic is classified as belonging to the library, we use the individual-based model to evaluate interventions.

However, if the epidemic is different from those in the library, we present a simulation optimization approach to estimate the parameters and forecast the epidemic curve in Chapter 5. The Nelder-Mead simplex method is used in optimization. The library is updated after each run through the procedure if the new epidemic is different from those in the library. The prediction procedure is repeated for the entire duration of the epidemic.

Chapter 2

Sensitivity Analysis of an Individual-Based Model for Simulation of Influenza Epidemics

Elaine Nsoesie [1], Richard Beckman[1], and Madhav Marathe[1,2]

[1] Network Dynamics and Simulation Science Laboratory,
Virginia Bioinformatics Institute, Virginia Tech,
Blacksburg, Virginia, USA

[2] Computer Science Department, Virginia Tech,
Blacksburg, Virginia, USA

Abstract

Individual-based epidemiology models are increasingly used in the study of influenza epidemics. Several studies on influenza dynamics and evaluation of intervention measures have used the same incubation and infectious period distribution parameters based on the natural history of influenza. A sensitivity analysis evaluating the influence of slight changes to these parameters (in addition to the transmissibility) would be useful for future studies and real-time modeling during an influenza pandemic.

In this study, we examined individual and joint effects of parameters and ranked parameters based on their influence on the dynamics of simulated epidemics. We also compared the sensitivity of the model across synthetic social networks for Montgomery County in Virginia and New York City with demographic and rural-urban differences. In addition, we studied the effects of changing the mean infectious period on age-specific epidemics. The research was performed from a public health standpoint using three relevant measures: time to peak, peak attack rate and total attack rate. We also used statistical methods in the design and analysis of the experiments.

The results showed that: (i) minute changes in the transmissibility and mean infectious period significantly influenced the attack rate; (ii) the mean of the incubation period distribution appeared to be sufficient for determining its effects on the dynamics of epidemics; (iii) the infectious period distribution had the strongest influence on the structure of the epidemic curves; (iv) the sensitivity of the individual-based model was consistent across social networks investigated in this study and (v) age-specific epidemics were sensitive to changes in the mean infectious period irrespective of the susceptibility of the other age groups. These findings suggest that small changes in some of the disease model parameters can significantly influence the uncertainty observed in real-time forecasting and predicting of the characteristics of an epidemic.

2.1 Introduction

Sensitivity analysis is the study of the contribution of different parameters to the uncertainty present in the outcome of a system [72, 114]. Various scientific fields use sensitivity and uncertainty analysis to: (i) highlight important and remove irrelevant data, (ii) optimize the design of a system and (iii) rank by importance the influence of various parameters on the behavior of a system [26, 59]. The scope of a sensitivity analysis procedure can be local or global. Local analysis aims to examine the effects of local deviations of a parameter or a chosen trajectory in the parameter space [54]. Alternatively, global analysis is used to evaluate the entire parameter space in addition to interactions between parameters to determine all of the system's critical points [26, 69]. Methods for sensitivity analysis can be either statistical or deterministic [72, 115]. However, selection of methods depends on the purpose and system under study. Typically, complex systems (models) are computationally expensive which tends to limit the scope of a sensitivity analysis.

In this study, we perform sensitivity analysis on a complex individual-based stochastic epidemiology model for the study of influenza epidemics. Individual-based models are increasingly used in the study of the dynamics of infectious diseases and evaluation of methods for controlling the spread [45]. For a few examples, see [92], [37] and [8]. These models capture human-to-human disease transmission by creating synthetic populations with time-varying contact networks [17]. The level of detail used in these models increases the complexity but also enables the model to more realistically capture the heterogeneity present in the natural system [45]. Although the level of realism is beneficial, the behavior of the systems can be challenging to explore analytically due to the large number of parameters [72]. In addition, validation of parameters used in these models tends to depend on qualitative comparison of model behavior and expert opinion [71].

In several studies, the sensitivity analysis of individual-based models of epidemic dynamics have been used to evaluate the effect of disease parameters on public policy related questions. For example, for a model aimed at simulating Smallpox epidemics, the sensitivity analysis could focus on model assumptions relating to how changes in individual behavior after infection might influence the observed outcomes such as in the study by Burke et al. [25]. For individual-based models used in the study of influenza-like epidemics, the sensitivity analysis could focus on changes in interventions and response strategies such as in the studies by Halloran et al. [67] and Germann et al. [57]. In this study we aim to explore the sensitivity of an individual-based epidemiology model to changes in the assumptions made regarding the characteristics of the disease. The disease model is one of the two major components of the individual-based model. The other is a state-of-the-art behavioral model which consists of synthetic populations and time-varying social networks. There have been several studies validating the structural aspects of the individual-based model [14, 47, 67] and the sufficiency of the amount of detail used in its development [46]. However, there have not been any studies exploring the epidemiological and mathematical assumptions relating to the underlying process describing disease transmission.

2.1.1 Parameters

We perform sensitivity analysis on the parameters of a networked **S**usceptible-**E**xposed-**I**nfectious-**R**ecovered (SEIR) disease model. The four disease states in the SEIR model (Figure 2.1) are used in describing within host disease progression and between host influenza transmission in the social network [17]. To simplify the disease process, three parameters are used: transmissibility, incubation period distribution and infectious period distribution. The transmissibility is the diffusion intensity of a disease through a population. The transmissibility is usually measured using the reproductive number - the number of secondary cases for each primary case. The incubation period is the interval during which infected individuals cannot spread the disease and usually lasts between 1-4 days for seasonal influenza [32]. The infectious period duration is the period during which infected individuals can transmit the disease to susceptible individuals. During typical seasonal influenza epidemics, infectious individuals can "shed the virus a day before onset through 5-10 days after illness onset" [32]. In this model, the incubation and infectious periods are described using discrete probability distributions since individuals in the population tend to have different incubation and infectious period durations based on their age and health status. The initial (base case) parameters based on the natural history of seasonal influenza have been used in several studies [46, 61, 67, 103]. The base case incubation period distribution is defined as follows: $t_{E\rightarrow I} = 1, 2$, or 3 days with probability $0.3, 0.5$ or 0.2, respectively. This implies an infected individual can have an incubation period duration of $1, 2$, or 3 days with probability $0.3, 0.5$ or 0.2. Likewise, the infectious period distribution is given by: $t_{I\rightarrow R} = 3, 4, 5$ or 6 days with probability $0.3, 0.4, 0.2$ or 0.1, respectively. To our knowledge, there is no defined standard for performing sensitivity analysis on parameters which are discrete probability distributions, especially not in individual-based models. Therefore, we use a combination of statistical methods and present a sensitivity analysis study which provides a framework for future studies.

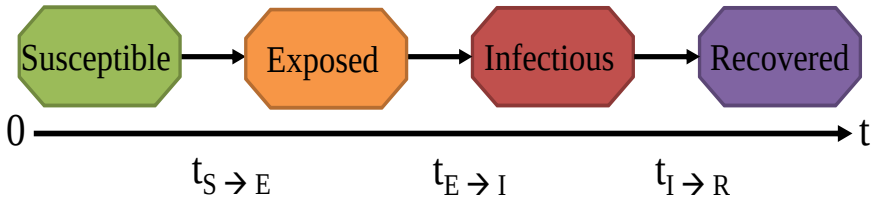


Figure 2.1: **The SEIR disease model used in describing disease progression within the agent-based model.** Individuals move through four health states. Susceptible individuals become exposed to the disease due to contact with an infected individual. After being exposed, an individual becomes infectious. An infectious person recovers at the end of the infectious period. Recovered individuals can no longer spread the disease.

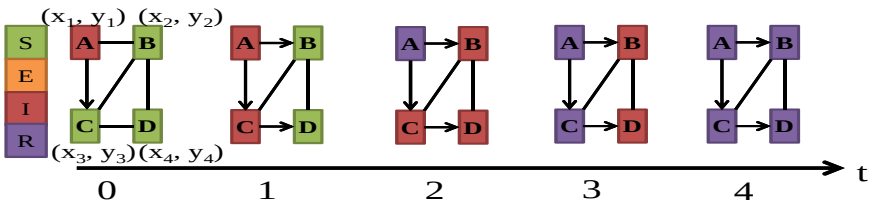


Figure 2.2: **Example of SEIR model describing between host disease transmission.** There are four individuals/nodes in the contact network and five edges. Different colors indicate different health states. Also each node is randomly assigned an incubation and infectious period (x_i,y_i) sampled from a discrete distribution. For simplicity, each of the nodes in this example has an incubation period of 0 days and an infectious period of 2 days. On day 0, node A is infectious while all other nodes are susceptible. On day one, node C is infected due to contact with node A and on day two, node A recovers, while nodes B, C and D are infectious. There are no susceptible nodes after day two. On day three, node C recovers. Finally on day four, all nodes recover. Unlike this example, the networks used in this study have approximately 76000 and 20 million nodes.

2.1.2 Aims and Relevance

This study is motivated by the need to improve methods for real-time modeling and predicting during a pandemic [90]. The usefulness of real-time modeling was illustrated during the 2009 influenza pandemic using both compartmental models [105] and individual-based models [38]. To improve real-time epidemic modeling using individual-based models, we need an in-depth knowledge of the effects of the disease parameters on the dynamics of predicted epidemics. We therefore explore the following aims: (i) evaluate individual and joint effects, and rank parameters based on influence on simulated epidemics and (ii) compare the sensitivity of the model across age groups and social networks with demographic differences. Studies have indicated that differences in the transmission of the 2009 H1N1(A) virus in various regions was partly due to differences in population demographics [95, 106]. In addition, several studies have suggested that school children tend to influence the propagation of influenza [13, 96, 111, 113]. Both observations are further investigated in this study under different parameter combinations. Comparison of the sensitivity across networks with demographic and urban-rural differences is essential since this would indicate whether results observed for one social network are reproducible in another. The experiments and analysis are expected to further our knowledge of how to improve real-time modeling of epidemics using such models.

A good understanding of how the disease parameters influence the dynamics of simulated epidemics would aid in the prediction of the epidemic curve and estimation of parameters during an epidemic of a novel influenza virus. The epidemic curve for the purpose of this study is the daily number of infected for the duration of an epidemic. There are several possible approaches for real-time estimation of disease parameters during an epidemic [101, 104, 105]. However, uncertainty in the data collected during an epidemic can result in extremely unreliable results [90]. In addition to improving input data used in prediction, a study of how minute changes in the model parameters affect the predicted outcomes would

be invaluable. For instance, [103] proposed a real-time epidemic curve prediction method based on matching surveillance data for an ongoing epidemic to epidemics simulated using parameters from previous outbreaks. If the new epidemic cannot be matched to any of the simulated cases, then a combination of expert opinion and search algorithms are used in suggesting new parameters. A sensitivity analysis study would enable easy assessment of the initial values and selection of the parameter space for the search algorithm.

To accomplish these aims, we explore the space of possible incubation and infectious period distributions by generating distributions with the same mean and also by perturbing the probabilities of the base case distribution. This process is further discussed in the problem definition. We use mono-factorial and full factorial designs to study the effects of each parameter and joint effects due to interactions between parameters. We also use principal components clustering, analysis of variance, and Pearson correlation to determine the level of sensitivity in the model [72]. Moreover, we use the epidemic curve as the main outcome measure. These procedures and reasons for selecting them are discussed in later sections.

The complexities of this study lie in quantifying the sensitivity of an individual-based model to changes in parameters which are discrete probability distributions. As shown in Figure 2.2, each individual in the population is randomly assigned an incubation and infectious period duration sampled from a probability distribution. For each replicate of an epidemic simulation, individuals receive new samples from the incubation and infectious period distribution. Variations in these parameters add to the stochasticity present in the model. Unlike most studies where the difficulty of sensitivity analysis is introduced by the number of parameters, in this study the complexity lies in the model and type of parameters. To our knowledge, no previous studies have focused on quantifying the sensitivity of an individual-based model to changes in parameters which are discrete probability distributions.

2.1.3 Overview of the Individual-based Model

The process of developing the individual-based model is described briefly since it is not a novel contribution of this work. A detailed description of the individual-based model can be found in [17]. The individual-based epidemiology model consists of a social network and a disease model. The disease model is described in the next section. The model includes representation of each individual in a population along with activity schedules describing their movements. Contacts that occur between individuals as they move about their daily schedule are represented using time-varying contact networks. These contacts can result in disease transmission depending on the duration, type of contact and health state of the individuals. The main steps involved in the construction of the individual-based model are: creation of the synthetic population; activity and location assignment; the definition of an infectious disease model and interventions both pharmaceutical and non-pharmaceutical used in controlling the propagation of the disease.

The synthetic population is created using demographic information, survey and land use

data. Synthetic individuals are defined with specific sets of demographic variables and assigned to households based on US census data provided in SF3 and PUMA (Public Use Microdata Area) files [7]. Each individual is assigned to a household. Some of the demographic information available for each individual in the synthetic population includes age, education level, and household income. The synthetic populations are created to represent the real population as realistically as possible, while maintaining confidentiality. A census collected at the block level of the synthetic population is statistically indistinguishable from the census data [14]. Individuals in the synthetic population interact with each other and their environment to produce time-varying social contact networks. Further information on the creation of the synthetic social network can be found in [14], [116] and [117].

In addition to having specific demographic information, each node in the synthetic population is also allocated activities based on thousands of responses to a time-use survey for a specific geographical region. As expected, there will be differences in the time-use survey collected in different geographical regions such as New York and Montgomery County due to demographic differences. The National Household Transportation survey is used in this model to create the activity templates. Activities can include shopping, work, daycare, etc. Individuals in the synthetic populations are matched to the survey households using decision trees based on demographic variables. Activities are then assigned realistic geographical locations using a gravity model based on land-use data [14]. In addition, each activity is assigned a start and end time, resulting in a minute-by-minute schedule for each synthetic individual. Currently, this modeling approach is considered the *de facto* standard for travel demand models in transportation science [8].

The social contact network results from interactions between individuals in the synthetic population based on their activity schedules. Individuals are represented as vertices in a graph and edges are used to describe interactions between individuals and locations (Figure 2.2). Since individuals can visit a location more than once, multiple edges can be used to describe these visits. The modeling approaches used in each step of creating the individual-based model can be found in [8]. Additional information can also be found in [19] and [20].

Disease Model

Using a computational model such as the basic SEIR model (Figure 2.1), disease transmission is explained within the previously described network. This implies that each individual moves through four disease states (susceptible, exposed, infectious and recovered), and transmission is dependent on the contact between two individuals in the susceptible and infectious states. The transition between disease states is probabilistic and timed (e.g. represented by a probability distribution), and can also depend on the demographics of an individual (such as their age, work and health status). As individuals go about their different activities (such as shopping, and work) they come in contact with other individuals. Through this process, the disease can be transmitted from an infected individual to a susceptible individual.

See (Figure 2.2) for an example illustrating disease transmission between nodes in a basic network.

For the disease model used in this study, the probability that an infectious person i infects a susceptible person j is given by:

$$Pr(w(i,j)) = 1 - (1-\tau)^{w(i,j)} \tag{2.1}$$

where τ is the probability of transmission per unit of contact time between persons i and j [17]. *w(i,j)* is the contact duration between i and j measured in seconds. The SEIR model is one of the simplest disease models which can be used for this individual-based model. Additional information can be added to the disease model to better capture different infectious diseases. In addition, intervention options such as vaccination, antiviral and social distancing are included in the model to control disease spread. Single interventions or a combination of interventions can be introduced either at the start or during a simulated epidemic.

Simulations are run by randomly selecting a number of people to introduce the disease into the population. During each simulation, we keep track of information on all contacts, duration of contact, and which contacts result in infection. Information on the contacts resulting in disease transmission, the vulnerability of the nodes, the epidemic size and epidemic curve are some of the information used in studying the dynamics of an epidemic. See [17] for more details on the individual-based model.

2.1.4 Problem Definition

The following formulation is used in defining the problem. Given a distribution Γ: a vector with finite support $(p_1, p_2, p_3, \ldots, p_k)$ where $\sum_{i=1}^{k} p_i = 1$: k is the number of days and $p_i's$ are the probabilities of observing each day. In terms of the incubation (infectious) period distribution, $(p_1, p_2, p_3, \ldots, p_k)$ are the probabilities that an infected (infectious) individual will have an incubation (infectious) period of k days.

There are several possible techniques for generating new distributions depending on the aim of the study. A likely procedure would involve placing a probability distribution such as the beta density [134] over Γ. New distributions can be created by changing the shape parameters. This implies that a new distribution for the incubation and infectious period is defined by systematically generating new vectors, $(q_1, ...q_k)$ such that $\sum_{i=1}^{k} q_i = 1$. If this process is carried out naively, an infinite number of possible distributions can be generated. A simple approximation is to use a step size d to perturb one q_i to another. The mean of the distribution is shifted by a few days using such perturbations. This method can be used to study how changes in the mean of the incubation and infectious periods affect the behavior of the model.

Table 2.1: Table of notations

Notation	Description
t	time
n, k	counts
τ	unit of contact time
p, g	probability
x_i, y_i	factor
$z(x_i)$	epidemic curves

An alternative to the previously described procedure would involve defining the mean duration of incubation or infectiousness while randomly generating new distributions with different variances. This technique would enable the study of the effects of the variance of the incubation and infectious period distributions on the dynamics of the epidemics. In addition, the random generation of distributions would result in distributions with different shapes which might not be epidemiologically relevant for influenza. However, this allows for a mathematically exploration of the parameter space. It also enables the applicability of this procedure to models of other infectious diseases with parameters that are distinct from those based on the natural history of influenza. In this study, we use both perturbations of the $q_i s$ and random generation around the same mean. We selected these procedures based on their simplicity and results from preliminary studies which indicated that these techniques are sufficient for investigating the aims in this study.

The above procedures are for the incubation and infectious period distributions only. For the transmissibility, which is a real number, values are selected to simulate epidemics similar to seasonal influenza, previous pandemics and more extreme epidemics.

To perform sensitivity analysis on these three quantities, we explore the mapping $[x_i, z(x_i)]$, $i = 1, 2, \ldots, I$, where x_i represents the parameters: transmissibility, infectious period distribution, and the incubation period distribution [72]. $z(x_i)$ are the epidemic curves resulting from different parameter combinations. See Table 2.1 for a summary of the notations used in this study.

2.2 Methods and Analysis

We used statistical methods and tests in the design and analysis of the experiments in this study. Since no statistical methods are universally accepted as infallible, we chose methods based on their applicability to the study [69]. As previously mentioned, for the study design we used factorial experiments [21]. Mono-factorial experiments were used in studying the sensitivity of the model to each of the parameters. A full factorial design was used in evaluating the influence of factor interactions on the observed outcomes. In addition, to

find underlying groupings within the collection of curves from all factorial experiments, we used principal components clustering [15]. The groupings of curves with similar structures aided in determining the influence of the different parameters on the shape and form of the epidemic curves. These procedures are described below.

2.2.1 Public Health Measures

As earlier stated, the epidemic curves were used as model outcomes $z(x_i)$. To facilitate comparison of the epidemic curves, we performed the analysis from a public health standpoint using three relevant measures: peak attack rate or peak infected proportion, time to peak or peaking time and total attack rate or total infected proportion as shown in Figure 2.3. The *attack rate = infected-counts/population size*. The peak attack rate indicates the time point at which there would be an increase in the need for public health resources such as nurses and hospitals during an epidemic. The time to peak (or peaking time) suggests the best time point for implementing control measures such as vaccines, antivirals, and sequestration. Lastly, the total attack rate can be used to quantify the disease's effect based on the morbidity and mortality rates [46, 109]. These measures are also useful for real-time forecasting and prediction of the impact of an epidemic.

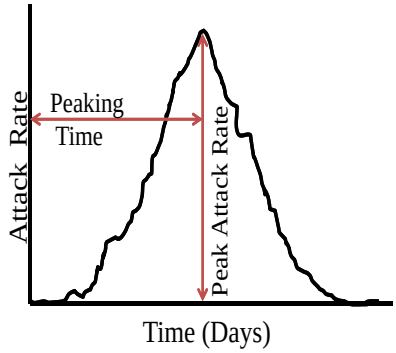


Figure 2.3: **Public health measures used in comparing epidemic curves:** peak attack rate, peaking time and total attack rate (proportion).

Analysis of variance (ANOVA) was used to evaluate differences in the mean of the public health measures given changes in the parameters. In its simplest form, ANOVA is an extension of the t-test. ANOVA is typically used to test whether the means of several groups are equal. The groups in this study were the different sets of epidemics/experiments resulting from changes in the parameters. We tested whether the means of the various peak attack rates, peaking times and total attack rates for the different sets of epidemics were equal. If a statistically significant difference was observed, we used a pairwise t-test with Bonferroni correction to find which pairs of experiments were statistically different.

2.2.2 Social Networks

We used two synthetic social networks in order to compare possible effects of demographic and rural-urban differences. The demographic information for the synthetic populations in the social networks were based on the population of Montgomery County in Virginia and New York city with approximately 76000 and 20 million synthetic individuals respectively. The same disease model parameters were used in simulating epidemics over the synthetic social networks. As previously mentioned, random sampling was used in the assignment of the incubation and infectious period to the nodes in the network. Random assignment was selected instead of targeted assignment because it is simpler to implement, and increases the flexibility of designing and running experiments. In addition, random assignment of these parameters has been used in most studies [8, 37, 46, 61, 103].

2.2.3 Factorial Experiments

In the mono-factorial design, we varied a single parameter and held the others fixed. In the first set of experiments, we altered the incubation and infectious period parameters by shifting the mean by a single day. The main reason for doing this was to evaluate how changes in the mean infectious and incubation period distribution would affect the dynamics of simulated epidemics. This was investigated for two reasons: (i) the natural history of infection for the 2009 H1N1(A) pandemic was similar to that for seasonal influenza [40] and (ii) typically the mean incubation and infectious period are reported during an influenza epidemic, not the entire distribution.

We varied each of the parameters over five values and studied the effects of these changes to the dynamics of the simulated epidemics. First, using five transmissibility values, one incubation period distribution and one infectious period distribution, we simulated five sets of epidemics. The five transmissibility values were in the range: 4.6×10^{-5} to 8.6×10^{-5} per sec/contact time. As earlier mentioned, these values were selected to simulate epidemics similar to previous pandemics and more extreme outbreaks. Next, to evaluate the effects of the infectious period distribution on the epidemic curve, we experimented with five infectious period distributions, a fixed transmissibility and one incubation period distribution. The mean of the infectious period was altered across two to six days. Recall that the mean of the base case infectious period distribution was four days. So the new distributions either had a mean that was greater or less by a day or two compared to the base case. Finally, to evaluate the effects of the incubation period, we generated five incubation period distributions while holding the transmissibility and infectious period distribution fixed. The mean incubation period fluctuated between one to five days. All epidemics were simulated over both social networks. In addition, each epidemic was replicated twenty-five times. These experiments are labeled as Exp.1-3 in Table 2.2. The results from this set of analysis led to the second set of experiments.

In the second set of experiments, we further assessed the effects of the mean and variance of the incubation and infectious period distributions on the dynamics of an epidemic. To evaluate the effects of the variance of the infectious period duration on the dynamics of the simulated epidemics, we randomly created thirteen infectious period distributions with mean infectious period duration of four days. Using two similar incubation period distributions and one transmissibility, we simulated twenty-six sets of epidemics. Each epidemic was replicated twenty-five times. These experiments are labeled Exp.5 in Table 2.2. Next, to understand the influence of the variance of the incubation period duration on the dynamics of simulated epidemics, we randomly generated twelve distributions. Each of the incubation period distributions had a mean of approximately two days (Exp.4 in Table 2.2). Using these twelve unique incubation period distributions, one transmissibility and one infectious period distribution, we simulated twelve epidemics with twenty-five replicates each. The mean of the incubation and infectious period distributions were set at two and four days respectively because those were the means of the base case distributions [67]. We analyzed how the variance of these distributions affected the dynamics of the epidemics. The simulated epidemics were compared based on the results from an ANOVA and t-test with Bonferroni correction for multiple comparisons. We tested for significant differences in the mean of the total attack rate, peak attack rate and time to peak.

Using a full factorial design, we varied each of the parameters across three levels, resulting in twenty-seven combinations/experiments. The parameters used in the full factorial experiments were generated by shifting the mean of the base case distributions by a single day. This resulted in infectious period distributions with means of three, four and five days. The incubation period distributions had a mean of one, two and three days. This variation in the parameters was done for the same reason as the first set of experiments. Each parameter was defined at three levels as shown in Exp.6 in Table 2.2. The parameter levels were labeled: (t1, t2, t3), (inc1, inc2, inc3) and (inf1, inf2, inf3). The epidemics each had fifty replicates and were simulated for a duration of three hundred and fifty days so as to accommodate epidemics with durations longer than the typical influenza season in the United States. The epidemics were analyzed using ANOVA and clustering with principal components.

As previously stated, due to observations made during the 2009 H1N1 pandemic, we also evaluated the sensitivity of the model based on disease spread within age groups. The four age groups were: pre-schoolers, school-agers, adults and seniors. The epidemics were simulated using the parameters for Exp.3 in Table 2.2 because the full factorial analysis indicated that the infectious period distribution had the strongest influence on the epidemics compared to the other parameters. There was one transmissibility, one incubation period distribution and five infectious period distributions. The infectious period distributions had different mean durations ranging from two to six days. We evaluated how differences in the mean infectious period duration affected the dynamics of the age-specific epidemics. For each of the age-groups, we compared the time to peak, total and peak attack rates. We performed this analysis under three scenarios. In the first case, we assigned the same susceptibility to all age-groups. In the second scenario, children and elderly were allowed a higher susceptibility.

In the third case, only children had a higher susceptibility compared to all other age groups. These settings were modeled based on observations made during seasonal influenza epidemics and the 2009 H1N1 pandemic. In addition, several studies have suggested that school-age children are highly susceptible to infectious disease spread due to regular incidence of close proximity interactions [113]. We therefore studied how the epidemic dynamics for each age group (especially children) varied with changes in the infectious period distribution and age-specific susceptibility.

All experiments were implemented in the individual-based model under the base case scenario: no interventions were introduced to control the spread of the epidemic. To capture the heterogeneity present in the model and elucidate the influence of the social networks, each epidemic was replicated between 25 to 200 times. See Table 2.2 for a summary of the parameters used in these epidemics.

2.2.4 Principal Components Clustering

To uncover the parameter with the most significant influence on the structure of the epidemic curves, we used principal components clustering to find underlying groupings within the 1350 epidemic curves from the full factorial experiments. Typically, replicates of the same epidemic tend to have similar characteristics. Therefore, one would expect epidemics with the same parameters to fall into the same groups if clustered. However, this is not always the case due to the stochastic nature of the model and imperfection of clustering algorithms. In this study, we expected that by clustering the epidemic curves, we would observe patterns in the distribution of epidemics into different clusters based on the levels of the parameters. The parameter having the strongest influence on the structure of the epidemic curves should result in groupings based on different levels.

Epidemic curves can be viewed as time series since infected-counts are collected over fixed time intervals (e.g. on a daily or weekly basis). The daily infected-counts can be represented as a vector. The process of clustering based on principal component analysis was carried out as follows: first we estimated principal components based on the variance-covariance matrix of the vectors representing the epidemic curves. Next, we fit a linear regression equation to centered daily infected-counts for each epidemic curve to the nonlinear principal components. Lastly, the regression coefficients were clustered into groups using k-medoids, which is a robust version of the k-means algorithm. For additional information on this process see [15]. Several other clustering approaches could be used, however, the principal components clustering method was selected because the process captures curves with similar structures and the regression coefficients provide a description of the characteristics of curves within each cluster. The clustering was based on the first six principal components since those explained over 80% of the variance. Using additional components did not improve the clustering. We decided to use nine clusters after experimenting with different groupings. Increasing the number of clusters failed to provide better separation.

Table 2.2: The experimental description indicates the type of statistical design and the parameter under focus. We present the incubation and infectious period distributions as follows: $k : p_i$ where k is the day and p the probability.

Labels	Experiment Description	Transmissibility	Incubation Period Distribution	Infectious Period Distribution
Exp.1	Mono-factorial experiment with focus on slight changes to the transmissibility	4.6×10^{-5} 5.6×10^{-5} 6.6×10^{-5} 7.6×10^{-5} 8.6×10^{-5}	1:0.3 2:0.5 3:0.2	3:0.3 4:0.4 5:0.2 6:0.1
Exp.2	Mono-factorial experiment to evaluate the effect of changes to the mean of the incubation period distribution	6.0×10^{-5}	1:0.3 2:0.5 3:0.2 0:0.3 1:0.5 2:0.2 2:0.3 3:0.5 4:0.2 3:0.3 4:0.5 5:0.2 4:0.3 5:0.5 6:0.2	3:0.3 4:0.4 5:0.2 6:0.1
Exp.3	Mono-factorial experiment to investigate the effect of changes to the mean of the infectious period distribution	6.0×10^{-5}	1:0.3 2:0.5 3:0.2	3:0.3 4:0.4 5:0.2 6:0.1 2:0.3 3:0.4 4:0.2 5:0.1 1:0.3 2:0.4 3:0.2 4:0.1 4:0.3 5:0.4 6:0.2 7:0.1 5:0.3 6:0.4 7:0.2 8:0.1
Exp.4	Mono-factorial experiment to study the effect of fixing the mean while changing the variance and shape of the incubation period distribution	6.0×10^{-5}	1:0.3 2:0.5 3:0.2 1:0.2 2:0.5 3:0.3 1:0.3 2:0.4 3:0.3 1:0.325 2:0.45 3:0.225 1:0.25 2:0.40 3:0.35 1:0.3335 2:0.333 3:0.3335 1:0.35 2:0.3 3:0.35 1:0.4 2:0.2 3:0.4 1:0.375 2:0.25 3:0.375 1:0.333 2:0.334 3:0.333 1:0.2 2:0.4 3:0.4 1:0.233 2:0.3 3:0.467	3:0.3 4:0.4 5:0.2 6:0.1
Exp.5	Mono-factorial experiment to study the effect of fixing the mean while changing the variance and shape of the infectious period distribution	6.0×10^{-5}	0:0.223 1:0.405 2:0.372 0:0.123 1:0.605 2:0.272	2:0.1 3:0.2 4:0.3 5:0.4 2:0.15 3:0.2 4:0.3 5:0.2 6:0.15 2:0.1 3:0.2 4:0.4 5:0.2 6:0.1 3:0.2 4:0.6 5:0.2 3:0.13 4:0.74 5:0.13 2:0.2 3:0.2 4:0.2 5:0.2 6:0.2 3:0 4:1 5:0 3:0.17 4:0.66 5:0.17 3:0.1 4:0.8 5:0.1 3:0.4 4:0.2 5:0.4 3:0.5 4:0 5:0.5 3:0.25 4:0.5 5:0.25 3:0.333 4:0.334 5:0.333
Exp.6	Full factorial experiment to investigate the impact of interactions between parameters and rank parameters by influence	6.6×10^{-5}[t1] 7.6×10^{-5}[t2] 8.6×10^{-5}[t3]	1:0.3 2:0.5 3:0.2[inc1] 3:0.3 4:0.5 5:0.2[inc2] 2:0.3 3:0.5 4:0.2[inc3]	3:0.3 4:0.4 5:0.2 6:0.1 [inf1] 2:0.3 3:0.4 4:0.2 5:0.1 [inf2] 1:0.3 2:0.4 3:0.2 4:0.1 [inf3]

Table 2.3: Summary of the Various Components in the Analysis

Components	Description
Social Networks	Montgomery County (VA) and New York City
Factorial Experiments	Mono-factorial experiments and a full factorial experiment
Public Health Measures	Peak attack rate (peak infected proportion), time to peak (peaking time) and cumulative attack rate (total infected proportion)
Statistical Methods	Analysis of variance (ANOVA), Pearson Correlation, T-tests with Bonferroni adjustment and Principal Components Clustering

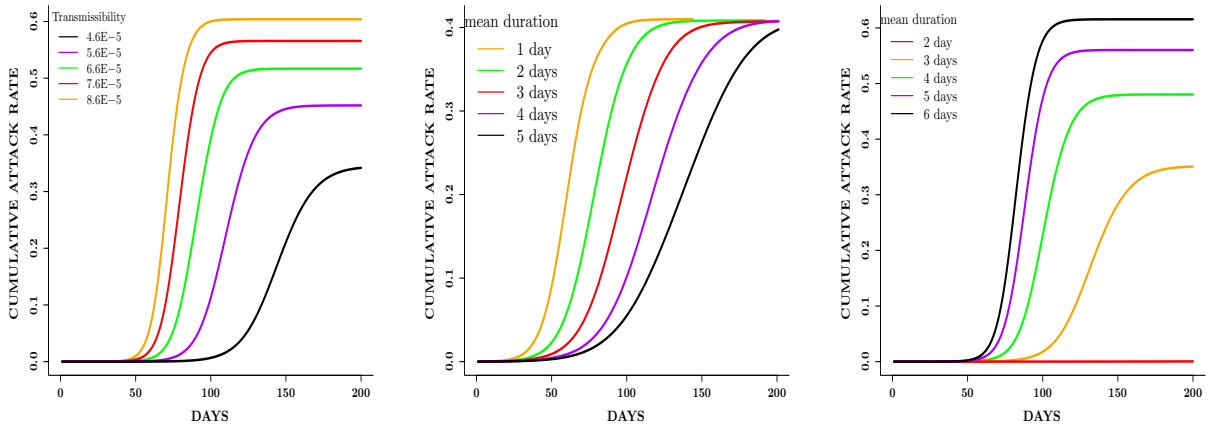
2.3 Results

The results are presented by outcomes. Recall that the aims of the study were to: (i) evaluate individual and joint effects, and rank parameters based on influence on simulated epidemics and (ii) compare the sensitivity of the model across age groups and social networks with demographic differences. The results are presented only for New York since the observations for New York were highly similar to that for Montgomery County as discussed in a later section.

