Time-varying Individual-level Infectious Disease Models

by

Lin Zhang

A Thesis
presented to
The University of Guelph

In partial fulfilment of requirements
for the degree of
Doctor of Philosophy
in
Mathematics and Statistics

Guelph, Ontario, Canada
© Lin Zhang, December, 2013
ABSTRACT

Time-varying Individual-level Infectious Disease Models

Lin Zhang
Advisor: Dr. Rob Deardon

University of Guelph, 2013

Individual-level models (ILMs) of infectious disease spread are a system of statistical models which can be used to model infectious disease transmission through a population in discrete-time. These models allow researchers to incorporate risk factors at the individual level; thus they are suited for modeling epidemics spatially. Individuals, here, may refer to people, animals, or plants, or aggregated units such as animals on a farm or students in a school. ILMs are usually fitted to data within a Bayesian statistical framework using Markov chain Monte Carlo (MCMC) methods.

Ideally, covariate data and the infection status of individuals over time would be used to obtain parameter estimates for the ILMs. However, owing to various practical reasons, there are often situations in which the collection of infectious disease data at the individual level is infeasible. Instead, infectious disease data is collected at a regional level (e.g. a level which actually consists of spatially aggregated sets of individual units), such as health units or census regions. Therefore, it is reasonable
to assume that the infectivity of such aggregated units varies as the status of infectiousness (i.e. the number/proportion of infectious individuals) within the aggregated unit changes.

In the thesis, ILMs are extended to allow for time-varying susceptibility, infectivity and contact functions. A series of time-varying infectivity ILMs (TVI-ILMs) are then developed for the problem of modeling disease spread at the regional level. A method of carrying out model comparison and assessment based on the use of probability scoring rules is also developed and explored. Finally, the TVI-ILMs are extended to allow for infectivity curves that are dependent on regional-level covariates. Models and methods are tested on a combination of simulated epidemic data, and data from the 2009 H1N1 influenza pandemic collected in Southern Ontario.
This thesis is dedicated to everyone who has supported me all the way since the beginning of my studies. Also, this thesis is dedicated to my family who has been a great source of motivation and inspiration. But especially, to my husband Bill, who believes the power of love.
ACKNOWLEDGEMENTS

First and most, I would like to thank my advisor, Dr. Rob Deardon, for not only accepting me as a graduate student, but also his patience and encouragement throughout the journey of my PhD studies. I truly appreciate his guidance and support, both academic and non-academic. To me, he is not only a well qualified advisor, but also a respectable and responsible scholar.

I would also like to express my sincere gratitude to my advisory committee members, Dr. Julie Horrocks and Dr. Oalf Berke. Their support and advice have been crucial to the accomplishment of my studies, and their comments have been beneficial to the completion of this thesis. Also, appreciations to Larry Banks, for being a constant source of technical support.

A huge gratitude to Dr. Paul McNicholas, for his kindness and unreserved help. I have always enjoyed his courses which ignited my passion as well as opened a new horizon for me in Statistics. To Dr. Zeny Feng, for her sense of humor, support and caring. I am also indebted to Xin Xin and Weiqiang Wang, who squeezed time from their busy schedules to help me in many ways.

Finally, I owe so much thanks to my family and friends. Without your love and support, I would never made it this far. To Bill, nothing more to express, but thanks for being the tireless "technical support" and "customer service" in everyday life.
# TABLE OF CONTENTS

ABSTRACT ................................................................. ............................... iv

DEDICATION ................................................................................. v

ACKNOWLEDGEMENTS ................................................................. vi

LIST OF TABLES ........................................................................... ix

LIST OF FIGURES .......................................................................... x

CHAPTER

1. Introduction ............................................................................. 1

1.1 Overview ............................................................................. 1

1.2 Stochastic infectious disease modeling ................................. 2

1.2.1 Individual-level Infectious Disease Modeling ............... 2

1.3 Bayesian Paradigm ............................................................. 3

1.3.1 Bayesian Inference for Individual-level Models ......... 4

1.3.2 Prior Distributions ......................................................... 5

1.3.3 Comments on the Bayesian Inference ...................... 7

1.4 Markov Chain Monte Carlo Methods ............................... 8

1.4.1 The Metropolis Hastings Algorithm ......................... 9

1.4.2 Gibbs sampling ............................................................ 10

1.5 Data-augmented MCMC .................................................. 11

1.6 Thesis Overview ............................................................... 12

2. Analyzing Spatially Aggregated Infectious Disease Data using Time-varying Individual Level Models ......................... 14

2.1 Introduction ....................................................................... 14
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Averaged Difference of Deviance Information Criterion (DIC)</td>
<td>28</td>
</tr>
<tr>
<td>2.2</td>
<td>Deviance Information Criterion (DIC) for fitted models</td>
<td>42</td>
</tr>
<tr>
<td>2.3</td>
<td>Estimated posterior means and associated 95% percentile interval</td>
<td>44</td>
</tr>
<tr>
<td>3.1</td>
<td>Three examples of Strictly Proper Scoring Rules</td>
<td>60</td>
</tr>
<tr>
<td>3.2</td>
<td>Estimated posterior means and associated 95% percentile interval</td>
<td>70</td>
</tr>
<tr>
<td>3.3</td>
<td>The averaged DIC values and averaged quadratic score (QS), spherical score (SS) and logarithmic score (LS) for the true fitted model EG and the seven misspecified models, with performance rankings in parentheses.</td>
<td>73</td>
</tr>
<tr>
<td>3.4</td>
<td>The averaged values of PMCC, CRPS, $\alpha.\text{IS}(\alpha = 0.1, 0.5)$, $\alpha.\text{QS}(\alpha = 0.75, 0.95)$ over the 10 trials for the true fitted model EG and the 7 misspecified models, with performance, rankings in parentheses.</td>
<td>77</td>
</tr>
<tr>
<td>3.5</td>
<td>The averaged values of PMCC, CRPS, $\alpha.\text{IS}(\alpha = 0.1, 0.5)$, $\alpha.\text{QS}(\alpha = 0.75, 0.95)$ over the 10 trials for the true EG model and the 7 misspecified models, with performance, rankings in parentheses, based on the posterior predictive distribution of quadratic score over time.</td>
<td>79</td>
</tr>
<tr>
<td>3.6</td>
<td>The DIC values and quadratic score (QS), spherical score (SS) and logarithmic score (LS) for the time-invariant infectivity CE and PNSI models and the time-varying infectivity EG and GS models, with performance, rankings in parentheses.</td>
<td>84</td>
</tr>
<tr>
<td>3.7</td>
<td>The values of PMCC, CRPS, $\alpha.\text{IS}(\alpha = 0.1, 0.5)$, $\alpha.\text{QS}(\alpha = 0.75, 0.95)$ for the time-varying infectivity EG and GS model based on the posterior predictive distribution of cumulative number of infected regions over time.</td>
<td>85</td>
</tr>
<tr>
<td>3.8</td>
<td>The values of PMCC, CRPS, $\alpha.\text{IS}(\alpha = 0.1, 0.5)$, $\alpha.\text{QS}(\alpha = 0.75, 0.95)$ for the time-varying infectivity EG and GS model based on the posterior predictive distribution of quadratic score over time.</td>
<td>87</td>
</tr>
<tr>
<td>4.1</td>
<td>Averaged Difference of Deviance Information Criterion (DIC) over the ten epidemics</td>
<td>113</td>
</tr>
</tbody>
</table>
### LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Simulated heterogeneous population with 100 square regions.</td>
<td>25</td>
</tr>
<tr>
<td>2.2</td>
<td>Predictive cumulative number of cases with posterior means (dotted line) and 95% percentile intervals (blue lines) and the observed cumulative number of cases (black line) at the regional level.</td>
<td>30</td>
</tr>
<tr>
<td>2.3</td>
<td>The estimated exponential distance kernel against $d_{ij}$ under the posterior means for the time-invariant infectivity individual-level model and the time-varying infectivity individual-level models.</td>
<td>31</td>
</tr>
<tr>
<td>2.4</td>
<td>The MCMC trace plots and marginal posterior distributions for Exponential Infectivity and Geometric Kernel Model.</td>
<td>41</td>
</tr>
<tr>
<td>3.1</td>
<td>A typical simulated heterogeneous population on a 10×10 grid.</td>
<td>68</td>
</tr>
<tr>
<td>3.2</td>
<td>The MCMC trace plots and marginal posterior distributions for Exponential Infectivity and Geometric Kernel Model (EG).</td>
<td>71</td>
</tr>
<tr>
<td>3.3</td>
<td>The quadratic score (QS), spherical score (SS), and logarithmic score (LS) with posterior means under the one-step ahead predictive distribution over time for the true EG model (black line), the misspecified GG model (green dotted line), the misspecified CG model (blue dotted line).</td>
<td>75</td>
</tr>
<tr>
<td>3.4</td>
<td>(a) The quadratic score lines with posterior means from the EG model (black line), the misspecified EE model (orange line), the misspecified WEMIP model (blue line) and the misspecified CEMIP model (green line). (b) The quadratic score lines with posterior means from the EG model (black line), the misspecified GNSI model (green line) and the CH model (blue line) over time, respectively.</td>
<td>76</td>
</tr>
<tr>
<td>3.5</td>
<td>Predictive cumulative number of cases with posterior means (dotted line) and 95% percentile intervals (blue lines) and the observed cumulative number of cases (black line) for all the fitted models over time.</td>
<td>90</td>
</tr>
</tbody>
</table>
3.6 Posterior prediction of the quadratic score (QS), spherical score (SS) and logarithmic score (LS) with posterior means and 95% percentile intervals (black lines) over time and the observed score line (dotted line) approximated from the observed epidemic under the corresponding fitted model. 91

3.7 (a) and (b) The predictive cumulative number of cases under the fitted EG and GS models with posterior means (dotted line) and 95% percentile intervals (blue lines) and the observed cumulative number of cases (black line) over time, respectively. (c) and (d) The posterior prediction of the quadratic score (QS) with posterior means (dotted line) and 95% percentile intervals (blue lines) over time and the observed score line (black line) approximated from the observed epidemic under the EG and GS models, respectively. 92

4.1 Simulated heterogeneous population with 100 square regions. 108

4.2 Predictive cumulative number of cases with posterior means (black line) and 95% percentile intervals (dotted lines) and the observed cumulative number of cases (black line) at the regional level. 115

4.3 The estimated exponential distance kernel against $d_{ij}$ under the posterior means for the time-invariant infectivity model (CE) and the time-varying infectivity models (EE, GE, EP and GP). 116
CHAPTER 1

Introduction

1.1 Overview

Statistical infectious disease modeling has emerged as a research field of great importance, since outbreaks of infectious disease have imposed terrible risks to both animal and public health, as well as economic growth (Anderson and May, 1991). For example, the 2009 H1N1 pandemic influenza outbreak resulted in over 300,000 laboratory cases and 3917 deaths in 191 countries and territories (WHO, 2009b). Further, the 2001 foot-and-mouth disease (FMD) in the UK led to more than four millions of animals being slaughtered due to the control measures (see, e.g., Anderson, 2002; Ward et al., 2004).

Infectious disease models have been widely used for providing essential insights into the dynamics of epidemic spread. Suitable models are able to facilitate the understanding of associated risk factors, which in turn can help to develop efficient control strategies (e.g. Tildesley et al., 2006).
1.2 Stochastic infectious disease modeling

Over the past decade, the use of stochastic models for infectious disease modeling has exploded. An early pioneering contribution can be found in Bailey (1975), in which several stochastic and deterministic models are covered, along with their applications to data analysis. Later on, Becker (1989) considers the statistical analysis of infectious disease data. Stochastic models tend to be favored over deterministic models for modeling infectious disease, largely because they incorporate the probability of epidemic spread from individual to individual, and, thus, can better reflect the inherent dynamics of epidemic spread (Becker, 1989).

Further, the inclusion of a spatial component to infectious disease modeling has drawn great attention, having demonstrated importance to understanding epidemic spread (e.g. Lawson, 2006). Such spatial information can be represented by, for example, an infection kernel often based on the distance between susceptible and infectious individuals (see, Neal and Roberts, 2004; Deardon et al., 2010). Alternatively, a network function or contact-based kernel, based on social network structure (Cauchemez et al., 2011) can be used. In the case of livestock disease a network based on animal movement data might be used (e.g. Heath et al., 2008).

1.2.1 Individual-level Infectious Disease Modeling

With the development of computational techniques, infectious disease modeling has advanced to incorporate more informative epidemic data, including covariates
and disease information at the level of individuals in a population. Such models are known as "individual-level" models or "individual-based" models. Their high flexibility for taking into account various individual-level risk covariates (e.g. Gibson, 1997) has meant that different forms of individual-level models have been applied in a variety of areas in infectious disease epidemiology. For example, Neal and Roberts (2004) proposed a two-level individual-level model to study the outbreak of a historic Hagelloch measles epidemic in German, with the factor of heterogeneity incorporated. The stochastic spatial-temporal model of Cook et al. (2007) models *Heracleum mantegazzianum* (Giant Hogweed) transmission in Britain, identifying several risk factors. Additionally, several researchers carried out studies based on individual-level models to investigate the spread of the 2001 foot-and-mouth disease (FMD) in the UK (see, e.g., Keeling et al., 2001; Deardon et al., 2010; Jewell et al., 2009b). In this thesis, we focus on individual-level models as discussed in Deardon et al. (2010), and extensions of the models (see Section 1.6).

1.3 Bayesian Paradigm

Statistical inference for the individual-level models (ILMs) introduced in Section 1.2.1 is typically formulated within a Bayesian statistical framework. In contrast to classical inference, Bayesian inference is based on information in the observed data, characterized by the likelihood function, in conjunction with prior information for parameters $\theta$. In a sense, Bayesian parameterization consists of updating our prior
knowledge about the parameter $\theta$ by using the observed data through the likelihood. In the classical approach, $\theta$ is considered an unknown, but fixed, parameter that is estimated by a point estimate; and inferences about $\theta$ (e.g. frequentist confidence intervals) are based upon the sampling distribution (Bernardo and Smith, 1994). However, under the Bayesian approach, $\theta$ is assumed to be a random variable, following some unknown distribution. In other words, Bayesian inference results in a distributional estimate of the parameter $\theta$, which is known as the posterior distribution (see, e.g., Casella and Berger, 2002; Gelman et al., 2004).

1.3.1 Bayesian Inference for Individual-level Models

Suppose $D$ is the observed individual-level infectious disease data. The goal of the Bayesian approach is to make inference on the posterior distribution of $\theta$, conditional on $D$ (e.g. Robert, 2001; Gelman et al., 2004). The posterior distribution, $\pi(\theta|D)$, is derived from Bayes’ theorem, wherein the likelihood function is multiplied by the prior distribution and then normalized:

$$
\pi(\theta|D) = \frac{\pi(D|\theta)\pi(\theta)}{\pi(D)}
$$

(1.3.1)

where $\pi(D|\theta)$ is the likelihood function of the ILMs; $\pi(\theta)$ is the prior density of $\theta$; and $\pi(D) = \int \pi(D|\theta)\pi(\theta)d\theta$, is the normalization constant. Further, $\pi(\theta|D)$ is, up
to a constant of proportionality, given by:

\[ \pi(\theta|D) \propto \pi(D|\theta)\pi(\theta) \]

### 1.3.2 Prior Distributions

In the Bayesian framework, the prior distribution of the parameter \( \theta \) is a probability distribution that describes uncertainty about \( \theta \) prior to data observed. It is beneficial to analyze how sensitive the posterior predictive results are to various priors (Casella and Berger, 2002). In this thesis, four broad types of priors are used:

- **Conjugate priors**: the prior distribution of \( \theta \) is a conjugate prior for the likelihood function \( \pi(D|\theta) \), if the posterior distribution, \( \pi(\theta|D) \), belongs to the same distributional family as the prior \( \pi(\theta) \). Then, the prior and posterior are called conjugate distributions. All members of the exponential family have conjugate priors (Gelman et al., 2004). For example, if the likelihood function is gamma, then choosing a gamma prior for the rate parameter will guarantee that the posterior distribution is also gamma.

- "Informative" priors: the prior distribution of \( \theta \) can incorporate strong prior belief about \( \theta \) if it is available. For example, if the current model is developed as a new version of a previous model with a similar structure, then the posterior conclusions of \( \theta \) obtained from the previous model may be used to inform the prior distribution of \( \theta \) for the current model. Alternatively, there are some sit-
uations in which informative prior information is based on an expert’s opinion, elicited into the form of a proper probability density function. This process is so-called "prior elicitation". More details can be found in Gelman et al. (2004).

• "Vague" priors: a vague prior, by contrast to an informative prior, is used as an (approximately) uninformative prior so that the posterior distribution is mainly based on the likelihood function, which represents the data observed. That is, prior information would have little influence on the posterior results. Such priors are not invariant to transformation, and as such, the term "non-informative" should be avoided for vague priors (for details on "non-informative" priors that are invariant to transformation, see Jeffreys (1961) or Bernardo (2005)).

• Weakly informative priors: Gelman (2006) proposed the use of weakly informative priors for the variance hyperparameter of a two-level hierarchical normal model. The intentions of introducing weakly informative priors in statistical inference are: first, to provide relatively weak prior knowledge; second, to provide sufficient prior information to avoid results that are strongly contradictive to our prior beliefs. Weakly informative priors are commonly applied in reality, wherein many natural constraints exist (Gelman, 2006). However, it is tricky to choose a weakly informative prior that satisfies the two contradictory objectives above. Thus, it is important to examine the posterior results if weakly informative priors are used for a model, ascertaining that the results are generally in accord with the our prior knowledge (Gelman, 2008). Finally, weakly
informative priors are usually assigned to hyperparameters within hierarchical modeling (Gelman, 2008).

1.3.3 Comments on the Bayesian Inference

A number of advantages of the Bayesian framework are discussed here. First, informative priors can be incorporated in Bayesian inference to inform the present model, through the use of prior knowledge obtained from previous analysis (however, this aspect does cause everlasting controversy between frequentists and Bayesians’). Second, the Bayesian approach follows from a decision-theoretic framework (Bernardo and Smith, 1994; Robert, 2001). Robert (2001) sets the primary objective of statistical inference as a decision-making procedure. Following the Bayesian approach, the optimal decision is the Bayesian decision which minimizes the posterior expected loss, conditional on the observed data. It is attractive due to its inadmissibility (Gelman et al., 2004). Third, the Bayesian framework is a natural framework to allow for uncertainty associated with the modeling procedure. This is especially beneficial for infectious disease modeling, since missing or uncertain data is nearly always a problem that afflicts statistical inference in epidemic systems (e.g. dates of infection, which as rarely observed) (Deardon et al., 2010). Under the Bayesian framework, missing data can be treated as parameters to be estimated, and then the joint posterior distributions of the missing data and model parameters can be sampled (Gelman et al., 2004). Another desirable property is the ability to make inference on highly complex
models, such as ILMs, via Monte Carlo techniques such as Markov chain Monte Carlo (MCMC) methods (see Section 1.4).

As mentioned previously, Bayesian inference is also criticized, mainly due to the inclusion of prior information in the modeling procedure. To deal with this criticism, Bayesians often use vague prior distributions so that the posterior distribution of $\theta$ is primarily influenced by the data observed (Gelman et al., 2004). Second, with respect to highly complex models (e.g. ILMs), Bayesian MCMC techniques for carrying out statistical inference may be time-consuming and computationally expensive. In epidemic systems this can occur, for example, when an epidemic outbreak takes place in a large population, or a large amount of missing data need to be imputed (Deardon et al., 2010).

1.4 Markov Chain Monte Carlo Methods

Markov chain Monte Carlo (MCMC) techniques can be used to simulate the posterior distribution to avoid the difficulty of integrating to, for example, calculate the normalization constant of Bayes’ theorem. MCMC methods are an example of Monte Carlo simulation methods, in which our goal is to build a Markov chain whose stationary distribution is the posterior distribution to be sampled from (Robert and Casella, 2004; Gamerman and Lopes, 2006). From this, Monte Carlo integration can be used to estimate a given integral, such as a posterior mean. Further discussion of MCMC can be found in Gamerman and Lopes (2006) and Brooks et al. (2011).
1.4.1 The Metropolis Hastings Algorithm

Metropolis-Hastings (MH) algorithm, was originally developed by Metropolis and co-authors in 1953, and then generalized by Hastings in 1970. It is one of the commonly used MCMC algorithms (Chib and Greenberg, 1995). In the MH algorithm, a proposal density, \( q(\theta' | \theta^{(t)}) \), is used for generating a candidate parameter value \( \theta' \), which is either accepted or rejected at iteration \( t \) of the MCMC chain. For Bayesian parameterization, a general algorithm is given as follows (Robert and Casella, 2004):

- At iteration \( t \), a new parameter value \( \theta' \) is generated via a proposal density \( q(\theta' | \theta^{(t)}) \).

- The acceptance probability \( \alpha(\theta'; \theta^{(t)}) \) is then calculated by:

\[
\alpha(\theta'; \theta^{(t)}) = \min\left(1, \frac{\pi(D | \theta') \pi(\theta') q(\theta^{(t)} | \theta')}{\pi(D | \theta^{(t)}) \pi(\theta^{(t)}) q(\theta' | \theta^{(t)})}\right)
\]  

(1.4.1)

- The candidate \( \theta' \), is then accepted with probability \( \alpha(\theta'; \theta^{(t)}) \), in which case \( \theta^{(t+1)} = \theta' \); otherwise, it is rejected with probability, \( 1 - \alpha(\theta'; \theta^{(t)}) \), and then \( \theta^{(t+1)} = \theta^{(t)} \).

One major advantage of the MH algorithm over some other sampling methods (e.g. rejection sampling) is that it is more efficient when dealing with high dimensional data (Robert and Casella, 2004). Theoretically, the choice of the proposal density \( q \) is of great flexibility. However, in the Bayesian framework, when \( q \) is poorly chosen, the rejection rate can be high, resulting in poor efficiency of the MCMC algorithm.
The most common proposal distributions used to generate Markov chains within the MH context are the random walk and the independence sampler. Details regarding these two sampling methods can be found in Appendix A.

1.4.2 Gibbs sampling

Gibbs sampling is an alternative MCMC method for generating a Markov chain. The underlying concept of Gibbs sampling is: given a parameter vector \( \Theta = (\Theta_1, ..., \Theta_p) \), a new realization of \( \Theta \) is simulated as a part of a Markov chain from the full conditional densities, \( \pi_1, ..., \pi_p \). For example, based on the full set of conditional densities, \( \theta_i | \theta_{-i} \) can be randomly drawn from the conditional distributions with densities \( \pi_i(\theta_i | \theta_{-i}) \), for \( i = 1, ..., p \), respectively, where \( \theta_{-i} = (\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_p) \) (Robert and Casella, 2004). The general algorithm, under the Bayesian framework is given as follows:

Given the current state \( \theta^{(t)} = (\theta_1^{(t)}, ..., \theta_p^{(t)}) \), at each iteration \( t \):

- \( \theta_1^{(t+1)} \) is generated from \( \pi_1(\theta_1 | \theta_{-1}^{(t)}, D) \)
- \( \theta_2^{(t+1)} \) is generated from \( \pi_2(\theta_2 | \theta_1^{(t)}, \theta_3^{(t)}, ..., \theta_p^{(t)}, D) \)
- ...
- \( \theta_p^{(t+1)} \) is generated from \( \pi_p(\theta_p | \theta_{-p}^{(t+1)}, D) \)

The Gibbs sampler is a special case of the MH algorithm wherein the full set of conditional densities are used as proposal densities in the MH algorithm. It appears to
be rather straightforward when the conditional densities ideally have a standard form. Then parameters can be easily sampled from the conditionals. Prior distribution can sometimes be chosen to make sure the conditional posteriors are of standard form, for example, using conjugate priors. However, if there is difficulty in sampling from the conditionals, the more general Metropolis-Hastings algorithm can be used.

1.5 Data-augmented MCMC

As described before, one of the advantages with the Bayesian framework is that missing data can be treated as parameters to be estimated via data augmented MCMC. Supposing $Z$ denotes the missing data (e.g. dates of infection in infectious disease data), the joint likelihood function $\pi(D, Z|\theta)$ is assumed to exist and the prior distributions, $\pi(Z)$ and $\pi(\theta)$, are assumed independent. Then, joint posterior distribution of $\pi(\theta, Z|D)$ that can be sampled from, up to a constant of proportionality, is given by Press et al. (2004):

$$\pi(\theta, Z|D) \propto \pi(D, Z|\theta)\pi(\theta)$$ (1.5.1)

However, implementation issues might arise when many extra parameters are introduced representing missing or uncertain data. First, there is high potential that complex relations exist between the missing data, or between the missing data and the model parameters. As a result, the efficiency of the Markov chain could be
quite low thus requires a large number of MCMC iterations to achieve stationarity of the chain (Press et al., 2004). Second, if using single-parameter updates (or a large number of block updates), as would be typically done, multiple likelihood calculations might be required at each MCMC iteration also leading to a large computational cost.