2.3.1 Finding 1. *The transmissibility and mean infectious period duration significantly affected the time to peak, peak attack rate and total attack rate. In contrast, an increase in the mean incubation duration did not significantly affect the total attack rate, but slightly influenced the time to peak and peak attack rate.*

Support: Figures 2.4 and 2.5 display mean epidemic curves from the first three mono-factorial experiments (Exp.1-3) in Table 2.2. Per Figure 2.4, increasing the transmissibility by a value of 1.0×10^{-5} per sec/contact time raised the total attack rate by $[6.3\% - 26.5\%]$. No overlap was observed between the ranges of the total attack rates from the five sets of epidemics. A pairwise comparison using the t-test revealed statistically significant differences ($P < 0.0001$) for all pairs.

Figure 2.4 also displays the results for the incubation and infectious period distributions based on the mean of the distributions. In general, raising the mean of the incubation period distribution did not significantly increase the variability observed in the total attack rates. This is because an increase in the incubation duration did not affect the infectious duration. Although it took a longer duration to become infectious, the time required to



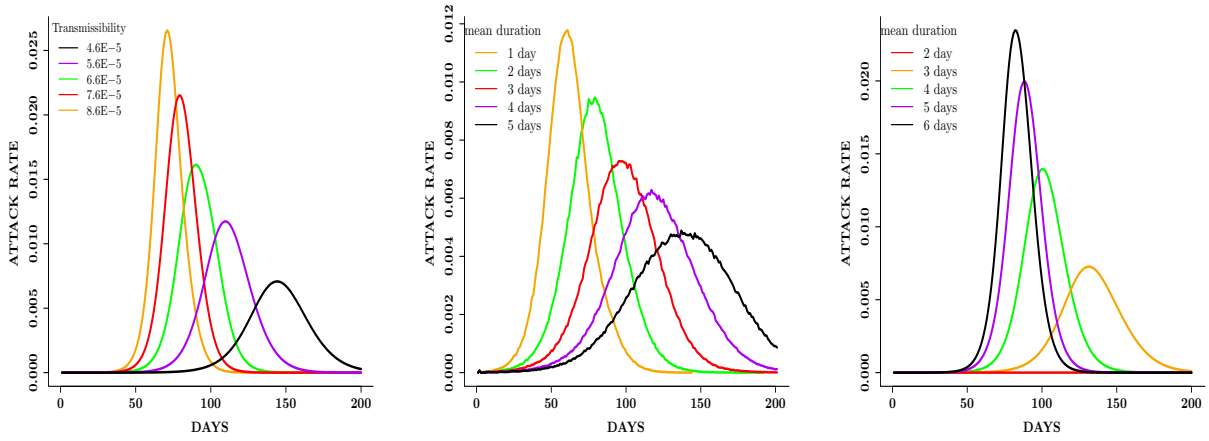
Transmissibility Incubation period Infectious period

Figure 2.4: **Results for mono-factorial experiments for each of the factors: transmissibility, incubation period distribution and infectious period distribution.** Each curve represents the mean cumulative attack rate based on 25 replicates of each experiment. The legends show the transmissibility, mean of the incubation and infectious period distributions. Simulations with mean infectious period of 2 days failed to become epidemics. Note that an increase in the transmissibility and the mean infectious period resulted in an increase in the total attack rate. On the contrary, changing the mean of the incubation period had a minimal effect on the total attack rate.

spread the disease was not affected, thereby resulting in similar total attack rates. In contrast, an increase in the mean of the infectious period raised the total attack rate. Since more individuals had a longer infectious period, the disease affected a larger proportion of susceptible individuals in the population.

The results for the time to peak and peak attack rate are displayed in Figure 2.5. Epidemics with a higher transmissibility tended to peak sooner, and had a higher morbidity rate compared to epidemics with a lower transmission rate. Pairwise t-tests to evaluate differences in the means of these measures suggested statistical significant differences ($P < 0.0001$). This indicated that the transmissibility had a major impact on the dynamics of an epidemic, as would be expected. Though the total attack rates for the incubation period experiments were not statistically significantly different per ANOVA, the mean time to peak and peak attack rates were statistically significantly different ($P < 0.0001$). This implied that changing the mean incubation period did not significantly affect the total number of people infected but rather altered the shape of the epidemic curves.

Furthermore, changes in the infectious period distributions affected the time to peak and peak attack rate in a similar manner as the transmissibility. The peak attack rate rose with an increase in the mean infectious period. Also, epidemics with a higher mean infectious period tended to peak earlier compared to epidemics with a lower mean infectious period. A pairwise comparison using the t-test indicated that the peak attack rate, time to peak and



Transmissibility Incubation period Infectious period

Figure 2.5: **Results for mono-factorial experiments for each of the factors: transmissibility, incubation period distribution and infectious period distribution.** Each curve represents the mean epidemic curve from 25 replicates of each experiment. Increasing the transmissibility, mean infectious period and reducing the mean incubation period shortened the peaking time and increased the peak attack rate.

total attack rates were statistically significantly different ($P < 0.0001$) for all pairs. These observations suggested that the mean infectious and incubation durations might have a role on the dynamics of the simulated epidemics. Therefore, we further investigated the effects of fixing the mean infectious and incubation periods as discussed in the next section.

2.3.2 Finding 2. *The mean of the incubation period distribution appeared to be the sole determinant of its effects on the epidemics. In contrast, the mean and variance of the infectious period distribution were needed to determine its influence on epidemic dynamics.*

Support: Figure 2.6 displays the mean epidemic curves resulting from simulated epidemics with similar mean incubation periods (Exp.4 in Table 2.2). The mean epidemic curves presented in Figure 2.6 are based on twenty-five replicates of each experiment. An ANOVA on the total attack rate indicated lack of statistically significant differences ($P = 0.514$). Similar results were observed for time to peak and peak attack rates.

Moreover, a plot of the variance of the incubation period distribution against the mean total attack rates (Figure 2.6) suggested a weak relationship with a Pearson correlation coefficient r of *-0.289*. A similar observation was also made between the variance of the incubation period distribution, mean time to peak *(r = -0.263)* and mean peak attack rates *(r = 0.146)* (Figure

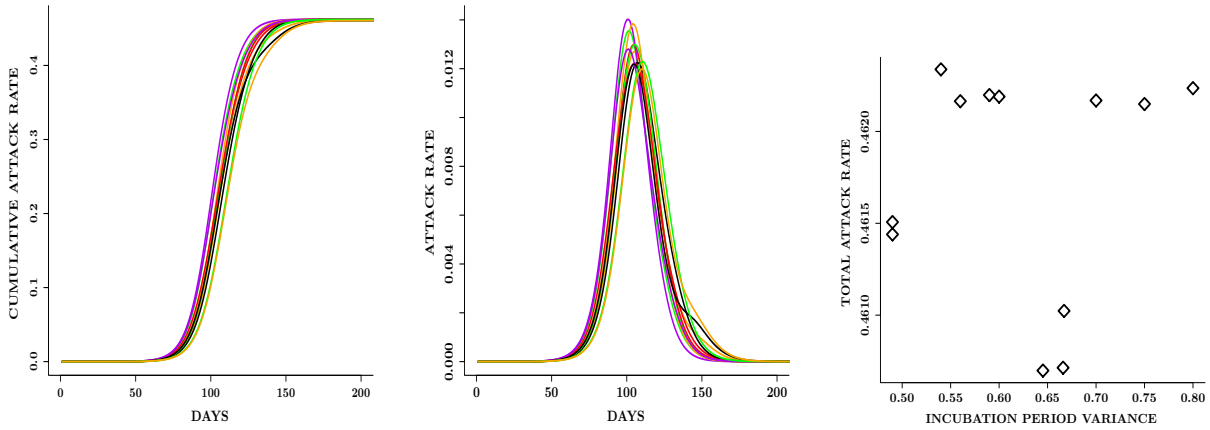


Figure 2.6: **Results from mono-factorial experiments focused on changes in the variance of the incubation period distribution.** The total attack rates were not statistically significantly different across all 12 epidemics. A weak negative relationship was observed between the variance of the distributions and total attack rates. These results indicated that epidemics with similar mean incubation periods had similar total attack rates even with differences in the variance of the distributions.

1 SI). This suggested that changes in the variance of the incubation period distribution did not have a strong influence on the total attack rate, time to peak and peak attack rates. These results also alluded to the idea that solely changing the incubation period distribution while holding the mean incubation period fixed might not significantly influence epidemic prediction. Furthermore, changing the base case incubation period, which has been used in several studies [13, 61, 103], might not affect the results in these studies if the mean of the incubation period distribution remained the same.

The results based on comparing epidemics resulting from infectious period distributions with the same mean values (Exp.5 in Table 2.2) are described in Figure 2.7 based on the mean epidemic curves. The epidemics appeared to have similar shapes. An ANOVA indicated that at least one of the mean time to peak, peak attack rates and total attack rates was statistically significantly different ($P < 0.0001$) from the others. Since there were two incubation period distributions, we compared epidemics with the same incubation period distribution. The results were also significantly different for all public health measures.

In addition, a plot of the variance of the infectious period distributions against these measures implied that the mean of the infectious period was not the sole influence on these measures. Rather, the total and peak attack rates decreased with an increase in the variance of the infectious period distribution (r = *-0.99* and *-0.99*). On the contrary, the time to peak increased with a raise in the variance of the infectious period distribution (r =*0.95*) (Figure 2 SI). These outcomes demonstrated that the dynamics of simulated epidemics might be sensitive to the mean, variance and shape of the infectious period distribution. These observations were contrary to that observed for the incubation period distribution where the

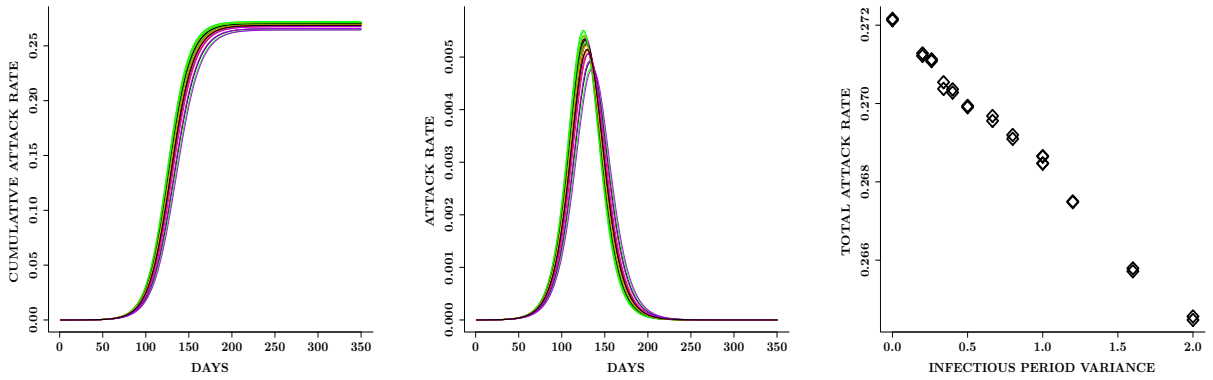


Figure 2.7: **Results from mono-factorial experiments focused on changes in the variance of the infectious period distribution.** The means of the total attack rate, time to peak and peak attack rate were all statistically significantly different based on an ANOVA and pairwise t-test. In addition, an increase in the variance resulted in a decrease in the total attack rate and peak attack rate. These observations suggest that epidemics with the same mean infectious period can have different dynamics.

mean of the distribution appeared to be the sole determinant of its effects on the epidemic outcome.

2.3.3 Finding 3. *Compared to the other parameters, the infectious period distribution exerted the strongest influence on the total attack rate and structure of the epidemic curves.*

Support: The epidemics simulated based on the full factorial design (Exp.6 in Table 2.2) failed to have unique shapes (Figure 3 SI). This indicated that a one-to-one mapping did not exist between the factor combinations and shape of the epidemic curves. In general, a combination of high transmissibility and long mean infectious periods resulted in epidemics which peaked sooner and had a higher peak attack rate. Epidemics with low transmissibility, long mean incubation periods and short infectious periods peaked later and had lower peak attack rates. Epidemics from all other factor combinations fell between. In addition, epidemics with the same transmissibility and infectious period distributions had similar mean total attack rates. For example, an ANOVA on the total attack rates of three sets of epidemics with the same transmissibility and infectious period distribution indicated a lack of statistical significant difference *($P = 0.871$).*

An ANOVA on the total attack rates, peak attack rates, and peaking times indicated that the mean of at least one of the experiments was statistically significantly different from the others ($P < 0.0001$). A pairwise comparison to find which experimental pairs were statistical significant different would have resulted in three hundred and fifty one comparisons. To simplify our analysis, we present the contributions to the variance of each of the measures

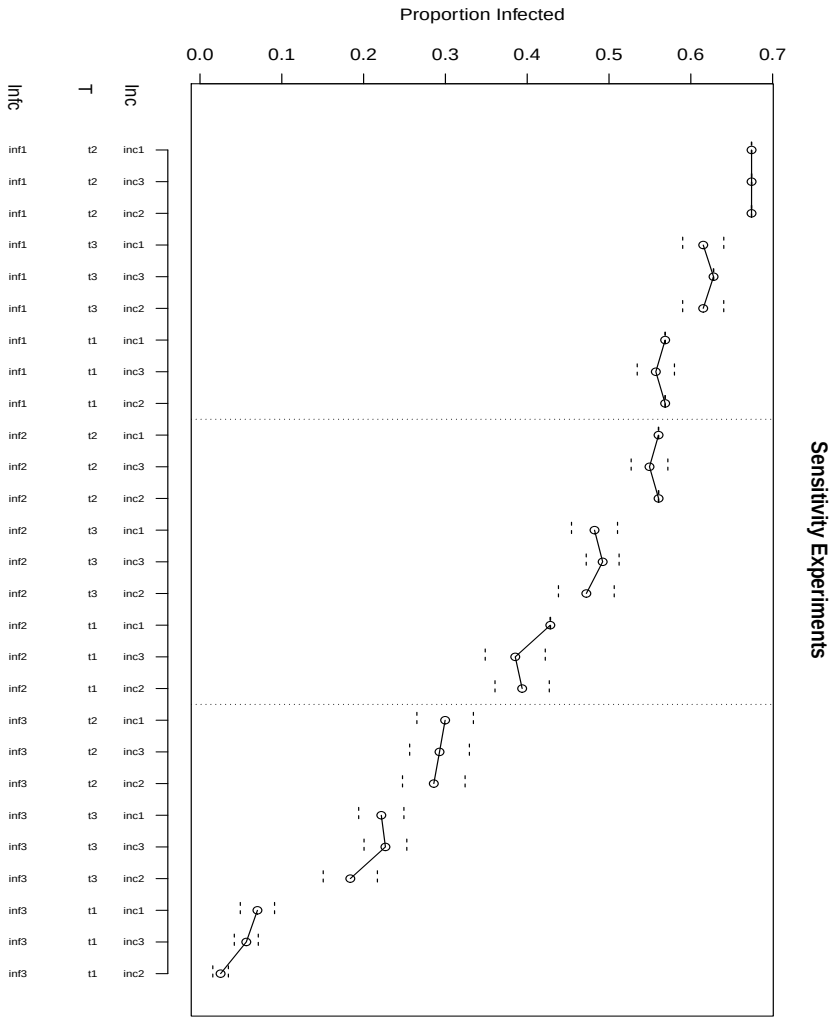


Figure 2.8: **A plot of the mean total attack rate (infected proportion) against all factor combinations.** t stands for transmissibility, infc is an abbreviation for infectious period and inc represents the incubation period. The parameters are labeled according to Table 2.

in Table 2.4 and investigate the factor with the strongest impact on the structure of the epidemic curves.

Per Table 2.4, the infectious period distribution explained the highest proportion of variance observed in the time to peak, peak and total attack rates. The parameters were ranked by variance explained in Figure 2.8. In terms of parameter combinations, transmissibility and infectious period distribution described the highest proportion of variance for the total attack rate and days to peak. On the other hand, the combination of the infectious and incubation periods explained the highest proportion of variance observed in the peak attack rate (Table 1 SI). However, the values noted for the interactions were much smaller than those for individual parameters. The results therefore indicated that the individual parameters explained most of the variance observed in the three public health measures.

Factor	Attack Rate	Peak Attack Rate	Days to Peak
Total Variance	25.36	6.48×10^{-2}	5.06×10^{6}
Transmissibility	2.659	0.0063	4.25×10^{5}
Infectious period	21.67	0.0511	3.23×10^{6}
Incubation period	0.008	0.0058	6.46×10^{5}

Table 2.4: **Analysis of contributions of the parameters to the variance observed in the total attack rate, peak attack rate, and peaking time.** These are based on the full factorial design. The first row shows the total variance in each of the three outcome measures, while the rows beneath display the variance explained by each factor. Note that the infectious period explained the highest proportion of variance observed in all three public health measures.

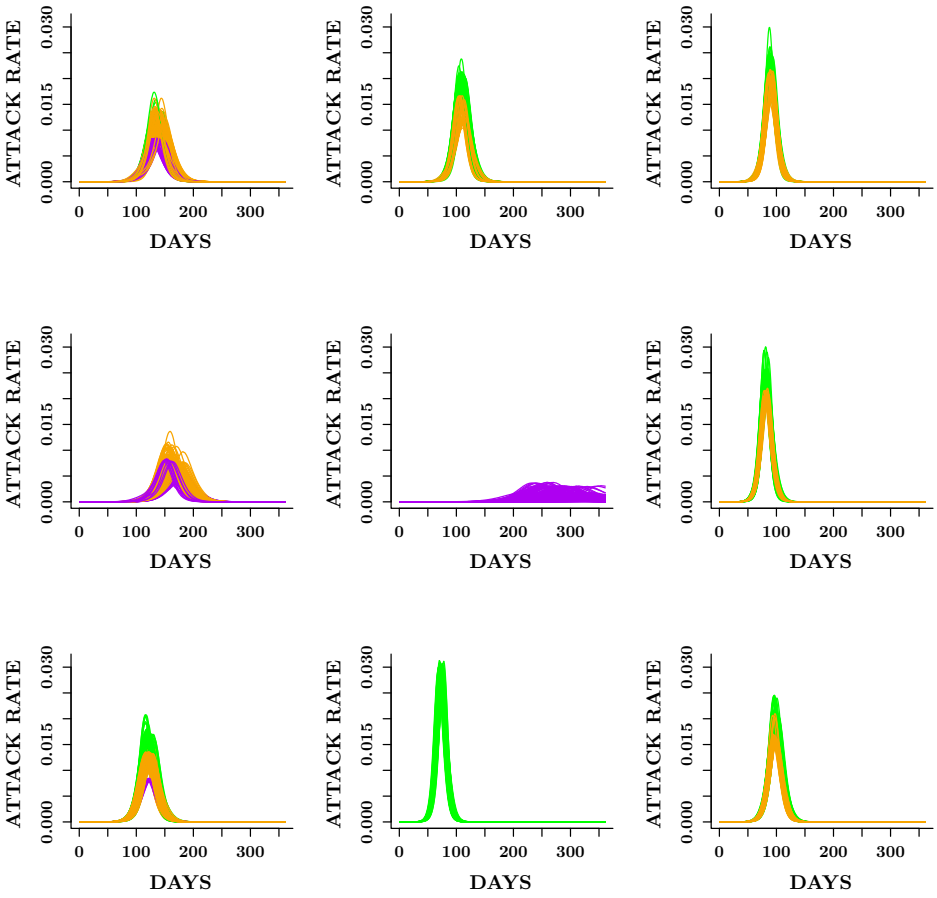


Figure 2.9: **The epidemic curves from all 27 factorial experiments grouped using principal components cluster analysis.** Note that curves in each cluster are grouped by time to peak and spread. Different colors are used to differentiate curves with different levels of the the infectious period distribution. Green, orange and purple are used to represent epidemics with mean infectious period of five, four and three days respectively.

To reaffirm the findings based on the ANOVA, we used principal components clustering to find underlying groupings within epidemic curves from all experiments. Epidemic curves from the twenty-seven experiments appeared to be randomly distributed across clusters, irrespective of parameter level (Figure 4 SI). Except for one cluster, there were more than two sets of parameter combinations in all the clusters. However, in most cases, the epidemic curves in the same clusters had mean infectious period distributions with a difference of one day. These observations suggested that the infectious period distribution had some influence on the clustering.

The clustering in Figure 2.9 was color coded based on the levels of the infectious period distribution. Note that except for two clusters, all others had one or two levels of the infectious period distribution. In most clusters, we observed groupings of epidemic curves with mean infectious periods of three and four days or four and five days. In one cluster, all epidemics had a mean infectious period of five days. Furthermore, note that the curves in Figure 2.9 are grouped by time to peak and spread, indicating that the clustering captured the structure of the epidemic curves.

We also visualized the clustering based on the levels of the transmissibility and incubation period distributions (Figures 5 and 6 SI). Unlike the groupings observed in the levels of the infectious period distribution, most of the clusters contained all levels of the incubation period distribution and the transmissibility. This indicated that there was no particular influence of these parameters on the shape of the epidemic curves. These results therefore alluded to the idea that the infectious period distribution exerted the strongest impact on the total attack rate and shape of the epidemic curves.

2.3.4 Finding 4. *The model sensitivity was consistent across social networks with demographic and rural-urban differences.*

Support: The same trends were observed across all experiments for both Montgomery County and New York. Although Montgomery County and New York have both demographic and rural-urban differences, these differences were not apparent in the epidemic outcomes. In Table 5.3, we present the Pearson correlation coefficients indicating the similarities between the results observed over the two social networks. The columns were labeled based on the notations used in Table 2.2. Note the high correlations signifying that the trend observed in the results was almost identical in most situations. For example, the column for Exp.1 showed that an increase in the transmissibility on average resulted in a shorter time to peak, an increase in the total and peak attack rates. This is because increasing the transmission value increased the probability of infections in the population.

The lowest correlations were observed for Exp.4, were we studied the effect of the mean and variance of the incubation period distribution on the three public health measures. The

	Experiments				
Measures	Exp.1	Exp.2	Exp.3	Exp.4	Exp.5
Attack rate	1.000	0.965	0.983	0.479	0.993
Peak attack rate	0.996	0.991	0.987	0.211	0.995
Peaking Time	0.996	0.999	0.998	0.751	0.921

Table 2.5: **Pearson correlation between the trends observed in the results for Montgomery County in Virginia and New York.** The labels of the experiments are based on the labeling used in Table 2.

conclusions drawn from this set of analysis indicated that the incubation period distributions resulted in epidemics with similar peak and total attack rates. There was no trend observed between the three public health measures and changes in the variances; therefore the lack of significant correlation. However, the overall observations were similar across the social networks. This therefore suggests that the results observed in this sensitivity study are not restricted to a particular network but can easily be applicable to others with similar assumptions.

2.3.5 Finding 5. *School-age children had the highest age-specific attack rates irrespective of mean infectious period and susceptibility of the other age groups.*

Support: In Figures 2.10 and 2.11, we display the mean epidemic curves for five sets of epidemics simulated under three scenarios. The five sets of epidemics were based on varying the mean infectious period between 2-6 days. Note that the outbreaks with mean infectious duration of two days were not displayed since the simulations failed to become epidemics. The three scenarios were based on: assigning the same disease susceptibility to all age groups (Figures 2.10(a) and 2.11(a)), assigning a higher susceptibility to school-age children and elderly (Figures 2.10(b) and 2.11(b)) and defining a higher susceptibility only for school-age children (Figure 2.10(c) and 2.11(c)). These will be called first, second and third scenarios for the remainder of this section. The discussions are focused on how these changes affect the school-age epidemics relative to the other groups.

School-age children had highest age-specific attack rates for all epidemics across all scenarios. Preschoolers had second highest attack rates. Adults had third highest attack rates and the elderly had the lowest attack rates. For example, under the first scenario, epidemics for which the population had a mean infectious period of five days, school-agers, preschoolers, adults and elderly had a mean age-specific total attack rate of approximately *90%, 65%, 47%* and *35%* respectively. Similar total attack rates were observed under the other two scenarios. Allowing school-age children to have a higher susceptibility compared to other age groups did not result in higher total attack rates compared to the case in which all age groups had the same susceptibility. These findings suggest that neither increasing the mean infectious

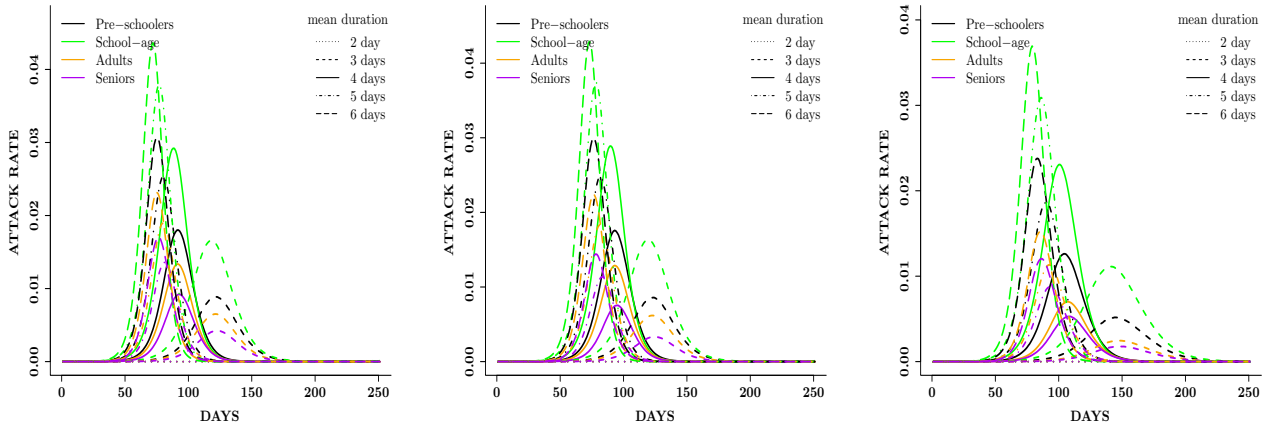
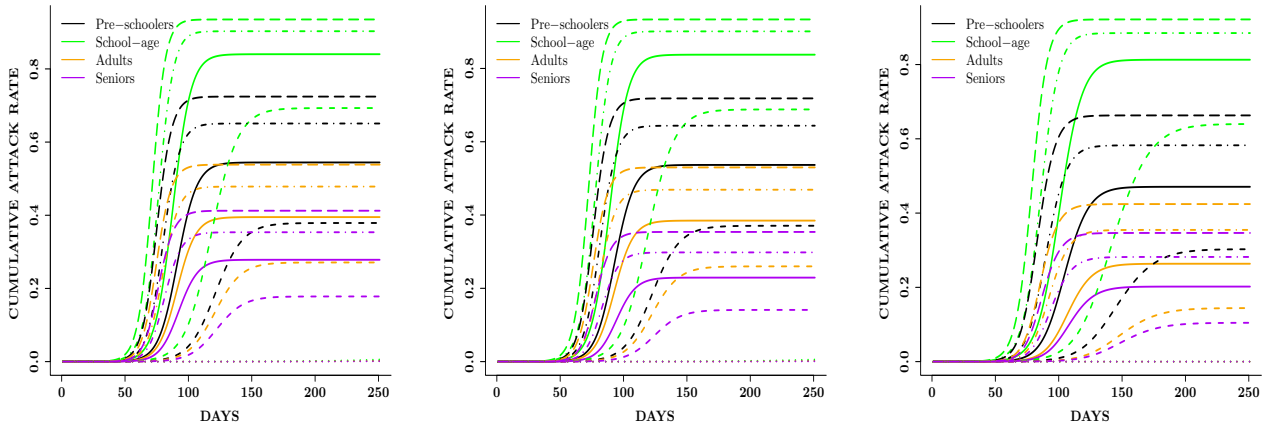


Figure 2.10: **Age-specific mean epidemic curves presented by mean infectious duration.** In (a), all age groups had the same disease susceptibility, (b) school-age children and elderly had a higher susceptibility and (c) school-age children had a higher susceptibility to the disease. Children had the highest mean peak attack rate irrespective of the mean infectious period duration and susceptibility of the other age groups.

period nor changing the susceptibility affected the trend of disease spread observed across age groups.

The overall trend of disease spread was similar across all three scenarios. The trends observed in the attack rates and time to peaks were similar for all sets of epidemics and age groups ($r \geq 0.90$). In addition, under the first scenario, the following pairs did not have statistical significantly different peaking times for epidemics with mean infectious duration of three days: (preschoolers, seniors), (preschoolers, adults) and (adults, seniors) with $P = 0.23, 1.00$ and *0.16* respectively. For epidemics with mean infectious duration of four days, we did not observe statistical significant differences in the following age group pairs: (preschoolers, seniors), (preschoolers, adults) and (adults, seniors) with *P=0.183, 1.00* and *0.094* respectively. The same pairs did not have statistically different peaking times (*P=1.00*) for epidemics with mean infectious periods of five and six days. Note that the school-age population was not included in any of the pairs with similar peaking times. On the contrary, all pairwise comparisons of the attack rates resulted in statistically significant differences across all age groups and epidemics ($P < 0.0001$).

Per Figures 2.10 and 2.11, longer mean infectious periods resulted in shorter times to peak and higher attack rates. However, the total and peak attack rates differed between scenarios for each epidemic. The attack rates for school-agers appeared to be more practically similar between scenarios (1) and (2), although a pairwise comparison of the means indicated statistical significant differences ($P < 0.0001$). Epidemics with mean infectious period durations of five and six days seemed to have more similar attack rates compared to epidemics with



(a) (b) (c)

Figure 2.11: **Age-specific mean cumulative attack rates presented by mean infectious duration.** In (a), all age groups had the same disease susceptibility, (b) school-age children and elderly had a higher susceptibility and (c) school-age children had a higher susceptibility to the disease. Children had the highest mean total attack rate across all epidemics irrespective of the mean infectious period duration and susceptibility of the other age groups.

mean infectious duration of three and four days. Lastly, the peaking time steadily shortened with increase in the mean infectious period.

2.4 Discussion

In this study, we performed sensitivity analysis on discrete probability distributions parameters for an individual-based model for influenza. The major findings in this study were: (i) minute changes in the disease parameters significantly increased the peak attack rate, total attack rate and time to peak. (ii) Knowing the mean of the incubation period distribution appeared to be sufficient in predicting its effects on the dynamics of a simulated epidemic. (iii) The characteristics of the infectious period distribution affected the total attack rate and structure of the epidemic curves. (iv) The sensitivity of the individual-based model was independent of the demographical aspects of the social networks. (v) Differences in age-group susceptibility did not influence the overall trend of disease severity observed within the population. These observations can aid in improving real-time epidemic modeling using similar models.

Two important measures to predict during an epidemic are the time to peak and the total attack rate [38]. However, to make accurate predictions of these measures, we need good estimates of parameters such as the transmissibility, incubation and infectious period. Typically, these parameters are estimated using household transmission data [31]. These are usually

observational studies and not experimental studies. Understanding that the models can be highly sensitive to small changes in the parameters can aid in mitigating some of the bias usually observed in observational studies. In addition, upon estimation of the parameters, intervention strategies can be evaluated to control the spread of the epidemic.

The uncertainty in real-time epidemic modeling and predictions in the early stages of an outbreak are partly due to the imperfection of incidence data [90]. However, incomplete knowledge of the effects of the parameters can also influence predicted outcomes. Since there is not a one-to-one mapping of epidemic characteristics to disease parameters, different parameter combinations can produce epidemics with similar characteristics. Therefore, both biological details on the virus and epidemiological data are needed to improve real-time epidemic modeling [90].

However, this study is not without limitations. The space explored for the parameters in the experiments was limited since in most cases, differences in the mean incubation and infectious period distributions lay between 1-3 days. More extreme differences might be observed if more extensive distributions were used. However, one can argue that using parameters which are similar provides a more strenuous analysis on the system, since parameters with significant influences would be more distinguishable from those with minute influences. In addition, assumptions about the disease model were simplified for this study. Additional layers can be included to describe different compartments of the population, such as infected asymptomatic and infected symptomatic individuals. Furthermore, future studies can also evaluate sensitivity in the presence of various intervention strategies and also with additional information such as school opening dates by region [36].

The results in this study add to the growing literature of real-time modeling of epidemics [101, 103, 104, 105]. These methods are essential for pandemic planning and improving public policy decisions. In addition, this study provides a framework from which future studies can build on for more complex sensitivity analysis on individual-based models with discrete probability distribution parameters. Network models based on urban transportation systems, ad hoc communication and computing systems, and public health which use a probabilistic structure to describe interaction between nodes can all benefit from this study [6]. Furthermore, the level of sensitivity of this model to slight changes in these parameters reaffirms the idea that studies about epidemics using individual-based models are suggestions and not precise predictions, which could benefit public health pandemic planning.

Appendix A

Chapter 2: Appendix

A.1 Figures

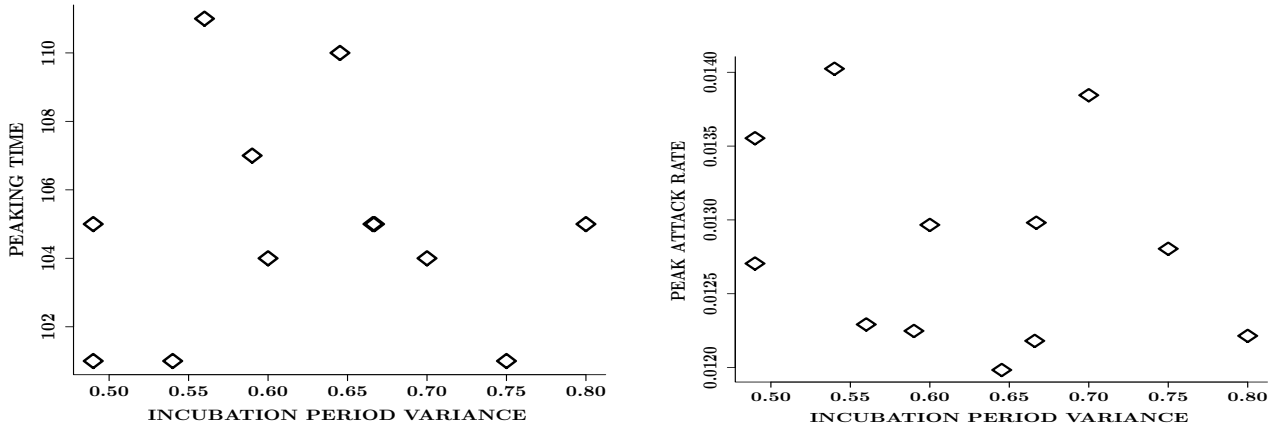


Figure A.1: **Results from mono-factorial experiments (Exp.4 in Table 2.2) focused on the incubation period distribution.** Plot of the variance of the incubation period distribution against the peak attack rate and time to peak.