1.6 Thesis Overview

Individual-level models (ILMs), as discussed in Deardon et al. (2010), are a framework of mechanistic statistical models that can be applied with the aim of modeling infectious disease transmission over time and space. One desirable property of these models is their ability to incorporate risk factors, on which the risk of epidemic spread depends, at the individual level (e.g. animals or people). Covariate data and the infection history of individuals throughout the population are necessary in order to obtain parameters estimates of the ILMs.

However, a common problem is that there are occasions in which the collection of such infectious disease data at the level of individuals is infeasible due to various practical reasons. Instead, the ILMs can be fitted to summarized infectious disease data (i.e. data at a level which consists of an aggregated set of individual units, such as a regional/farm level).

In such situations, it would seem reasonable to assume that the infectivity of each region varies over the course of infectiousness of a region as the actual number/proportion of infectious individuals in a region changes over time. In this sense,
the use of ILMs with time-invariant infectivity may be insufficient and ineffective in explaining the effect of time-varying nature of infectivity.

In this thesis, the general ILMs of Deardon et al. (2010) are extended to allow for time-varying infectivity, susceptibility and/or contact functions. Specifically, the use of time-varying infectivity ILMs (TVI-ILMs) are considered as a tool for modeling spatially aggregated data.

In Chapter 2, time-varying individual-level models (TV-ILMs) are presented, which incorporate homogeneous time-varying infectivity curves for regional-level transmission modeling. These time-varying infectivity models are applied to the epidemic data from the 2009 H1N1 influenza outbreak in Southern Ontario.

The use of probability scoring rules for comparing and validating individual-level models, and in particular the time-varying infectivity models of Chapter 2, is explored in Chapter 3. Here, a simulation study is carried out to assess the ability of these probability scoring rules to compare and assess model fit. This is done for the 2009 H1N1 Ontario data set as well.

In Chapter 4, two new time-varying infectivity models are proposed based on the ones described in Chapter 2, in which some parameters of the infectivity curve are dependent upon some regional-level covariate.

Finally, in Chapter 5, a discussion of the work carried out and detail of possible future work is given.
CHAPTER 2

Analyzing Spatially Aggregated Infectious Disease Data using Time-varying Individual Level Models

2.1 Introduction

Well-constructed mathematical models are of great use for gaining essential insights into the dynamics of infectious disease transmission. They can be useful for extending the understanding of potential risk factors associated with the spread of the epidemic, and for developing efficient control measures. For example, Cauchemez et al. (2011) investigated the importance of social networks on disease transmission and population structure in an analysis of an H1N1 influenza outbreak in an elementary school in April 2009. Lekone and Finkenstädt (2006) developed methods of inference for the use of a chain-binomial model to study 1995 Ebola outbreaks in the Democratic Republic of Congo. Infectious disease modeling has also been widely

---

1Chapter 2 has been submitted to, and a revision requested from, the peer-reviewed journal, Biostatistics.
employed in modeling the dynamics of animal epidemics, such as the 2001 foot-and-mouth disease in the UK (see, e.g., Keeling et al., 2001; Jewell et al., 2009b; Deardon et al., 2010).

Individual-Level Models (ILMs), as denoted by Deardon et al. (2010), are a system of models designed to model the spread of infectious disease when that spread depends on various individual-level risk factors. It is assumed that individuals in population are represented as discrete points in time and space. Such models have been applied to a variety of scenarios. For example, Vrbik et al. (2012) detailed the application of ILMs as a tool for analyzing data obtained from an experimental fire, where the "individuals" were a set of cells constituting a piece of burning wax paper.

There are also occasions when the "individual-level" at which we wish to model disease spread actually refers to an aggregated set of individual units. For example, in the aforementioned work considering the UK 2001 epidemic, modeling was carried out at the farm level rather than animal level, with the number and type of animals (e.g. sheep and cattle) on each farm being used as covariates in the model. Similarly, disease prevalence data may be also aggregated at some regional level—e.g. census regions or health units—so that "individual-level" modeling between regions can be carried out, perhaps with demographic and disease prevalence information on the individual regions being used to inform the model. In such situations, it is reasonable to assume that the infectivity, or transmissibility, of aggregated units varies over time as the infectious status of the individuals that make up that aggregated unit change.
In this Chapter, the ILMs as proposed by Deardon et al. (2010) are generalized to allow for time-varying susceptibility, infectivity, and/or contact functions. For example, it might be necessary to incorporate time-varying susceptibility to account for changing vaccination status in a region, time-varying infectivity to account for varying infectious levels in a region, and/or time-varying distance kernel/networks to account for changing weather or population movement. Here, the focus is on the issue of time-varying infectivity due to aggregation over regions. The infectivity is then assumed to vary over time, and to follow some discrete curves. Three discrete curves—discrete exponential, gamma, and Weibull—are proposed and examined in the time-varying infectivity models.

The objective of this Chapter is to show that such time-varying infectivity models can better fit to data when it has been spatially aggregated. A simulation study is first carried out, in which the individual-level data is collected under various under-reporting conditions, and is then aggregated to the regional level. Different forms of TV-ILMs, as well as a time-invariant ILM, are compared when fitted to the regional-level data under these underreporting conditions. In such scenarios, the capture of the spatial-temporal information with the TV-ILMs is scrutinized as well. Then, time-invariant models and time-varying models are applied to data from the 2009 H1N1 influenza outbreak in Southern Ontario. The modeling is carried out at the level of census regions in the Great Toronto Area (GTA). Here, models are compared using the Deviance Information Criterion (DIC) approach suggested by Spiegelhalter.
et al. (2002).

In Section 2.2 the general individual-level model (ILM) of Deardon et al. (2010) is introduced and a generalized form of time-varying ILMs is proposed, followed by a discussion of the Bayesian MCMC framework in which the models are fitted. In Section 2.3 the specific models are presented and the simulation study and the the models used to compare them are detailed. The 2009 H1N1 influenza study is presented in Section 2.4, while some potential future research of interest is then discussed in Section 2.5.

2.2 Individual-Level Model

2.2.1 The General ILM

Deardon et al. (2010) proposed a class of individual-level models (ILMs) for modeling infectious disease spread. These ILMs are placed in an SEIR compartmental framework (Anderson and May, 1991). Briefly, the status of an individual $i$ can be one of four categories: susceptible (S), exposed (E), infectious (I) and removed (R), at any given time point: if individuals are in the susceptible state this means they do not have the disease, but can contract it; if in the exposed state individuals are infected but not able to transmit the virus; if in the infectious state individuals can transmit the disease; in the removed state individuals have been removed from the population, possibly through recovery, acquired immunity, quarantine, or death. In addition, models can also be fitted within other compartmental frameworks (e.g.
SEIS).

Here, time $t$ is discretized so that time point $t$ represents an interval $[t, t + 1)$. Let $S(t)$, $E(t)$, $I(t)$, $R(t)$ denote the sets of individuals being susceptible, exposed, infectious and removed within time interval $[t, t + 1)$, respectively. Let $T$ denote the last observed day of the epidemic.

Now let $P_{it}$ be the probability of a susceptible individual $i$ being infected within the time interval $[t, t + 1)$ in continuous time. A general form of the ILM is given by Deardon et al. (2010) as:

$$P_{it} = 1 - \exp \left\{ -\Omega_S(i) \sum_{j \in I(t)} \Omega_T(j)k(i, j) \right\} - \varepsilon(i, t)$$  \hspace{1cm} (2.2.1)

where $\Omega_S(i)$ is a susceptibility function representing risk factors associated with susceptible individual $i$ contracting the disease; $\Omega_T(j)$ is a transmissibility function representing risk factors associated with infectious individual $j$ passing on the disease (such risk factors might include age, gender, the population size in different regions/farms, and environmental risk factors, for example); $k(i, j)$ is an infection kernel representing shared risk factors often based on the separation distance (e.g. Euclidean distance) between the susceptible and infectious individuals $i$ and $j$, respectively; and $\varepsilon(i, t)$ is a "spark" term, which can be introduced to account for some other random behaviors that are not well explained by the previous terms in the model (e.g. infection coming in from outside the study population).

The likelihood of the model is a product of all infection events and non-infection
events over the entire time period $t = 0, \ldots, T$, which is given by:

$$
\pi(S, E, I, R; \theta) = f(S, E, I, R|\theta) = \prod_{t=0}^{T} f_t(S, E, I, R|\theta)
$$

(2.2.2)

where, $S = \{S(t)\}_{t=1}^{T}$, $E = \{E(t)\}_{t=1}^{T}$, $I = \{I(t)\}_{t=1}^{T}$, $R = \{R(t)\}_{t=1}^{T}$; $\theta$ is vector of parameters that needs to be estimated; $f_t(S, E, I, R|\theta)$ is the product of the probability of all new infection events and of all non-infection events occurring in time interval $[t, t+1)$:

$$
f_t(S, E, I, R|\theta) = \left[ \prod_{i \in E(t+1) \setminus E(t)} P_{it} \right] \left[ \prod_{i \in S(t+1)} \left( 1 - P_{it} \right) \right].
$$

2.2.2 Time-varying Individual-level Models

In this Chapter, the ILMs of Deardon et al. (2010) are generalized to time-varying individual-level models (TV-ILMs), allowing for time-varying susceptibility and transmissibility functions and infection kernel. The generalized model has the form:

$$
P_{it} = 1 - \exp \left[ \{-\Omega_S(i, t) \sum_{j \in I(t)} \Omega_T(j, t) k(i, j, t)\} - \varepsilon(i, t) \right]
$$

(2.2.3)

where, $\Omega_S(i, t)$ is a susceptibility function of potential risk factors associated with susceptible individual $i$ contracting the disease during time interval $[t, t+1)$; $\Omega_T(j, t)$ represents the transmissibility function of potential risk factors associated with infectious individual $j$ passing on the disease in $[t, t+1)$; $k(i, j, t)$ is a time-varying
infection kernel representing shared risk factors often based on the separation distance (e.g. Euclidean distance), between the infectious and susceptible individuals \( i \) and \( j \) in \([t, t + 1]\).

### 2.2.3 Statistical Framework

The ILMs here are fitted within a Bayesian statistical framework. Missing or uncertain data is almost always a problem in epidemic systems, but in the Bayesian framework missing data can be treated as model parameters to be estimated (Gelman et al., 2004). Under a Bayesian approach, \( \theta \) is assumed to be a random variable, following some unknown distribution. Therefore, Bayesian inference results in a distributional estimate of parameter \( \theta \) which is known as the posterior distribution (Casella and Berger, 2002). For the ILMs, the posterior distribution, \( \pi(\theta|S, E, I, R) \), can be obtained by using information in the observed data, characterized by the likelihood function, to update the prior knowledge about the distribution of \( \theta \). \( \pi(\theta|S, E, I, R) \) is given by:

\[
\pi(\theta|S, E, I, R) = \frac{\pi(S, E, I, R|\theta)\pi(\theta)}{\pi(S, E, I, R)} \tag{2.2.4}
\]

where \( \pi(\theta) \) is the prior density of \( \theta \), and \( \pi(S, E, I, R) \) is a normalization constant.

Markov chain Monte Carlo (MCMC) techniques can be used to simulate the posterior distribution to avoid the difficulty of integrating, for example, to calculate the normalization constant of Bayes’ theorem. MCMC is a Monte Carlo simulation method in which the goal is to build a Markov chain whose stationary distribution is
the posterior distribution to be sampled from (Robert and Casella, 2004; Gamerman and Lopes, 2006).

One of the most widely used algorithms to produce such a Markov chain is the Metropolis-Hastings (MH) algorithm which was originally developed by Metropolis and co-authors in 1953, and then generalized by Hastings in 1970 (Chib and Greenberg, 1995). A general algorithm for this study is given as follows:

- At MCMC iteration \( k \), a proposed parameter value \( \theta' \) is generated via a proposal density \( q(\theta'|\theta^{(k)}) \).