A.2 Tables

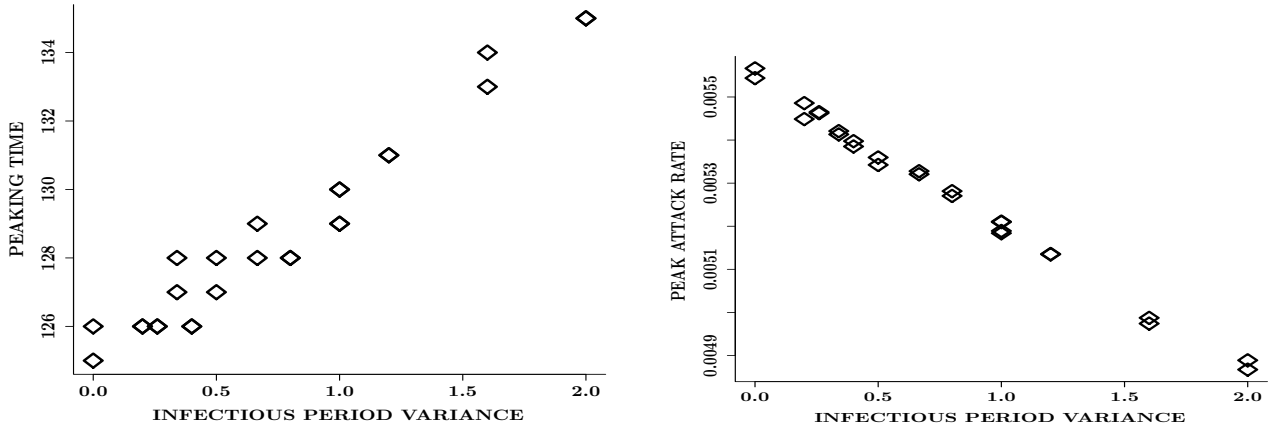


Figure A.2: **Results from mono-factorial experiments (Exp.5 in Table 2.2) focused on the infectious period distribution.** Plot of the variance of the infectious period distribution against the peak attack rate and time to peak.

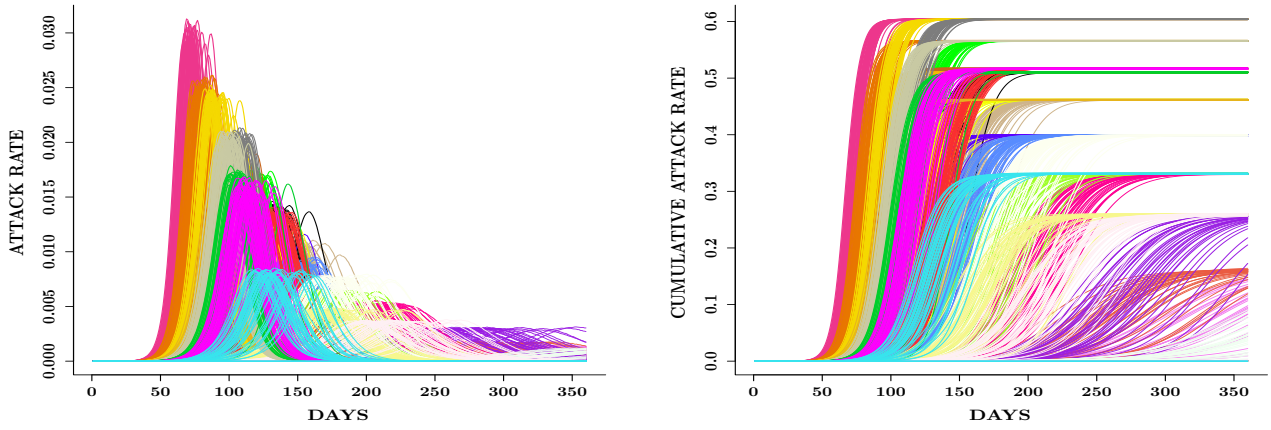


Figure A.3: **Epidemic curves and cumulative epidemic curves from 50 replicates of each full factorial experiment (Exp.6 in Table 2.2).** Each outbreak was simulated for duration of 350 days so as to accommodate outbreaks which are similar to seasonal influenza epidemics and influenza pandemics.

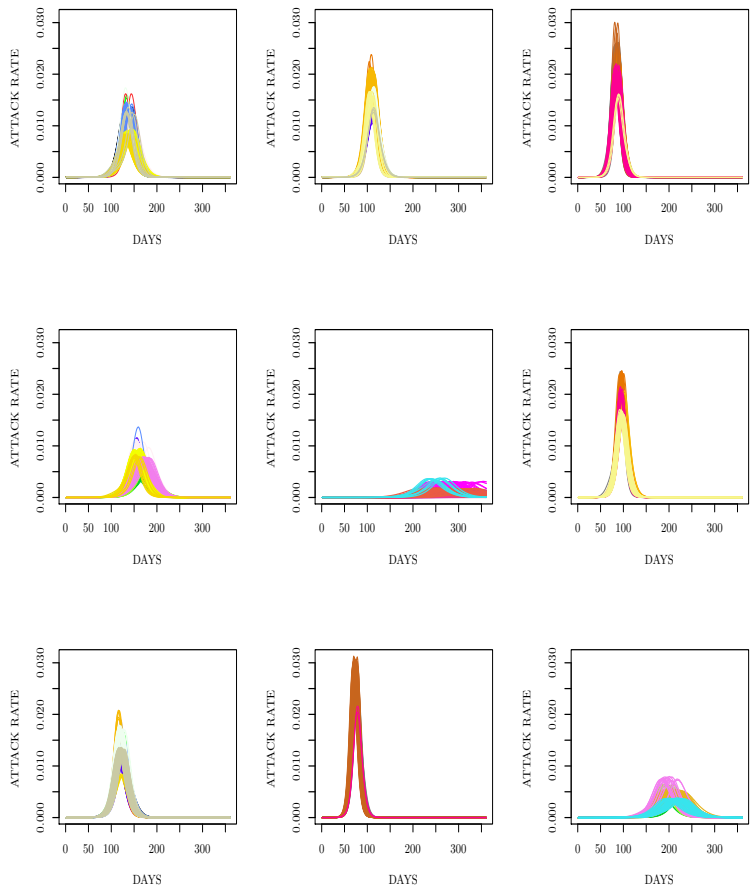


Figure A.4: **The epidemic curves from all 27 factorial experiments were grouped using principal components cluster analysis.** Note that curves in each cluster were grouped by time to peak and spread. Different colors are used to differentiate curves resulting from replicates of different epidemics.

Factor	Total Attack Rate	Peak Attack Rate	Days to Peak
Total Variance	25.36	6.48×10^{-2}	5.06×10^{6}
t	2.659	0.0063	4.25×10^{5}
infc	21.67	0.0511	3.23×10^{6}
inc	0.008	0.0058	6.46×10^{5}
t*infc	0.6744	0.0003	3.99×10^{5}
t*inc	0.0332	0.0003	1.07×10^{4}
infc*inc	0.0701	0.0008	5.61×10^{4}
t*inf*inc	0.0866	0.0000	4.73×10^{2}

Table A.1: **ANOVA on the total infected rate, peak infection rate, and peaking time.** The first row shows the total variance in each of the three outcome measures, while the rows beneath display the fraction of the total variance explained by each factor. t is used to represent transmissibility, inc abbreviates incubation period distribution and infc represents the infectious period distribution.

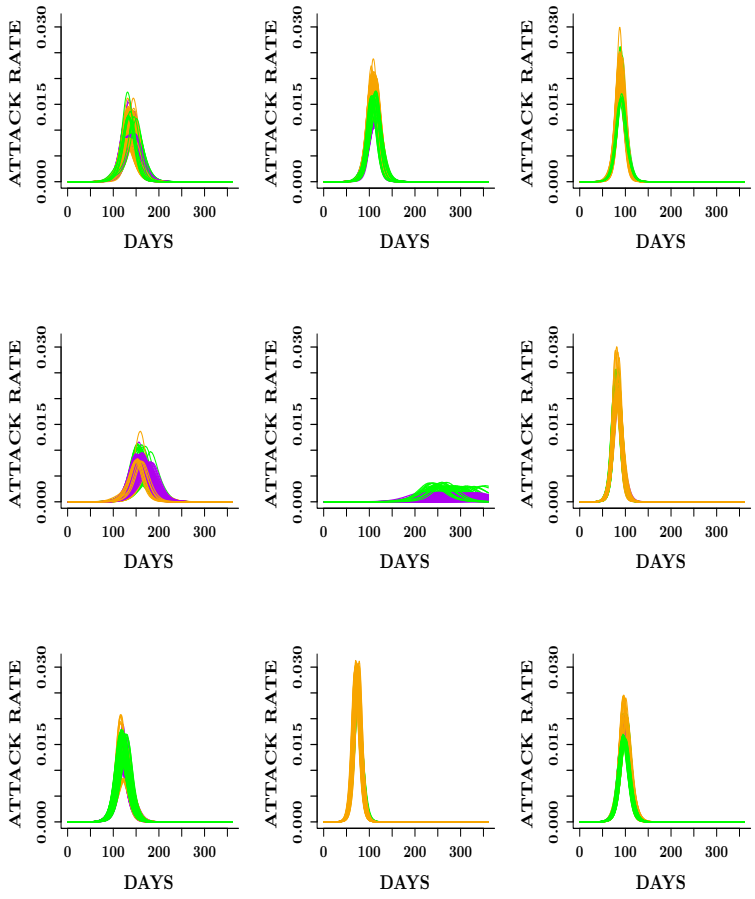


Figure A.5: **The epidemic curves from all 27 factorial experiments were grouped using principal components cluster analysis.** Note that curves in each cluster are grouped by time to peak and spread. Different colors are used to differentiate curves from different transmissibility levels.

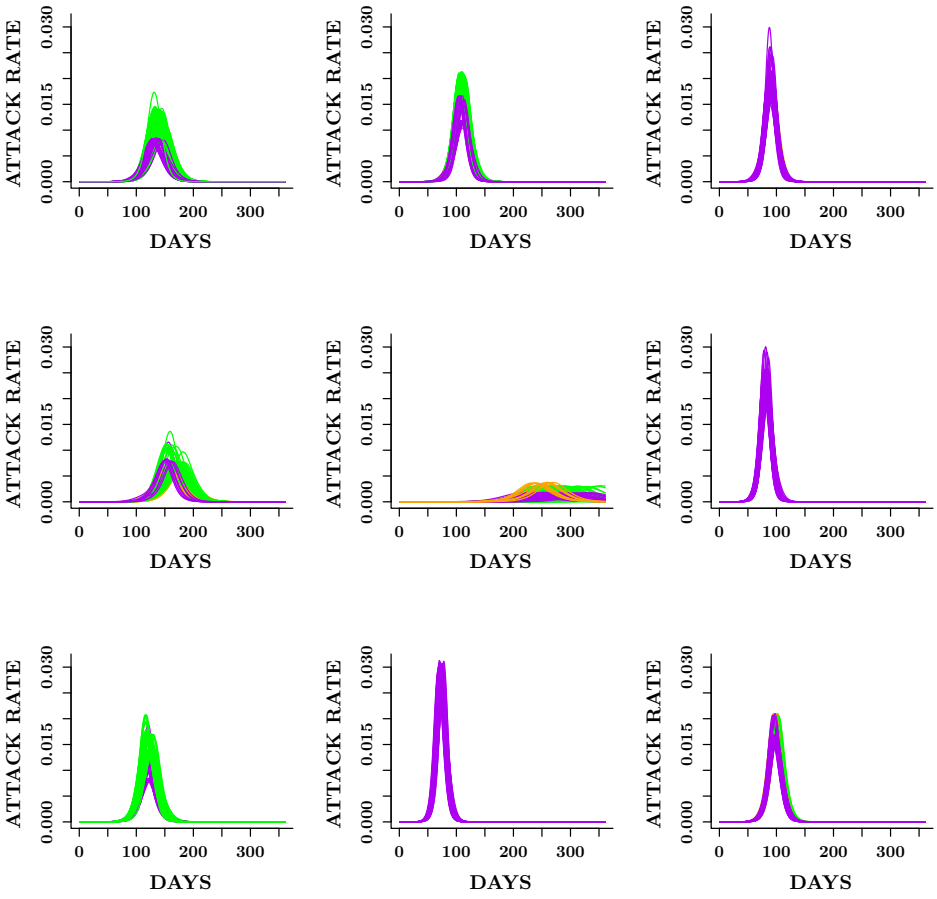


Figure A.6: **The epidemic curves from all 27 factorial experiments were grouped using principal components cluster analysis.** Note that curves in each cluster are grouped by time to peak and spread. Different colors are used to differentiate curves from different incubation period distribution levels.

Chapter 3

Prediction of the Epidemic Curve: A Classification Approach

Elaine Nsoesie [1], Richard Beckman[1], Madhav Marathe[1,2], and Bryan Lewis[1]

[1] Network Dynamics and Simulation Science Laboratory, Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, Virginia, USA

[2] Computer Science Department, Virginia Tech, Blacksburg, Virginia, USA

Abstract

Classification methods are widely used for identifying underlying groupings within datasets and predicting the class for new data objects given a trained classifier. This study introduces a project aimed at using a combination of simulations and classification techniques to predict epidemic curves and infer underlying disease parameters for an ongoing outbreak.

Seven supervised classification methods (random forest, support vector machines, nearest neighbor with three decision rules, linear and flexible discriminant analysis) were used in identifying partial epidemic curves from six agent-based stochastic simulations of influenza-like-illness (ILI) epidemics. The accuracy of the methods was compared using a performance metric based on the McNemar test.

The findings showed that: (1) assumptions made by the methods regarding the structure of an epidemic curve influences their performance i.e. methods with fewer assumptions perform best, (2) the performance of most methods is consistent across different individual-based networks for Seattle, Los Angeles and New York and (3) combining classifiers using a weighting approach does not guarantee better prediction.

3.1 Introduction

Epidemic curves are graphical representations of the incidence of a disease plotted over time and are useful for inferring the magnitude, incubation duration and other attributes of an outbreak. In this study, we seek to predict the epidemic curve for an ongoing outbreak using a combination of simulations and classification methods. Predicting the epidemic curve implies that given data up to day j, we seek to predict the number of daily infections in the future. Real-time prediction of an epidemic curve during a (global) disease outbreak could be invaluable to public health officials since it could aid in the postulation of disease transmission parameters for studying the dynamics of the outbreak and influence selected measures for containment [82, 94, 101, 104].

Infectious disease pandemics in recent years have increased interest in real-time forecasting and long-term prediction of epidemic curves and disease transmission parameters (see [82, 94, 101, 104] and [65] for a few examples). [94] introduced a simulation-based approached for the estimation of disease transmission parameters. The similarities between their method and ours are: (i) both approaches use simulations in the prediction of epidemic curves which allows for the estimation of missing data, (ii) given observed surveillance data for an ongoing epidemic, the surveillance data are matched to simulated samples using a metric and (iii) the analysis from both methods indicate that the choice of the metric affects the conclusions.

However, differences exist in the approach, the focus, and assumptions made in both studies. Some of the differences are: (i) the temporal SEIR model used by [94] assumes homogeneous mixing while the agent-based network approach captures the heterogeneity present in the spread of an infectious disease. (ii) Although not exploited in this study, the agent-based model has a representation of every pair of individuals connected in the network which enables analysis at the individual level and the prediction of the epidemic curve based on changes in individual behavior. (iii) Unlike [94], one of the focuses of this paper is to explore different classification methods with the purpose of finding the best ones for predicting outbreaks with different parameters and simulated over different networks. (iv) Also, contrary to [94], we do not use a Bayesian approach in this study.

The relevance of a simulation-based approach without likelihoods lies in the challenges faced by likelihood-based methods and alternative Bayesian approaches [94]. Most of the challenges are related to the availability of epidemic data and the tractability of the likelihood for large populations [41]. In addition, there are other potential advantages of using a simulation- and classification-based approach for predicting epidemic curves. In the event of a previously unobserved pandemic such as the recent 2009 H1N1(A), by comparing the data for the daily infected or influenza-like illness (ILI) cases to existing simulations of previous outbreaks, an initial model can be proposed for studying the spread of the disease. In addition, the agent-based epidemic modeling approach allows for easy introduction of behavioral changes that occur at the individual level which can affect the spread of a disease [45]. Although compartmental models could be used in simulating similar shaped epidemic curves as those

used in this study, agent-based models (ABM) are used instead, because the social networks focus on particular regions which imply that results observed for one region are not necessarily applicable to other regions due to spatial and demographic differences. More details on the ABM are presented in the Appendix.

In the long run the overall goal of this project is to develop an extensive digital library of tens of thousands of simulated outbreaks and during an outbreak of an infectious disease quickly match real world surveillance data to one or more (or possibly none) of these simulations. By matching the surveillance data to simulated cases, both the epidemic curve and underlying model parameters can be estimated. This paper represents the first step in this process. Here we compare eight supervised classification methods to determine those with a high accuracy rate in identifying the underlying disease parameter for an influenza epidemic. The supervised classification schemes are used in sequentially classifying twelve hundred partial epidemic curves (half used as a classification training set while the other half represent surveillance samples) from six agent-based stochastic simulations of influenza epidemics. The underlying disease transmission model is an SEIR model with three parameters: incubation period distribution, infectious period distribution and transmissibility.

Sequential classification implies that for each day j of an outbreak, a set of transmission parameters are proposed for describing the outbreak by assigning the epidemic curve to one of the epidemic clusters in the library. An epidemic cluster is represented by stochastic simulations from the same SEIR disease model parameterization. Supervised classification methods are used in this study since each classifier can be trained using data already available in the library. Ideally, the choice of the classification method can make a substantial difference in the sensitivity, specificity and accuracy of classification.

In a typical supervised classification scheme, given learning samples with known data classes, the goal is to build a classifier that can correctly predict the classes of new data objects. In this study, the new data objects (partial epidemic curves) do not have the complete information available in the learning samples. The sparse nature of the partial curves is likely to affect the performance of the classification techniques. The complexity of this study therefore lies in the incompleteness of information in the new data objects.

The two main aims of this study are to perform a systematic comparison of eight classification methods (three nearest neighbor methods, support vector machines (SVM), linear discriminant analysis (LDA), flexible discriminant analysis (FDA), random forests (RF) and a combined classifier) to find which methods perform best in correctly identifying partial epidemic curves for six epidemics and evaluate whether the performance of these methods differs by social networks. There are two main assumptions in this study. (i) Epidemic curves represent the counts of daily-infected cases. (ii) Epidemic curves in the test set are assumed to be described by one of the sets of disease transmission parameters in the data library.

However, the latter would not always be the case if surveillance data from a real outbreak is used. In general, the epidemic curve for an ongoing outbreak can be assigned to one, several or none of the clusters in the library. If the epidemic curve is assigned to one of

the clusters, then the underlying model for that epidemic cluster can be used in modeling the outbreak. The aim is not to find a specific model for the epidemic but rather a set of possible models since in most scenarios uncertainty in prediction increases with sparsity of the data [94]. In addition, the set of possible parameters can be extremely large, which is one of the reasons for presenting this special case. To make this method applicable to the case where an epidemic curve cannot be assigned to any of the clusters in the library, an iterative scheme can be used whereby a combination of expert opinion and search methods are applied over multiple iterations to propose a set of parameters possibly describing the outbreak. New clusters of epidemics can be created based on the newly proposed parameters and the epidemic curve can be reclassified until a good fit is found.

The epidemics are simulated under the assumption of a novel virus with little or no prior immunity. Simulations could be conducted that further change the shape of the epidemic curves by including more details about individual behavior and immunity that improve the realism. To test out the classification scheme these curves are sufficient. One could argue that by maintaining fewer parameters, this further "differentiates" the curves thereby providing a strenuous exploration of the classification scheme's performance since in essence it has fewer parameters to perform the classification on. In addition, the use of simulated data enables a thorough evaluation of the performance of the classification methods under a controlled setting. The outbreaks are simulated across social networks for metropolitan regions surrounding New York, Los Angeles and Seattle with population sizes of approximately 20 million, 16 million and 3.2 million respectively. For the purpose of this paper, these metropolitan regions are referred to as Seattle, Los Angeles and New York.

The rest of this paper is organized as follows: the proposed approach is presented in the next section, the methods are discussed in section 3.3, the results are presented in section 3.4, the conclusions follow in section 3.5 and the Appendix contains a description of the supervised learning methods, the ABM, and a compartmental model. Initial results for this study were published in the 2010 Joint Statistical Meeting Proceedings [102].

3.2 Approach

Let the vector $\boldsymbol{X} =< x_1, x_2, \ldots, x_t >$ represent an epidemic curve where t is the number of observed infections over time measured in days and $x_t s$ are the daily-counts of infected. Suppose in the early stages of an outbreak, a partial curve $\boldsymbol{Y} =< y_1, y_2, \ldots, y_d >$ of duration d is observed. Given that the model parameters underlying the new outbreak are unknown, the prediction of future daily infected counts would be difficult.

A possible solution would involve using a classification approach. Based on the shape of the partial epidemic curve, several possible disease transmission parameters (e.g. incubation period, infectious period, serial interval etc.) can be hypothesized. Thousands of epidemic curves can be simulated based on the hypothesized parameters and the simulated epidemics

Name	Transmissibility	Incubation Period (day probability)	Infectious Period (day probability)
Catastrophic *	0.00006	1 0.3 2 0.5 3 0.2	3 0.3 4 0.4 5 0.2 6 0.1
Mildly Catastrophic **	0.000083	0 0.20 1 0.45 2 0.35	2 0.66 3 0.33 4 0.01
Strong	0.000042	same as *	same as *
Mildly Strong	0.0000365	same as *	same as *
Moderate	0.0000581	same as **	same as **
Mild	0.0000333	same as *	same as *

Table 3.1: Parameters used in simulating the epidemics in this study. Catastrophic flu infects about 50% of the population, while strong, mildly strong, and mild flu infect approximately 30%, 20% and 10% of the population respectively. Each infected individual has a randomly assigned probability of having a specific incubation or infectious duration. For example, for catastrophic flu, each infected individual have a probability of 0.3, 0.5 or 0.2 of having an incubation period duration of 1, 2, or 3 days respectively.

can be organized in a library with clusters representing epidemics with the same parameters. Using a supervised classification method, the complete epidemic curve can be estimated by assigning the partial epidemic curve into the cluster with the most similar full curves.

In general, a partial epidemic curve observed during an outbreak can be assigned to one, several or none of the clusters in the library. If the partial curve is assigned to one (or several) of the clusters, then this can be used as a starting point for modeling the outbreak. If the partial epidemic curve is not assigned to any of the clusters in the library, a new set of disease models can be hypothesized based on a search for possible model parameters that could be used to describe the outbreak.

In order to find the best supervised classification method that consistently outperforms other methods in predicting the correct epidemic cluster early on in an outbreak, we compare the accuracy of seven supervised classification methods in addition to a combined classifier. The epidemic curves used in the classification are simulated using six SEIR model parameterizations. The parameters are shown in Table 3.1. For the purposes of this paper, these outbreaks are called catastrophic, mildly catastrophic, strong, mildly strong, moderate and mild flu epidemics.

The incubation and infectious period distributions used in the catastrophic model (Table 3.1) are based on a consensus by three research groups and was initially used in a study by [67]. The same parameters have been used in several other studies (e.g. [46] and [61]). The parameters used in modeling the mildly catastrophic outbreaks have not been published but are based on the serial interval proposed by [31]. The distributions of the incubation and infectious durations are obtained by adjusting the previously described joint distributions of the incubation and infectious period distributions in [67] to match the serial interval in [31].

The transmissibility parameters are selected so as to explore outbreaks ranging from what is observed during normal influenza seasons to more extreme outbreaks in order to provide a thorough examination of the sensitivity of the proposed method. Samples of epidemic curves simulated for Seattle are given in Figure 3.1. There are 200 replicates for each simulated epidemic; half are used as a training set and the other half are used as surveillance samples for the test set. Although all the epidemics are simulated for a duration of 365 days most have a total duration of less than 210 days.

In the classification scheme, the data in the training set are used to "learn" the methods in order to obtain the optimal classifier on each day $j = 1, \ldots, d$. Each partial epidemic curve in the test set is then used as an input sequence $< x_1, x_2, \ldots, x_d >$ into each of the trained classifiers and a single epidemic cluster label is returned as output. The accuracy of each classifier is estimated by epidemic (catastrophic, mildly catastrophic, strong, mildly strong, moderate and mild flu epidemics) based on the number of correct classifications on each day. This is a combination of a time series prediction and a sequence classification problem since each epidemic curve can be viewed as a time series. [43] discusses these classification problems in detail.

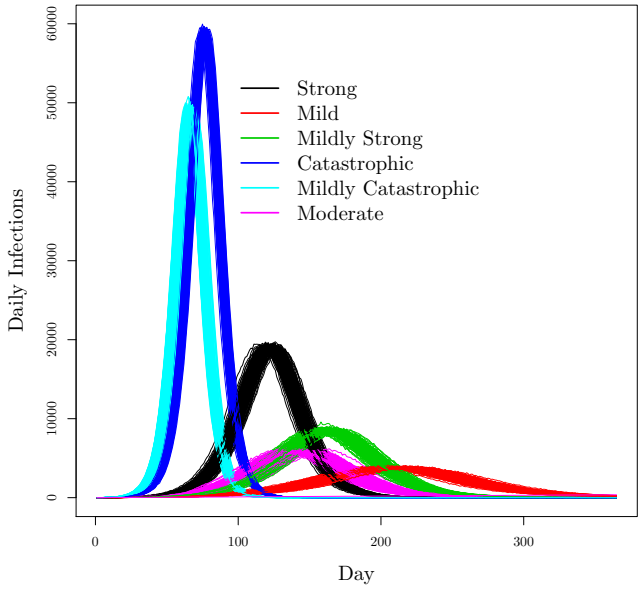


Figure 3.1: Sample epidemic curves simulated using the disease models in Table 3.1 for Seattle. Each group of epidemic curves represents 200 stochastic simulations. Although the curves appear smooth, a closer look would reveal they are not.

3.3 Methods

An agent-based modeling approach is used to simulate the outbreaks investigated in this study. This ABM has previously been used to study the transmission dynamics of an infectious agent through individual populations and to evaluate the effectiveness of control strategies over specific populations [17, 61]. For discussions and studies using similar models see [8] and [47]. The creation of this ABM involves two major steps: the creation of a social contact network from a state-of-the-art behavioral model and a computational model for disease transmission. To enable the readability of this paper, these methods are discussed in the Appendix.

3.3.1 Classification Methods

The supervised classification methods used in this study are support vector machines (SVM), random forest (RF), nearest neighbor methods: nearest mean (Mean), nearest median (Median), minimum distance (Minimum), linear discriminant analysis (LDA) and flexible discriminant analysis (FDA). There are several professed advantages to using each of the supervised learning techniques. The advantages of random forest include efficiency on large databases, high accuracy and estimation of importance variables [70]. Likewise, support vector machines have been shown to achieve a high accuracy rate across various data types and also tend to perform well on high dimensional data [70]. Discriminant analysis methods perform well due to the simplicity of the methodology and can provide low-dimensional views of high dimensional data [70]. Nearest neighbor methods are relatively easy to implement and are highly adaptive [74]. A brief and detailed discussion of these methods can be found in the Appendix and in [70] respectively.

The supervised classification methods are selected in a manner that allows exploration of different types of classification methods from machine learning and statistics: tree based methods (random forest), distance based methods (nearest neighbor), probabilistic classification methods (LDA) and maximum margin classifiers (support vector machine). These methods have been shown to perform differently based on the performance criteria and the data structure [29].

In some cases, the selection of a single method from all supervised learning methods investigated might not prove to be the ideal choice since "potentially valuable information may be wasted by discarding the results of less-successful classifiers" [127]. A pooled classifier might be useful in situations where no classification method is likely to consistently outperform the others or the surveillance data contains a large amount of noise and are high dimensional [127].

Several methods have been proposed for combining classifiers. The most popular of these is simple averaging of the output from each of the classification methods [127]. Weighted

averaging with different definitions of how to calculate the weights is an extension of this method [127]. Rank-based combiners, voting schemes, order statistics combiners, and belief functions are other methods for pooling classifiers [35, 127].

Both simple and weighted voting classification schemes are used in this study. In a simple voting scheme, the prediction from each of the classification methods receives a single vote and the final classification is based on a majority of votes. The simple voting scheme is simple and easy to implement. Weighted voting is an alteration to simple voting, which involves associating each classifier's vote with a weight. The weights increase the influence of better classifiers on predictions and are calculated based on the performance of each classification method on a validation set [35]. The weights in this study are defined as linear, polynomial and exponential and are given by $2\varepsilon_{i,j}{}^{-1}$, $2\varepsilon_{i,j}{}^{-3}$, and $\exp(2\varepsilon_{i,j}{}^{-1})$ respectively. The weights are calculated on each day of classification and there are six misclassification errors on each day since there are six disease clusters. $\varepsilon_{i,j}$ is the mean of the six misclassification errors on day j for method i where $i=1,\ldots,7$ and $j=5,\ldots,t$ where t is the duration of the epidemic.

An additional set of 600 simulated epidemics from all parameter sets is used as a validation set in estimating the weights for each method on each day. The classification procedure using the combined weighted voting scheme is described below. The first half of the algorithm excluding the calculation of weights is the basic approach for classification based on the individual methods.

Algorithm 1 Classification

Inputs: epidemic curves on day j
Outputs: predicted cluster on day j

Read in all epidemic curves in training set
Loop over days
for (each day j) **do**
  Train each supervised learning method using all curves in the training set
  for (each epidemic in the test set) **do**
    Predict the epidemic cluster using each method
  end for
  Estimate error $\varepsilon_{i,j}$: mean misclassification error on day j for method i
end for
For weighted classifiers
Using validation data set estimate weights: $2\varepsilon_{i,j}{}^{-1}$, $2\varepsilon_{i,j}{}^{-3}$, and $\exp(2\varepsilon_{i,j}{}^{-1})$
for (each day j) **do**
  for (each epidemic in the test set) **do**
    Predict the epidemic cluster using each method and consider each prediction as a single vote
    Assign appropriate weights to each method
    Sum the weighted votes assigned to each cluster
    Predict the epidemic cluster based on the majority weighted vote
  end for
end for

3.3.2 Performance Accuracy Metric

The McNemar test [48] is used in a pair-wise comparison of the methods. The McNemar test is selected because it deals with the lack of independence between data samples and has been shown to be less prone to Type I error compared to other statistical tests for comparing classification methods [42]. The McNemar test evaluates the null hypothesis of equality between p_b and p_c, where p_b is the number of test samples misclassified by classifier A but not by B and p_c is number of test samples misclassified by classifier B but not by A. The McNemar test is based on the hypotheses:

H_{0_a}: The error rate of method A is greater than or equal to the error rate of method B on day j.
H_{0_b}: The error rate of method B is greater than or equal to the error rate of method A on day j.

The McNemar test compares accuracy of the methods at a single time point. However, since the epidemic curves are classified over several time points, an accuracy metric which evaluates the method performance over time is needed. For the purposes of this paper, performance accuracy is defined under two categories: "better" and "consistent" in order to find which methods perform best in identifying epidemic curves from each parameter set and which method consistently outperform other methods across all epidemics. The performance accuracy metric is defined as follows:

Method A is *significantly better* than method B if there is statistically significant evidence for rejecting the null hypothesis on more than 50% of the days.
Method A is *consistent* if it performs better than most methods in the identification of epidemic curves from all clusters. This implies that there can be more than one consistent method.

The significance level for the test is set at $\alpha = 0.05$. The analysis is focused on the first few days of the outbreaks since the aim of the study is to find a method with a high accuracy rate at the early stages of an outbreak.

3.3.3 Chi-Square Tests

The Chi-square test is used in evaluating the null hypothesis of no association between the performance of the methods and social networks for which the outbreaks are simulated across. Both the Pearson Chi-square and Likelihood Chi-square tests are considered. These tests are selected because they can be used for both nominal and ordinal data [2].

Table 3.2: Summary of the components of the experimental design

Components	**Number**	**Names**
Social Networks	Three	Seattle, Los Angeles and New York
Classification Methods	Seven	Linear and Flexible Discriminant Analysis, Nearest Mean Method, Nearest Median Method Minimum Distance Method, Random Forest, Support Vector Machines
Combined Classifiers	Four	Simple voting classifier, linear, exponential, and polynomial weighted voting classifiers
Influenza epidemics	Six	Catastrophic, Mildly Catastrophic, Strong, Mildly Strong, Moderate and Mild
Statistical Methods	Two	McNemar and Chi-square tests

A summary of the experimental design for this study is given in Table 3.2.

3.4 Results

The results are presented under four subsections. The first two sections, 3.4.1 and 3.4.2, answer the question of which methods perform best in identifying epidemic curves from each parameter set and which methods are consistent across all epidemics. Section 3.4.3 discusses the performance of the combined classifiers, while section 3.4.4 examines the performance of the methods across different social networks.

3.4.1 Daily Accuracy of the Classification Methods

As shown in Figure 3.1, 200 epidemic curves are simulated for Seattle using each set of parameters in Table 3.1. Epidemics based on the same assumptions are also simulated for New York and Los Angeles. Day 1 of the epidemics represents the day on which the first infected case is observed. Six hundred epidemics are randomly assigned to the training set and six hundred are assigned to the test set for each of the study regions.

The comparative performance of each of the supervised classification methods in correctly identifying the epidemic cluster for all six hundred epidemic curves in the test set are given in Figure 3.2 for Seattle. The results are presented for days five to eighty-one since the accuracy of most of the methods remain stable across all epidemics after day eighty-one. Each of the subfigures in Figure 3.2 represents the accuracy in the identification of epidemics from a single SEIR parameterization. The results are presented by day of peak, i.e. the outbreak with the earliest peak (mildly catastrophic) based on Figure 3.1 is presented first.

The eight classification methods compared in Figure 3.2 are: support vector machines (SVM), random forest (RF), nearest neighbor methods (nearest mean (Mean), nearest median (Me-

dian) and minimum distance (Minimum)), linear discriminant analysis (LDA), flexible discriminant analysis (FDA) and the combined simple voting classifier (Combined). Only the combined simple voting classifier is shown here since there is not much difference between "simple voting" and "weighted voting" as discussed later in section 3.4.3. In addition, the results are presented only for Seattle since similar results are observed for New York and Los Angeles as discussed later in section 3.4.4.

For the two catastrophic flu epidemics (Figure 3.2), it is easy to make an early identification irrespective of the classification method. However, as the epidemic curves "get closer", the complexity of correct identification increases. In such cases methods such as LDA should not be used. The prediction accuracy rate is unstable for LDA and FDA in the classification of mild, mildly strong and strong epidemics. The performance of the nearest mean and median methods are also unstable in the identification of strong and moderate epidemics. In contrast, random forest, the simple voting classifier and the minimum distance methods appear to perform well across all epidemics.

The results shown in Figure 3.2 indicate that all methods have over a 50% accuracy rate in identifying mildly catastrophic on the first day of classification. Figure 3.2 also shows that the minimum distance method is the only method which achieves an accuracy above 50% in the identification of catastrophic epidemics on day 5. The high accuracy of the minimum distance method can be explained by the low variability between clusters at the start of the epidemics. With the minimum distance method, an observed outbreak can be matched to more than one of the clusters.

Most methods exhibit instability in the identification of both strong and moderate epidemics. However, the accuracy rate is much higher on day eighty-one across all methods in the identification of strong epidemics compared to moderate epidemics. The instability in the classification can be explained by the overlap between epidemic curves for mildly strong, strong and those of mild and moderate epidemics.

The accuracy of the methods in the identification of mild and mildly strong epidemics are important because they are the last to peak which increases the complexity of identification. Except for the two discriminant analysis methods, all other methods perform relatively well in the identification of both mild and mildly strong epidemics. However, the methods achieved a higher accuracy by day eighty-one in identifying mild epidemics relative to mildly strong epidemics although the mild epidemics peak after the mildly strong.

The results observed in Figure 3.2 indicate that "time to peak" affects the accuracy of random forest, support vector machines and nearest neighbor methods since curves which peak early are most easily identifiable. Figure 3.2 also suggests that in choosing methods which are likely to perform well across all epidemics, linear and flexible discriminant analysis methods should be avoided.

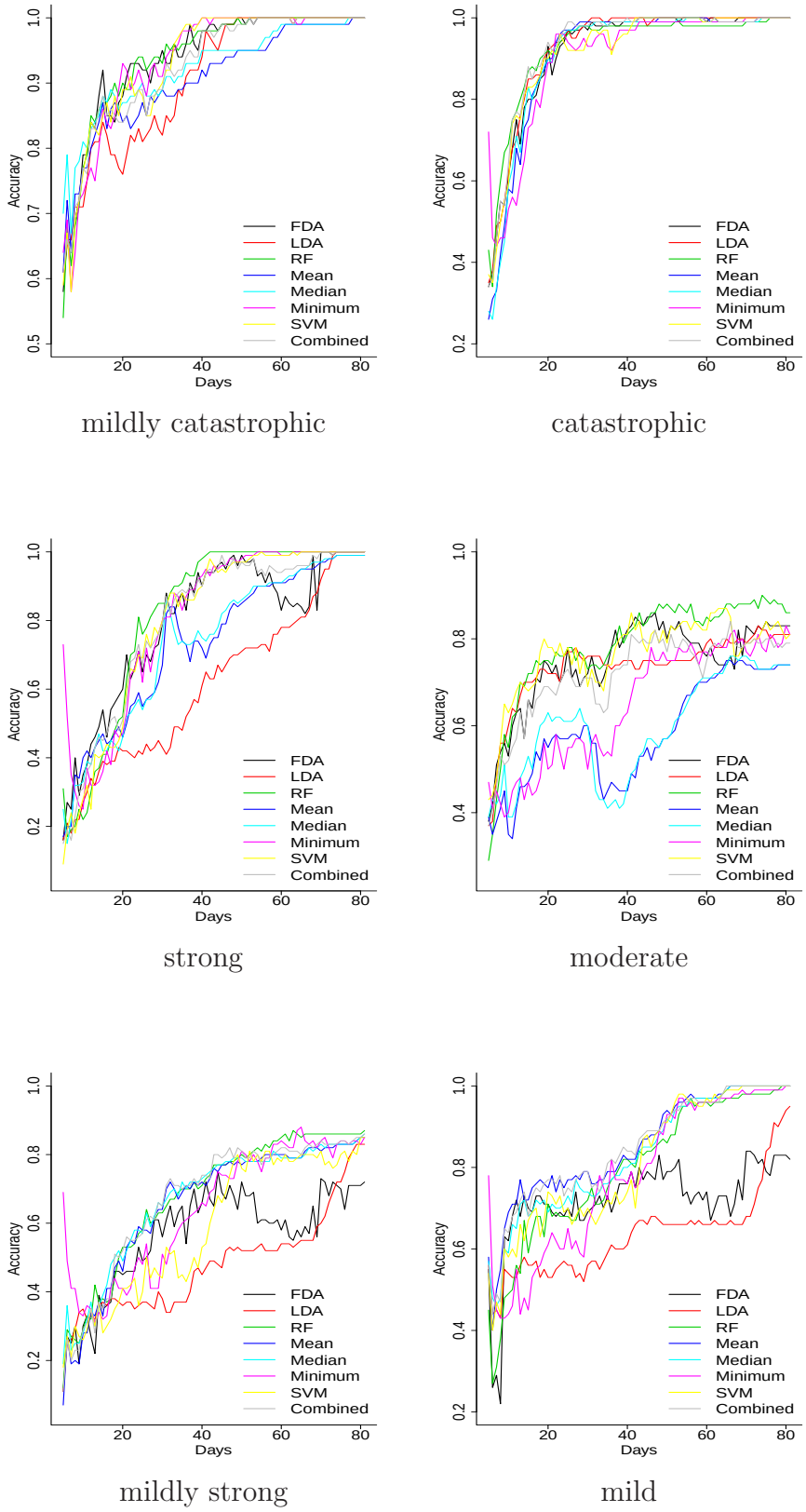


Figure 3.2: The daily accuracy of eight classification methods. Results are presented for Seattle.

3.4.2 Consistency of Classification Methods

The observations in Figure 3.2 are further investigated by performing a pair-wise comparison of the methods using the McNemar test and the performance accuracy metric. The methods are ranked from one to three for each epidemic, where three is the most preferable method and one is the least preferable method. The rankings are displayed on the heatmaps in Figure 3.3. The rankings 1, 2, and 3 are represented with the colors royal blue, dark green, and orange respectively.

Based on Figure 3.3, RF is ranked as the most preferable method for five out of the six epidemics for Seattle. Minimum is ranked as the most preferable method for three out of the six epidemics. Each of the other methods except LDA is ranked twice as the most preferable method. Although, SVM has fewer best rankings than Minimum, it is not ranked as least preferable for any of the epidemics. This could suggest that SVM is a more consistent method than Minimum. Similar results are observed for Los Angeles. However, for New York RF, SVM, FDA and LDA are ranked as least preferable in the classification of mild flu. Nevertheless, RF and SVM perform significantly better than LDA and FDA. SVM and RF are included in this group because the nearest neighbor methods perform better.

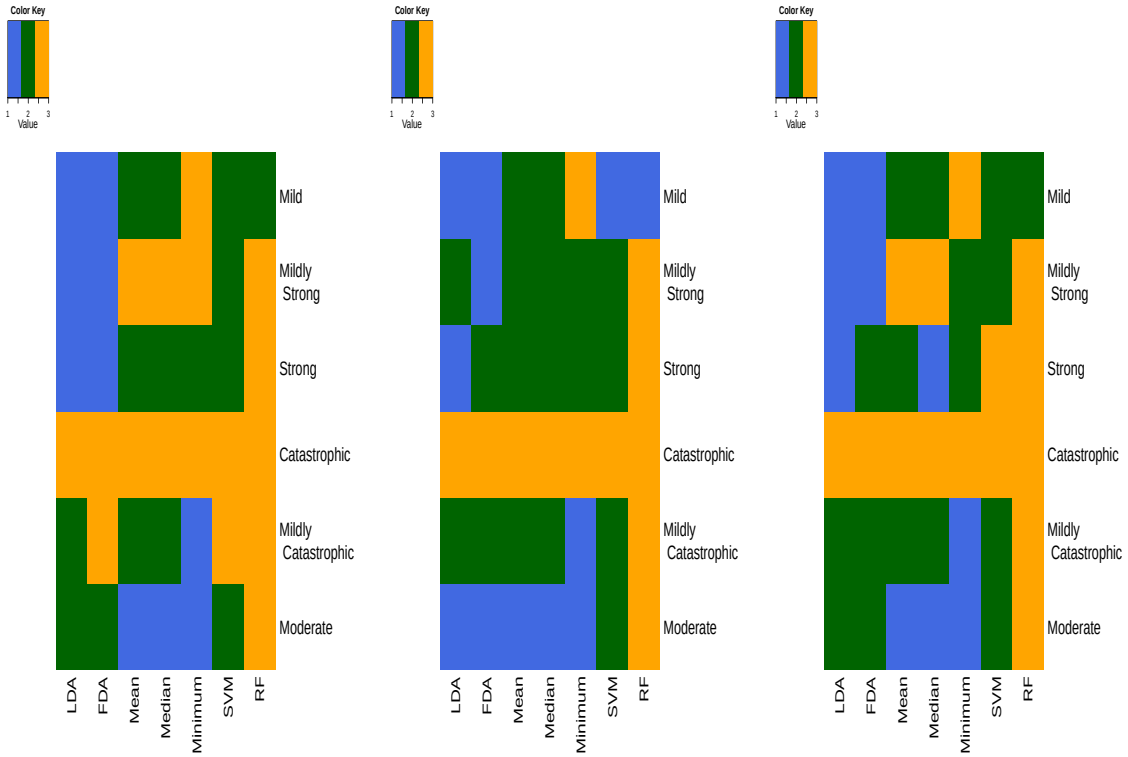
The nearest neighbor methods perform significantly better than all other methods in the identification of mild epidemics across all social networks. All the methods perform well in the classification of catastrophic epidemics, while RF is best for the classification of mildly strong, strong, mildly catastrophic and moderate epidemics. Comparisons of rankings across social networks are further discussed in section 3.4.4.

Based on Figures 3.2 and 3.3, RF appears to be the most consistent method. SVM is also consistent, ranking as the least preferable method only once. FDA and LDA are ranked as the least preferable methods more often than others which suggests they should be avoided.

3.4.3 Combined Classification Weighting Schemes

The simple voting combined classifier does not appear to perform better than the individual classification methods in the daily identification of epidemics curves for the six epidemics (Figure 3.2). In most cases, the classifier seems to capture the mean accuracy rate of all seven classification methods since it is influenced by both successful and less-successful classifiers.

In addition to the simple voting combined classifier, weighted voting classifiers were also proposed as discussed in section 3.3. Figure 3.4 shows the results of the comparison of the simple voting classifier to the weighted voting classifiers for all six epidemics. The combined classifiers are represented as follows: simple voting classifier by “simple voting”, weighted linear voting by “weighted-linear”, weighted polynomial voting by “weighted-poly” and weighted exponential voting by “weighted-expo”. In addition, the combined classifiers are also compared to the consistent methods (RF, SVM) from section 3.4.2.



(a) Seattle (b) New York (c) Los Angeles

Figure 3.3: Ranking of methods by epidemic and by region based on results of the McNemar test and the performance accuracy metric. 1 represents the least preferable and 3 represents the most preferable method for each epidemic. The color code is 1: royal blue, 2: dark green, and 3: orange.

There appear to be few differences in the classification of the catastrophic, mildly catastrophic and strong epidemics (Figure 3.4). Differences between the methods are more apparent for the mildly strong, moderate and mild epidemics. These differences are further investigated using the McNemar test and the performance accuracy metric.

The results from the McNemar tests indicate that none of the methods perform better in the classification of catastrophic and mildly catastrophic, which is reinforced by Figure 3.4. In addition, RF and SVM perform significantly better than all combined classifiers in the classification of moderate epidemics. In the classification of moderate and mild epidemics, none of the methods perform significantly better than others. Similar results are observed for New York and Los Angeles. Based on these results, we can conclude that in most cases, the combined classifiers perform as well as random forest and support vector machines.

3.4.4 Different Social Networks

The Chi-square test for independence is used in testing whether the performance of each of the classification methods is independent of the social networks over which the epidemics are simulated. Both the Pearson and Likelihood Chi-square tests indicate that there is no statistically significant evidence to reject the hypothesis of independence with p-values in the range [0.38, 1.00]. These results suggest that the accuracy of the eight classification methods in identifying epidemic curves from each of the six stochastic simulated epidemics is independent of the social networks over which the epidemics are simulated.

In addition, the best methods for identifying all six epidemics are also similar and the most consistent method across all epidemics is RF for all social networks (see Figure 3.3). However, there are a few differences in the rankings of the methods by epidemic across the social networks. For Seattle, RF and all nearest neighbor methods are ranked best in the classification of mildly strong epidemics. For New York, only RF is ranked best, while for Los Angeles, RF, Median and Mean are ranked best in the classification of mildly strong epidemics. Another difference can be observed in the classification of mildly catastrophic outbreaks. Only RF is ranked best for Seattle and Los Angeles, while for New York, RF, SVM and FDA are ranked best. These differences in rankings are however minute and are also subject to our definition of the accuracy metric.

There are several possible reasons why the results observed for the three regions are similar. Similarities in the area under the epidemic curve (total attack rate), the time to peak and peak infection rate could imply that the shape and form of the simulated epidemic curves are alike across regions. An analysis of variance on these three measures gave the following results: the attack rates are statistically significantly different $(P < 0.00001)$ across all regions, the peak infection rates are statistically significantly different $(P < 0.00001)$ across all regions and except for mildly catastrophic epidemics, the times to peak are also statistically significantly different across all regions. A comparison of the times to peak of mildly catas-

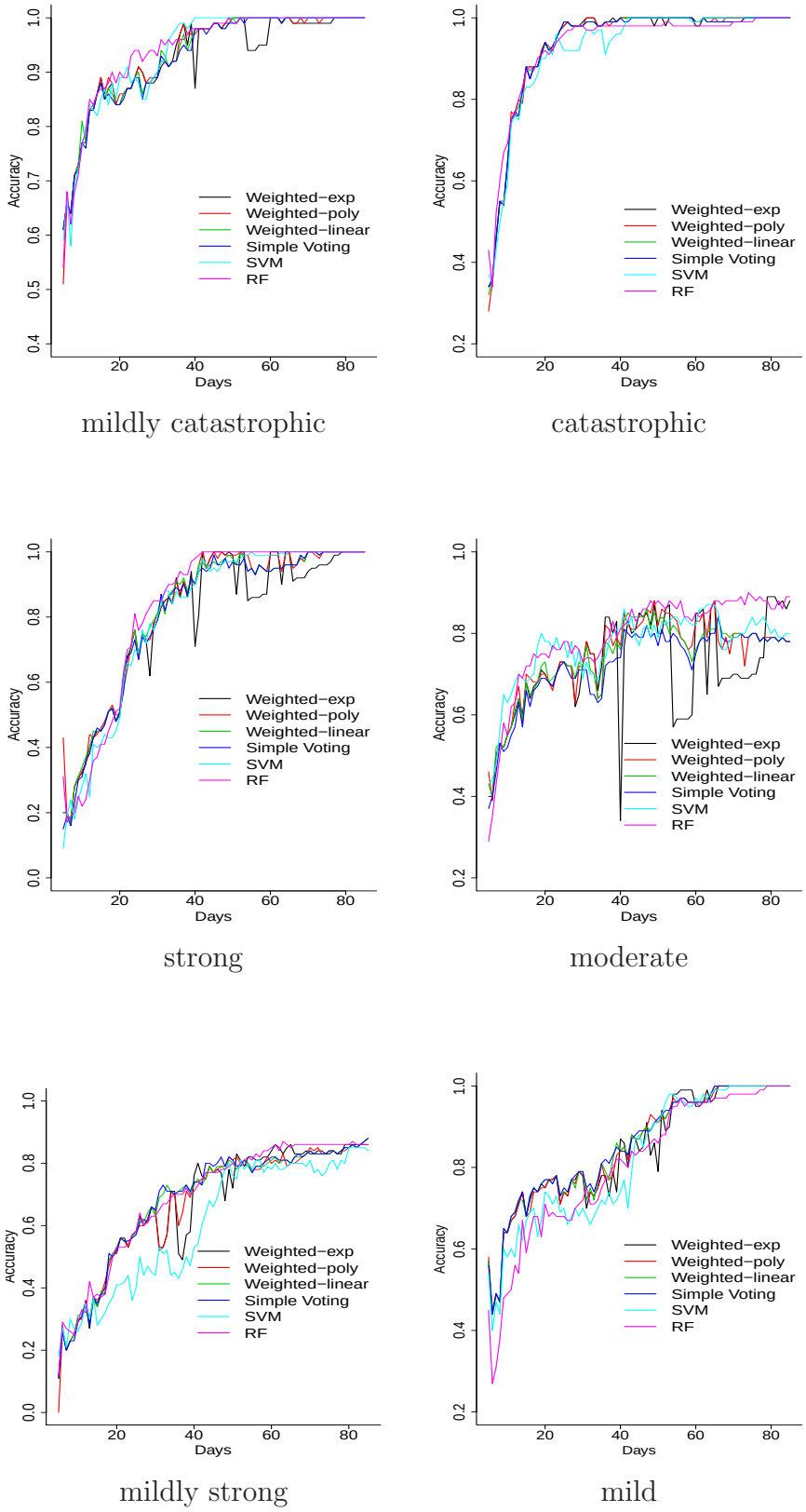


Figure 3.4: The performance of the combined classification schemes. Results are presented for Seattle.

trophic epidemics for Los Angeles and Seattle suggested that there is no statistical evidence to reject the null hypothesis that the times to peak are the same *(P = 0.21)*.

However, although the epidemic curves from the same parameter sets appear to be different across the different regions, the epidemic curves for each region peak in the same sequence (mildly catastrophic, catastrophic, strong, moderate, mildly strong and mild epidemics). This therefore suggests that the similarity in performance might be due to the sequence in which the epidemics reach their peaks across the three social networks.

The performance of the methods are also tested on a basic compartmental SEIR model representing a "null" network. Four parameterizations are used to simulate epidemics and the daily infected are used in constructing the epidemic curves. The SEIR model and results are presented in the Appendix. Similar conclusions are drawn from the "null" network as from the three previously discussed social networks; random forest and SVM are the most preferable methods.

3.5 Discussion

This paper serves as the first step in a project where the overall goal is to create a digital library of simulated outbreaks to be used to classify real world surveillance data to assist in estimating infectious disease characteristics. The main aims of this paper were to perform a systematic comparison of seven supervised classification methods in addition to a combined classifier, in order to find which methods perform best in identifying outbreaks from six parameterizations of influenza epidemics and whether the performance of these methods were affected by the social networks across which the outbreaks were simulated.

The correct identification of the underlying transmission parameters for an ongoing outbreak would provide an estimate of the epidemic curve under the assumption that no intervention methods have been applied to control the spread of the disease. Different epidemic curves with different transmission parameters are expected to have different shapes. However, this is not always apparent at the early stages of the outbreaks, which can make the differentiation of outbreaks with different transmission parameters difficult. For instance, in our study, mild outbreaks were misclassified as mildly strong and strong outbreaks at the start of the outbreaks.

The results show that multiple classification techniques can be used for predicting epidemic curves and inferring transmission parameters. However, random forest, which is easy to implement and has few parameters, would be the most preferable method. The consistency and high accuracy achieved by random forest in this problem could be due to its lack of assumptions regarding the structure of the data and the methodology, which involves classification based on the majority vote. In contrast, linear discriminant analysis had the poorest performance out of all classification techniques examined. LDA assumes that all outbreak clusters have a common covariance matrix, which is not consistent with the expectation that

input variables belong to different classes. In addition, LDA uses a single class centroid per class, which can be inadequate in some cases [70].

The results in this analysis demonstrate the complexities associated with dealing with sparse data in classification. In addition, combining classification methods using a voting scheme does not always perform better than the individual classification methods as is usually assumed. The error introduced by less accurate methods heavily affects the performance of the combined classifier whether weights are placed on the more accurate methods or not. Therefore, instead of combining several methods with different assumptions about the data, simpler and more efficient methods such as random forest can be used in the prediction of an epidemic curve.

Some of the limitations of our method deal with the assumption that the observed outbreak can be assigned to at least one of the clusters in our library. A method that systematically identifies an outbreak if it does not belong to any of the clusters and then proposes possible parameters for modeling the outbreak would be more beneficial. Also, based on the healthcare infrastructure of a particular region or city, differences exist in methods used for disease surveillance and selection of interventions. Therefore, the proposed method would need to be adapted to the specific scenario under study by adjusting the details of the ABM.

In the next step, the methods used in this study can be extended to include a probabilistic framework to measure uncertainty in our classification and to extend the analysis to different intervention scenarios and subpopulations such as age groups. We think the results in this study are promising and reinforce the idea that a combination of simulations and classification methods can be used in the prediction and estimation of infectious disease transmission parameters.

3.6 Acknowledgments

We thank Dr. Scotland Leman and members of the Network Dynamics and Simulation Science Laboratory (NDSSL) for their suggestions and comments. This work has been partially supported by NSF Netse Grant CNS-1011769, DTRA R&D Grant HDTRA1-0901-0017, DTRA CNIMS Grant HDTRA1-07-C-0113, and NIH MIDAS project 2U01GM070694-7.

Appendix B

Chapter 3: Appendix

B.0.1 Computational Epidemiology Model

A detailed description of the agent-based model (ABM) used in simulating the outbreaks used in this study is discussed in [17]. The ABM belongs to a class of models called network based epidemiology models that uses a representation of the population that includes each individual and their minute-by-minute movements. Their interactions with other agents are used to generate a dynamic social network. These networks are then in turn used to simulate epidemics and study the effects of changes in individual behavior and public policy on the propagation of an outbreak [8]. Although neither changes in individual behavior nor public policy are directly explored in this study, it is extremely easy in these models to change individual behaviors, like keeping children home from school or in general limiting the number of non-essential activities of specific members of the population. However the purpose of this study is to assess the performance of classification techniques given partial epidemic curves. When the full procedure is implemented in the future, simulations will include epidemic curves that have these characteristics.

The creation of the agent-based epidemic model used in this study entails two major steps, the first of which consists of the creation of a social contact network from a state-of-the-art behavioral model. This involves creating synthetic populations and time varying social networks. Synthetic individuals and households, located in specified geographical regions (such as Los Angeles), each with a set of demographic variables are created using an iterative proportional fit to joint demographic distributions from the 2000 US census data provided in SF3 and PUMA (Public Use Microdata Area) files [7]. For example, a list of demographic information such as household income, family size, age, education etc. are available for each of the approximately 16 million individuals in the Los Angeles region.

The synthetic populations are created to produce realistic features and demographics while preserving the confidentiality of the original data sets. An node represents an individual in the synthetic population. Each node is placed in a household with other synthetic individuals

and each household is geographically located such that a census aggregated to the block level of the synthetic population would be statistically identical to the real census data [14]. Additional information can be found in [14], [116] and [117].

Next, each individual in the synthetic household is allotted activities by time of day based on several thousand responses to an activity or time-use survey for a specific region. The National Household Transportation Survey was used in creating the activity templates assigned to each household. The time-use or activity survey is expected to vary by region given factors such as the geographical location and age composition of the population. Presently, this modeling approach is considered the *de facto* standard in transportation science and is called activity based travel demand models [8]. See [19] and [20] for additional information.

Using a decision tree based on demographic information (such as number of people in a household, number of children etc.), each household in the synthetic population is matched to a survey household. Each activity for each synthetic person is then assigned an appropriate real location based on a gravity model and land-use data [14]. The addresses of locations are obtained from Dun and Bradstreet's Strategic Database Marketing Records. The activities for each household are assigned to actual locations based on "the distance from the previous activity and its *attractiveness* a measure of how likely that the activity happens there - number of employees, school enrollment, square feet of retail shopping etc." [46].

In addition to specific assignment of activities, the time at which each activity starts and ends is also included. This leads to each individual in each household having a minute-by-minute schedule for each day. Synthetic individuals in the population interact with each other based on their minute-by-minute schedule to produce realistic contact graphs where vertices represent individuals and edges represent contacts between individuals [8]. Individuals mimic the behaviors of real people by participating in everyday activities such as eating, socializing, shopping etc. and multiple edges can be used between each person and the locations representing their frequency of visits. The modeling approaches used in the ABM as presented in [8] are given in Table B.1.

Next, a computational model is developed to represent disease within individuals and the transmission between individuals in the synthetic population. The transition from one disease state (susceptible, exposed, infectious and removed) to another is probabilistic and timed (e.g. it may be represented by the distribution of the infectious period). The transition between states can also depend on the attributes of the people (e.g. age, health status etc.) and the type of contact (e.g casual, intimate etc.). For the disease model used in this study, the probability that an infectious person i infects a susceptible person j is given by:

$$Pr(\text{person } i \text{ infects person } j) = 1 - exp(-f(S, T_{i,j})) \tag{B.1}$$

where $f(S, T_{i,j})$ is a nonnegative monotone increasing function of S, the severity of the disease (called transmissibility) and $T_{i,j}$, the contact time between persons i and j. The disease model is combined with the information in the network to study an infectious disease. At each time step of a simulation, each of the nodes is either: susceptible, exposed, infectious,

Table B.1: Models and modeling approaches used in ABM

Models	**References**
Urban Population Mobility Models	[10], [20], [125], and [124]
Natural Disease History	[3], [44], [67], [73], and [93]
Transmission Models	[67], [73], and [93]
Social Network Models	[47], [67], [100]
Types of Interventions	[50], [51], [66], and [67]

and removed. Time is divided into units based on days and the state of each individual is noted at the start of each day.

To run a simulation experiment, a population (contact network), characteristics of a disease and initial conditions (such as duration) are specified. In this study, the social networks studied were based on Seattle, Los Angeles, New York and surrounding metropolitan regions. The disease characteristics were based on influenza epidemics. For each simulated outbreak, several realizations of the stochastic process of disease propagation are computed. Intervention options such as vaccination, antiviral and social distancing can be applied during the outbreak to control its propagation. Each simulation is seeded with a randomly selected set of initially infected individuals. An epidemic curve, the vulnerability of different individuals in the network, epidemic size, number of new exposures on each day etc. can be explored at the end of each simulation of a disease outbreak.

Several studies have been implemented to validate specific components of the model and the general approach. See [14], [47], and [67] for structural validity of these models.

B.0.2 Analysis on a "Null" Network

The epidemics for the "null" network are simulated using a stochastic compartmental SEIR model for an epidemic. The SEIR model is based on the discretized stochastic SEIR model given by [89]. S(t), E(t), I(t) and R(t) represent the number of susceptible, exposed, infectious and removed individuals respectively at time t. Given the initial number of individuals in each compartment, the model can be specified as follows:

$$\begin{aligned} S(t+h) &= S(t) - S_E(t) \\ E(t+h) &= E(t) + S_E(t) - E_I(t) \\ I(t+h) &= I(t) + E_I(t) - I_R(t) \\ S(t) + E(t) + I(t) + R(t) &= N \end{aligned} \tag{B.2}$$

where $S_E(t) \sim Bin(S(t), P(t))$ is the number of susceptible persons who become exposed and $P(t) = 1 - exp[-\frac{\beta(t)}{N} hI(t)]$. $E_I(t)$ represents the number of exposed individuals who become infected and $E_I(t) \sim Bin(E(t), pc)$ where pc= *1- exp(-*θ *h).* $I_R(t)$ is the number of cases removed from infectious compartment at time t and $I_R(t) \sim Bin(I(t), p_R)$ where p_R = *1 - exp (-*γ *h).* The time-dependent transmission rate, mean incubation and mean infectious periods are given by $\beta(t)$, $1/\theta$ and $1/\gamma$ respectively.

The counts of daily infected are used in constructing the epidemic curves. Epidemics are simulated from four parameterizations of the SEIR model with two sets of mean incubation and infectious periods, and two transmission rates. The mean incubation and infectious periods are based on the distributions in Table 3.1.

The methods are tested on all four sets of epidemic curves and ranked based on the performance metric as shown in Figure B.1. Random forest and SVM are ranked best, while the nearest neighbors are ranked "acceptable" two out of four times. In agreement with the results in 3.4.4, this also suggests that the discriminant analysis methods should be avoided. As previously mentioned, the ABM is used instead of the simple compartmental SEIR model due to the overall goal of the project. In addition, the ABM results can be specific to a particular region while the "null" network results are generalized.

B.0.3 Classification Techniques

In this section, we present a brief introduction into the classification methods used in our analysis.

Random Forest

Bagging which stands for bootstrap aggregating is a method for decreasing the variance of an estimated prediction function by combining several predictors [22, 70]. Random Forest is an extension of bagging. It involves growing several de-correlated trees, which are then averaged [70]. Trees in a random forest are identically distributed and have the same expectation.

The random forest algorithm for classification as stated in [70] follows:

1. For b = 1 to B:
(a) Draw a bootstrap sample Z* of size N from the training data

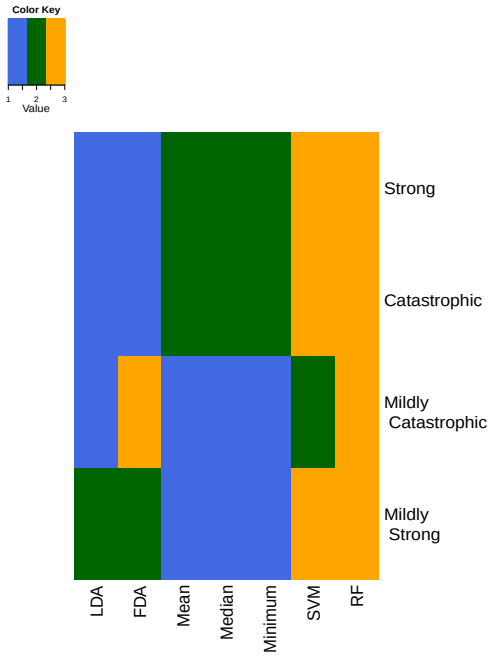


Figure B.1: Ranking of methods by epidemic based on results of the McNemar test and the performance accuracy metric. 1 represents the least preferable and 3 represents the most preferable method for each epidemic. The color code is 1: royal blue, 2: dark green, and 3: orange.

(b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached

i. Select m variables at random from the p variables

ii. Pick the best variable/split-point among the m

iii. Split the node into two daughter nodes

2. Output the ensemble of trees $(T_b)_1{}^B$

To make a prediction at a new point x: Let $C_b(x)$ be the class prediction of the b_{th} random-forest tree. Then $C^B{}_{rf}(x)$ = majority vote $(C_b(x))^B{}_1$

There are several professed advantages to random forests. Random forest runs efficiently on large databases in a short amount of time, provides estimates of important variables in classification and can be used in unsupervised classification [23]. Random forest also calculates an out-of-bag (oob) error rate based on its out-of-bag samples feature which involves constructing random forest predictor for an observation "by averaging only those trees corresponding to bootstrap samples in which the observation did not appear" [70].

The random forest technique was modeled using the randomForest package in R [81]. In order to improve the classification rate, for each model fitted to the training data, the number of variables randomly sampled from candidates at each split was set to a value which produced

the minimum error rate with 500 trees grown.

Linear Discriminant Analysis

In linear discriminant analysis (LDA), the data are projected onto a low dimensional vector space to maximize the ratio of between-class variance to within-class variance, thereby, obtaining maximal discrimination between classes. The density for each class is modeled as a multivariate Gaussian and the classes are assumed to have a common covariance matrix [70].

Consider the problem of classifying an object $x_i = (x_{i1}, ..., x_{id})^T$ based on d dimensions to one of k formerly defined classes. The linear discriminant function (classification score) is given by:

$$\varphi_{\mathrm{k}}(\mathrm{x}) = \mathrm{x}^{\mathrm{T}}\Sigma^{-1}\mu_{\mathrm{k}} - \frac{1}{2}\mu_{\mathrm{k}}{}^{\mathrm{T}}\Sigma^{-1}\mu_{\mathrm{k}} + \ln(\pi_{\mathrm{k}}) \tag{B.3}$$

where Σ is the common covariance, π_k is the prior probability and μ_k is the mean vector for class k. The parameters of the Gaussian distributions are estimated using the training data:

$$\begin{aligned} \widehat{\Sigma} &= \frac{\sum_{k=1}^{K}\sum_{g_i=k}(x_i-\mu_k)(x_i-\mu_k)^T}{(N-K)} \\ \widehat{\pi_k} &= \frac{N_k}{N} \\ \widehat{\mu_k} &= \frac{1}{N_k}\sum_{g_i=k} x_i \end{aligned} \tag{B.4}$$

where N is the number of objects in the training data and N_k is the number of objects in class k. Object x_i is classified into the class with the smallest $\varphi_k(x)$ [70, 133].