- The acceptance probability \( \alpha(\theta';\theta^{(k)}) \) is then calculated by:

\[
\alpha(\theta';\theta^{(k)}) = \min \left( 1, \frac{\pi(S,E,I,R|\theta')\pi(\theta')q(\theta^{(k)}|\theta')}{\pi(S,E,I,R|\theta^{(k)})\pi(\theta'|\theta^{(k)})q(\theta'|\theta^{(k)})} \right) \tag{2.2.5}
\]

- The candidate \( \theta' \), is then accepted with probability \( \alpha(\theta';\theta^{(k)}) \), in which case \( \theta^{(k+1)} = \theta' \); otherwise, it is rejected with probability, \( 1 - \alpha(\theta';\theta^{(k)}) \), and then \( \theta^{(k+1)} = \theta^{(k)} \).

### 2.3 Simulation Study

In this study, the use of TV-ILMs incorporating time-varying transmissibility/infectivity will be explored when data have been aggregated at, say, a regional level. The primary objective of the simulation study is to ascertain how much spatial-temporal information can be captured with the TV-ILMs under various underreporting condi-
tions when data have been spatially aggregated. It is also concerned with the question of whether TV-ILMs incorporating time-varying infectivity better fit the data than ILMs with time-invariant (constant) infectivity. Here individual-level data is simulated and then aggregated into regional levels, whilst incorporating an underreporting mechanism. A time-invariant infectivity model and three different time-varying infectivity models are fitted to the regional-level data. For the purpose of this study the susceptibility function and infection kernel are assumed to be time-invariant; i.e., \( \Omega_S(i, t) = \Omega_S(i) \) and \( k(i, j, t) = k(i, j) \).

2.3.1 Models for simulation study

**Model 1. Constant Infectivity and Exponential Kernel Model (CE).**

The first model is a time-invariant ILM. Here, \( \Omega_S(i)\Omega_T(j) = \alpha \) (e.g., \( \Omega_S(i) = \alpha \) and \( \Omega_T(j) = 1 \) for each \( i \) and \( j \)). It is assumed that the infection kernel \( k(i, j) = k(d_{ij}) = \exp\{-\beta d_{ij}\} \), an exponential distance kernel, where \( \beta \) is the exponential decay parameter of the infection kernel and \( d_{ij} \) is the Euclidean distance between susceptible individual \( i \) and infectious individual \( j \). The sparks term, \( \varepsilon(i, t) = \varepsilon \), is introduced to allow for purely random infection. The exponential kernel model (CE) is given by:

\[
P^{(CE)}_{it} = 1 - \exp\left[-\left(\alpha \sum_{j \in I(t)} \exp\{-\beta d_{ij}\}\right) - \varepsilon\right]
\]  

(2.3.1)

**Model 2. Exponential Infectivity and Exponential Kernel Model (EE).**
Here, consider a TV-ILM in which the transmissibility is allowed to vary over time and follows a discrete exponential curve. Additionally $\Omega_S(i, t) = 1$ and $k(i, j, t) = k(d_{ij}) = \exp\{-\beta d_{ij}\}$ for each $i$ and $j$. The probability of a susceptible individual $i$ becoming infected within the time interval $[t, t + 1)$ is thus given by:

$$P_{it}^{(EE)} = 1 - \exp\left[-\left(\sum_{j \in I(t)} \Delta_1(t; A_1, \lambda_1, \tau_j) \exp\{-\beta d_{ij}\}\right) - \varepsilon\right]$$

(2.3.2)

where,

$$\Omega_T(j, t) = \Delta_1(t; A_1, \lambda_1, \tau_j) = A_1 \exp\{-\lambda_1(t - \tau_j)\},$$

$\tau_j$ is the time point at which individual $j$ becomes infectious, and $A_1$ is a scale parameter.

**Model 3. Gamma Infectivity and Exponential Kernel Model (GE).**

Here a model with a different time-varying transmissibility is considered; specifically, $\Omega_T(j, t)$ is modified to follow a discrete gamma curve. The model is therefore:

$$P_{it}^{(GE)} = 1 - \exp\left[-\left(\sum_{j \in I(t)} \Delta_2(t; A_2, \phi_2, \lambda_2, \tau_j) \exp\{-\beta d_{ij}\}\right) - \varepsilon\right]$$

(2.3.3)

where

$$\Omega_T(j, t) = \Delta_2(t; A_2, \phi_2, \lambda_2, \tau_j) = A_2(t - \tau_j)^{\phi_2-1} \exp\{-\lambda_2(t - \tau_j)\},$$
and $A_2$ is a scale parameter.

**Model 4. Weibull Infectivity and Exponential Kernel Model (WE).**

Another alternative is to use a discrete Weibull curve, $\Delta_3$, as the time-varying transmissibility. Thus, the model is as follows:

$$
P_{it}^{(WE)} = 1 - \exp \left[ - \left( \sum_{j \in I(t)} \Delta_3(t; A_3, \phi_3, \lambda_3, \tau_j) \exp\{-\beta d_{ij}\} \right) - \varepsilon \right] \quad (2.3.4)
$$

where,

$$
\Omega_T(j, t) = \Delta_3(t; A_3, \phi_3, \lambda_3, \tau_j) = A_3 \left[ \frac{t - \tau_j}{\lambda_3} \right]^{\phi_3-1} \exp \left[ - \left( \frac{t - \tau_j}{\lambda_3} \right)^{\phi_3} \right],
$$

and $A_3$ is the scale parameter.

### 2.3.2 Epidemic Simulation

As stated previously, the purpose of this simulation study is to compare different TV-ILMs, as well as the time-invariant ILM, when fitted to data aggregated to a regional level under various underreporting conditions. The study is carried out in the following manner.

First, the population is assumed to have five clusters. The spatial locations of 1,000 individuals are randomly generated from truncated bivariate Gaussian clusters.
within a 40 × 40 unit square area. The Gaussian mixture model used is:

\[ f(x|\theta) = \sum_{i=1}^{5} \frac{1}{\sigma_i} g(x|\mu_i, \Sigma_i) \]

where, \( x = (x_1, x_2)^T : 0 \leq x_1 \leq 40, 0 \leq x_2 \leq 40 \); and \( \mu_1 = (8.5, 9.0)^T \), \( \Sigma_1 = \begin{pmatrix} 9.25 & 0 \\ 0 & 9.50 \end{pmatrix} \); \( \mu_2 = (31.5, 9.0)^T \), \( \Sigma_2 = \begin{pmatrix} 10.0 & 0 \\ 0 & 9.5 \end{pmatrix} \); \( \mu_3 = (9.0, 31.0)^T \), \( \Sigma_3 = \begin{pmatrix} 10.5 & 0 \\ 0 & 9.5 \end{pmatrix} \); \( \mu_4 = (20.5, 20.0)^T \), \( \Sigma_4 = \begin{pmatrix} 9.5 & 0 \\ 0 & 10.5 \end{pmatrix} \); \( \mu_5 = (31.0, 31.0)^T \), \( \Sigma_5 = \begin{pmatrix} 9.5 & 0 \\ 0 & 10.0 \end{pmatrix} \); The simulated population is shown in Figure 2.1.

Second, five epidemics are simulated through this population from the constant

![Simulated Individuals](image)

**Figure 2.1:** Simulated heterogeneous population with 100 square regions.

infectivity and exponential kernel model (CE). In each epidemic simulation, the ILM is placed within an SIR framework with a latent period of 0 days, and the infectious period, \( \gamma \), is 8 days for every individual. At time \( t = 1 \), all the individuals are assumed to be susceptible except one that is randomly chosen to be infectious. The
model parameters are set to be $\alpha = 0.5$, $\beta = 1.2$ and $\varepsilon = 0$. A typical simulated epidemic results in 986 infections and terminated at $T = 49$. Further, three under-reported individual-level data sets are obtained for each of the five epidemics with 25%, 50% and 75% reporting rates, respectively. In other words, only 25%, 50% and 75% cases are randomly recorded, respectively, among the total individual infection events; a data set with a 100% reporting rate is also analyzed.

Third, for each epidemic, each data set is aggregated at what is termed the "regional-level". 100 regions are defined, obtained by evenly dividing the $40 \times 40$ unit square area into 100 square cells (see Figure 2.1). The centre of each region is used to represent its regional coordinates. Additionally, the time point of the first individual being infected in a region is time which indicates the beginning of infection for a region, but here the time point at which the epidemic dies out for the region is assumed unknown. In terms of the modeling procedure, the mean infectious period for each infected region, $\gamma$, is thus unobserved and treated as an unknown parameter. Thus, the parameter vectors for the CE, EE, GE and WE models are given by $\Theta_1 = (\alpha, \beta, \varepsilon, \gamma)$, $\Theta_2 = (A_1, \lambda_1, \beta, \varepsilon, \gamma)$, $\Theta_3 = (A_2, \phi_2, \lambda_2, \beta, \varepsilon, \gamma)$, and $\Theta_4 = (A_3, \phi_3, \lambda_3, \beta, \varepsilon, \gamma)$, respectively.

2.3.3 Model Fitting Procedure

Each of the four models of Section 2.3.1 are fitted to the regional-level data. That means, individuals $i$ and $j$ in the models are actually regions $i$ and $j$, respectively.
The random walk Metropolis Hastings (MH) Markov chain Monte Carlo (MCMC) algorithm, as described in Section 2.2.3, is applied to sample realizations from the respective posterior distributions. In the modeling procedure, the parameters $\alpha$, $A_1$, $A_2$, and $A_3$ are transferred to be on the exponential-scale in all the fitted models (i.e., $\alpha = \exp\{\alpha\}', A_1 = \exp\{A_1\}', A_2 = \exp\{A_2\}', A_3 = \exp\{A_3\}'\}$). These transformations are found to allow for more efficient MCMC mixing. Vague prior distributions (independent positive half-normal distributions with mode 0 and variance 10,000) are chosen for all parameters.

With an attempt to improve the efficiency of the MCMC chain, correlated parameters are updated in blocks and the proposal densities are tuned with correlation structure that matches the underlying posterior distributions. For example, in the GE model, parameters $A_2', \phi_2, \lambda_2, \beta, \varepsilon$ are stronger correlated, and the unknown mean infectious period, $\gamma$, is not highly correlated with any of these parameters. Therefore, these correlated parameters are updated in one block, and then $\gamma$ is updated in a single parameter update. Here, for all models, uniform proposals are used for such single parameter update, and multivariate normal distributions are used as proposal distributions for all blocks of correlated parameters. After a burn-in period of 2,000 iterations, the Markov chain is run for another 1,000,000 iterations. In all the cases, convergence is verified visually for different starting values.
Table 2.1: Averaged Difference of Deviance Information Criterion (DIC)

<table>
<thead>
<tr>
<th>Infectivity Profile</th>
<th>Model</th>
<th>25% reporting</th>
<th>50% reporting</th>
<th>75% reporting</th>
<th>100% reporting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time-invariant Inf.</td>
<td>Constant Inf. (CE)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Gamma Inf. (GE)</td>
<td>-30.947</td>
<td>-45.636</td>
<td>-40.519</td>
<td>-36.866</td>
</tr>
<tr>
<td></td>
<td>Weibull Inf. (WE)</td>
<td>-25.815</td>
<td>-32.222</td>
<td>-21.500</td>
<td>-19.233</td>
</tr>
</tbody>
</table>

2.3.4 Results

The posterior means obtained for each individual simulated epidemic are not of great interest and so are not reported here. The posterior means, of course, differ substantially from the true parameter values, as the true model describes epidemic spread at the individual level, and the fitted models describe epidemic spread at the aggregated regional level. Instead, the fit of the various models is compared via the Deviance Information Criterion (DIC) suggested by Spiegelhalter et al. (2002). The model with the smallest DIC is to be preferred according to this criterion (Spiegelhalter et al., 2002). Table 2.1 contains the averaged difference of DIC values over the five simulated epidemics, for each of the four reporting rates tested. In this table, the time-invariant ILM (the CE model) is arbitrarily chosen as the reference, or baseline level, and the averaged difference of DICs between each of the TV-ILMs and the CE model are calculated for each reporting setting. So, for example, for 25% reporting, the DIC values of the EE model, the GE model, and the WE model are 33.574, 30.947, and 25.815 lower than that of the CE model, respectively; for other settings, the DIC values of TV-ILMs are all smaller than that of the corresponding CE models. For each of the regional-level data based on different reporting rates, the TV-ILMs
appear to perform much better than the corresponding time-invariant ILM (the CE model). These results seem to indicate that such TV-ILMs can better account for the time-varying nature of the infectious regions. Comparing three TV-ILMs, the results show that the average DIC values of the WE models are all higher than corresponding EE and GE models for each of the four reporting conditions. This indicates that the WE models may fit worse than the other two TV-ILMs. However, the GE models seem to perform slightly better than the corresponding EE models for 50%, 75%, and 100% reporting rates, where the absolute averaged differences of DIC values of the GE models are higher. This would suggest that the extra flexibility offered by the 3-parameter gamma curve or the 2-parameter exponential curve offers a tangible benefit.