The linear discriminant analysis was modeled using the lda function in the MASS package in R [81]. Equal prior probabilities were set for the six data class memberships.

Flexible Discriminant Analysis

"LDA can be performed by a sequence of linear regressions, followed by classification to the closest class centroid in the space of fits" [70]. This leads to a generalization of LDA where the linear regression fits are replaced by more flexible nonparametric fits such as multivariate adaptive regression splines (MARS) introduced by [55]. See [55] for more information on MARS. The nonparametric fits results in a more flexible classifier which is expected to result in better classification than LDA since the underlying relationship between variables are not assumed but derived from basis functions.

Support Vector Machines

Classification using support vector machines involves two steps: mapping the data into a predetermined high-dimensional space via a kernel function (linear, Gaussian, radial and polynomial) and finding the hyperplane that maximizes the margin between the data classes. The overall aim of an SVM is to find the hyperplane that maximizes separation and minimizes misclassifications.

Consider the training data with two predictor variables and two separable classes. Say there are n objects in each sample given by $(\overrightarrow{x_1},y_1),\ldots,(\overrightarrow{x_n},y_n)$. Finding an optimal hyperplane for the data, involves finding the weight $\overrightarrow{w}$ and bias (threshold) b to construct two margins, which provide maximum separation between the two classes.

$$\begin{aligned} \overrightarrow{w^T}\overrightarrow{x_i}+b \geq 1, y_i &= +1 \\ \overrightarrow{w^T}\overrightarrow{x_i}+b \leq 1, y_i &= -1 \end{aligned} \tag{B.5}$$

where $\overrightarrow{w}=(w_1,...,w_n)^T$ is a vector of size n.

The optimal hyperplane would maximize the distance between the closest points between classes while separating the two classes [128]. The vectors (or points) near the hyperplane are called support vectors. Any given hyperplane can be expressed as: $\overrightarrow{w}\overrightarrow{x}+b=0$ and width of the margin is: $\gamma=\frac{2}{||w||}$. So the optimal separating hyperplane can be found by maximizing γ such that (B.5) holds true. The case discussed above is the ideal case. In cases of non-linearly separable problems, the optimal hyperplane is extended to include a penalty term for misclassifications [70].

The support vector machine was modeled using the e1071 package in R [81]. The classification was implemented using a linear kernel and the cost of misclassification was set to the value resulting in the minimum error rate.

Nearest Neighbor

The nearest neighbor procedure is defined as follows: for each new object, find the k most similar objects in the training set based on a distance metric. The cluster to which the new instance is assigned is the most frequent out of all k nearest objects.

Three nearest neighbor approaches were defined for the purposes of this paper. The nearest neighbor classifiers were: nearest median, nearest mean and minimum distance approaches. Given a partial epidemic curve, the squared Euclidean distance was calculated between the partial epidemic curve and the median (mean) curve of each of the epidemic clusters. The partial curves was assigned to the epidemic cluster with the nearest median (mean) curve. The minimum distance approach is based on the typical nearest neighbor approach; a new

epidemic curve is compared to all curves in the library and then assigned to the cluster with the closest curve.

The nearest median and mean approaches are less computationally intensive than the usual nearest neighbor method since the new object is not compared to every object in the training set.

Chapter 4

A Dirichlet Process Model for Prediction of Epidemic Curves

Elaine Nsoesie [1], Scotland Leman[2] and Madhav Marathe [1,3]

[1] Network Dynamics and Simulation Science Laboratory, Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, Virginia, USA

[2] Department of Statistics, Virginia Tech, Blacksburg, Virginia, USA

[3] Computer Science Department, Virginia Tech, Blacksburg, Virginia, USA

Abstract

We present a Dirichlet process model for predicting the influenza epidemic curve. The Dirichlet process is a nonparametric Bayesian approach that enables the matching of current influenza activity to simulated and historical epidemic curves, identify epidemics different from those observed in the past and predict the time point at which an epidemic is expected to peak. This study is part of a project aimed at using a combination of simulation and classification techniques to predict epidemic curves during an influenza outbreak.

The method is illustrated using simulated influenza epidemics from an individual-based model. We compare the accuracy to that of the tree based classification technique random forest, which has been shown to achieve high accuracy in the early prediction of epidemics using a classification approach.

The results indicate the following: (i) the method's accuracy consistently improves over time, (ii) the best predictions are made during or after the exponential growth phase of the epidemic and (iii) the time to peak for most of the epidemics in this study can be accurately predicted several days before the peak.

4.1 Introduction

Influenza pandemics result from novel influenza viruses circulating in the human population for which there is little or no pre-existing immunity [34]. The 1918 pandemic resulted in an estimated 20-50 million deaths worldwide [83]. A similar pandemic today will result in higher morbidity and mortality rates due to factors such as ease of travel and population density [11, 113]. Several methods have been suggested to improve preparedness for the next influenza pandemic. These include methods for forecasting the epidemic curve (see [101] and [104] for a few examples). The epidemic curve is defined in this study as the daily counts of infected persons for the duration of the epidemic. A range of models at the individual and population level have been used in forecasting the influenza epidemic curve [38, 65, 101, 104, 105]. In this paper, we introduce a Bayesian nonparametric clustering approach for predicting the epidemic curve and illustrate its performance using simulated data from an individual-based model.

Formerly, Jiang et al. [82] published a Bayesian network for predicting the daily-infected counts of an influenza outbreak based on the assumed severity, duration and the daily counts of patients reporting to the emergency department with a respiratory illness. Additional methods were proposed during the 2009 pandemic. One of such methods employed a discrete time stochastic likelihood-based model for forecasting epidemic dynamics in Japan [101]. Similarly, Ong et al. [105] proposed a Bayesian and stochastic compartmental model for real-time epidemic monitoring and forecasting during the pandemic in Singapore. In contrast to likelihood and Bayesian methods, a simpler approach using classification techniques was proposed for predicting the epidemic curve and inferring underlying disease model parameters [103]. This classification approach was based on the idea of matching new epidemics to historical and simulated influenza epidemic curves. However, the major shortcoming to this approach is that it was fully supervised, which cannot extend to the case of correctly predicting unanticipated disease trends.

In this study we present a semi-supervised clustering technique which does not only seek to classify epidemics similar to those observed in the past, but also seeks to identify novel outbreaks. The forecasting problem is explored under the following scenario: during an epidemic of a novel influenza virus, several possible parameters sets are proposed for modeling the new epidemic. The parameters investigated in this study are the transmissibility (typically represented using the reproduction number), and the infectious and incubation period distributions. These parameters are based on a **S**usceptible, **E**xposed, **I**nfectious and **R**ecovered (SEIR) disease model as shown in Figure 4.1. Using the hypothesized parameters, several possible scenarios describing the ongoing epidemic are stochastically simulated. We create a library of epidemics consisting of the simulated epidemics and historical epidemic data (Figure 4.2). Epidemics with the same parameters are grouped to form an epidemic cluster. The parameters for the new epidemic are estimated by comparing the partial epidemic curve based on daily infected cases to the epidemics in our library. In the event that the new epidemic cannot be assigned to any of the clusters in our library, a combination

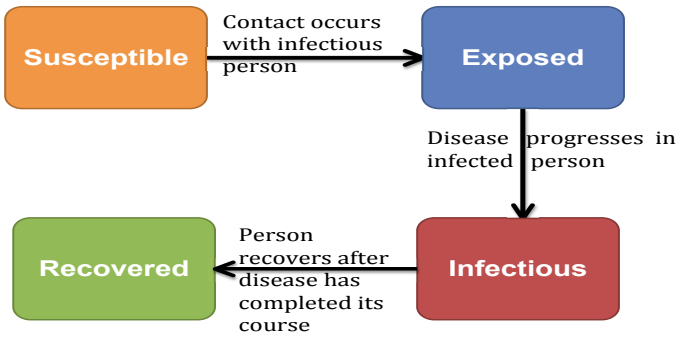


Figure 4.1: SEIR model

of expert opinion and search algorithms are used to propose new sets of parameters. The simulation and classification process is repeated for each day j as data is updated. This study completes the first part of a two-step procedure for predicting the epidemic curve using this approach. The first step of the procedure involves identifying current epidemics with parameters similar or different from those in our epidemic library (Figure 4.2). The second step involves searching for new sets of parameters to model the ongoing epidemic.

An individual-based model [17] is used in simulating the epidemics in the library. Individual-based models for influenza are computational epidemiology models built on synthetic populations, with detailed representation of entities (such as humans) and their environment [17, 37, 41, 45]. Disease is transmitted through contacts between susceptible and infectious individuals. Individuals move through four disease states (Figure 4.1): susceptible, exposed, infectious and recovered. Recovered individuals remain in the population but can no longer spread the disease due to immunity. To classify epidemics stochastically simulated using the individual-based model and to infer transmission parameters, we explore the clustering properties of a semi-supervised Dirichlet process model. The Dirichlet process is a non-parametric Bayesian procedure that presents a good solution to this problem since novel outbreaks can be identified. Since Ferguson [52] formalized the Dirichlet process as a prior over distributions, there have been several extensions in terms of inference and applications [58]. The Dirichlet process has been proposed as a solution to finding the number of spatial activation patterns in fMRI images [87], the modeling of unknown number of topics across several corpora of documents [121], clustering population genetics data [110], detecting positive selection in protein-coding DNA sequences [79] etc. Although the Dirichlet process has been used in several studies, to our knowledge it has not been used in a procedure aimed at forecasting the epidemic curve.

There are several advantages to using this simulation and classification approach for epidemic curve prediction. (i) The method captures the temporal trend of the epidemic curve and estimates the expected peaking time. (ii) Using an individual-based model for simulating epidemics implies that in addition to forecasting the epidemic curve, the effects of various intervention methods and changes in individual behavior can be explored. Furthermore,

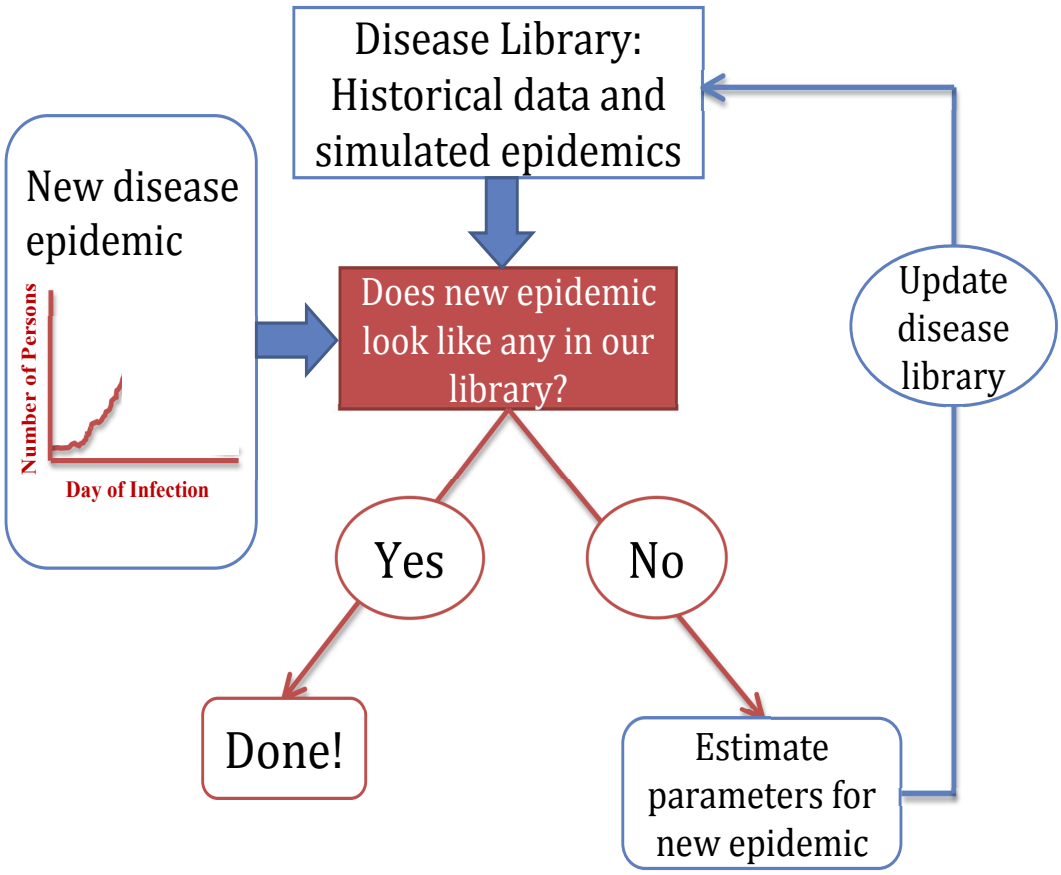


Figure 4.2: Summary of methodology. We develop a library of past and simulated epidemics. Given surveillance data for a current epidemic, we compare the partial surveillance epidemic curve to those in the library. The novel epidemic is either assigned to a case in the library or identified as being different from those in the library. If the epidemic is different from those in the library, we estimate the model parameters, forecast the epidemic curve and update the library.

the severity of the epidemic can be estimated under different scenarios by implementing changes to the individual-based model. (iii) The proposed approach does not only predict the epidemic curve for an outbreak with similar characteristics to the epidemics in our library but can also identify novel outbreaks. Note, that the focus of this initial study is to establish the accuracy and usefulness of the method.

4.2 Epidemic Simulation

The individual-based model used in this study falls into a class of computational epidemiology models used in studying disease transmission and methods for control. The individual-based model has been used in several studies in evaluating the effects of various intervention methods on the spread of influenza [17, 61, 103]. The construction of the individual-based model involves the creation of a state-of-the-art behavioral model and a model for disease transmission. To create the behavioral model, a synthetic population for a specific geographic region is constructed using United States census data [7]. Each individual is assigned a set of demographic variables such as age, household size, household income etc. Households in the synthetic population are located such that a census of the synthetic population is statistically identical to the real census data at the block level [14]. Each synthetic individual is also assigned a schedule based on data from an activity survey [8]. Synthetic individuals come in contact with other individuals at different activity locations resulting in a dynamic social contact network through which disease transmission occurs.

The individual-based model is not the focus of this study since similar epidemics can be simulated using a differential equations SEIR model. However, the individual-based model is used in this study due to the overall premise of the project to not only forecast the epidemic curve but also to investigate the effects of intervention methods and changes in individual behavior to the spread of the disease. Details of the individual-based model are summarized in the Appendix. A detailed description can also be found in [17].

The data used in this study consists of three epidemic clusters (Figure 4.3). Each cluster contains 115 epidemic curves. These clusters are named strong, severe and catastrophic based on the total number of persons infected during the epidemic. Epidemics in each cluster are stochastically simulated using the same parameters. The incubation and infectious period distributions are based on parameters used to model seasonal influenza epidemics [67] and the 2009 H1N1(A) pandemic [103]. However the transmissibilities are selected to produce epidemics more severe than seasonal influenza since milder epidemics are less likely to have a significant impact on the population. For epidemics in the catastrophic and strong clusters, infected individuals in the synthetic population have an incubation period duration of 1, 2 or 3 days with probabilities 0.3, 0.5 and 0.2 respectively. Infected individuals also have an infectious period duration of either 3, 4, 5 or 6 days with probabilities 0.3, 0.4, 0.2 and 0.1. The second set of incubation and infectious period distributions are obtained by adjusting the previously described distributions to match the serial interval [31] for the 2009 H1N1

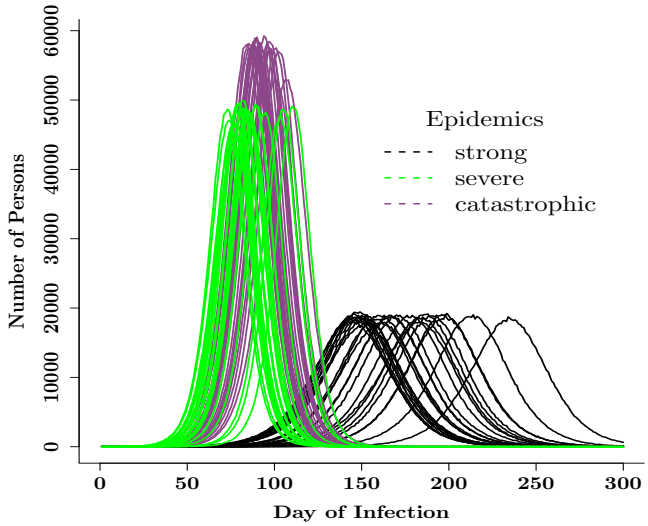


Figure 4.3: Sample epidemic curves for three epidemics

(A) pandemic. This results in infected individuals in the severe epidemic cluster having an incubation period of 0, 1, or 2 days with probabilities 0.20, 0.45 and 0.35 respectively. The infectious period can be 2, 3 or 4 days with probabilities 0.66, 0.33 and 0.01 respectively. The transmissibilities are set at 4.2E-5, 8.3E-5 and 6.0E-5 per sec/contact time for strong, severe and catastrophic epidemics respectively. The epidemics are simulated over a single synthetic social network representing the population of Seattle and surrounding metropolitan region (hereafter referred to as Seattle).

Per Figure 4.3, the complexity in this study lies in differentiating these epidemic curves during the early stages of the epidemics. Figure 4.3 also suggests that the accuracy of predicting the epidemic curves should improve over time. Although this is expected, we wish to propose a procedure which identifies epidemics in their early stages. Accurate predictions would be invaluable to public health officials.

4.3 Selection of Parametric Distribution to Model Data

During most infectious disease outbreaks, the number of new cases increases until it peaks and then declines thereafter [4]. Under the assumption that the epidemic curve has a single peak, we model the daily infected-counts using parametric models based on statistical distributions. An alternative to this would involve using a time series model or a mixture of various parametric models, which might better capture the shape of the epidemic curves. However, we propose first using a simpler model before attempting to use a more complex model, which would require a longer time frame for implementation and is not guaranteed

Table 4.1: Overall fit

Distributions	Mean of mean absolute error	Variance of mean absolute error
Negative binomial	0.00070	2.56e-09
Poisson	0.00167	5.18e-07
Weibull	0.00118	2.25e-08
Normal	0.00058	3.18e-09
Pareto	0.00747	1.02e-06
GEV	0.00122	4.63e-08
Cauchy	0.00163	5.81e-09

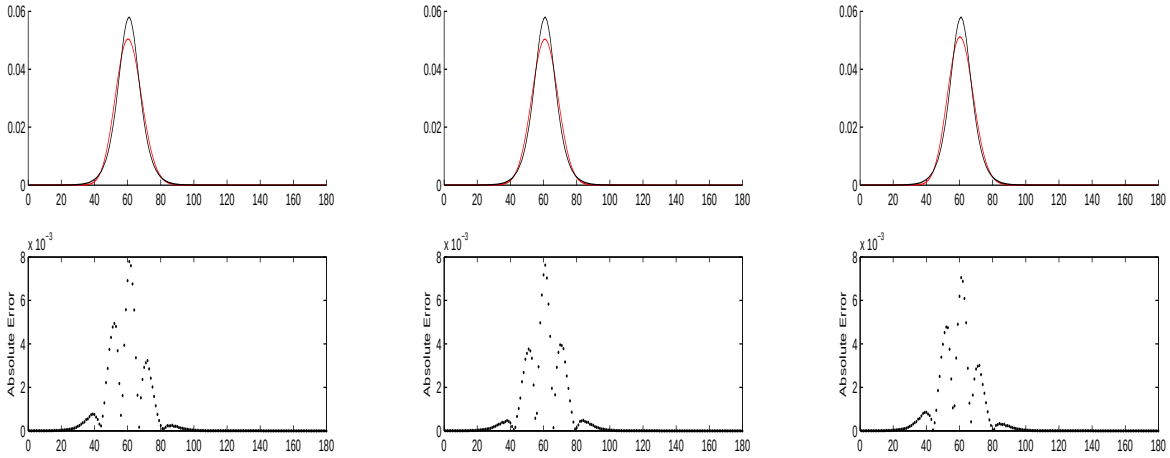
to perform better.

We fit samples of the data to seven parametric models: negative binomial, Poisson, Weibull, normal, Pareto, generalized extreme values (GEV) and Cauchy. The parametric models are selected to allow the exploration of different discrete and continuous distributions that appear to capture different aspects of the epidemic curves. The normal distribution is selected because in most cases, epidemic curves appear to be slight deviations of the normal distribution. In addition, epidemic curves in this study are daily counts of infected persons over time, which implies discrete distributions such as the Poisson and negative binomial could be used for modeling the data. On the other hand, the Weibull, Pareto and Cauchy are heavy tailed distributions which could capture the heavy tails observed in some epidemics. Finally, the GEV is a three parameter (location, scale and shape) family distribution, which could be used to capture the deviations in the shape of the epidemic curves.

To illustrate the fits of the data to the selected distributions, we do the following. (i) Estimate the parameters for each epidemic curve using the maximum likelihood procedure. (ii) The estimated parameters are then used to draw samples from the parametric distribution with the same length (number of days) as the epidemic curves. (iii) The epidemic curve is normalized and compared against the normalized samples from the parametric distribution. (iv) The absolute error is estimated for each day and the mean absolute error is estimated across all days. The analysis is performed for each of the epidemic curves in the epidemic library. The average of the mean and variance of the mean absolute error are taken across all epidemic curves. The results are shown in Table 4.1.

The normal, negative binomial and Weibull distributions have the best fit to the complete data for all epidemics based on the mean of the mean absolute error of the fitted data to the observed data as shown in Table 4.1. The normal, negative binomial and Weibull are ranked best based on the variance of the mean absolute error. The GEV distribution results in fits similar to that of the Weibull with estimated $k < -0.5$. This implies that the GEV distribution converges to the reversed Weibull distribution.

The best fits for each epidemic cluster are different in some cases from what is observed in Table 4.1. However, in most cases, the normal and negative binomial distributions are ranked



Negative Binomial Normal Poisson

Figure 4.4: A sample fit of a randomly selected epidemic curve to three parametric distributions. The x-axis represents the days and the y-axis is the normalized number of infected persons relative to the cumulative infected. The black curve is the true epidemic curve and the red is the fitted curve. Using the epidemic data, the maximum likelihood is used in estimating the parameters for each of the distributions. The normalized epidemic data is plotted against samples from the distributions based on the estimated parameters.

best based on the measures in Table 4.1. For the third best distribution, the selection lay between Weibull, the GEV and Poisson for most epidemic curves. The Poisson distribution is selected as the third parametric distribution for this analysis, because the beta distribution is used as a conjugate prior, enabling a closed form estimation of the posterior and predictive distributions. See Figure 4.4 for example fits for normal, negative binomial and Poisson illustrated for a randomly chosen epidemic curve. Note how similar the fits are for this particular epidemic curve; this is not always the case. The fit of the parametric distributions varies depending on the shape of the epidemic curve.

4.4 Methodology

Each epidemic curve belongs to an epidemic cluster. Each cluster has a different set of parameters for the same family of distributions such as the normal distribution. Note that these parameters are different from those used in the disease model (transmissibility, incubation and infectious period) but do have the same meaning. The number of possible clusters and the parameters describing each curve are considered to be random variables. The prior probability distribution for the number of clusters is described by a Dirichlet process prior. Since the number of mixture components for the Dirichlet process can be countably infinite, it can be used for identifying epidemics with parameters sets which are distinct from those

in our epidemic library. The Dirichlet process is described in the proceeding section.

4.4.1 Dirichlet Process Models

The Dirichlet process (DP) represents a process prior in nonparametric Bayesian mixture models. Here nonparametric implies that distributions from the Dirichlet process have an infinite number of parameters [120]. The Dirichlet process can be defined as follows:

Let H be a distribution over a measurable space Γ and α a positive real number. For any finite measurable partition $B_1,\ldots,B_r$ of Γ, the random vector $(G(B_1),\ldots,G(B_r))$ has a finite-dimensional Dirichlet distribution with base measure H and concentration parameter α given by $G \sim DP(\alpha,H)$ if:

$$(G(B_1,\ldots,G(B_r)) \sim Dir(\alpha H(B_1),\ldots,\alpha H(B_r)) \tag{1}$$

The two parameters H and α denote the mean: $E[G(B)] = H(B)$ and inverse variance: $V[G(B)] = H(B)(1-H(B))/(\alpha+1)$ of the DP respectively for any measurable partition B $\subset \Gamma$. Since α represents the inverse variance, when α is large which implies a small variance, the DP is concentrated around its mean. As $\alpha \to \infty$ for any measurable B, $G(B) \to H(B)$. Even with a small H, draws from a DP are discrete since G is discrete with countably infinite point masses [18].

A simple property of a finite-dimensional Dirichlet distribution is that the sum of the probabilities of disjoint partitions is also a joint Dirichlet distribution whose parameters are sum of the parameters of the original Dirichlet distribution [58]. This property also holds true for the Dirichlet process. In addition, samples from a DP are discrete which leads to the observation of ties useful for clustering. The clustering property of the Dirichlet process can be best described using the Chinese restaurant process.

Chinese Restaurant Process

The Chinese Restaurant Process (CRP) can be described as follows: consider a restaurant with infinitely many tables and infinite number of customers can be seated at each table. Each customer enters the restaurant and selects a table to sit on. In general, the $(n+1)^{st}$ customer would sit at an occupied table with probability proportional to the number n_k of customers at that table or sit at a new table with probability proportional to α. This can be represented as:

$$X_n|X_1,\ldots,X_{n-1} \sim \begin{cases} X_k{}^* & \text{with probability } \frac{n_k}{\alpha+n-1} \quad \text{j} = 1, \ldots, \text{k} \\ \text{new draw from G} & \text{with probability } \frac{\alpha}{\alpha+n-1} \end{cases} \tag{2}$$

The tables can be thought of as clusters and customers can be represented using integers $1,2,\ldots,$. An important aspect of CRP, is the fact that most Chinese restaurants have round tables. This implies that with n customers in the restaurant, the tables would define both a distribution over permutations of n and a distribution over partitions n.

The expected number of tables m among the n customers is given by:

$$\mathbb{E}[m|n] = \sum_{k=1}^{n} \frac{\alpha}{\alpha+k-1} \in O(\alpha \log \text{ n}) \tag{3}$$

Where $\alpha/(\alpha+k+1)$ is the probability that the k^{th} customer takes a new table. Note that the number of tables grows logarithmically in the number of observations. A large α will result in a large number of tables a priori.

4.4.2 Semi-supervised Dirichlet Process Model

Using the CRP representation of the Dirichlet process, the classification problem can be formulated as follows: think of each epidemic curve as a customer. The reference type for each disease cluster represents the tables. Initially we already know that there are at most (k+1) tables at each time point since a new epidemic can either be classified to a pre-existing epidemic cluster or to a new one. Using data in the epidemic library, we assign epidemic curves from each of the epidemic clusters to the tables. Each table sits members from the same epidemic cluster. Given surveillance data for a new epidemic, the posterior probability of belonging to each of the tables is calculated. The epidemic curve is assigned either to an existing table or to a new table or cluster.

Initially, we developed a DP model for each of the three previously selected parametric distributions: normal, Poisson and negative binomial. The parameters for each of the models were estimated using Gibbs sampling, which is a Markov Chain Monte Carlo (MCMC) [30] procedure. However, we decided to use the normal model for several reasons. (i) All model fits were reasonably close. (ii) Parameters of the normal model were easy to interpret. (iii) The model fit well to the data and had less variability.

At the start of the normal DP procedure, we estimate the parameters for each of the clusters using a slice sampling method. Slice sampling is also an MCMC method which enables

random sampling from probability distributions [97]. The parameters are uniformly sampled from the area under the density function. See Neal [97] for additional information. This procedure is equivalent to parameter estimation in a nonlinear regression framework. The nonlinear regression model relating the daily infected counts $\boldsymbol{y}$ to the time vector $\boldsymbol{x}$ is given by:

$$y = f(\theta, x) + \varepsilon \tag{4}$$

θ is the vector of parameters and ε is the random error. Here $f(\theta, \boldsymbol{x})$ is a normal distribution with parameters $\theta = \phi, \mu, \sigma$ given by:

$$f(\theta, x) = \phi e^{\sum_{j=1}^{t} \frac{-(x_j - \mu)^2}{2\sigma^2}} \tag{5}$$

ϕ is the scale parameter, μ is the mean representing the peak day and σ^2 is the variance. Under the reference prior for $p(\gamma) \propto \frac{1}{\gamma}$ and $p(\mu) \propto 1$, the posterior distribution from which the slice sampler draws samples of the parameters is given by:

$$p(\theta, \gamma) = \gamma^{(N*t-2)/2} e^{-(\gamma/2) \sum_{i=1}^{N} \sum_{j=1}^{t} (y_j - f(\theta, x_j))^2} \tag{6}$$

N is the number of epidemic curves in each cluster, ϕ is the residual variance and t represents the day on which the epidemic is predicted.

The procedure is implemented in four steps. (i) Assign epidemics in the library to clusters based on disease model parameters. (ii) For each day j, infer normal model parameters for each cluster using the slice sampling procedure. (iii) Given a new curve on day j, calculate the posterior predictive probability of the curve belonging to each of the clusters in the library. (iv) Assign curve to one of the clusters in the library or create a new cluster. We perform the prediction procedure independently for each of the three hundred epidemics in our test set.

4.5 Results

The results are organized into two sections. In section 4.5.1, we discuss the results from classifying epidemic curves similar to those in the epidemic library. The simulated epidemics

are divided into two sets: one set represents data in the epidemic library and the other is used as a test set. The classification process is fully supervised. We compare the accuracy of prediction to that of random forest. The methodology of random forest; a tree based classification method is summarized in the Appendix. Random forest has been shown to be efficient and accurate in the prediction of epidemic curves using supervised classification [103]. We also discuss the performance of the Dirichlet process model in predicting the timing to the peak. In section 4.5.2, we present the results from classifying epidemics different from those in our library.

4.5.1 Accuracy of Dirichlet Process (DP) Model

We perform the classification on ten day intervals starting from day 10. The test set contains a hundred curves from each epidemic cluster: strong, severe and catastrophic. In Figure 4.5, we present the cumulative accuracy of identifying epidemics from all clusters. We discuss the accuracy of prediction from day 10 to 300; the epidemic duration.

Per Figure 4.5, the accuracy of both methods improves with additional data. However, random forest performs better than the DP model; reaching 96% accuracy by day 300. The performance of random forest is likely due to its classification scheme. In this study, each epidemic curve is classified 1000 times. The final prediction is based on a voting process [70]. The epidemic cluster to which the new epidemic is classified the majority of the times is assigned the epidemic curve. Random forest also requires a shorter computation time compared to the DP model. This is because rather than considering every data point in the library and test set, the classification is based on a sample of size $\sqrt{j}$ from each curve where j is the day of prediction.

However, the methodology of random forest fails to take into account the shape and temporal structure of the epidemic curve, which in reality defines a time series. This implies that the data on day 20 can be moved to day five and the performance of random forest will not be affected. In addition, classification algorithms such as random forest can suffer from the curse of dimensionality; the accuracy of the method depreciates as the dimensionality of the data increases. However this is not an issue in this study. In certain instances, the accuracy of random forest in identifying severe epidemics drops for a few days. We observe this instability between days ten and fifty for severe epidemics. The appropriate explanation of this instability could lie in the shape of the epidemic curves. As seen in Figure 4.3, epidemics from all clusters tend to overlap till about day thirty. Before day thirty, most of the accuracy in random forest can be attributed to chance. For example on day 20, 49% of severe epidemics are correctly classified, 34% are incorrectly classified as catastrophic, while 17% are classified as strong. However, around day thirty when the strong epidemics are more distinguishable, similar to the DP model, most of the misclassification tends to occur between severe and catastrophic epidemics. For instance on day forty, only 6% of severe epidemics are misidentified as strong epidemics.

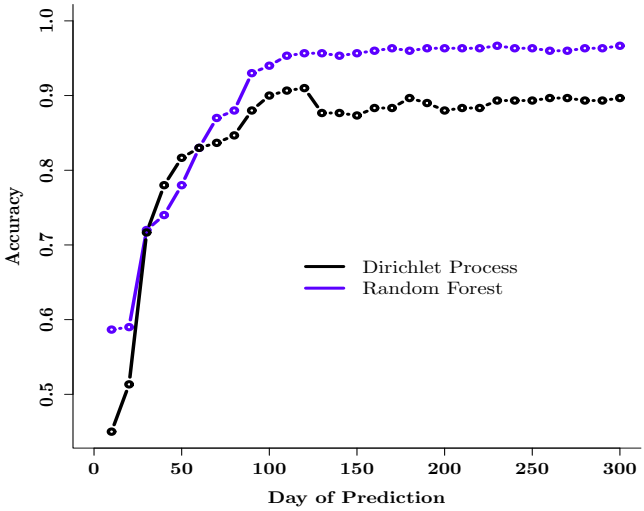


Figure 4.5: The accuracy of predicting the epidemic curves in the test set. The accuracy is defined as the proportion of correctly predicted epidemic curves.

Although random forest achieves a higher accuracy in this study, the DP model provides more details regarding the structure of epidemic curve. The DP model does not only capture the trend of the epidemic but also estimates the day on which the epidemic will peak. We discuss this in Section 4.5.1. Furthermore, the DP model can also identify epidemics different from those in our library (see Section 4.5.2). Similar to random forest, some of the inconsistencies in the accuracy of the DP model could be attributed to the shape of the epidemic curves. Also, the stochasticity inherent in MCMC methods could influence the day to day predictions. Lastly, the normal distribution might not fully capture the noise observed in the tails of the epidemic curves. However, since there are benefits to using both methods in prediction, combining the two approaches might be beneficial.

Overall, both methods perform well in identifying epidemics similar to those in our library. The results are more reliable and consistent during or after the exponential growth phase of the epidemics. This agrees with other published studies which indicate that the accuracy of forecasting methods tend to be sensitive to the time point at which forecasting occurs [65, 101, 105]. This is especially true at the early stages of the epidemic .

Prediction of Peaking Time

As previously stated the mean μ of the normal distribution used in modeling the epidemic curves represent the day on which the epidemics peak. The mean peaking day for strong epidemics is day 165. The catastrophic epidemics on average tend to peak on day 93. The severe epidemics have a mean peak day of 82. We calculate the 95% credible interval around

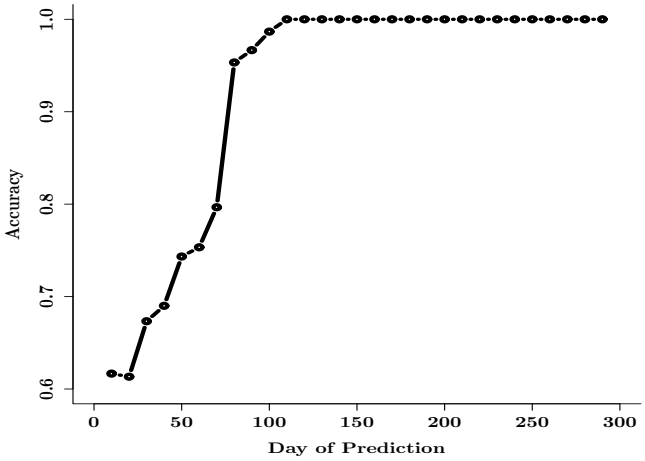


Figure 4.6: The accuracy of predicting novel epidemic curves.

μ to gauge when the model correctly predicts the expected peak day. For strong epidemics, the mean peak day is correctly identified within the 95% credible interval [163 - 166] on about day 120, which is at least six weeks before the mean peak is observed. On the contrary, for catastrophic epidemics, the mean peak falls within the 95% credible interval [92–98] on day 90, which is three days from the average peak day. However, for severe epidemics, the peak is identified much earlier on day 60, which is approximately three weeks from the peak day. The 95% credible interval is from days [72 - 87]. Note that accurate prediction of the peak day differs by epidemic.

4.5.2 Identification of Novel Epidemics

We test the method's accuracy in identifying epidemics different from those in our library. We reclassify the epidemic curves in our library as novel outbreaks. First, we remove the cluster from the library. Using the same procedure described in the methods section, we predict the full epidemic curves based on the partial curve. Since the information for that cluster no longer exist in the library, the model should identify the epidemics as novel outbreaks.

The number of new clusters created by the DP model is dependent on the choice of α. To select an appropriate value for α, we perform a sensitivity analysis by perturbing the value of α and measuring the accuracy of prediction. We set α at 0.001 after comparing the error rates of values above and below.

The accuracy based on identifying 300 novel epidemics is shown in Figure 4.6. Note that the accuracy of the method improves over time. A significant increase in accuracy is observed between days 50 and 100. This increase could be due to the peaking of the severe and catastrophic epidemics between days 80 and 90. All predictions made after day 110 are at

least 95% accurate.

The accuracy of identifying novel epidemics is influenced by the parameters and how distinct the epidemic is to those in the library. Particularly, the peaking time of each epidemic relative to the others appear to influence the accuracy of prediction.

4.6 Discussion

In this study we present a Dirichlet process model for predicting the epidemic curve. The focus of this initial study is to establish the accuracy and usefulness of the method. We focus on identifying outbreaks with high transmissibilities since these are likely to result in epidemics or pandemics with high mortality rates. The method achieves over 80% accuracy by day 80, which is about when some of the severe epidemics peak. In addition, we can also predict the time to the peak before the peak and identify novel epidemics with a high accuracy.

We acknowledge that there are several limitations to using the proposed model. For instance, using a parametric model to describe the epidemic curve might not be the most suitable option since the shape of the epidemic curve easily varies from one epidemic to another. In addition, more than one peak can be observed during an epidemic due to a second wave of infections. This could be triggered by certain changes in the environment such as the opening of schools. However, this study provides an initial step in this prediction process and appears sufficient for predicting epidemic curves with a single peak.

During a pandemic, public health official will profit most from a method which provides timely and accurate predictions. The proposed approach like most MCMC methods is timely and computationally expensive. Additional analysis is required to improve the accuracy and efficiency of the method. Most of the time cost in this procedure is encountered in the slice sampling step. Slice sampling is simple to implement and effective. However, the cost of the procedure can be a drawback. Several modifications have been suggested that employ computational resources by making the procedure parallel, while also improving the methodology of the sampling scheme [123]. In future studies, we will explore ways to make the method more applicable to real-time forecasting of the epidemic curve.

The accuracy of the predictions in this study depend both on good parameter estimates and the shape of the epidemic curve on the day on which predictions are made. However, since the proposed method is relatively simple, additional details can be added to improve efficiency, capture the time component and update epidemiological details on a current epidemic. Finally, the proposed method shows promise and could be easily improved for better performance.

Appendix C

Chapter 4: Appendix

C.0.1 Computational Epidemiology Model

A detailed description of the individual-based model (ABM) used in simulating the epidemics used in this study is discussed in [17]. The ABM belongs to a class of models called network based epidemiology models that uses a representation of the population that includes each individual and their minute-by-minute movements. Their interactions with other agents are used to generate a dynamic social network. These networks are then in turn used to simulate epidemics and study the effects of changes in individual behavior and public policy on the propagation of an epidemic [8]. Although neither changes in individual behavior nor public policy are directly explored in this study, it is extremely easy in these models to change individual behaviors, like keeping children home from school or in general limiting the number of non-essential activities of specific members of the population. However the purpose of this study is to assess the performance of classification techniques given partial epidemic curves. When the full procedure is implemented in the future, simulations will include epidemic curves that have these characteristics.

The creation of the individual-based epidemic model used in this study entails two major steps, the first of which consists of the creation of a social contact network from a state-of-the-art behavioral model. This involves creating synthetic populations and time varying social networks. Synthetic individuals and households, located in specified geographical regions (such as Seattle), each with a set of demographic variables are created using an iterative proportional fit to joint demographic distributions from the 2000 US census data provided in SF3 and PUMA (Public Use Microdata Area) files [7]. For example, a list of demographic information such as household income, family size, age, education etc. are available for each of the individuals in the Seattle region.

The synthetic populations are created to produce realistic features and demographics while preserving the confidentiality of the original data sets. An edge represents an individual in the synthetic population. Each node is placed in a household with other synthetic individuals

Table C.1: Models and Modeling Approaches used in ABM

Models	**References**
Urban Population Mobility Models	[10], [20], [125], and [124]
Natural Disease History	[3], [44], [67], [73], and [93]
Transmission Models	[67], [73], and [93]
Social Network Models	[47], [67], [100]
Types of Interventions	[50], [51], [66], and [67]

and each household is geographically located such that a census aggregated to the block level of the synthetic population would be statistically identical to the real census data [14]. Additional information can be found in [14], [116] and [117].

Next, each individual in the synthetic household is allotted activities by time of day based on several thousand responses to an activity or time-use survey for a specific region. The National Household Transportation Survey was used in creating the activity templates assigned to each household. The time-use or activity survey is expected to vary by region given factors such as the geographical location and age composition of the population. Presently, this modeling approach is considered the *de facto* standard in transportation science and is called activity based travel demand models [8]. See [19] and [20] for additional information.

Using a decision tree based on demographic information (such as number of people in a household, number of children etc.), each household in the synthetic population is matched to a survey household. Each activity for each synthetic person is then assigned an appropriate real location based on a gravity model and land-use data [14]. The addresses of locations are obtained from Dun and Bradstreet's Strategic Database Marketing Records. The activities for each household are assigned to actual locations based on "the distance from the previous activity and its *attractiveness* a measure of how likely that the activity happens there - number of employees, school enrollment, square feet of retail shopping etc." [46].

In addition to specific assignment of activities, the time at which each activity starts and ends is also included. This leads to each individual in each household having a minute-by-minute schedule for each day. Synthetic individuals in the population interact with each other based on their minute-by-minute schedule to produce realistic contact graphs where vertices represent individuals and edges represent contacts between individuals [8]. Individuals mimic the behaviors of real people by participating in everyday activities such as eating, socializing, shopping etc. and multiple edges can be used between each person and the locations representing their frequency of visits. The modeling approaches used in the ABM as presented in [8] are given in Table 3.

Next, a computational model is developed to represent disease within individuals and the transmission between individuals in the synthetic population. The transition from one disease state (susceptible, exposed, infectious and removed) to another is probabilistic and timed (e.g. it may be represented by the distribution of the infectious period). The transition between states can also depend on the attributes of the people (e.g. age, health status etc.) and the type of contact (e.g casual, intimate etc.). For the disease model used in this study, the probability that an infectious person i infects a susceptible person j is given by:

$$Pr(\text{person } i \text{ infects person } j) = 1 - exp(-f(S, T_{i,j})) \tag{C.1}$$

where $f(S, T_{i,j})$ is a nonnegative monotone increasing function of S, the severity of the disease (called transmissibility) and $T_{i,j}$, the contact time between persons i and j. The disease model is combined with the information in the network to study an infectious disease. At each time step of a simulation, each of the nodes is either: susceptible, exposed, infectious, and removed. Time is divided into units based on days and the state of each individual is noted at the start of each day.

To run a simulation experiment, a population (contact network), characteristics of a disease and initial conditions (such as duration) are specified. In this study, the social network studied was based on Seattle and surrounding metropolitan regions. The disease characteristics were based on the influenza epidemics. For each simulated epidemic, several realizations of the stochastic process of disease propagation are computed. Intervention options such as vaccination, antiviral and social distancing can be applied during the epidemic to control its propagation. Each simulation is seeded with a randomly selected set of initially infected individuals. An epidemic curve, the vulnerability of different individuals in the network, epidemic size, number of new exposures on each day etc. can be explored at the end of each simulation of a disease epidemic.

Several studies have been implemented to validate specific components of the model and the general approach. See [14], [47], and [67] for structural validity of these models.

C.0.2 Random Forest

Bagging which stands for bootstrap aggregating is a method for decreasing the variance of an estimated prediction function by combining several predictors [22, 70]. Random Forest is an extension of bagging. It involves growing several de-correlated trees, which are then averaged [70]. Trees in a random forest are identically distributed and have the same expectation.

The random forest algorithm for classification as stated in [70] follows:

1. For b = 1 to B:
(a) Draw a bootstrap sample Z* of size N from the training data

(b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached

i. Select m variables at random from the p variables

ii. Pick the best variable/split-point among the m

iii. Split the node into two daughter nodes

2. Output the ensemble of trees $(T_b)_1{}^B$

To make a prediction at a new point x: Let $C_b(x)$ be the class prediction of the b_{th} random-forest tree. Then $C^B{}_{rf}(x)$ = majority vote $(C_b(x))^B{}_1$

There are several professed advantages to random forests. Random forest runs efficiently on large databases in a short amount of time, provides estimates of important variables in classification and can be used in unsupervised classification [23]. Random forest also calculates an out-of-bag (oob) error rate based on its out-of-bag samples feature which involves constructing random forest predictor for an observation "by averaging only those trees corresponding to bootstrap samples in which the observation did not appear" [70].

The random forest technique was modeled using the randomForest package in R [81]. In order to improve the classification rate, for each model fitted to the training data, the number of variables randomly sampled from candidates at each split was set to a value which produced the minimum error rate with 500 trees grown.

Chapter 5

A Simulation Optimization Approach to Epidemic Forecasting

Elaine Nsoesie [1], Richard Beckman [1], Kalyani Nagaraj[1], Sara Shashaani[1] and Madhav Marathe[1,2]

[1] Network Dynamics and Simulation Science Laboratory, Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, Virginia, USA

[2] Computer Science Department, Virginia Tech, Blacksburg, Virginia, USA

Abstract

We introduce a simulation optimization approach for real-time forecasting of the epidemic curve. This study represents the final step of a project aimed at using a combination of simulation, classification, statistical and optimization techniques to predict the epidemic curve and infer underlying model parameters for an ongoing outbreak.

An individual-based model is combined with the Nelder-Mead simplex method to estimate model parameters and forecast the epidemic curve. The method is used to forecast epidemics simulated over synthetic social networks representing Seattle and surrounding metropolitan regions, and Montgomery County in Virginia. Furthermore, using estimated influenza incidence data from September 2009, we forecast the baseline epidemic curve for the 2009 H1N1(A) pandemic in Los Angeles.

The results suggest the following: (i) combining the proposed approach with good monitoring systems is likely to result in timely and accurate predictions, (ii) as expected, predictions improve as data is sequentially updated and (iii) retrospective predictions made for the 2009 pandemic agree with published reports for Los Angeles county.

5.1 Introduction

Influenza continues to be one of the most important human infectious diseases; responsible for thousands of deaths in the United States per year. In April of 2009, a novel influenza A virus emerged in Mexico and the United States. Although the 2009 H1N1 influenza pandemic was relatively milder than expected, the emergence of the novel virus reinforced the need to improve tools for analyzing surveillance data and forecasting for decision making during a pandemic [90]. Mathematical and computational models are used as tools to aid pandemic planning. Specifically, individual-based epidemiology models are useful in evaluating the possible effectiveness and economic impact of different response strategies [8, 37, 47, 67, 119].

This study extends the application of the individual-based epidemiology model to forecasting of the epidemic infection curve (hereafter referred to as the epidemic curve). The epidemic curve is defined as the daily or weekly number of cases observed for the duration of the epidemic [130]. We seek to predict the time at which the epidemic peaks, the number of infected individuals at the peak and the cumulative infected counts. These measures provide a summary of the epidemic curve and are important to public health officials. An accurate prediction of these measures at a regional level would enable local public health officials to evaluate intervention strategies and make educated decisions during an influenza epidemic [38, 46, 109].

Real-time prediction of the epidemic curve requires a combination of good monitoring systems and adequate assumptions about the disease model parameters [38, 105]. Conventional methods for monitoring influenza-like illness (ILI) data and acute respiratory tract infections from general practices, family doctor and government clinics are being used in many countries [27, 39, 49, 53, 56, 78, 105, 122]. These methods were also used to monitor influenza activity during the 2009 H1N1 outbreak [5, 77, 85].

Several methods have been proposed for real-time modeling and forecasting of epidemic dynamics [38, 65, 101, 104, 105]. Hall et al. [65] proposed using a deterministic compartmental model to estimate epidemic dynamics. Their method was used to retrospectively predict the amplitude and durations of three pre-2006 influenza pandemic events in England and Wales. They used regression techniques to fit a time-series disease incidence curve obtained from a traditional differential equation epidemiology model to the mortality and influenza-like illness (ILI) data for the three epidemics. This technique required estimation of nine parameters, including the reproduction number. The model also assumed knowledge of the disease natural history (the incubation and infectious periods of the disease) from detailed epidemiological studies in the early stages of the pandemic.

Hsieh and Cheng [75] demonstrated the use of a variation of a single-equation Richards model to estimate outbreak severity. Their method used a power-law logistic equation to estimate parameters based on the epidemic curve. The method was applied to the multiphase 2003 severe acute respiratory syndrome (SARS) outbreak in Toronto. Hsieh [76] also employed their model to estimate epidemic curve parameters for the 2009 H1N1 influenza pandemic

in six countries in the southern hemisphere.

Similarly, Nishiura presented a discrete time stochastic model for forecasting the 2009 H1N1 pandemic [101]. To retrospectively forecast the 2009 H1N1 pandemic in Japan, a likelihood-based approach was used in parameter estimation. Ohkusa et al. [104] also used a simple SIR model for prediction during the pandemic.

In contrast, Ong et al. [105] described a real-time system to both monitor and forecast different epidemic outcome measures in Singapore during the 2009 H1N1 pandemic. The surveillance system collected data on ILI instances from twenty three participating general practice and family doctor clinics in Singapore. Since H1N1 had low hospitalization and mortality rates, the study did not use hospital and fatality data. The forecasting model developed was a stochastic compartmental model with particle filtering for real-time epidemic incidence prediction. ILI data collected at general practice and family doctor clinics in Singapore was refitted each day to provide sequential updates on predictions.

All previously discussed approaches to prediction use either a variation of the differential equation epidemic model or a region-dependent disease transmission model, or both, making it difficult to model for changes in human mobility and interaction patterns . Chao et al. [38] used a stochastic epidemic simulation model, which includes descriptions of interactions between individuals (with demographic information) at different mixing groups (schools, homes, work etc.). Predictions of the characteristics of the 2009 influenza pandemic in addition to the potential effects of interventions were made for Los Angeles county using data for the region.

The stochastic epidemic simulation model used by Chao et al [38] is similar to that used in this research. Both models seek to represent individuals and interactions between individuals. However, there are differences in the data sources, the method of constructing the networks and some of the assumptions in the disease model. There are also differences in the manner in which the models are used in prediction. In this study we present an approach which combines an individual-based model and an optimization technique to recursively estimate model parameters and forecast the epidemic curve as data is sequentially updated during an epidemic. This approach has not been studied previously.

5.1.1 Approach

The problem can be formerly defined as follows: given surveillance data $x_1, \ldots, x_t$ for a current epidemic, we seek to predict $x_{t+1}, \ldots, x_n$. x_j represents the number of new cases on day j, t indicates the day on which prediction is made and n is the expected duration of the epidemic.

This study represents the final step of a project aimed at using a combination of simulation, classification, statistical and optimization techniques to predict the epidemic curve and infer underlying disease model parameters (Figure 5.1). First, we build a library of past and simulated epidemics. Simulated epidemics are replicated several times. Using a classification

approach, we propose a parameter set at time t based on available data up to time t. We use random forest; a supervised classification method to assign the new epidemic to an existing cluster in our library. The efficacy of random forest was illustrated in [103]. If the match suggested by random forest is deemed suitable, then the parameters of the epidemic in the library are used in modeling the new outbreak. On the contrary, if none of the epidemics in the library is deemed a good match, then we recursively apply a combination of simulation and optimization methods to propose new parameters.

In this study, we focus on the event that the epidemic cannot be classified to any of the cases in the library (Figure 5.1). During an epidemic, ILI or other forms of surveillance data can be obtained from sources such as the United States Centers for Disease Control and Prevention (CDC), FluNet, Distribute Project, etc. Given the availability of surveillance data, we describe the approach in five steps. (i) We define initial parameters for our optimization algorithm by sampling from the epidemics in the library. (ii) Using the individual-based model, we simulate an epidemic using each of the initial parameter sets. (iii) We input the simulated epidemics into our optimization method, which then proposes new parameters based on whether our objective function is minimized. (iv) We repeat the simulation optimization process till the algorithm converges. (v) The optimal parameters given at convergence are added to the library and also used to forecast the epidemic curve. The optimal parameters at time t are used to initialize the process at time $t+1$.

We use the Nelder-Mead simplex method for optimization. The method is discussed in a later section and detailed in the Appendix. The method proposes optimal parameters that minimize the objective function representing the difference between the daily newly infected counts of the partial surveillance epidemic curve and the simulated instances. The procedure is repeated each day for the duration of the epidemic. Nonetheless, predictions made before the peak of the epidemic are most preferable. Upon identification of a parameter set for modeling the outbreak, the individual-based model is used to investigate the effectiveness of various intervention measures and the effects of changes in individual behavior during the epidemic [17, 45]. However control measures are not presented in this study. In this preliminary study, we present predictions made under the baseline scenario and focus on epidemics with a single peak. Nevertheless, the methods can be applied to study situations in which a second peak (wave) is observed during an epidemic.

5.1.2 Disease Model and Parameters

The three parameters estimated in this study are the disease transmissibility, incubation and infectious period distributions (see Table 5.1 for definitions). The transmissibility of a disease is typically represented using measures such as the reproduction number or the household secondary attack rate [108, 135]. The attack rate is the cumulative infection incidence observed within a population over the span of an epidemic. If the time of infection is known, the incubation duration can be derived. The infectiousness typically differs for

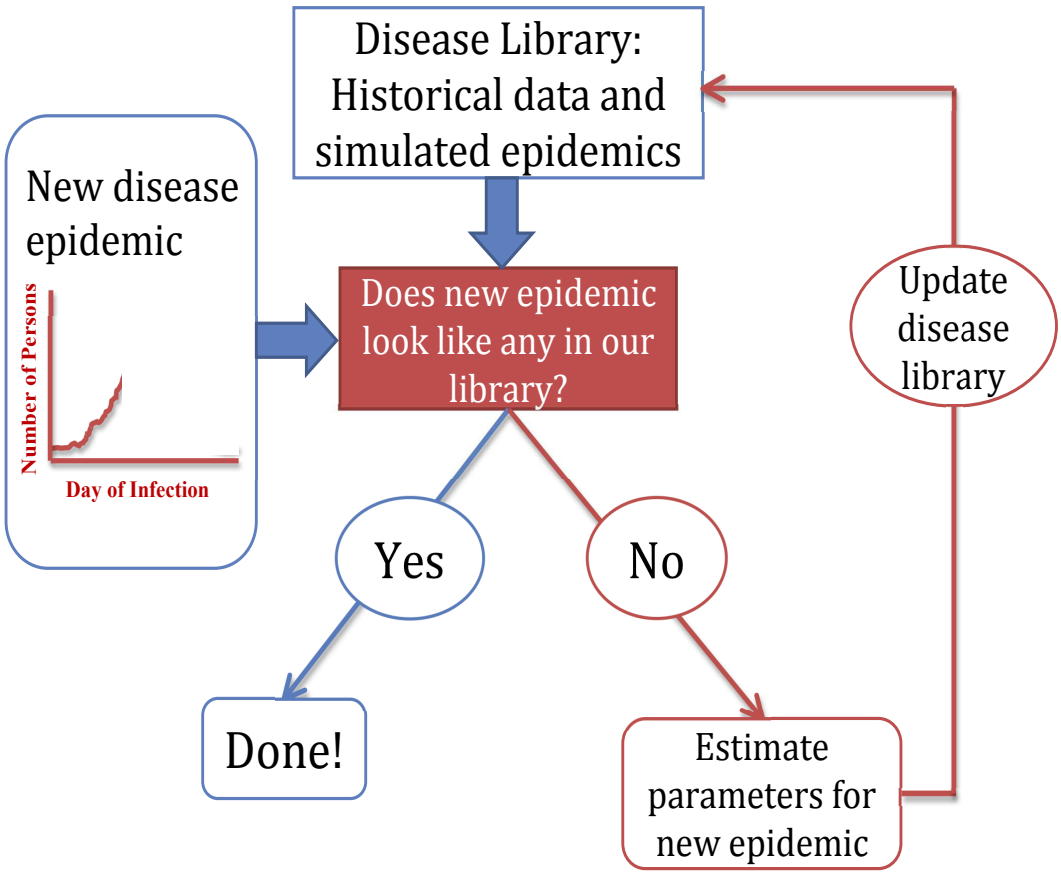


Figure 5.1: Summary of methodology. We develop a library of past and simulated epidemics. Given surveillance data for a current epidemic, we compare the partial surveillance epidemic curve to those in the library. The novel epidemic is either assigned to a case in the library or identified as being different from those in the library. If the epidemic is different from those in the library, we estimate the model parameters, forecast the epidemic curve and update the library.

different individuals due to factors such as age, symptoms and health state [47]. The incubation and infectious period parameters are therefore represented using discrete probability distributions.

Parameters	**Definitions**	**Example**
Transmissibility	The rate at which disease propagates through the population	4.6E-5 per sec/ unit of contact time
Incubation Period	Duration between infection and onset of symptoms	0:0.0 1:0.3 2:0.5 3:0.2
Infectious Period	Period during which infected persons can shed the virus	2:0.0 3:0.3 4:0.4 5:0.2 6:0.1

Table 5.1: The incubation (infectious) period is defined as follows: $k : p_k$ where k is the duration and p_k is the probability that an infected (infectious) individual will have an incubation (infectious) period of k days. The disease transmissibility is given as the probability of infection per unit of contact time between a susceptible and infectious individual in the network.

These parameters are part of the disease model. The individual-based model consists of a dynamic social network and a disease model as discussed in a later section. In order to estimate disease model parameters, we make the following assumptions: (i) the **S**usceptible, **E**xposed, **I**nfectious and **R**ecovered (SEIR) model is sufficient to describe within host disease progression and between host disease transmission. (ii) The possible durations of the incubation and infectious periods are fixed as shown in Table 5.1. We therefore focus on estimating the probabilities of observing each incubation (infectious) duration in the network. (iii) The network is assumed to remain unchanged during the course of the epidemic implying new individuals do not enter or leave the synthetic population.

The proposed method is tested on simulated surveillance epidemic data. The epidemics are simulated over synthetic social networks representing Seattle and surrounding metropolitan regions, and Montgomery County (MC) in Virginia. Studying simulated epidemics for regions with demographic and rural-urban differences enables a thorough illustration of the methods' performance. In addition, using CDC estimated influenza incidence during the 2009 H1N1(A) pandemic, we predict the epidemic curve for Los Angeles and surrounding metropolitan regions under the scenario that no vaccinations were introduced to control the spread of the disease. The aims of this study are therefore to: (i) forecast the epidemic curve by predicting the time to peak, peak infected counts and total infected counts, (ii) compare predictions made across different social networks, and (iii) apply the method to a real epidemic by retrospectively predicting the baseline epidemic curve for the 2009 H1N1 pandemic.

5.2 Methods

There are two parts to the simulation optimization procedure: the individual-based model and the Nelder-Mead simplex method. The Nelder-Mead simplex algorithm is used to propose new parameters. The parameters are then used in simulating epidemics using the individual-based model. This process is repeated till the algorithm converges as discussed in the proceeding section.

5.2.1 Modified Nelder-Mead Simplex Method

We use the Nelder-Mead simplex method in the optimization process. The Nelder-Mead method was selected after comparing its performance (accuracy, computational time and cost) to Simulated Annealing [88] and the classical stochastic root finding approach in Robbins and Monro [112]. The method serves as an illustration that similar optimization techniques can be used in combination with simulations to solve the problem of forecasting the epidemic curve. The Nelder-Mead method is also easy to implement and modify. To enable readability of this paper, we present a summary of the method in this section and additional details in the Appendix.

Nelder-Mead simplex is a direct search method that attempts to minimize function of real variables using only function evaluations without any derivatives. The minimized objective function representing differences in the daily infected counts is given by:

$$SSQ = \sum_{j=1}^{t} (y_j(\theta^{true}) - z_j(\theta))^2 \tag{5.1}$$

$z_j(\theta)$ and $y_j(\theta^{true})$ represent the simulated and true epidemic curves respectively. j indicates a single day and t is the day on which the epidemic curve is predicted. θ represents the vector of parameters. Each parameter set contains a disease transmissibility value, an incubation period and infectious period distribution. The range of possible days for the incubation and infectious period distributions are fixed as shown in Table 5.1. These ranges are based on parameters used in published studies for seasonal influenza [67] and the serial interval of the 2009 pandemic [31, 135].

The algorithm proposes new values for the transmissibility, in addition to the probability values $p_k's \in \theta$ for the incubation and infectious period distributions (Table 5.1). The probabilities must be non-negative and sum to one independently for the incubation and infectious periods. We therefore modify the Nelder-Mead algorithm by introducing conditions which reinforce this requirement. See the Appendix for more information on the modified algorithm.

Each parameter set and its relative SSQ value corresponds to a vertex in a simplex. During the optimization process, the Nelder-Mead algorithm proceeds through recursive updates of

the simplex vertices via a series of four basic operations: reflection, expansion, contraction and shrinkage. At each step of the Nelder-Mead algorithm, one of the formerly mentioned operations is used to generate a new parameter set that replaces a vertex in the simplex representing the parameter set with the worst SSQ value. After each update, epidemics are simulated using the new parameters and the objective function is evaluated. The next appropriate operation is selected based on the ranking (smallest to largest) of the new SSQ value relative to the values at the other vertices.

For a function of m variables (parameter sets), Nelder-Mead maintains $m+1$ vertices forming a polytope. Since there is a single transmissibility value, four possible incubation period durations and five possible infectious period durations (Table 5.1), we therefore need eleven initial parameter sets. An example of initial parameters used in this procedure are shown in Table D.1 in the Appendix. The dimension of the polytope always remains the same; containing $m+1$ vertices. The algorithm stops if the relative tolerance is met. The relative tolerance which represents the relative difference between the vertex with the maximum SSQ and that with the minimum SSQ is defined as:

$$RelTol = \frac{(max(SSQ) - min(SSQ))}{min(SSQ)} \tag{5.2}$$

After carefully studying the convergence of the algorithm and trying several relative tolerance values, we fix the relative tolerance at 0.5. The parameter set with the smallest SSQ at convergence is used in forecasting the epidemic curve. See the Appendix and references [16, 99] for additional details on the Nelder-Mead simplex method.

5.2.2 Individual-based Model

Individual-based network models in epidemiology have recently garnered much attention for their advantage of being able to closely mimic realistic social networks over traditional differential equation-based disease models that assume homogeneous mixing [9, 47]. The individual-based model used in the simulations was formerly described in [17]. This and similar models have been used in several published studies [8, 47, 61, 103]. Since the creation of the individual-based model is not a novel aspect of this work, we present a brief description. Additional details are presented in the Appendix.

In brief, the model is divided into two parts: a time varying social contact network and a disease model describing within host disease progression and disease transmission between individuals. The synthetic social contact networks are generated from a hierarchical composition of data-driven stochastic processes:

(1) The baseline populations are synthesized based on socio-demographic statistics from the United States Census.

(2) Mobility patterns from a nationwide household survey and land use data are used to estimate contact networks for different regions.

In addition to demographic information, each individual is assigned an activity schedule based on responses to a national travel survey. Individuals come in contact at different activity locations such as school, work, and daycare, resulting in disease transmission between infected and susceptible individuals.

To run a simulation experiment, a population (contact network), characteristics of a disease and initial conditions (such as duration) are specified. Each simulated outbreak is replicated several times to capture different realizations of the stochastic process of disease propagation through the network.

Compartmental models or other aggregated models can be used in place of the individual-based model. However, the model already exists and is available. The model enables the design and implementation of studies focused on disease spread at the subpopulation level: age groups and localized up to the county level. In addition, large populations with millions of synthetic individuals can be easily assessed. See the Appendix for additional information.

5.3 Data

As previously mentioned, we test the accuracy of the proposed method using both simulated surveillance epidemic data from an individual-based model and CDC estimated H1N1 influenza incidence counts in September 2009.

5.3.1 Synthetic Epidemic Data

To test the prediction process, we generate synthetic influenza surveillance epidemic data for Seattle and their surrounding metropolitan regions, and MC in Virginia. The synthetic populations consist of approximately 3.2 and 0.16 million individuals for Seattle and MC respectively. These regions are selected due to population and demographic differences.

Each epidemic representing a surveillance sample is simulated for 180-days or approximately 25-weeks for each of the synthetic networks. The epidemics are simulated using incubation and infectious period parameters which have been used in several published studies [46, 67, 103]. The simulated epidemics has a mean infectious period of 4-days, mean incubation period of 2-days and transmissibility (6.00E-5 per sec/contact time) significantly higher than that of seasonal influenza. Each epidemic is seeded by randomly selecting five individuals in the population to initially infect. The epidemic curve is noted at the end of each simulation.

We test our forecasting approach by predicting the epidemic curve at different time points during the epidemic. Specifically, we predict the epidemic curve on days 14, 21 and 28.

We evaluate accuracy based on the predicted timing to peak, peak infected and cumulative infected counts. In addition, Pearson correlation coefficient and root mean squared error (RMSE) are used in evaluating similarities in the temporal trend and difference between the predicted and true epidemic curve respectively.

The accuracy of the prediction process depends not only on the Nelder-Mead algorithm but also on the objective function, and uncertainty in the available surveillance data. The amount of noise or error in the data would mask the signal of the true curve, thereby increasing the difficulty of prediction.

5.3.2 2009 H1N1 Pandemic in Los Angeles

Recall that there are two parts to the simulation optimization procedure: the individual-based model and the Nelder-Mead simplex method. For each simulation, we select initial parameters for both the individual-based model and the Nelder-Mead algorithm. In order to retrospectively predict the 2009 pandemic in Los Angeles, we need to decide on the number of individuals to initially infect in the individual-based model. We apply the procedure used by Chao et al [38] to initialize our individual-based model as follows. The CDC reported an estimated 3 million individuals were infected in the United States within a six-week period spanning July to September. This implies approximately 71,000 individuals were infected per day. We assume that the number of infected individuals in the Los Angeles region was proportional to that of the United States. Since our social network has approximately 16 million synthetic individuals for Los Angeles and surrounding metropolitan regions (hereafter referred to as Los Angeles) this results in approximately 15,800 infected persons within the six week period. This therefore implies roughly 3800 infected persons per day. As in [38], infected individuals in our model can have influenza infectious periods of up to 6-days. Based on this, we seed our model with three times the daily infected number or 11,400 initially infected.

We use the parameters in Table D.1 to initialize the Nelder-Mead algorithm. Instead of an epidemic curve, we have the total number of infected persons for a six-week period. Rather than minimizing the SSQ as discussed previously, we minimize the difference between the estimated total infected counts during the six-week period and first six-weeks of the simulated instances. By setting our estimated cumulative infected counts for the six-week period at 15,800 for Los Angeles, we apply the simulation optimization procedure to find the best set of parameters that minimize the difference between the simulated and the estimated infected counts. We aim to predict the timing of the epidemic curve in the absence of vaccination and other control measures.