The model fit is also assessed by checking the posterior predictive distributions of the epidemic curves for four different reporting rates. Specially, Figure 2.2 shows the means and associated 95% percentile intervals, of the posterior predictive distribution of cumulative cases over time along with the observed cumulative number of cases (epidemic 1 shown as an example). For all four reporting rates, the TV-ILMs obviously fit better than the associated time-invariant ILM except that the posterior predictive distribution of cumulative cases with posterior mean for the EE model under 50% reporting seems slightly deviate from the observed cumulative number of cases.

For all the four fitted models, spatial information is detectable despite of the data
Figure 2.2: Predictive cumulative number of cases with posterior means (dotted line) and 95% percentile intervals (blue lines) and the observed cumulative number of cases (black line) at the regional level.
aggregation. This can be seen in Figure 2.3, which shows the estimated exponential distance kernel under the posterior means for each model in each reporting scenario. The graphics show that the fitted distance kernels vary over relatively short distances for 25% reporting, 50% reporting, 75% reporting and 100% reporting, respectively. The WE model has a lower rate of decay, allowing for greater long distance infection ratio compared to the other three models at each reporting rate. Then, the spatial rate of decay detected at 25% reporting seems lower than the others, and the results for another three reporting scenarios are quite similar.

2.4 Case Study

Recent outbreaks of H1N1 influenza in 2009 have exhibited a terrible cost on public health (WHO, 2009b). In this study, different forms of ILMs and TV-ILMs are
applied to data from the 2009 H1N1 pandemic observed in the Greater Toronto Area (GTA). This data was recorded at the level of census regions. The main aim of this study is to explore the spatial-temporal dynamics of the epidemic and whether these dynamics can be modelled in a simple manner. So here, the census regions over the GTA are treated as "individuals" similar to the manner described in Section 2.3.1.

The detection of spatial information via such modeling may be complicated by following reasons. First, misspecification of infection times or removal times in the data may have a potential influence on the extractable spatial information. Second, the underreporting problem may lead to the situation where the number of reported cases is a poor indication of true infectivity for regions; for such problems, it is difficult to measure infectivity of a region over time. Third, and perhaps most obviously, data aggregation at the level of census regions may destroy spatial information. Here, the first two problems may be addressed by applying data augmented and reversible jump MCMC (see, for example, Jewell et al., 2009a). However we may wish to avoid such an approach due to its high computational demands. Therefore, the study here is focused on examining the use of TV-ILMs incorporating exponential infectivity, gamma infectivity, and Weibull infectivity at the regional level, respectively. Hopefully such models provide a better fit to the data than their time-invariant equivalences. It might also be hoped that such modeling at a regional level might be less sensitive to errors in recorded infection times and underreporting, while avoiding the massive computational burden that would be required by a full data augmented and/or
reversible jump MCMC analysis.

2.4.1 Influenza Data

*Individual-Level Data*

H1N1 influenza is characterized by flu-like symptoms which includes sudden onset of high fever, cough, headache, muscle and joint pain, severe malaise, sore throat and runny nose (*WHO*, 2009a). Once exposed, individuals go through a latent period of approximately 2 days after which they become infectious for a period that is estimated to be 8 days (*WHO*, 2009a). Here, the individual-level data regard the first wave of the 2009 H1N1 outbreak in the Greater Toronto Area (GTA) (the data are available at Dr. Dongmei Chen’s lab in Queen’s University). The data consist of hospitalization cases recorded from April to June 2009, and the associated information contain individuals’ postcodes, age, gender, onset-date (recorded/estimated by clinicians), data-collecting date and lab-testing date. A total of 1435 individual cases are identified; however, the dates of symptom onset for 1002 cases are not documented in the given time series. In this study, the lab-testing date is used to refer the approximate date of individual being exposed or infectious. No information in the data set allows for the identification of individual cases.

*Regional-Level Data*

The individual-level data are combined with 2006 census population data by matching
the postcode coordinates of infected individuals to the census region in which there was an infection. Dr. Dongmei Chen of Queen’s University provides the regional-level data, which are combined in the following way. Since the location data for each infectious individual are given only by a six digit postcode, an x-y location is defined for each individual based upon the centroid of their postcode region. This individual is then allocated to the particular census tract in which their postcode centroid falls. Of course, this can introduce an error, since there may be occasions where an individual’s actual residence is located in a different census tract to the one in which their postcode centroid is located. For this study, however, this error is assumed to have a negligible effect – this is discussed further in Section 2.5.

A total of 1003 regions are contained in the study region, and 642 regions had infection recorded within them. The date of the first individual being infected in a region is used as the date which indicates the beginning of infection for a region, and the date of the last individual being removed is used to indicate the end of the "infectious period" for this region. Since this information varies among regions, the infectious periods for regions are different. Other information, including number of infectious cases at each time point, population size, coordinates of centroid of the regions and region size, are recorded in the regional-level data as well.
The 2009 H1N1 influenza dynamics is modeled at the level of census regions with models that fall into one of two categories: time-invariant ILMs and TV-ILMs. Here, follow the assumption that the latent period and the infectious period for each person in a region are approximately 2 and 8 days, respectively (WHO, 2009a). The individual-level data has been aggregated within particular regions in which there are infection events so that the number of cases for those regions is recorded at each time point. In time-invariant ILMs, the number of infectious individuals can be incorporated to indicate the infectivity for each particular region. Once again, in the TV-ILMs, consider three different curves, these being the discrete exponential, gamma, and Weibull curves.

2.4.2.1 Time-invariant Infectivity Models

The ILMs described here are placed within an SEIS model framework, which means, regions go through the states from susceptible (S), exposed (E), infectious (I) and back to susceptible (S) again immediately when there are no exposed or infectious individuals being recorded in the regions. Five time-invariant ILMs are considered.

1. Constant Infectivity and Exponential Kernel Model (CE)

The first model is the same as described in equation (2.3.1), containing an exponential kernel model with time-invariant infectivity. Here, $\varepsilon(i, t) = \varepsilon = 0$, and the model is
given by:

\[ P_{it}^{(CE)} = 1 - \exp \left[ -\alpha \sum_{j \in I(t)} n_{jt} \exp \{ -\beta_1 d_{ij} \} \right] \]  

(2.4.1)

where, \( n_{jt} \) is the number of infectious cases for infectious regions \( j \) at time \( t \), and \( d_{ij} \) is the Euclidean distance between the centroid of susceptible region \( i \) and infectious region \( j \). The parameter vector is given by \( \theta = (\alpha, \beta_1) \).

2. Constant Infectivity and Geometric Kernel Model (CG)

In this second model, the exponential kernel is replaced with a geometric kernel \( k(i, j, t) = d_{ij}^{-\beta_2} \), a simple power law. \( \beta_2 \) is the spatial parameter.

\[ P_{it}^{(CG)} = 1 - \exp \left[ -\alpha \sum_{j \in I(t)} n_{jt} d_{ij}^{-\beta_2} \right] \]  

(2.4.2)

Here, the parameter vector \( \theta = (\alpha, \beta_2) \).

3. Population-size and Exponential Kernel Hybrid Model (PE)

Here, the CE model is extended to account for the population size of region. The model becomes:

\[ P_{it}^{(PE)} = 1 - \exp \left[ -\alpha m_{ri} \sum_{j \in I(t)} n_{jt} \exp \{ -\beta_1 d_{ij} \} \right] \]  

(2.4.3)
where $m_{ri}$ is the number of population recorded in the census for susceptible region $i$. Once again, the parameter vector is given by $\theta = (\alpha, \beta_1)$.

4. *Population-size Model with Non-spatial Information (PNSI)*

In the previous models, spatial information represented through the infection/distance kernel is incorporated. For the purpose of exploring whether regional-level spatial information has a significant effect on the spread of the H1N1 epidemic, a model with no spatial component is therefore considered. Specifically, the distance kernel is, thus, set to 1 in the model. The probability of a susceptible individual $i$ becoming infected within the time interval $[t, t + 1)$ is then given by:

\[
P_{it}^{(PNSI)} = 1 - \exp\left[ -\alpha m_{ri} \sum_{j \in I(t)} n_{jt} \right]
\] (2.4.4)

Here, the parameter vector is univariate with $\theta = \alpha$.

5. *Population-size and Nearest Neighbor Kernel Model (PNN)*

An alternative to the exponential kernel or geometric kernel is the nearest neighbor (NN) kernel. In this model, the infection kernel is now an indicator function $1(d_{ij} \leq r)$ in which $r$ denotes the neighborhood radius in kilometers. Here, the model is as follows:

\[
P_{it}^{(PNN)} = 1 - \exp\left[ -\alpha m_{ri} \sum_{j \in I(t)} n_{jt} 1(d_{ij} \leq r) \right]
\] (2.4.5)
Here, the parameter vector is $\theta = (\alpha, r)$.

### 2.4.2.2 Time-varying Infectivity Models

The ILM framework described here is now placed within an SEIR model framework as described in Section 2.2.1. A region becomes infectious when the first person goes through the latent period and becomes infectious; and the epidemic terminates in that region when the last person observed goes through the infectious period and is then removed from the disease. Five TV-ILMs are also considered. Once again, for the purpose of this study, the susceptibility function and infection kernel are assumed to be time-invariant; i.e., $\Omega_S(i, t) = \Omega_S(i)$ and $k(i, j, t) = k(i, j)$.

#### 6. Exponential Infectivity and Exponential Kernel Model (EE)

Here is a model in which the number of infectious individuals, $n_{jt}$, is replaced by a discrete exponential curve $\Delta_1(t)$. This model is the same as the EE model described in equation (2.3.2), but here $\varepsilon(i, t) = \varepsilon = 0$. The model becomes:

$$P_{it}^{(EE)} = 1 - \exp \left[ - \sum_{j \in I(t)} \Delta_1(t; A_1, \lambda_1, \tau_j) \exp\{-\beta_1 d_{ij}\} \right] \tag{2.4.6}$$

The parameter vector is $\theta = (A_1, \lambda_1, \beta_1)$.

#### 7. Exponential Infectivity and Geometric Kernel Model (EG)
Again, the exponential distance kernel is replaced by a geometric distance kernel.

\[
P^{(EG)}_{it} = 1 - \exp \left[ - \sum_{j \in I(t)} \Delta_1(t; A_1, \lambda_1, \tau_j) d_{ij}^{-\beta_2} \right] \tag{2.4.7}
\]

where, the parameter vector is given by \( \theta = (A_1, \lambda_1, \beta_2) \).

8. Exponential Infectivity and Spark Term Model (ES)

As described in Section 2.2.2, the spark term \( \varepsilon(i, t) \) is generally introduced to represent random behavior which cannot be explained by the main terms of the model. Here consider a model in which \( \varepsilon(i, t) = \varepsilon \) and the probability of a susceptible individual \( i \) becoming infected within the time interval \( [t, t+1) \) is defined as:

\[
P^{(ES)}_{it} = 1 - \exp \left[ - \left( \sum_{j \in I(t)} \Delta_1(t; A_1, \lambda_1, \tau_j) \right) - \varepsilon \right] \tag{2.4.8}
\]

Here, the parameter vector becomes: \( \theta = (A_1, \lambda_1, \varepsilon) \).

9. Gamma Infectivity and Spark Term Model (GS)

Here, the time-varying infectivity is modified to follow a discrete gamma curve (see equation 2.3.3). Combining with the spark term, the model is therefore:

\[
P^{(GS)}_{it} = 1 - \exp \left[ - \left( \sum_{j \in I(t)} \Delta_2(t; A_2, \phi_2, \lambda_2, \tau_j) \right) - \varepsilon \right] \tag{2.4.9}
\]
Here, \( \theta = (A_2, \phi_2, \lambda_2, \varepsilon) \) is the parameter vector to be estimated.

10. **Weibull Infectivity and Spark Term Model (WS)**

Another alternative is to use a discrete Weibull curve as the time-varying infectivity \( \Delta_3(t) \) (see equation 2.3.4). Here, the probability of a susceptible individual \( i \) becoming infected within the time interval \([t, t+1]\) is given as:

\[
P_{it}^{(WS)} = 1 - \exp \left[ -\left( \sum_{j \in I(t)} \Delta_3(t; A_3, \phi_3, \lambda_3, \tau_j) \right) - \varepsilon \right] \tag{2.4.10}
\]

\( \theta = (A_3, \phi_3, \lambda_3, \varepsilon) \) is the parameter vector.

**2.4.3 Model Fitting Procedure**

All the models are fitted within a Bayesian framework via MCMC techniques as described in Section 2.2.3. Positive half-normal prior distributions with mode 0 and a large variance of 10,000 are assigned to all the parameters.

Once again, the random walk Metropolis Hasting (MH) algorithm is employed for simulating from the posterior distribution of the parameters. For the time-invariant ILMs, uniform proposals are utilized to update parameters. For the TV-ILMs, correlated parameters are updated in one block and the multivariate normal proposal densities tuned to have correlation structure that matches the underlying posterior distributions. For example, in the EG model (see equation 2.4.6), the parameter \( A_1, \lambda_1 \) and \( \beta_1 \) are correlated, and thus are then updated together in one block.
2.4.4 Results

The random walk MH algorithm is run for 200,000 iterations after a burn-in period of 2,000 iterations for all the models and convergence is verified visually for different starting values. Figure 2.4 illustrates typical MCMC trace plots and marginal posterior distributions for Exponential Infectivity and Geometric Kernel Model (EG), as an example.

Table 2.2 shows the DIC results for the ten fitted models. It is seen that the five TV-ILMs obviously fit better than the other five time-invariant ILMs. The DICs for the CE and CG models are 7579.84 and 7576.41, respectively. The DICs for the EE and EG models are 6149.17 and 6146.61, respectively. Therefore, the


<table>
<thead>
<tr>
<th>Model</th>
<th>DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant Infectivity &amp; Exponential Kernel (CE)</td>
<td>7579.84</td>
</tr>
<tr>
<td>Constant Infectivity &amp; Geometric Kernel (CG)</td>
<td>7576.41</td>
</tr>
<tr>
<td>Population-size &amp; Exponential Hybrid (PE)</td>
<td>7385.65</td>
</tr>
<tr>
<td>Population-size Model with Non-spatial Information (PNSI)</td>
<td>7380.40</td>
</tr>
<tr>
<td>Population-size &amp; Nearest Neighbor Kernel (PNN)</td>
<td>7380.37</td>
</tr>
<tr>
<td>Exponential Infectivity &amp; Exponential Kernel (EE)</td>
<td>6149.17</td>
</tr>
<tr>
<td>Exponential Infectivity &amp; Geometric Kernel (EG)</td>
<td>6146.61</td>
</tr>
<tr>
<td>Exponential Infectivity &amp; Sparks Term (ES)</td>
<td>6118.91</td>
</tr>
<tr>
<td>Gamma Infectivity &amp; Sparks Term (GS)</td>
<td>5948.15</td>
</tr>
<tr>
<td>Weibull Infectivity &amp; Sparks Term (WS)</td>
<td>5921.83</td>
</tr>
</tbody>
</table>

performances for the exponential kernel and the geometric kernel are quite similar in terms of the DIC. Then, the time-invariant ILMs incorporating population size are observed to have slightly lower DIC values. However, the DIC values for both the PNSI and ES models are noticeably smaller than the models with exponential kernel or geometric kernel, which implies the spatial effect seems not to increase the goodness-of-fit significantly. Of the time-invariant infectivity models, the Nearest Neighbor (NN) kernel model seems to fit roughly as well as the PNSI model with a DIC value of 7380.37. Of the time-varying infectivity models, the lowest DIC value is given by the Weibull infectivity and spark term model (WS). The estimated posterior means and associated 95% percentile interval for the fitted models are reported in Table 2.3.
2.5 Discussion

In this Chapter the ILMs of Deardon et al. (2010) are generalized to allow for time-varying susceptibility, infectivity, and/or contact functions. Here, however, the main focus was the issue of time-varying infectivity due to aggregation over regions, in which three discrete curves—discrete exponential, gamma, and Weibull—were examined. A simulation study was carried out, and then various models applied to data from the H1N1 epidemic as observed in the Greater Toronto Area (GTA).

In the simulation study carried out, for all the four reporting rates, it is observed that the TV-ILMs performed better than the associated time-invariant ILMs in terms of DIC values. In general, it is also observed that the TV-ILMs with gamma infectivity (the GE models) outperformed the corresponding TV-ILMs with exponential infectivity (the EE models) or Weibull infectivity (the WE models). The exception to this is that the EE model fit appeared slightly better than the GE model at 25% reporting rate. Model comparison, carried out using the posterior predictive distribution of the cumulative number of infected regions, also seemed to show that the TV-ILMs generally perform better in terms of accuracy and precision than the associated time-invariant ILMs for all four reporting rates. Then, it does appear that the issue of underreporting at the level of individuals may have an influence on the model fit at the regional level for both time-invariant ILM and TV-ILMs. It is noticed that the model fit at 25% reporting seemed to perform much worse than those of other reporting settings. In addition, it is seen that both time-invariant ILM and TV-ILMs
Table 2.3: Estimated posterior means and associated 95% percentile interval

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Posterior Mean</th>
<th>95% PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE</td>
<td>α</td>
<td>$1.104 \times 10^{-4}$</td>
<td>$(9.570 \times 10^{-5}, 1.328 \times 10^{-4})$</td>
</tr>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>$4.355 \times 10^{-3}$</td>
<td>$(1.811 \times 10^{-4}, 1.148 \times 10^{-2})$</td>
</tr>
<tr>
<td>CG</td>
<td>α</td>
<td>$1.131 \times 10^{-4}$</td>
<td>$(9.768 \times 10^{-5}, 1.347 \times 10^{-4})$</td>
</tr>
<tr>
<td></td>
<td>$\beta_2$</td>
<td>$5.265 \times 10^{-3}$</td>
<td>$(4.923 \times 10^{-4}, 8.213 \times 10^{-3})$</td>
</tr>
<tr>
<td>PE</td>
<td>α</td>
<td>$2.785 \times 10^{-8}$</td>
<td>$(2.204 \times 10^{-8}, 3.518 \times 10^{-8})$</td>
</tr>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>$1.109 \times 10^{-2}$</td>
<td>$(3.574 \times 10^{-3}, 2.137 \times 10^{-2})$</td>
</tr>
<tr>
<td>PNSI</td>
<td>α</td>
<td>$2.076 \times 10^{-8}$</td>
<td>$(1.929 \times 10^{-8}, 2.229 \times 10^{-8})$</td>
</tr>
<tr>
<td>PNN</td>
<td>α</td>
<td>$2.076 \times 10^{-8}$</td>
<td>$(1.959 \times 10^{-8}, 2.202 \times 10^{-8})$</td>
</tr>
<tr>
<td></td>
<td>$r$</td>
<td>$234.796$</td>
<td>$(98.703, 475.791)$</td>
</tr>
<tr>
<td>EE</td>
<td>$A_1$</td>
<td>$4.166 \times 10^{-4}$</td>
<td>$(1.291 \times 10^{-4}, 8.490 \times 10^{-4})$</td>
</tr>
<tr>
<td></td>
<td>$\lambda_1$</td>
<td>$0.197$</td>
<td>$(0.148, 0.253)$</td>
</tr>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>$6.123 \times 10^{-3}$</td>
<td>$(1.291 \times 10^{-4}, 8.490 \times 10^{-4})$</td>
</tr>
<tr>
<td>EG</td>
<td>$A_1$</td>
<td>$7.583 \times 10^{-4}$</td>
<td>$(4.073 \times 10^{-4}, 1.286 \times 10^{-3})$</td>
</tr>
<tr>
<td></td>
<td>$\lambda_1$</td>
<td>$0.196$</td>
<td>$(0.152, 0.244)$</td>
</tr>
<tr>
<td></td>
<td>$\beta_2$</td>
<td>$0.237$</td>
<td>$(0.050, 0.439)$</td>
</tr>
<tr>
<td>ES</td>
<td>$A_1$</td>
<td>$1.836 \times 10^{-4}$</td>
<td>$(1.532 \times 10^{-4}, 2.165 \times 10^{-4})$</td>
</tr>
<tr>
<td></td>
<td>$\lambda_1$</td>
<td>$0.101$</td>
<td>$(0.065, 0.142)$</td>
</tr>
<tr>
<td></td>
<td>$\varepsilon$</td>
<td>$5.017 \times 10^{-3}$</td>
<td>$(4.495 \times 10^{-3}, 5.600 \times 10^{-3})$</td>
</tr>
<tr>
<td>GS</td>
<td>$A_2$</td>
<td>$1.186 \times 10^{-6}$</td>
<td>$(4.637 \times 10^{-7}, 2.444 \times 10^{-6})$</td>
</tr>
<tr>
<td></td>
<td>$\phi_2$</td>
<td>$23.027$</td>
<td>$(21.307, 23.870)$</td>
</tr>
<tr>
<td></td>
<td>$\lambda_2$</td>
<td>$6.008$</td>
<td>$(5.534, 6.324)$</td>
</tr>
<tr>
<td></td>
<td>$\varepsilon$</td>
<td>$2.600 \times 10^{-3}$</td>
<td>$(1.822 \times 10^{-3}, 3.458 \times 10^{-3})$</td>
</tr>
<tr>
<td>WS</td>
<td>$A_3$</td>
<td>$3.086 \times 10^{-5}$</td>
<td>$(7.944 \times 10^{-6}, 5.335 \times 10^{-5})$</td>
</tr>
<tr>
<td></td>
<td>$\phi_3$</td>
<td>$17.214$</td>
<td>$(12.590, 21.770)$</td>
</tr>
<tr>
<td></td>
<td>$\lambda_3$</td>
<td>$4.702$</td>
<td>$(3.103, 5.655)$</td>
</tr>
<tr>
<td></td>
<td>$\varepsilon$</td>
<td>$2.734 \times 10^{-3}$</td>
<td>$(1.512 \times 10^{-3}, 3.625 \times 10^{-3})$</td>
</tr>
</tbody>
</table>
could detect spatial information despite data aggregation at the level of regions.

In the H1N1 case study, for both time-invariant ILMs and TV-ILMs, the geometric distance kernel and exponential distance kernel were seen to perform quite similar in terms of DIC. It is observed that the time-invariant ILM with population size included (the PE model) had a slightly lower DIC than that of the CE model. However, the TV-ILMs with population size incorporated (which were also fitted) have not been included in this study since they did not show a lower DIC value compared to the associated TV-ILMs without accounting for population size. In term of model comparison, the TV-ILMs have noticeably lower DIC values than those of time-invariant ILM. However, the inclusion of a spatial kernel dose not seem to increase the fit, indicating that spatial information is not detectable, at least when data is aggregated at the level of census regions. It may be, however, that spatial patterns in the epidemic are observable within different census regions.

Some issues that might be explored in the future are now discussed. In the H1N1 case study in this Chapter, the latent period and the infectious period for each individual are assumed to be constant, and these data are then aggregated into the level of census regions. For TV-ILMs, a region becomes infectious when the first person goes through the latent period and becomes infectious; and that region is removed from the epidemic when the last person observed goes through the infectious period and is removed from the population. These assumptions about the latent and/or infectious period, however, could be relaxed. This may be tackled by data augmented
and/or reversible jump MCMC techniques (see, e.g., Jewell et al., 2009a; Cauchemez and Ferguson, 2011; Neal and Roberts, 2004), where variables of latent and infectious periods are treated as unknown parameters that need to be estimated; hence, the joint posterior distribution of the missing data and model parameters is sampled via MCMC technique. Reversible jump MCMC might also be used to incorporate the existence of hidden infections in the inference.

However, implementation difficulties can arise when a large number of extra parameters is introduced. Primarily, there is high potential that complex correlated relationships exist between the missing data, or between the missing data and the model parameters. As a result, the efficiency of the MCMC can be low so that a large number of iterations might be required to achieve stationarity of the chain (Gamerman and Lopes, 2006), which substantially increases the computational burden.

As previously mentioned, a potential spatial error may exist in the data due to the fact that individuals were allocated to census tracts according to the centroid of their postcode region. Of course, an individual may reside in a different census tract to that of the centroid of their postcode region. Here, this effect was assumed to be negligible, but it would be possible to allow for such spatial measurement error in future work. For example, Deardon et al. (2012) look at the problem of individual-level spatial measurement error in the recorded locations of both susceptible and infected individuals by introducing Gaussian spatial random effects. Such an approach could be adapted to model measurement error in the infectious population here. Alternatively,
a "nearest-neighbour" random effect could be introduced, allowing a proportion of infectious individuals recorded as being in each census tract to be shifted to neighbouring census tracts.