As with all methods, our assumptions can be criticized on several grounds. However, these assumptions appear to be sufficient for demonstrating the performance of the proposed approach. In addition, the difficulty of obtaining accurate estimates of the incidence curve of an influenza epidemic due to unreported cases limits the available data needed to test the

proposed method [62].

5.4 Results

As recently stated, the aims of this study are to: (i) forecast the epidemic curve by predicting the timing to the peak, peak infected counts and overall infected counts, (ii)compare predictions made for synthetic social networks with differences in population size and demographics and (iii) forecast the baseline epidemic curve for the 2009 H1N1 pandemic in Los Angeles.

For each day j, the epidemic curve is forecasted using data from the start of the epidemic to day j. We discuss predictions made during the first four weeks of each epidemic. As previously stated, we use the parameter set with the minimum SSQ value at convergence to forecast the epidemic curve.

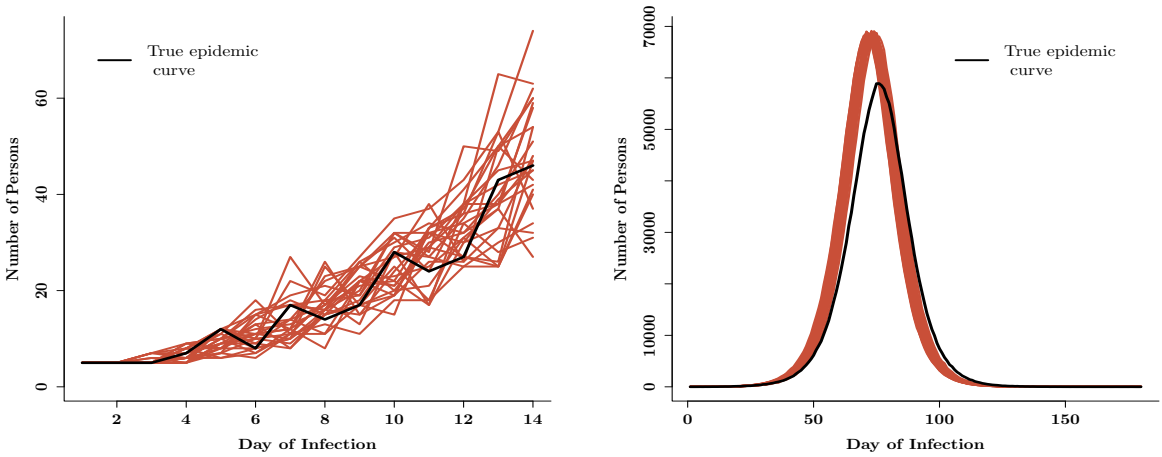
Additionally, although the same parameters are used in simulating the epidemics for each of the social networks, the shape of the epidemic curve, daily counts and magnitude of the epidemics differ. This suggests that predictions made for one region are not necessarily applicable to another. We therefore present results for each of the synthetic social networks independently. We also present baseline predictions for the 2009 H1N1(A) pandemic in Los Angeles.

5.4.1 Synthetic Epidemic Data

Seattle

The forecasted curves on days 14, 21 and 28 are given in Figure 5.3 for Seattle. The predicted epidemics are replicated 25 times. Per Figure 5.2(A), replicates of the predicted epidemic overlap with the true curve. This observation is also made on days 21 and 28. However, Figure 5.2(B) shows that the predicted curve deviates from the true epidemic curve after day 14. This would suggest that several possible parameter sets could result in epidemics similar to the true curve up to day 14. However, several deviations are possible after this day. One would expect that as the epidemic nears its peak, the possible space of parameters which could result in epidemics similar to the true curve up to time t would be minimized. On day 14, the epidemic is only two weeks old and the trend is not yet easily distinguishable.

In general, the prediction better captures the true trend and daily infected counts as the epidemic nears its peak. This is supported by a drop in the mean RMSE from 4394.63 on day 14 to 1721.26 on day 28 indicating improved similarity between the true and predicted curves. In addition, the Pearson correlation coefficient between the true and predicted curves is $\geq 92.8\%$ on day 14 and $\geq 96.1\%$ on day 28.



(A) (B)

Figure 5.2: True and predicted epidemic curves for Seattle. The black curve is the true epidemic and the colored curves are 25 replicates of the predicted epidemic. (A) Predicted and true curves shown from day 1 to 14. Note that the predicted curves cover the true data up to day 14. (B) Predicted epidemic on day 14.

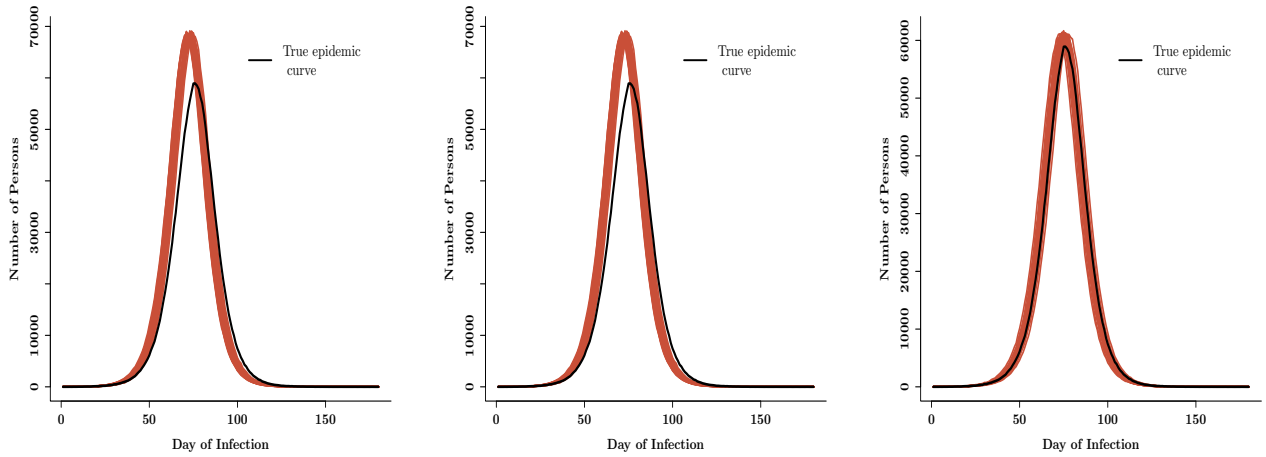
The correct time to peak is rightly forecasted within the range given in Table 5.3 on all days. However, the peak infected counts and total infected counts are overestimated. The lower range of forecasted peak infected counts on day 28 is approximately 59600, which is similar to the true value of 58912. Additionally, an estimated 1.7 million persons are infected for the duration of the epidemic, and the predicted value on day 28 is approximately 1.71 million. These observations seem to indicate that with additional data, these measures would be correctly predicted. Also note that day 28 is less than halfway to the epidemic's peak.

The results in Figure 5.3 allude to the idea that predictions tend to improve as epidemics near the peak. In addition, the accuracy of prediction tends to be sensitive to the time point at which forecasting occurs as has been noted in other studies[65, 101, 105].

MC in Virginia

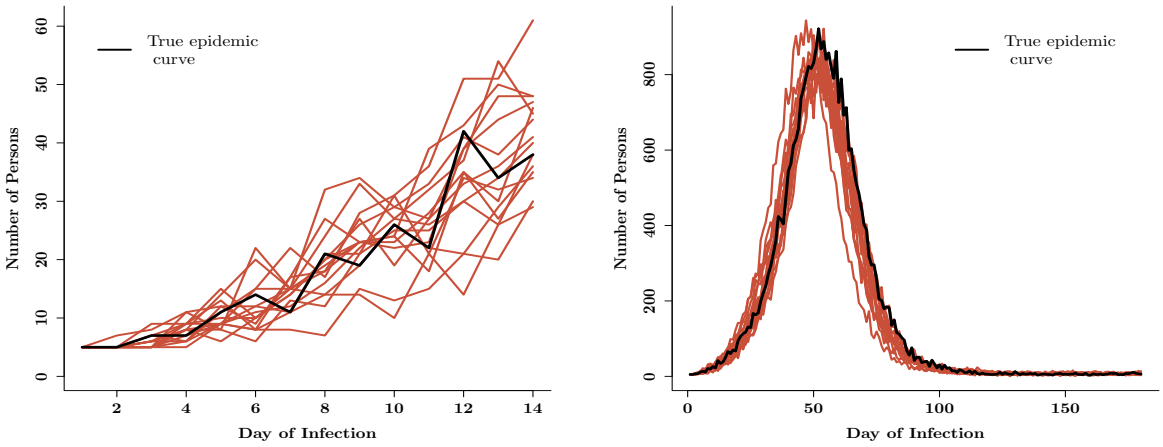
The true and predicted epidemic curves on days 14, 21 and 28 are given in Figures 5.4 and 5.5 for MC. Each predicted epidemic is replicated 25 times. The partial epidemic curves in Figures 5.4(A) show the available data on which predictions are made on day 14. Per Figure 5.4, the true epidemic curve falls within the range of predicted outcomes.

Similar to the observations for Seattle, the mean RMSE between the true and predicted curves is reduced from 42.46 to 31.48 on days 14 and 28 respectively. In addition, the Pearson correlation coefficients between the true and predicted curves also improves from values $\geq 91.0\%$ on day twenty-six to values $\geq 97.5\%$ on days 14 and 28 respectively. These outcomes agree with the assumption that predictions improve as the epidemic progresses.



(A) Day 14 (B) Day 21 (C) Day 28

Figure 5.3: True and predicted epidemic curves for Seattle. The black curve is the true epidemic. The colored curves are predictions made on days 14, 21 and 28. Predictions tend to improve as the epidemic progresses.

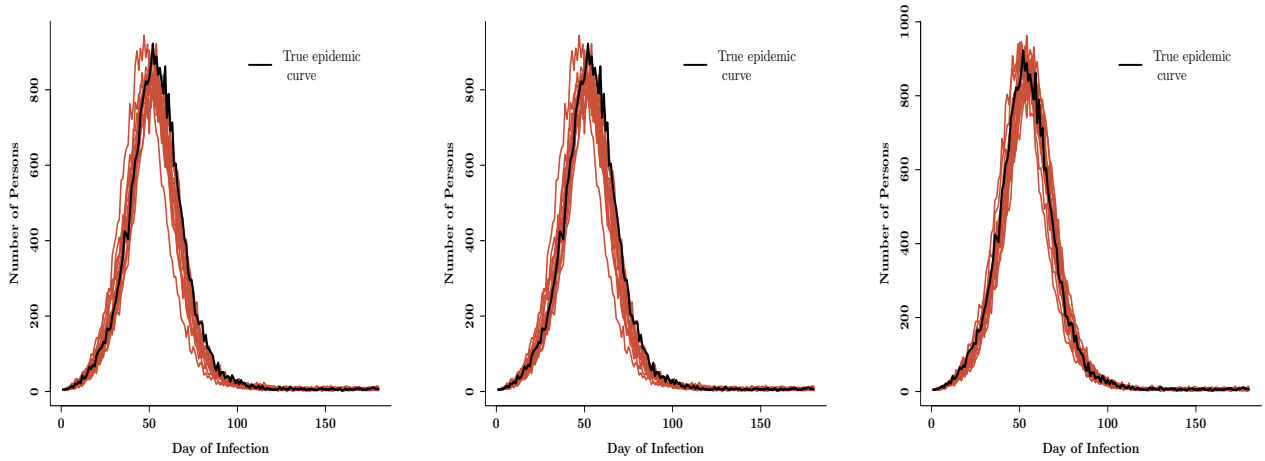


(A) (B)

Figure 5.4: True and predicted epidemic curves for MC in Virginia. The black curve is the true epidemic and the colored curves are 25 replicates of the predicted epidemic. (A) Predicted and true curves shown from day 1 to 14. Note that the predicted curves cover the true data up to day 14. (B) Prediction of epidemic on day 14.

The number of infected individuals at the peak and day of peak are correctly predicted. Peak infected counts are predicted on day 14 to fall between 814 and 944, which includes the true value 922. Furthermore, the true epidemic peaks on day 52 and epidemics predicted on day 14 peaked between days 47 and 54. Predictions made on days 21 and 28 also correctly

capture the peaking day and peak number of infected persons. However, the total infected counts is underestimated on all three days. The range of predicted cumulative number of infected persons on day 28 is approximately 30 thousands to 31 thousands. However, the true total infected is estimated at 32 thousands.



(A) Day 14 (B) Day 21 (C) Day 28

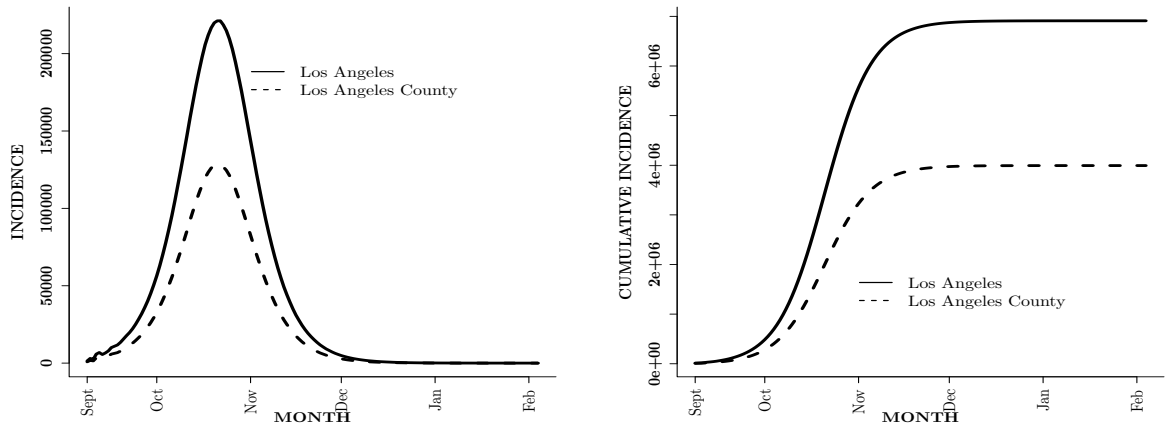
Figure 5.5: True and predicted epidemic curves for MC in Virginia. The black curve is the true epidemic. The colored curves are predictions made on days 14, 21 and 28. Predictions improve as the epidemic progresses.

The time to the peak appears to be the easiest measure to predict for both Seattle and MC. These observations suggest that by the fourth week of the epidemic, the forecasting procedure is able to correctly predict the timing to the peak and peak infected rate of the epidemic. However with additional data, the expected magnitude of the epidemic can also be predicted.

5.4.2 2009 H1N1(A) pandemic for the Los Angeles

Figure 5.6 plots the simulated daily and cumulative infection incidence for Los Angeles and Los Angeles County for the months of September to February. Los Angeles county is part of Los Angeles and surrounding metropolitan regions. The infection incidence declined towards February reflecting the usual end of the influenza season. The epidemic peak is forecasted to have occurred towards the end of October. The Los Angeles county department of public health report based on influenza-like illness data indicated that the pandemic peaked towards the end of October to early November [1]. In addition, the mean illness attack rate based on 15 replicates is predicted to be 21.6% for the entire Los Angeles region and 21.7% for Los Angeles county. The estimated illness attack rate of 21.7% for Los Angeles county is similar to the 21.4% estimated by Chao et al. [38]. The illness attack rate is estimated based on the

assumption that 67% of infected individuals become symptomatic as is expected for seasonal influenza epidemics [28, 68]. The peak infected rate is predicted at 1.36% and 1.37% for the entire Los Angeles region and Los Angeles county respectively. Recall that these predictions are made under the assumption that no vaccinations were introduced to control the spread of the epidemic.



(A) (B)

Figure 5.6: Predicted time to peak of the 2009-2010 influenza A(H1N1) epidemic for Los Angeles and surrounding metropolitan regions and Los Angeles county. (A) Simulated infection incidence of pandemic H1N1. The solid lines shows the predicted daily incidence in the Los Angeles region. The dotted lines displays the forecasted daily incidence in Los Angeles county. The months are labeled at the beginning of the month except for September since we had data available for early September. (B) Simulated cumulative infection incidence curves.

5.5 Discussion

In this study, we presented a method for forecasting the epidemic curve. Recall the objectives of this study were to: (i) predict the three measures: time to peak, peak infected counts and total infected counts, (ii) compare predictions made across different social networks and (iii) apply the method to a real epidemic by retrospectively predicting the baseline epidemic curve for the 2009 H1N1 pandemic. This study is the final step in a procedure aimed at predicting the epidemic curve using a combination of simulations, classification, statistical and optimization techniques.

The following observations were made. (i) The proposed procedure could accurately forecast the three public health measures during the first four weeks of the epidemic in most cases. The peaking time appeared to be the easiest measure to predict while the cumulative infected counts was the most difficult to predict. Given the assertion that the highest incidence of influenza infection is typically observed during the first three weeks [129], these predictions are

invaluable. (ii) There were differences in predictions made across different social networks. This suggests observations made for one social network are not necessary applicable to another and therefore reinforces the need for community-based predictions [38]. By providing predictions for a particular region, informed decisions can be made at a regional level on how to best control the epidemic especially given unavailability of vaccinations. In addition, demographic differences have been suggested to influence epidemic spread and transmissibility [95, 106]. The disease might spread differently through a population with a larger proportion of elderly compared to a population with a larger proportion of children. (iii) Lastly, the predictions made using the estimated incidence data agreed with published studies and observations made during the 2009 pandemic. In this particular situation, several of the assumptions used in initializing the procedure were based on the paper by Chao et al. [38]. These assumptions do not appear to be generalizable. However, in this case, they were sufficient for illustrating the method.

We observed similarities and differences in the predictions made across the different social networks. Apart from the obvious differences in population size, variations in demographical distributions could also influence the predicted outcomes. As shown in Table 5.2, the proportion of each age group within each population is different. The age groups are defined as follows: <5 years old are preschoolers, 5 to 18 year olds are classified as school age, 19 to 64 year olds are adults and persons older than 64 are considered seniors.

Age Groups	**Social Networks**	
	Seattle	MC in Virginia
Preschool(%)	6.8	6.1
School-age(%)	20.3	10.5
Adults(%)	63.1	74.4
Seniors(%)	9.8	9.0

Table 5.2: Distribution of population across age groups

Differences in demographics could impact disease spread across the population. Positive correlations have been found between the attack rate and the proportion of children within a population [106]. Several studies have also suggested that school children tend to impact the spread of influenza [13, 96, 111, 113]. The percentage of the population consisting of children is approximately 27.1% for Seattle compared to 16.6% for MC in Virginia. On the contrary, MC has the highest proportion of adults 74.4% compared to approximately 63.1% for Seattle. Exactly how these differences in demographics influence the disease spread and consequently the prediction process is not easily quantifiable.

Timely and accurate predictions of disease incidence are difficult to obtain during an influenza epidemic. Only a small percentage of incidence data is collected during an epidemic. Typically, influenza-like illness and symptom surveillance data are used to observe timing and other characteristics of an epidemic. Goldstein et al. [62] proposed a method for estimating incidence data from symptom surveillance data. However, due to the scarcity of the

necessary data, the method was fully illustrated only on synthetic data and only partially illustrated on real outbreak data. Methods for estimating incidence data from other forms of surveillance data would be invaluable for this procedure. More recently, search engine query data and social media data have been suggested to be used in conjunction with traditional surveillance epidemic data for estimating influenza activity [24, 60].

Several other issues arise when dealing with surveillance data. Some of these issues involve decisions on how to initialize the epidemic model, how many new cases to introduce into the population during the epidemic and how to model data affected by non-pharmaceutical interventions. Unlike the simulated epidemics where we know the initial number of infected cases, during an epidemic this information is not always available. Some amount of calibration would be needed to estimate the true incidence of disease whether influenza incidence or influenza-like illness data is used.

As already demonstrated in several published studies, a combination of good monitoring systems and modeling would improve real-time epidemic forecasting considerably [38, 90, 105]. If the collected data has few uncertainties, then the predictions would most likely have few uncertainties. The results indicate that the procedure in most cases tends to overestimate rather than underestimate. Since the number of reported cases is typically significantly less than the true incidence of the epidemic, the procedure has a good chance of estimating the true epidemic curve. These predictions suggest that using a combination of this procedure with good incidence surveillance data can result in beneficial results during the early stages of an influenza outbreak.

In addition to good monitoring systems, model assumptions can also influence predicted outcomes. One of these includes the assumption that the susceptibility is the same across age groups. This is not always the case. As observed during seasonal influenza epidemics and the 2009 H1N1 pandemic, different age groups tend to have different susceptibility to the disease. If the virulence of a new influenza strain is estimated early on then, such information can be used to improve the performance of the procedure.

Limitations in the optimization algorithm can also influence performance. In this study we use only a single optimization algorithm after comparing its performance to two other algorithms. In future studies, we would compare several algorithms to see if a single method is sufficient or whether a combination of different methods would produce better results. Also, the initial sets of parameters are crucial to the performance of the method. If initial selected parameters are similar to the true parameter, then the time to convergence would likely be shorter than if the initial parameters were farther from the true parameters. Furthermore, a study comparing the effects of different objective functions would be beneficial.

The results in this study are meant to serve as an illustration that similar combination of methods can be used for prediction. The results are promising and indicate this approach would perform well given the right model assumptions and good surveillance data. Since no approaches have proved infallible, this would be a reasonable method to consider for real-time prediction of the influenza epidemic curve.

Social Networks		Day	Simulated Epidemic Data: Time to Peak True	Time to Peak Median	Time to Peak Range	Peak Infected Count True	Peak Infected Count Median	Peak Infected Count Range	Overall Infected Count True	Overall Infected Count Median	Overall Infected Count Range
Seattle	True		76			59k			1.7M		
	Predicted	14		73	71–76		68k	67k–69k		1.8M	1.81M–1.83M
	Predicted	21		73	71–76		68k	67k–69k		1.8M	1.81M–1.83M
	Predicted	28		75	73–77		61k	60k–62k		1.7M	1.71M–1.71M
MC	True		52			922			32k		
	Predicted	14		51	47–54		852	814–944		29k	29k–30k
	Predicted	21		51	47–54		852	814–944		29k	29k–30k
	Predicted	28		54	49–60		913	804–963		31k	30k–31k

Table 5.3: Summary of prediction by social network and day. Each predicted epidemic was replicated fifty times. The reported values are based on a summary of six repetitions of the experiment.

Appendix D

Chapter 5: Appendix

D.0.1 Modified Nelder-Mead Simplex Method

Nelder-Mead simplex is a direct search method that attempts to minimize function of real variables using only function evaluations without any derivatives. The Nelder-Mead algorithm proceeds through recursive updates of the simplex vertices via a series of four basic operations: reflection, expansion, contraction and shrinkage. For a function of m variables, Nelder-Mead maintains $m+1$ vertices forming a polytope. The m variables represent m parameters sets.

At every step of the algorithm, one of the above-mentioned operations is used to generate a new parameter set that replaces a vertex in the simplex representing parameters with the worst objective value. If no better parameter set is found, all vertices are drawn halfway toward the current best vertex. This requires new simulation runs to evaluate the objective function for the updated parameter set. The dimension of the polytope always remains the same; containing $m+1$ vertices. The algorithm converges if the relative difference between the best and worst function values in the new polytope is less than the defined relative tolerance. The method is illustrated in Algorithm 2. Based on [16, 99] best values for the Nelder-Mead parameters are $\rho = 1$ (reflection coefficient), $x_i = 2$ (expansion coefficient), $\gamma = 0.5$ (contraction coefficient), and $\sigma = 0.5$ (shrinkage coefficient).

Recall that SSQ is given by:

$$SSQ = \sum_{j=1}^{t} (y_j(\theta^{true}) - z_j(\theta))^2 \tag{D.1}$$

$z_j(\theta)$ and $y_j(\theta^{true})$ represent the simulated and true epidemic curves respectively. j indicates a single day and t is the day on which the epidemic curve is predicted. θ represents the vector of parameters. Each parameter set contains a disease transmissibility value, an

Algorithm 2 Nelder-Mead Simplex Algorithm

Input: Initial parameters, and surveillance epidemic curve
Output: Optimal parameter set

START:
for $x_i \in X$: initial set of parameters in the form of vertices of size m **do**
evaluate the objective function
end for
sort and find the best (x_l), the worst (x_h) and the second worst (x_s) set of parameters
find the centroid of the polytope ($\bar{x}$) based on the best m parameters
repeat
conduct REFLECTION ($x_r = (1+\rho)\bar{x} - \rho x_h$) and evaluate $f(x_r)$
if $f(x_l) < f(x_r) < f(x_s)$ **then**
$x_h = x_r$
else if $f(x_r) < f(x_l)$ **then**
conduct EXPANSION ($x_e = (1+\rho x_i)\bar{x} - \rho x_i x_h$) and evaluate $f(x_e)$
if $f(x_e) < f(x_r)$ **then**
$x_h = x_e$
else
$x_h = x_r$
end if
else if $f(x_s) < f(x_r) < f(x_h)$ **then**
conduct *outside* CONTRACTION ($x_c = (1+\rho\gamma)\bar{x} - \rho\gamma x_h$) and evaluate $f(x_c)$
if $f(x_c) < f(x_h)$ **then**
$x_h = x_c$
else
for all of the set of parameters except x_l SHRINK ($x_i = x_l + \sigma(x_i - x_l)$) and evaluate $f(x_i)$
end if
else
conduct *inside* CONTRACTION ($x_c = (1-\gamma)\bar{x} + \gamma x_h$) and evaluate $f(x_c)$
if $f(x_c) < f(x_h)$ **then**
$x_h = x_c$
else
for all of the set of parameters except x_l SHRINK ($x_i = x_l + \sigma(x_i - x_l)$) and evaluate $f(x_i)$
end if
end if
until converged

incubation period and infectious period distribution. Each parameter set contains a disease transmissibility value, an incubation period and infectious period distribution.

The estimated parameters are the transmissibility and probabilities $p_k's$ corresponding to each of the incubation and infectious periods. The probability values must be non-negative and sum up to one independently for the incubation and infectious period distributions. We therefore modify the Nelder-Mead algorithm by introducing conditions which reinforce this requirement. As earlier summarized, the Nelder-Mead algorithm works in the following way. (i) $m+1$ parameters, θ's are used to initialize the algorithm. (ii) Using the $m+1$ parameter sets, $m+1$ $z_t(\theta)$s are simulated. (iii) The SSQs for each of these $z_t(\theta)$s is computed. (iv) Then the point with the largest value of SSQ is identified and the next point to try is on the hyperplane passing through the remaining m points. In this case, we want $p_i's \in \theta$ such that $0 \leq p_i \leq 1$ and $\sum_{i=1}^{I} p_i = 1$. Therefore, we do the following: when the algorithm asks for a computation of a new SSQ with a new set of p_i, we first let $p_i = 0$ for any $p_i < 0$. Next, we let the values of p_i that are used to compute the function $z_t(\theta)$ be $p_i^{new} = \frac{p_i}{\sum p_i}$ independently for the incubation and infectious periods.

Several modifications have been proposed specifically for stochastic scenarios. Barton et al. [12] illustrated that when differences in the objective function values are dominated by stochastic perturbations, this can lead to inappropriate termination and substantial error in the estimated optimum. One of the modifications proposed in [12] is $S9+RS$, which was shown to result in statistically significant improvements compared to the original algorithm. In the $S9+RS$ modification, we resample the best point during the shrink operation. In addition, this requires setting $\sigma = 0.9$ instead of 0.5.

Another prominent modification RSS/RMS was proposed by Humphrey et al [80]. Humphrey et al [80] proposed restarting the algorithm in three phases. In the first phase, the algorithm is run in the original setting. In the second and third phases, the shrink coefficient σ is increased by 0.2. This modification is based on the idea that during the first exploration the simplex tends not to reach its optimum. At the start of the second and third phase, the optimum solution from the previous phase forms the base of the new simplex. The other vertices are generated by multiplying the base by a step vector. Additionally for the second and third phases, [80] proposes using the full step length vector and the half multiplication of the step length vector correspondingly.

Neddermeijer et al. [98] compared both of the above modifications and concluded that in general $S9+RS$ modification outperforms RSS modification. We therefore use the $S9+RS$ modification in our code and implemented the Nelder-Mead algorithm using the Perl programming language.

Transmissibility	Incubation Period				Infectious Period				
6.00e-05	0:0.00	1:0.30	2:0.50	3:0.20	2:0.00	3:0.30	4:0.40	5:0.20	6:0.10
8.30e-05	0:0.20	1:0.45	2:0.35	3:0.00	2:0.66	3:0.33	4:0.01	5:0.00	6:0.00
7.00e-05	0:0.10	1:0.10	2:0.45	3:0.35	2:0.56	3:0.33	4:0.01	5:0.10	6:0.00
2.63e-05	0:0.1005	1:0.1543	2:0.3633	3:0.3819	2:0.444	3:0.344	4:0.1296	5:0.0563	6:0.0261
5.80e-05	0:0.00	1:0.40	2:0.25	3:0.35	2:0.66	3:0.23	4:0.01	5:0.10	6:0.00
9.10e-05	0:0.00	1:0.50	2:0.15	3:0.35	2:0.36	3:0.33	4:0.01	5:0.30	6:0.00
5.00e-05	0:0.20	1:0.10	2:0.35	3:0.35	2:0.36	3:0.53	4:0.01	5:0.10	6:0.00
4.20e-05	0:0.00	1:0.30	2:0.35	3:0.35	2:0.46	3:0.33	4:0.20	5:0.01	6:0.00
6.76e-05	0:0.1173	1:0.4487	2:0.3513	3:0.0827	2:0.7331	3:0.2588	4:0.0081	5:0	6:0
4.35e-05	0:0.0941	1:0.3433	2:0.3681	3:0.1945	2:0.6106	3:0.2196	4:0.0253	5:0.0789	6:0.0656
6.66e-05	0:0.01	1:0.21	2:0.43	3:0.35	2:0	3:0.35	4:0.34	5:0.22	6:0.09

Table D.1: Parameters used in initializing the forecasting procedure. The incubation (infectious) period is defined as follows: $k : p_k$ where k is the duration and p_k is the probability that an infected (infectious) individual will have an incubation (infectious) period of k days. The disease transmissibility is given as the probability of infection per unit of contact time between a susceptible and infectious individual in the network.

D.0.2 Computational Epidemiology Model

A detailed description of the individual-based model (ABM) used in simulating the outbreaks used in this study is discussed in [17]. The ABM belongs to a class of models called network based epidemiology models that uses a representation of the population that includes each individual and their minute-by-minute movements. Their interactions with other agents are used to generate a dynamic social network. These networks are then in turn used to simulate epidemics and study the effects of changes in individual behavior and public policy on the propagation of an outbreak [8]. Although neither changes in individual behavior nor public policy are directly explored in this study, it is extremely easy in these models to change individual behaviors, like keeping children home from school or in general limiting the number of non-essential activities of specific members of the population.

The creation of the individual-based epidemic model used in this study entails two major steps, the first of which consists of the creation of a social contact network from a state-of-the-art behavioral model. This involves creating synthetic populations and time varying social networks. Synthetic individuals and households, located in specified geographical regions (such as Los Angeles), each with a set of demographic variables are created using an iterative proportional fit to joint demographic distributions from the 2000 US census data provided in SF3 and PUMA (Public Use Microdata Area) files [7]. For example, a list of demographic information such as household income, family size, age, education etc. are available for each of the approximately 16 million individuals in the Los Angeles region.

The synthetic populations are created to produce realistic features and demographics while preserving the confidentiality of the original data sets. A node represents an individual in the synthetic population. Each node is placed in a household with other synthetic individuals and each household is geographically located such that a census aggregated to the block level of the synthetic population would be statistically identical to the real census data [14]. Additional information can be found in [14], [116] and [117].

Next, each individual in the synthetic household is allotted activities by time of day based on several thousand responses to an activity or time-use survey for a specific region. The National Household Transportation Survey is used in creating the activity templates assigned to each household. The time-use or activity survey is expected to vary by region given factors such as the geographical location and age composition of the population. Presently, this modeling approach is considered the *de facto* standard in transportation science and is called activity based travel demand models [8]. See [19] and [20] for additional information.

Using a decision tree based on demographic information (such as number of people in a household, number of children etc.), each household in the synthetic population is matched to a survey household. Each activity for each synthetic person is then assigned an appropriate real location based on a gravity model and land-use data [14]. The addresses of locations are obtained from Dun and Bradstreet's Strategic Database Marketing Records. The activities for each household are assigned to actual locations based on "the distance from the previous

Table D.2: Models and Modeling Approaches used in ABM

Models	**References**
Urban Population Mobility Models	[10], [20], [125], and [124]
Natural Disease History	[3], [44], [67], [73], and [93]
Transmission Models	[67], [73], and [93]
Social Network Models	[47], [67], [100]
Types of Interventions	[50], [51], [66], and [67]

activity and its *attractiveness* a measure of how likely that the activity happens there - number of employees, school enrollment, square feet of retail shopping etc." [46].

In addition to specific assignment of activities, the time at which each activity starts and ends is also included. This leads to each individual in each household having a minute-by-minute schedule for each day. Synthetic individuals in the population interact with each other based on their minute-by-minute schedule to produce realistic contact graphs where vertices represent individuals and edges represent contacts between individuals [8]. Individuals mimic the behaviors of real people by participating in everyday activities such as eating, socializing, shopping etc. and multiple edges can be used between each person and the locations representing their frequency of visits. The modeling approaches used in the ABM as presented in [8] are given in Table D.2.