Additionally, census tract-level covariates that may display spatial heterogeneity between census tracts and potentially have an effect on transmission dynamics (e.g. age distribution, transportation network, and land-use pattern) could be included in the ILM in future work. The development of some other time-varying infectivity models, in which parameters of the time-varying infectivity curves are dependent upon some of these covariates, as well as covariates such as population size within a region, could be explored as well. Finally, some other TV-ILMs incorporating time varying susceptibility and/or contact functions, which account for varying infection levels in a region and/or population movement, could be examined.
CHAPTER 3

Comparing Bayesian Statistical Models of Infectious Disease Outbreaks via Probability Scoring

3.1 Introduction

Outbreaks of infectious disease, such as animal and plant disease, can exhibit a terrible cost on animal welfare, agricultural economy, and even public health. For example, during the 2001 foot-and-mouth disease (FMD) in the UK, more than 6 million animals were slaughtered across more than 10,000 farms, causing an economic cost of approximate £8 billion (Ward et al., 2004). Also, over 300,000 laboratory cases and 3917 deaths were confirmed in 191 countries and territories from the 2009 H1N1 pandemic influenza outbreak (WHO, 2009b).

Therefore, statistical infectious disease modeling has been emerging as an impor-

---

1Chapter 3 has been submitted to the peer-reviewed journal, Biometrics.
tant tool for gaining essential insights into the dynamics of infectious disease trans-
mission, and can help understanding associated risk factors and developing efficient
control strategies. For example, Lekone and Finkenstädt (2006) developed methods
of inference for the use of a chain-binomial model to study a 1995 Ebola outbreak
in the Democratic Republic of Congo. Statistical models have also been constructed
for modeling the dynamics of the 2001 foot-and-mouth disease (FMD) outbreak in
the UK (see, e.g., Keeling et al., 2001; Chis-Ster and Ferguson, 2007; Deardon et al.,
2010).

Individual-level models (ILMs) of infectious disease spread, as defined by Deardon
et al. (2010), are a system of models that can be used, to model the spatial-temporal
spread of infectious disease, incorporating various individual-level risk factors. Deardon
et al. (2010) applied these ILMs to the 2001 FMD outbreak carrying out the
modeling at the level of individual farms. In such a case, the associated risk factors
may include: geographic distance between animals; trading relationships; type and
number of animals on the farm; time, species, and number, of recent animal acquisi-
tions; biosecurity measures imposed, and etc..

Key to the profitable use of such models is the statistical task of fitting the ILM
to observed data, and validating that the model provides a sufficiently good fit to
that data. Failure to do this means predictions or conclusions drawn from the model
(e.g. how to control an outbreak) could be unreliable. Infectious disease models such
as ILMs are generally fitted within a Bayesian statistical frame using Markov chain
Monte Carlo. In a Bayesian context, a variety of methods such as the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002), Akaike’s Information Criterion (Akaike, 1974), Bayes factors, the predictive model choice criterion (PMCC) (Gelfand and Ghosh, 1998), and/or reversible jump MCMC (Neal and Roberts, 2004), can be applied to aid in the process of model comparison. Additionally, plots of the posterior predictive distribution of quantities of interest have been widely used as evidence of model adequacy (Gelman et al., 2000).

In this Chapter, the focus is on methods of comparing and assessing different forms of competing ILMs. The approach is to develop the tool of probability scoring rules (see, e.g., Matheson and Winkler, 1976; Winkler, 1996; Gneiting et al., 2007) to ascertain how well the model fits the data over time, based on the posterior predictive distributions obtained from MCMC output.

Winkler (1996) and Gneiting and Raftery (2007) define a probability scoring rule as a summary measure for evaluating probabilistic forecasts. It works by rewarding a numerical score based on the predictive distribution and on the event observed. There are essentially two roles of scoring rules: one is to encourage honest and careful forecasts; another is to provide a summary measure and to compare competing probabilistic forecasts. Under the Bayesian framework, Dawid (2006) relates probability score to utility, following the principle that the expected utility is maximized by the Bayesian decision. The application of probability scoring rules can be found in various scenarios, including count data (Czado et al., 2009), point forecast (Gneiting, 2011),
and time series data (Gneiting et al., 2007). Gschlößl and Czado (2005) also apply scoring rules, in a Bayesian framework, for modeling claim frequency and claim size in insurance.

Here, in the infectious disease transmission modeling context, the probability of an infection event, or a non-infection event (i.e. a non-infected individual remaining non-infected), occurring is modeled for each individual of the population at a discrete time point (representing a given time interval in continuous time). Therefore, a numerical score can be assigned treating the model as an expert attempting to predict infections or non-infections. The data is then used to verify the success of the model for each individual at each time point.

A simulation study is conducted to identify a probability scoring method sensitive to poor fit, both to the overall epidemic data set, and when considering the course of the epidemic over time. This can be accomplished by comparing the score-based measure of fit to situations where a misspecified model is fitted to data generated from some other model. Model comparison based upon three strictly proper scoring rules is then compared to one of the aforementioned approaches to model comparison, such as the DIC.

Model diagnostic or adequacy is often considered using plots of posterior predictive distribution of the cumulative number of infection events (or epidemic curve). The use of some other proper scoring rules is investigated based on these posterior predictive distributions and the observed number of cumulative cases, and the results
are compared to those of the PMCC. Here, the use of scoring rules is further proposed as a model fit criterion. Posterior predictive realizations of data are repeatedly simulated from the fitted model and the probability scores are calculated for each of the simulated epidemics as time progresses, giving an estimate of the posterior predictive distributions of the scores over time. Deviation of the scores over time for the original data is then checked for, being taken as evidence that the model fit is insufficient at certain times of the epidemic if found present.

Using the ILMs, data from the 2009 H1N1 influenza outbreak in Southern Ontario is analyzed. The modeling is carried out at the level of census regions in the Great Toronto Area (GTA). Then, the use of probability scoring rules is compared to the DIC approach in the evaluation of the fitted ILMs. And also, the posterior predictive distributions of the cumulative number of infected regions and the probability scores are simulated for model diagnostic.

This Chapter is organized as follows. An introduction of the individual-level models (ILMs) of Deardon et al. (2010) and a generalized time-varying individual-level model (TV-ILMs) are included in Section 3.2. In Section 3.3 the probability scoring rules and its relationship to a convex real-valued function are introduced, and then some examples of proper scoring rules are provided. The specific models and the simulation study are detailed in Section 3.4. The results regarding model comparison and model adequacy are presented in Section 3.5. The 2009 H1N1 influenza study with the use of scoring rules are detailed in Section 3.6, followed by a discussion about
some potential future research of interest in Section 3.7.

### 3.2 Individual-Level Model

Deardon et al. (2010) proposed a class of individual-level models (ILMs) for modeling epidemic transmission over time and space, based on various individual-level risk factors. The compartmental framework in which the ILMs of Deardon et al. (2010) are formulated is known as a susceptible-exposed-infectious-removed (SEIR) framework (Anderson and May, 1991), widely used in infectious disease epidemiology. In an SEIR framework, the status of an individual $i$ can be one of four states: susceptible (S), exposed (E), infectious (I) and removed (R), at any given time point. The susceptible state means that individuals are not infected, but can contract it. Being in the exposed state means that individuals are infected but not yet able to transmit the virus. The latent period is followed by the infectious state, indicating the individuals have become infectious and can transmit the disease. The final removed state implies removal from the population, either through recovery, acquired immunity, quarantine, or death. Models can also be fitted within other compartmental frameworks (e.g. SEIS, SIR, or SI); however, the focus of this Chapter is the SEIR framework.

Here, time is discretized so that time point $t$ represents a time interval $[t, t + 1)$. Let $S(t), E(t), I(t), R(t)$ denote the sets of individuals being susceptible, exposed, infectious and removed at $[t, t + 1)$, respectively. Let $T$ denote the last time point of
the observed epidemic.

### 3.2.1 The General ILMs

Let \( p_{it} \) represent the probability of a susceptible individual \( i \) becoming infected within \([t, t+1)\). A general form of the epidemic ILM is given by Deardon et al. (2010) as:

\[
p_{it} = 1 - \exp \left[ \{-\Omega_S(i) \sum_{j \in I(t)} \Omega_T(j)k(i,j)\} - \varepsilon(i,t) \right]
\]

(3.2.1)

where \( \Omega_S(i) \) is a susceptibility component containing risk factors associated with susceptible individual \( i \) contracting the disease; \( \Omega_T(j) \) is a transmissibility component containing risk factors associated with infectious individual \( j \) passing on the disease (such risk factors might include age, gender, the population size in different regions/farms, and environmental risk factors); \( k(i,j) \) is an infection kernel representing shared risk factors often based on the separation distance (e.g. Euclidean distance) between individuals \( i \) and \( j \), respectively; and \( \varepsilon(i,t) \) is a "spark" term, which is included to explain some other random infections that are not sufficiently explained by the previous terms in the model (e.g. infection from outside the study population).

An infection event is described as a new infection (if a susceptible individual \( i \) transits from the state \( S \rightarrow E \)) occurring within time interval \([t, t + 1)\), and a non-infection event indicates a susceptible individual \( i \) remaining infected within \([t, t + 1)\). The likelihood of the model is a product of all infection events and non-infection
events over the entire time period $t = 0, \ldots, T$, which is given by:

$$\pi(S,E,I,R;\theta) = f(S,E,I,R|\theta) = \prod_{t=0}^{T} f_t(S,E,I,R|\theta)$$

(3.2.2)

where, $S = \{S(t)\}_{t=1}^{T}$, $E = \{E(t)\}_{t=1}^{T}$, $I = \{I(t)\}_{t=1}^{T}$, $R = \{R(t)\}_{t=1}^{T}$ represent the sets of $S(t)$, $E(t)$, $I(t)$, $R(t)$ from the beginning to the last observed time point of the epidemic, respectively; $\theta$ is vector of parameters that are to be estimated; $f_t(S,E,I,R|\theta)$ is the product of the probability of all new infection events and of all non-infection events occurring at time $t$:

$$f_t(S,E,I,R|\theta) = \left[ \prod_{i \in E(t+1) \setminus E(t)} p_{it} \right] \left[ \prod_{i \in S(t+1)} \left( 1 - p_{it} \right) \right].$$

### 3.2.2 Time-varying Individual-level Models

Here, the ILMs of Deardon et al. (2010) are extended to allow for time-varying susceptibility, infectivity and infection kernel. The generalized time-varying individual level model (TV-ILM) is as follows:

$$p_{it} = 1 - \exp \left[ \left\{ -\Omega_S(i,t) \sum_{j \in I(t)} \Omega_T(j,t) k(i,j,t) \right\} - \varepsilon(i,t) \right]$$

(3.2.3)

where, $\Omega_S(i,t)$ represents a susceptibility function of potential risk factors related to susceptible individual $i$ contracting the disease during time interval $[t, t+1]$; $\Omega_T(j,t)$ is the transmissibility function of potential risk factors related to infectious individual
j passing on the disease in $[t, t + 1]$; $k(i, j, t)$ is a time-varying infection kernel representing shared risk factors often based on the separation distance (e.g. Euclidean distance), between individuals $i$ and $j$ in $[t, t + 1]$.

### 3.3 Probability Scoring Rules

Winkler (1969) proposed probability scoring rules as a tool to evaluate the performance of probabilistic forecasts. A scoring rule works by rewarding a numerical value according to the predictive distribution and the events actually observed (Gneiting et al., 2007). For infectious disease modeling, ILMs function by modeling the probability of an infection event, or a non-infection event occurring, for each member of the population in a given time interval. Therefore, the ILM can be treated as an expert trying to predict infections or non-infections, and be rewarded a numerical score for its prediction. Then, probability scores can be calculated based on the posterior predictive distribution under the ILM and the infection or non-infection event observed for a susceptible individual $i$ over time.

#### 3.3.1 Proper Probability Scoring Rules

Scoring rules considered in this Chapter all have positive orientation; in other words, a larger score is preferred and a probabilistic forecaster wishes to achieve maximal score. Let $G(P, x)$ denote a probability score in which $P$ represents a probabilistic forecast and $x$ is an event observed. $G(P, \cdot)$ is any extended real-valued
function. Assume then, that if $R$ is the underlying true distribution of the events, a scoring rule $G$ is said to be proper if

$$G(P, R) \leq G(R, R) \quad (3.3.1)$$

for all $P$ and $R$. Here, $G(P, R)$ is the expected score of $G(P, \cdot)$ under $R$. If the expected score is maximized only when $P = R$, the scoring rule $G$ is said to be strictly proper (Czado et al., 2009; Gneiting and Raftery, 2007). That is, a maximal expected score is obtained if and only if the probabilistic forecast $P$ represents the true distribution of the events, $R$. In this sense, one role of strictly proper scoring rules is to encourage honest and coherent predictions by the probabilistic forecaster. This is a fundamental property for the application of scoring rules to scientific and operational forecast assessment (Czado et al., 2009).