Next, a computational model is developed to represent disease within individuals and the transmission between individuals in the synthetic population. The transition from one disease state (susceptible, exposed, infectious and removed) to another is probabilistic and timed (e.g. it may be represented by the distribution of the infectious period). The transition between states can also depend on the attributes of the people (e.g. age, health status etc.) and the type of contact (e.g casual, intimate etc.). For the disease model used in this study, the probability that an infectious person i infects a susceptible person j is given by:

$$Pr(\textit{person } i \textit{ infects person } j) = 1 - exp(-f(S, T_{i,j})) \tag{D.2}$$

where $f(S, T_{i,j})$ is a nonnegative monotone increasing function of S, the severity of the disease (called transmissibility) and $T_{i,j}$, the contact time between persons i and j. The disease model is combined with the information in the network to study an infectious disease. At each time step of a simulation, each of the nodes is either: susceptible, exposed, infectious, and removed. Time is divided into units based on days and the state of each individual is noted at the start of each day.

To run a simulation experiment, a population (contact network), characteristics of a disease

and initial conditions (such as duration) are specified. In this study, the social networks studied are based on Montgomery County in Virginia, Seattle, Los Angeles and surrounding metropolitan regions. The disease characteristics are based on seasonal influenza epidemics but with a higher transmissibility. For each simulated outbreak, several realizations of the stochastic process of disease propagation are computed. Intervention options such as vaccination, antiviral and social distancing can be applied during the outbreak to control its propagation. Each simulation is seeded with a randomly selected set of initially infected individuals. An epidemic curve, the vulnerability of different individuals in the network, epidemic size, number of new exposures on each day etc. can be explored at the end of each simulation of a disease outbreak.

Several studies have been implemented to validate specific components of the model and the general approach. See [14], [47], and [67] for structural validity of these models.

Chapter 6

Concluding Remarks

A summary of the main findings in each chapter is presented in Section 6.1. Directions for future research are discussed in Section 6.2.

6.1 Summary of Findings

As mentioned in the introduction, the methodology presented in this dissertation is discussed in Figure 6.1. We start by building a library of epidemics consisting of past outbreaks and simulated epidemics. Each epidemic is represented in the library based on the epidemic curve and associated model parameters. Given surveillance data during an epidemic, we compare the structure of the partial epidemic curve to those in the library. We summarize the major findings for forecasting the epidemic curve and estimating parameters as observed in chapters 2 to 5.

In Chapter 2, we examined how minute changes in the disease model parameters affect the dynamics of simulated epidemics. The results both reaffirmed published findings and provided new insights into how the prediction process can be improved by making changes to these parameters. The infectious period and transmissibility significantly influenced the attack rate as expected. The sensitivity of the model was consistent across different social networks investigated. Additionally, age group specific epidemics were sensitive to changes in the mean infectious period irrespective of the susceptibility of the other groups.

In Chapter 3, we compared seven supervised classification methods for the prediction of the epidemic curve. Random forest; a tree base classification approach was found to be the most consistent in identifying epidemics similar to those in our disease library. Random forest achieved the highest accuracy in predicting epidemics from different models and was robust across synthetic social networks.

In Chapter 4, we extended the classification approach to prediction by proposing a Dirichlet

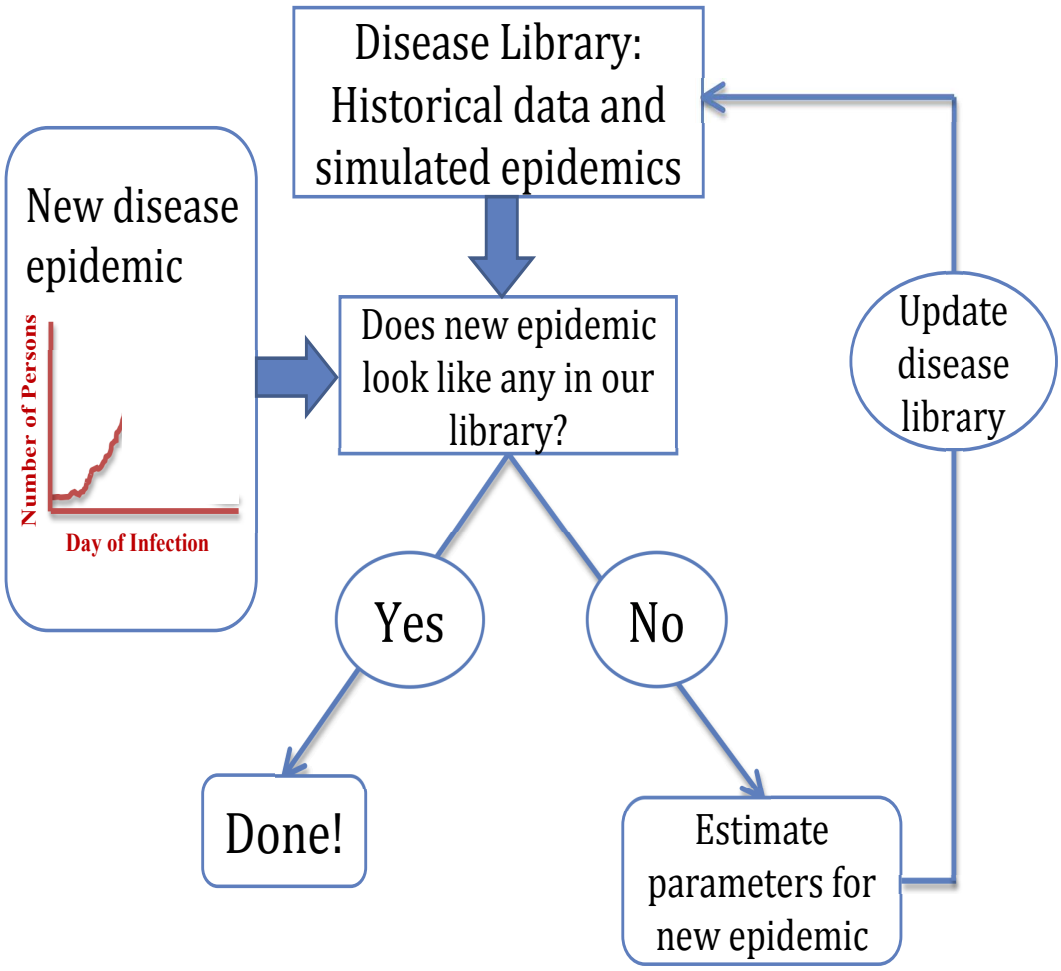


Figure 6.1: Summary of methodology

process model for predicting the epidemic curve. The Dirichlet process model was compared to random forest. Although the Dirichlet process model was able to predict most epidemics investigated before their peaks, the random forest method performed better. However, the Dirichlet process model provided additional benefits which are lacking in random forest. The model correctly identified most epidemic curves different from those in the library and also predicted the average time to peak.

In Chapter 5, we proposed a method for estimating parameters to model epidemics different from those in our library. The simulation optimization approach was tested on simulated data. The method was also illustrated by retrospectively predicting the 2009 H1N1(A) pandemic in Los Angeles.

6.2 Directions for Future Research

Forecasting of epidemic dynamics is an exciting research area with invaluable relevance to public health. The approach presented in this dissertation is not limited to influenza but can be applied to other infectious diseases or prediction problems with similar properties. Each of the methods presented in chapters 3-5 can be used jointly or independently for epidemic forecasting and parameter estimation under certain assumptions. However, further studies aimed at improving each of the methods would be useful. In each chapter we discussed additional research needed to improve the methods and make them more relevant for epidemic prediction.

In addition, we would need to integrate the forecasting methods within the epidemic modeling environment to enable prediction of the epidemic curve during a novel influenza outbreak. The modeling environment consists of different epidemic modeling software for studying infectious diseases and methods of intervention including vaccines, antivirals, school closure and other social distancing measures. Furthermore, we would like to investigate how these methods can be cued using both syndromic and social media based surveillance data.

There are also many fun and interesting research questions ranging from how to improve the epidemiology assumptions in the individual-based models to methods for reconciling differences in predictions made using different epidemic models. Other related research topics include: whether local surveillance data is better than data for an entire country, how to improve estimation of disease parameters from surveillance data, estimating severity and disease burden, and how to integrate data from different surveillance sources for prediction. I hope that I will have the opportunity to explore some of these research topics in the future.

Bibliography

[1] Acute communicable disease control program, los angeles county department of public health. *Influenza Watch Los Angeles County [newsletter]*, 4, 2010.

[2] Alan Agresti. *Categorical Data Analysis*. John Wiley & Sons, New York, NY, 2nd edition, 2002.

[3] Norman Bailey. *The Mathematical Theory of Infectious Diseases and its Applications*. Griffin, London, 1975.

[4] Norman T J Bailey. Some stochastic models for small epidemics in large populations. *Applied Statistics*, 13:9–19, 1964.

[5] MG Baker, N. Wilson, QS Huang, S. Paine, L. Lopez, D. Bandaranayake, M. Tobias, K. Mason, GF Mackereth, M. Jacobs, et al. Pandemic influenza A (H1N1) v in New Zealand: the experience from April to August 2009. *Eurosurveillance*, 14(34), 2009.

[6] C. Barrett, S. Eubank, V. Kumar, and M. Marathe. Understanding large scale social and infrastructure networks: A simulation based approach. *SIAM news: The Mathematics of Networks*, 2004.

[7] C L Barrett, R Beckman, K Berkbigler, K Bisset, K Bush, K Campbell, S Eubank, K Henson, J Hurford, D Kubicek, M Marathe, P Romero, J Smith, L Smith, P Speckman, P Stretz, G Thayer, E Van Eeckhout, and M Williams. TRANSIMS: Transportation analysis simulation system. *Technical Report, LA-UR-00-1725, Los Alamos National Laboratory Unclassified Report*, 3, 2001.

[8] Chris Barrett, Keith Bisset, Jonathan Leidig, Achla Marathe, and Madhav Marathe. Economic and social impact of influenza mitigation strategies by demographic class. *Epidemics*, 3(1):19–31, 2011.

[9] Chris Barrett, S. G. Eubank, and J. P. Smith. If smallpox strikes portland. *Scientific American*, 292:54–61, 2005.

[10] Christopher Barrett, Richard Beckman, Maleq Khan, V. S. Anil Kumar, Madhav Marathe, Paula Stretz, Tridib Dutta, and Bryan Lewis. Generation and analysis of

large synthetic social contact networks. In *Winter Simulation Conference*, WSC '09, pages 1003–1014, 2009.

[11] Christopher L. Barrett, Keith R. Bisset, Stephen G. Eubank, Xizhou Feng, and Madhav V. Marathe. Episimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks. In *Proceedings of the 2008 ACM/IEEE conference on Supercomputing*, SC '08, pages 37:1–37:12, Piscataway, NJ, USA, 2008. IEEE Press.

[12] Russell R. Barton and John S. Ivey, Jr. Nelder-mead simplex modifications for simulation optimization. *Management Science*, 42:954–973, November 1996.

[13] Nicole E. Basta, Dennis L. Chao, M. Elizabeth Halloran, Laura Matrajt, and Ira M. Longini. Strategies for pandemic and seasonal influenza vaccination of schoolchildren in the united states. *American Journal of Epidemiology*, 170(6):679–686, 2009.

[14] Richard Beckman, Keith Baggerly, and Michael Mckay. Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, 30(6):415–429, November 1996.

[15] Richard Beckman, Garima Chaturvedi, and Bryan Lewis. Clustering principal components to find groupings in a collection of curves. Technical Report 07-043, NDSSL, VBI, Virginia Tech, Blacksburg, Virginia, 2008.

[16] S. Bera and I. Mukherjee. Performance analysis of nelder-mead and a hybrid simulated annealing for multiple response quality characteristic optimization. *Proceedings of the International MultiConference of Engineers and Computer Scientists*, 3, 2010.

[17] Keith Bisset, Jiangzhuo Chen, Xizhou Feng, V S Anil Kumar, and Madhav Marathe. Epifast: a fast algorithm for large scale realistic epidemic simulations on distributed memory systems. In *Proceedings of the 23rd international conference on Supercomputing*, ICS '09, pages 430–439, 2009.

[18] D. Blackwell and J. B. Macqueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1:353–355, 1973.

[19] John Bowman. Activity-based disaggregate travel demand model system with activity schedules. *Transportation Research Part A: Policy and Practice*, 35(1):1–28, 2001.

[20] John Bowman, Mark Bradley, Yoram Shiftan, T Keith Lawton, and Moshe Ben-Akiva. Demonstration of an activity based model system for Portland. In *Proceedings of the 8th World Conference on Transport Research*, 1998.

[21] George E. P. Box, William G. Hunter, J. Stuart Hunter, and William G. Hunter. *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building.* John Wiley & Sons, June 1978.

[22] Leo Breiman. Bagging predictors. *Machine Learning*, 24:123–140, August 1996.

[23] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, October 2001.

[24] J.S. Brownstein, C.C. Freifeld, and L.C. Madoff. Digital disease detection – harnessing the web for public health surveillance. *New England Journal of Medicine*, 360(21):2153–2157, 2009.

[25] Donald Burke, Joshua Epstein, Derek Cummings, Jon Parker, Kenneth Cline, Ramesh Singa, and Shubha Charkravarty. Individual-based Computational Modeling of Smallpox Epidemic Control Strategies. *Academic Emergency Medicine*, 13(11):1142–1149, 2006.

[26] Dan G. Cacuci, Mihaela Ionescu-Bujor, and Ionel Michael Navon. *Sensitivity And Uncertainty Analysis: Applications to Large-Scale Systems.* Chapman & Hall/CRC, New York, volume ii edition, 2005.

[27] F. Carrat, A. Flahault, E. Boussard, N. Farran, L. Dangoumau, and A.J. Valleron. Surveillance of influenza-like illness in France. The example of the 1995/1996 epidemic. *Journal of Epidemiology and Community Health*, pages 32–38, 1998.

[28] Fabrice Carrat, Elisabeta Vergu, Neil M. Ferguson, Magali Lemaitre, Simon Cauchemez, Steve Leach, and Alain-Jacques Valleron. Time lines of infection and disease in human influenza: A review of volunteer challenge studies. *American Journal of Epidemiology*, 167(7):775–785, 2008.

[29] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 161–168, 2006.

[30] George Casella and Edward I. George. Explaining the Gibbs Sampler. *The American Statistician*, 46(3):167–174, 1992.

[31] Simon Cauchemez, Christl A Donnelly, Carrie Reed, Azra C Ghani, Christophe Fraser, Charlotte K Kent, Lyn Finelli, and Neil M Ferguson. Household transmission of 2009 pandemic influenza A (H1N1) virus in the United States. *New England Journal of Medicine*, 361(27):2619–2627, 2009.

[32] CDC. Centers for disease control and prevention: Weekly flu reports, April 2010.

[33] CDC. Updated CDC estimates of 2009 H1N1 influenza cases, hospitalizations and deaths in the united states, April 2009 ï£¡ April 10, 2010, Accessed on March 16, 2012 2010.

[34] CDC. Types of influenza viruses, 2011.

[35] Philip Chan and Salvatore Stolfo. A comparative evaluation of voting and meta-learning on partitioned data. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 90–98. Morgan Kaufmann, 1995.

[36] Dennis L. Chao, M. Elizabeth Halloran, and Ira M. Longini. School opening dates predict pandemic influenza a(h1n1) outbreaks in the united states. *Journal of Infectious Diseases*, 202(6):877–880, 2010.

[37] Dennis L. Chao, M. Elizabeth Halloran, Valerie J. Obenchain, and Ira M. Longini. FluTE, a Publicly Available Stochastic Influenza Epidemic Simulation Model. *PLoS Computational Biology*, 6(1):e1000656+, January 2010.

[38] Dennis L Chao, Laura Matrajt, Nicole E Basta, Jonathan D Sugimoto, Brandon Dean, Dee Ann Bagwell, Brit Oiulfstad, M Elizabeth Halloran, and Ira M Longini. Planning for the control of pandemic influenza a (h1n1) in Los Angeles county and the United States. *American Journal of Epidemiology*, 173(10):1121–1130, 2011.

[39] H.J. Clothier, J.E. Fielding, and H.A. Kelly. An evaluation of the Australian Sentinel Practice Research Network (ASPREN) surveillance for influenza-like illness. *Communicable diseases intelligence*, 29(3):231, 2005.

[40] Benjamin J. Cowling, Kwok H. Chan, Vicky J. Fang, Lincoln L. H. Lau, Hau C. So, Rita O. P. Fung, Edward S. K. Ma, Alfred S. K. Kwong, Chi-Wai Chan, Wendy W. S. Tsui, Ho-Yin Ngai, Daniel W. S. Chu, Paco W. Y. Lee, Ming-Chee Chiu, Gabriel M. Leung, and Joseph S. M. Peiris. Comparative Epidemiology of Pandemic and Seasonal Influenza A in Households. *New England Journal of Medicine*, 362(23):2175–2184, June 2010.

[41] R Deardon, S P Brooks, B T Grenfell, M J Keeling, M J Tildesley, and N Savill. Inference for individual level models of infectious diseases in large populations. *Statistica Sinica*, 20:239–261, 2010.

[42] Thomas Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:1895–1923, 1998.

[43] Thomas Dietterich. Machine learning for sequential data: A review. In *Proceedings of the Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition*, pages 15–30, 2002.

[44] L Elveback, J Fox, E Ackerman, A Langworthy, M Boyd, and L Gatewood. *American Journal of Epidemiology*, 103(2):152–165, 1976.

[45] Joshua Epstein. Modelling to contain pandemics. *Nature*, (7256):687, aug 2009.

[46] S Eubank, C Barrett, R Beckman, K Bisset, L Durbeck, C Kuhlman, B Lewis, A Marathe, M Marathe, and P Stretz. Detail in network models of epidemiology: are we there yet? *Journal of Biological Dynamics*, 4(5):446–455, 2010.

[47] Stephen Eubank, Hasan Guclu, V S Anil Kumar, Madhav Marathe, Aravind Srinivasan, Zoltan Toroczkai, and Nan Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 2004.

[48] Brian Everitt. *The Analysis of Contingency Tables*. Chapman & Hall, London, 2nd edition, 1992.

[49] IM Falcão, HR de Andrade, AS Santos, MT Paixão, and JM Falcão. Programme for the surveillance of influenza in Portugal: results of the period 1990-1996. *Journal of epidemiology and community health*, 52:39S, 1998.

[50] Neil Ferguson, Derek Cummings, Simon Cauchemez, Christophe Fraser, Steven Riley, Aronrag Meeyai, Sopon Iamsirithaworn, and Donald Burke. Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature*, 43:209–214, 2005.

[51] Neil Ferguson, Derek Cummings, Christophe Fraser, James Cajka, Philip Cooley, and Donald Burke. Strategies for mitigating an influenza pandemic. *Nature*, 442:448–452, 2006.

[52] T S Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[53] DM Fleming and AJ Elliot. Lessons from 40 years' surveillance of influenza in England and Wales. *Epidemiology and infection*, 136(07):866–875, 2008.

[54] H. Christopher Frey and Sumeet R. Patil. Identification and Review of Sensitivity Analysis Methods. *Risk Analysis*, 22(3):553–578, 2002.

[55] Jerome Friedman. Multivariate adaptive regression splines. *Annals of Statistics*, 19:1–67, 1991.

[56] P. Gates, K. Noakes, F. Begum, R. Pebody, and D. Salisbury. Collection of routine national seasonal influenza vaccine coverage data from GP practices in England using a web-based collection system. *Vaccine*, 27(48):6669–6677, 2009.

[57] Timothy C. Germann, Kai Kadau, Ira M. Longini, and Catherine A. Macken. Mitigation strategies for pandemic influenza in the United States. *Proceedings of the National Academy of Sciences*, 103(15):5935–5940, April 2006.

[58] S Ghosal. *The Dirichlet process, related priors and posterior asymptotics*. Cambridge University Press, 2010.

[59] Vincent Ginot, Sabrina Gaba, Rémy Beaudouin, Franck Aries, and Hervé Monod. Combined use of local and anova-based global sensitivity analyses for the investigation of a stochastic dynamic model: application to the case study of an individual-based model of a fish population. *Ecological Modelling*, 193:479ï£¡491, 2006.

[60] J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2008.

[61] E Goldstein, A Apolloni, B Lewis, J Miller, M Macauley, S Eubank, M Lipsitch, and J Wallinga. Distribution of vaccine/antivirals and the "least spread line" in a stratified population. *Journal of the Royal Society Interface*, 7(46):755–764, 2010.

[62] Edward Goldstein, Benjamin J. Cowling, Allison E. Aiello, Saki Takahashi, Gary King, Ying Lu, and Marc Lipsitch. Estimating incidence curves of several infections using symptom surveillance data. *PLoS ONE*, 6(8):e23380, 08 2011.

[63] Nicholas C Grassly and Christophe Fraser. Mathematical models of infectious disease transmission. *Nature Reviews Microbiology*, 6(6):477–487, 2008.

[64] Volker Grimm, Uta Berger, Finn Bastiansen, Sigrunn Eliassen, Vincent Ginot, Jarl Giske, John Goss-Custard, Tamara Grand, Simone K Heinz, and Geir Huse. A standard protocol for describing individual-based and agent-based models. *Ecological Modelling*, 198(1-2):115–126, 2006.

[65] I M Hall, R Gani, H E Hughes, and S Leach. Real-time epidemic forecasting for pandemic influenza. *Epidemiology and Infection*, 135(3):372–385, 2007.

[66] E Halloran, I Longini, M Cowart, and A Nizam. Community interventions and the epidemic prevention potential. *Vaccine*, 20(27ï£¡28):3254–3262, 2002.

[67] M Elizabeth Halloran, Neil Ferguson, Stephen Eubank, Ira Longini, Derek Cummings, Bryan Lewis, Shufu Xu, Christophe Fraser, Anil Vullikanti, Timothy Germann, Diane Wagener, Richard Beckman, Kai Kadau, Chris Barrett, Catherine Macken, Donald Burke, and Philip Cooley. Modeling targeted layered containment of an influenza pandemic in the United States. *Proceedings of the National Academy of Sciences*, 2008.

[68] M. Elizabeth Halloran, Frederick G. Hayden, Yang Yang, Ira M. Longini, and Arnold S. Monto. Antiviral effects on influenza viral transmission and pathogenicity: Observations from household-based trials. *American Journal of Epidemiology*, 165(2):212–221, 2007.

[69] D. M. Hamby. A comparison of sensitivity analysis techniques. *Health Physics*, 68:195–204.

[70] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. 2009.

[71] Brian Heath, Raymond Hill, and Frank Ciarallo. A survey of agent-based modeling practices (january 1998 to july 2008). *Journal of Artificial Societies and Social Simulation*, 12(4):9, 2009.

[72] J. C. Helton. Uncertainty and Sensitivity Analysis for Models of Complex Systems. In Frank Graziani, editor, *Computational Methods in Transport: Verification and Validation*, volume 62 of *Lecture Notes in Computational Science and Engineering*, chapter 9, pages 207–228. Springer Berlin Heidelberg, 2008.

[73] Herbert W Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42:599–653, 2000.

[74] C Holmes and N Adams. A probabilistic nearest neighbour method for statistical pattern recognition. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2), 2002.

[75] Y. Hsieh and Y. Cheng. Real-time forecast of multiphase outbreak. *Emerging Infectious Diseases*, 12(1):122, 2006.

[76] Y.H. Hsieh. Pandemic influenza A (H1N1) during winter influenza season in the southern hemisphere. *Influenza and Other Respiratory Viruses*, 4(4):187–197, 2010.

[77] QS Huang, D. Bandaranayake, LD López, R. Pirie, M. Peacey, R. Hall, J. Bocacao, B. Adlam, V. Hope, M. Croxson, et al. Surveillance for the 2009 pandemic influenza A (H1N1) virus and seasonal influenza viruses, New Zealand, 2009. *MMWR Morb Mortal Wkly Rep*, 58(33):918–21, 2009.

[78] Q.S. Huang, L. Lopez, and B. Adlam. Influenza surveillance in New Zealand in 2005. *NZ Med J*, 120(1256):U2581, 2007.

[79] J Huelsenbeck, S Jain, S Frost, and S Pond. A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. *PNAS*, 103:6263–6268, 2006.

[80] David G. Humphrey and James R. Wilson. A revised simplex search procedure for stochastic simulation response-surface optimization. *INFORMS Journal on Computing*, 12:751–759, 2000.

[81] Ross Ihaka and Robert Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, (3):299–314, 1996.

[82] Xia Jiang, Garrick Wallstrom, Gregory Cooper, and Michael Wagner. Bayesian prediction of an epidemic curve. *Journal of Biomedical Informatics*, 42:90–99, February 2009.

[83] Niall P A S Johnson and Juergen Mueller. Updating the accounts: global mortality of the 1918-1920 ï£¡spanishï£¡ influenza pandemic. *Bulletin of the History of Medicine*, 76(1):105–115, 2002.

[84] M. J. Keeling and L. Danon. Mathematical modelling of infectious diseases. *British Medical Bulletin*, 92(1):33–42, 2009.

[85] H. Kelly, K. Grant, et al. Interim analysis of pandemic influenza (H1N1) 2009 in Australia: surveillance trends, age of infection and effectiveness of seasonal vaccination. *Euro Surveill*, 14(31):1–5, 2009.

[86] 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.

[87] S Kim and P Smyth. Hierarchical Dirichlet processes with random effects. *Advances in Neural Information Processing Systems (NIPS)*, 2006.

[88] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by Simulated Annealing. *Science*, 220, 4598:671–680, 1983.

[89] Phenyo Lekone and Bärbel Finkenstädt. Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. *Biometrics*, 62(4):1170–1177, 2006.

[90] Marc Lipsitch, Lyn Finelli, Richard T Heffernan, Gabriel M Leung, and Stephen C Redd. Improving the evidence base for decision making during a pandemic: the example of 2009 influenza A/H1N1. *Biosecurity and bioterrorism biodefense strategy practice and science*, 9(2):89–115, 2011.

[91] A.L. Lloyd and S. Valeika. Network models in epidemiology: An overview. *In: Complex Population Dynamics: Nonlinear Modeling in Ecology, Epidemiology and Genetics, B. Blasius, J. Kurths and L. Stone (eds.), World Scientific*, ‘2007.

[92] I Longini, A Nizam, S Xu, K Ungchusak, W Hanshaworakul, D Cummings, and E Halloran. Containing pandemic influenza at the source. *Science*, 309(5737):1083–1087, 2005.

[93] I Longini, A Nizam, S Xu, K Ungchusak, W Hanshaworakul, D Cummings, and E Halloran. Containing pandemic influenza at the source. *Science*, 309(5737):1083–1087, 2005.

[94] Trevelyan McKinley, Alex Cook, and Robert Deardon. Inference in epidemic models without likelihoods. *The International Journal of Biostatistics*, 5, 2009.

[95] Stefano Merler and Marco Ajelli. The role of population heterogeneity and human mobility in the spread of pandemic influenza. *Proceedings of the Royal Society B: Biological Sciences*, 277(1681):557–565, 2010.

[96] A. S. Monto, F. M. Davenport, J. A. Napier, and Jr T. Francis. Effect of vaccination of a school-age population upon the course of an A2/Hong Kong influenza epidemic. *Bull World Health Organ.*, 41:537ï£¡42, 1969.

[97] Radford Neal. Slice sampling. *Annals of Statistics*, 31:705–767, 2000.

[98] H. Gonda Neddermeijer and Gerrit J. Van Oortmarssen. Adaptive extensions of the nelder and mead simplex method for optimization of stochastic simulation models. Technical report, 2000.

[99] J. A. Nelder and R. Mead. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313, January 1965.

[100] M Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[101] Hiroshi Nishiura. Real-time forecasting of an epidemic using a discrete time stochastic model: a case study of pandemic influenza (H1N1-2009). *BioMedical Engineering Online*, 10, 2011.

[102] Elaine Nsoesie, Richard Beckman, and Madhav Marathe. Estimation of an epidemic curve during an outbreak: A classification approach. *Proceedings of the Joint Statistical Meetings, Section on Statistics and Epidemiology*, pages 5177–5191, 2010.

[103] Elaine O. Nsoesie, Richard Beckman, Madhav Marathe, and Bryan Lewis. Prediction of an epidemic curve: A supervised classification approach. *Statistical Communications in Infectious Diseases*, 3(5), 2011.

[104] Y Ohkusa, T Sugawara, K Taniguchi, and N Okabe. Real-time estimation and prediction for pandemic A/H1N1(2009) in Japan. *Journal of infection and chemotherapy*, 2011.

[105] Jimmy Boon Som Ong, Mark I-Cheng Chen, Alex R. Cook, Huey Chyi Lee, Vernon J. Lee, Raymond Tzer Pin Lin, Paul Ananth Tambyah, and Lee Gan Goh. Real-time epidemic monitoring and forecasting of H1N1-2009 using influenza-like illness from general practice and family doctor clinics in singapore. *PLoS ONE*, 5(4):e10036, 04 2010.

[106] Lulla Opatowski, Christophe Fraser, Jamie Griffin, Eric De Silva, Maria D Van Kerkhove, Emily J Lyons, Simon Cauchemez, and Neil M Ferguson. Transmission characteristics of the 2009 H1N1 influenza pandemic: Comparison of 8 southern hemisphere countries. *PLoS Pathogens*, 7(9), 2011.

[107] C.W. Potter. A history of influenza. *Journal of Applied Microbiology*, 91(4):572–579, 2001.

[108] B. Pourbohloul, L.A. Meyers, D.M. Skowronski, M. Krajden, D.M. Patrick, R.C. Brunham, et al. Modeling control strategies of respiratory pathogens. *Emerging infectious diseases*, 11(8):1249, 2005.

[109] Hazhir Rahmandad and John Sterman. Heterogeneity and network structure in the dynamics of diffusion: Comparing agent-based and differential equation models. *Management Science*, 54:998–1014, May 2008.

[110] Brian J Reich and Howard D Bondell. A spatial dirichlet process mixture model for clustering population genetics data. *Biometrics*, 67(2):381–390, 2011.

[111] Thomas A. Reichert, Norio Sugaya, David S. Fedson, W. Paul Glezen, Lone Simonsen, and Masato Tashiro. The Japanese Experience with Vaccinating Schoolchildren against Influenza. *New England Journal of Medicine*, 344(12):889–896, 2001.

[112] Herbert Robbins and Sutton Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[113] Marcel Salathé, Maria Kazandjieva, Jung Woo Lee, Philip Levis, Marcus W. Feldman, and James H. Jones. A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences*, 2010.

[114] Andrea Saltelli, Paola Annoni, Ivano Azzini, Francesca Campolongo, Marco Ratto, and Stefano Tarantola. Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Computer Physics Communications*, 181(2):259–270, February 2010.

[115] Andrea Saltelli, Stefano Tarantola, Francesca Campolongo, and Marco Ratto. *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*. Halsted Press, New York, NY, USA, 2004.

[116] P Speckman, K Vaughn, and E Pas. Generating household activity-travel patterns (HATPs) for synthetic populations. *Transportation Research Board 1997 Annual Meeting*, 1997a.

[117] P Speckman, K Vaughn, and E Pas. A continuous spatial interaction model: Application to home-work travel in Portland, Oregon. *Transportation Research Board 1997 Annual Meeting*, 1997b.

[118] Jeffery K. Taubenberger and David M. Morens. 1918 Influenza: the Mother of All Pandemics. *Emerging Infectious Diseases*, 12(1), January 2006.

[119] Claudia Taylor, Achla Marathe, and Richard Beckman. Same influenza vaccination strategies but different outcomes across us cities? *International Journal of Infectious Diseases*, 14(9):e792 – e795, 2010.

[120] Y W Teh. Dirichlet processes. *Machine Learning Summer School Tutorial and Practical Course*, 2007.

[121] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[122] W.W. Thompson, L. Comanor, and D.K. Shay. Epidemiology of seasonal influenza: use of surveillance data and statistical models to estimate the burden of disease. *Journal of Infectious Diseases*, 194(Supplement 2):S82–S91, 2006.

[123] Matthew M. Tibbits, Murali Haran, and John C. Liechty. Parallel multivariate slice sampling. *Statistics and Computing*, 21(3):415–430, July 2011.

[124] TRB. *Transportation Research Board annual meetings*, 1998-2006.

[125] TRBC. *5th-9th Biennial National Academies Transportation Research Board Conferences on Application Of Transportation Planning Methods*, 1995-2003.

[126] Antoni Trilla, Guillem Trilla, and Carolyn Daer. The 1918 âĂIJspanish fluâĂİ in spain. *Clinical Infectious Diseases*, 47(5):668–673, 2008.

[127] Kagan Tumer and Joydeep Ghosh. Linear and order statistics combiners for pattern classification. In *Combining Artificial Neural Nets*, pages 127–162. Springer-Verlag, 1999.

[128] Vladimir Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[129] C Viboud, T Tam, D Fleming, A Handel, M A Miller, and L Simonsen. Transmissibility and mortality impact of epidemic and pandemic influenza, with emphasis on the unusually deadly 1951 epidemic. *Vaccine*, 24:6701ï£¡6707, 2006.

[130] Jacco Wallinga and Peter Teunis. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of Epidemiology*, 160(6):509–516, 2004.

[131] WHO. World health report 2007: A safe future: global public health security in the 21st century.

[132] WHO. Who handbook for journalists: Influenza pandemic, Accessed on April 5, 2012 Updated December 2005.

[133] W Wu, Y Mallet, B Walczak, W Penninckx, D Massart, S Heuerding, and F Erni. Comparison of regularized discriminant analysis linear discriminant analysis and quadratic discriminant analysis applied to NIR data. *Analytica Chimica Acta*, 329(3):257–265, 1996.

[134] Yang Yang, M. Elizabeth Halloran, Michael J. Daniels, Ira M. Longini, Donald S. Burke, and Derek A. T. Cummings. Modeling competing infectious pathogens from a bayesian perspective: Application to influenza studies with incomplete laboratory results. *Journal of the American Statistical Association*, pages 1310–1322, 2010.

[135] Yang Yang, Jonathan D Sugimoto, M Elizabeth Halloran, Nicole E Basta, Dennis L Chao, Laura Matrajt, Gail Potter, Eben Kenah, and Ira M Longini. The transmissibility and control of pandemic influenza a (h1n1) virus. *Science*, 326(5953):729–733, 2009.