Another role of scoring rules is to provide numerical measure and to rank competing probabilistic forecasts. Czado et al. (2009) calculate the mean score to measure the quality of fit; say

$$\mathcal{G} = \frac{1}{n} \sum_{i=1}^{n} G(P_i, x_i) \quad (3.3.2)$$

where, $P_i$ and $x_i$ are the $i$th predictive distribution and the $i$th event observed, respectively.
3.3.2 Proper Scoring Rules and Convex Functions

In order to develop a proper scoring rule $G$, Gneiting and Raftery (2007) propose and prove a general theorem which relates a proper scoring rule $G$ to a convex, real-valued function $H$. The theorem states that if,

$$G(P, x) = H(P) - \int H^*(P, x) \, dP(x) + H^*(P, x)$$  \hspace{1cm} (3.3.3)

where, $H^*(P, .)$ is a sub tangent of $G$ at the probabilistic forecast $P$, and $H$ is strictly convex, then $G$ is a strictly proper scoring rule. This theorem can be stated simply as: (strictly) proper scoring rule can be derived from any (strictly) convex, real-valued function, $H(P)$ (Winkler, 1996).

Categorical Variables

Savage Representation (Savage, 1971). Suppose the sample $x_1, ..., x_n$ are mutually exclusive events, and $p_1, ..., p_n$ are the corresponding probabilistic forecasts. For categorical variables, following Gneiting and Raftery (2007), a scoring rule $G$ is proper if and only if

$$G(p, x_i) = H(p) - \langle H'(p), p \rangle + H'(p, x_i), \text{ for } i = 1, \ldots, n.$$  \hspace{1cm} (3.3.4)

where, $p$ refers to the probability vector $(p_1, ..., p_n)$; $H$ is a convex function; $H'(p)$ is a subgradient of $H$ at $p$; and $\langle \cdot, \cdot \rangle$ represents the standard scalar product. This
representation is a special case of equation (3.3.3) for categorical forecasts.

**Binary Variables**

The Savage representation (see equation 3.3.4) can be further simplified and applied to a binary situation. Consider a sample space $\Omega = \{1, 0\}$, and let $p$ denote the probability that the event will occur. A scoring rule $G$ assigns the forecaster a probability score $G(p, 1)$ for the event to occur and $G(p, 0)$ for its complement, which is as follows (Gneiting and Raftery, 2007):

$$G(p, 1) = H(p) + (1 - p)H'(p),$$

$$G(p, 0) = H(p) - pH'(p)$$

where, $H$ is also a convex function and $H'(p)$ is a subgradient of $H$ at the point $p$.

### 3.3.3 Examples of proper scoring rules

Note that, all the probability scoring rules considered in this Chapter are positively oriented; in other words, the model with the higher score is preferred. For the ILMs, the overall score-based goodness-of-fit measure is averaged over epidemic time,

$$\bar{G} = \frac{1}{T} \sum_{t=1}^{T} \bar{G}(t)$$

where $T$ denotes the last day of the observed infection; and $\bar{G}(t)$ represents the average score calculated under posterior means under the one-step ahead predictive
Table 3.1: Three examples of Strictly Proper Scoring Rules

<table>
<thead>
<tr>
<th>Scores</th>
<th>Categorical variable</th>
<th>Binary variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quadratic Score (QS)</td>
<td>( QS(p, x_i) = 2p_i - \sum_{j=1}^{n} p_j^2 )</td>
<td>( QS(p, 1) = 1 - (1 - p)^2 )</td>
</tr>
<tr>
<td></td>
<td>( H(p) = \sum_{j=1}^{n} p_j^2 - 1 )</td>
<td>( QS(p, 0) = 1 - p^2 )</td>
</tr>
<tr>
<td>Spherical Score (SS)</td>
<td>( SS(p, x_i) = \frac{p_i}{\left(\sum_{j=1}^{n} p_j^2\right)^{1/2}} )</td>
<td>( SS(p, 1) = p(1 - 2p + 2p^2)^{-1/2} )</td>
</tr>
<tr>
<td></td>
<td>( H(p) = \left(\sum_{j=1}^{n} p_j^2\right)^{1/2} )</td>
<td>( SS(p, 0) = (1 - p)(1 - 2p + 2p^2)^{-1/2} )</td>
</tr>
<tr>
<td>Logarithmic Score (LS)</td>
<td>( LS(p, x_i) = \log(p_i) )</td>
<td>( LS(p, 1) = \log(p) )</td>
</tr>
<tr>
<td></td>
<td>( H(p) = \sum_{j=1}^{n} p_j \log(p_j) )</td>
<td>( LS(p, 0) = \log(1 - p) )</td>
</tr>
</tbody>
</table>

distribution from time point \( t \) to \( t + 1 \),

\[
G(t) = \frac{1}{N_t} \sum_{i,t} G(p_{it}, x_{it} \mid E[\theta \mid S, E, I, R])
\]

here, \( N_t \) is the number of all new infection events and all non-infection events from time \( t \) to \( t + 1 \), and \( E[\theta \mid S, E, I, R] \) is the posterior mean of the parameters.

Several probability scoring rules are discussed in *Gneiting and Raftery* (2007). Table 3.1 shows three examples of strictly proper scoring rules for probability forecasts of categorical variable and the respective convex, real-valued function \( H(p) \), and the examples for probability forecasts of binary variable.

Further, three proper scoring rules are also investigated: the continuous ranked probability score (CRPS), the interval score at level \( \alpha \) (\( \alpha.\text{IS} \)) and the quantile score at level \( \alpha \) (\( \alpha.\text{QS} \)). Let \( F \) denote the predictive cumulative distribution function, then the continuous ranked probability score (CRPS) is defined as,

\[
\text{CPRS}(F, x) = -\int_{-\infty}^{\infty} (F(z) - 1 \cdot (x \leq z))^2 \, dz
\]
where \((F(z) - 1(x \leq z))^2\) is equivalent to the Brier scores for the probabilistic prediction of the binary variable at the real-valued threshold \(z\) (Hersbach, 2000; Gneiting and Raftery, 2007). Further, the CRPS can be alternatively represented as, according to Székely and Rizzo (2005):

\[
CRPS(F, x) = \frac{1}{2} E_F |X - X'| - E_F |X - x|
\]

where \(X\) and \(X'\) denote independent replicated values from a predictive distribution function \(F\).

The quantile score at level of \(\alpha\) \((\alpha\text{.QS})\) is given by Gneiting and Raftery (2007), as:

\[
\alpha\text{.QS}(r_\alpha, x) = (x - r_\alpha)(1(r_\alpha \geq x) - \alpha)
\]

where \(r_\alpha\) is the quantile at the level \(\alpha\) of the predictive distribution.

Analogous to the quantile score, suppose \(l\) and \(u\) be the respective \(\frac{\alpha}{2}\) and \(1 - \frac{\alpha}{2}\) predictive quantiles, then the interval score \((\alpha\text{.IS})\) is defined as:

\[
\alpha\text{.IS}(l, u, x) = \alpha(u - l) - 2(l - x)1(x < l) - 2(x - u)1(u < x)
\]

There is no restrict rule as to how to choose a scoring rule in any given situation, when the decision problem is not uniquely defined (Czado et al., 2009). Here, some characteristics relevant to different scoring rules are briefly reviewed. For instance, the logarithmic score \((\text{LS})\) is subject to the predictive distribution \(P\) through the
probability assigned to the observed event alone (Winkler, 1996; Czado et al., 2009).
A scoring rule with this characteristic is called a local scoring rule (Winkler, 1996).
According to this, since the spherical and quadratic scores are dependent on the whole
probability distribution, they are not local scoring rules. Further, for situations aside
from a simple binary event, the LS is the exclusive strictly proper scoring rule which
is local (Winkler, 1996). In addition, it has a close relationship with the predictive
deviance which is defined as (Czado et al., 2009):

\[ D(P, x) = -2 \log p_x + C. \]

The CRPS, \(\alpha.IS\) and \(\alpha.QS\) are straightforward to implement when considering
posterior predictive distribution from MCMC (Gschlößl and Czado, 2005).

### 3.4 Simulation Study

The primary objective here is to develop the tool of proper probability scoring
rules to compare different forms of competing ILMs and to assess how well the ILMs
fit the epidemic data over time, as well as globally. Here, epidemics are simulated
from a geometric spatial model with exponentially time-varying infectivity, and then
fit the true generated model and various misspecified models to the simulated data.
In this study, for the time-varying individual-level models (TV-ILMs), our analysis
is restricted to the time-varying infectivity, and three discrete curves, the discrete
3.4.1 Model Description

The eight, mostly spatial ILMs used in this study, are first described.

Model 1. Exponential infectivity and Geometric Kernel Model (EG)

In the first model, consider a time-varying infectivity $\Omega_T(j, t)$ (defined in Section 3.2.2) given by a discrete exponential curve, and constant susceptibility, $\Omega_S(i, t) = \Omega_S(i) = 1$. The probability of a susceptible individual $i$ contracting diseases at time $t$ is given by:

$$p^{(EG)}_{it} = 1 - \exp \left( - \sum_{j \in I(t)} \Delta_1(t; A_1, \lambda_1, \tau_j) d_{ij}^{-\beta_1} \right)$$ (3.4.1)

where, $\Delta_1(t)$ is an exponential function:

$$\Omega_T(j, t) = \Delta_1(t; A_1, \lambda_1, \tau_j) = A_1 \exp \{-\lambda_1(t - \tau_j)\}$$

$\tau_j$ is the time point at which individual $j$ becomes infectious; $A_1$ is a scale parameter; $\lambda_1$ is a rate parameter; the distance kernel, $k(i, j) = d_{ij}^{-\beta_1}$, in which $\beta_1$ is the spatial parameter and $d_{ij}$ is the Euclidean distance between susceptible individual $i$ and infectious individual $j$. The parameter vector is given by $\theta = (A_1, \lambda_1, \beta_1)$.

Model 2. Exponential Infectivity and Exponential Kernel Model (EE)

The same time-varying infectivity $\Delta_1(t)$ as in (3.4.1) is used in this model, but here the
infection (distance) kernel is changed to the exponential kernel \( \exp\{-\beta_2 d_{ij}\} \), where \( \beta_2 \) is the exponential decay parameter of the infection kernel. Therefore, the model becomes:

\[
p^{(EE)}_{it} = 1 - \exp\left[ - \sum_{j \in I(t)} \Delta_1(t; A_1, \lambda_1, \tau_j) \exp\{-\beta_2 d_{ij}\} \right]
\]  

(3.4.2)

Here, the parameter vector is given by \( \theta = (A_1, \lambda_1, \beta_2) \).

**Model 3. Gamma Infectivity and Geometric Kernel Model (GG)**

Here, the time-varying infectivity is modified to follow a discrete gamma curve, and once again a geometric distance kernel is used. The gamma kernel model (GG) is given by:

\[
p^{(GNSI)}_{it} = 1 - \exp\left[ - \sum_{j \in I(t)} \Delta_2(t; A_2, \phi_2, \lambda_2, \tau_j) d_{ij}^{-\phi_1} \right]
\]  

(3.4.3)

where,

\[
\Omega_T(j, t) = \Delta_2(t; A_2, \phi_2, \lambda_2, \tau_j) = A_2(t + 1 - \tau_j)^{\phi_2 - 1} \exp\{-\lambda_2(t + 1 - \tau_j)\}
\]

\( A_2 \) is a scale parameter, \( \phi_2 \) is a shape parameter and \( \lambda_2 \) is a rate parameter. \( \theta = (A_2, \phi_2, \lambda_2, \beta_1) \) is the parameter vector to be estimated.

**Model 4. Gamma Infectivity Model with Non-spatial Information (GNSI)**

Once again using the discrete gamma infectivity curve, a model with no spatial ele-
ment (i.e. \( k(d_{ij}) = 1 \)) is given by:

\[
\begin{align*}
    p_{it}^{(GNSI)} &= 1 - \exp \left[ - \sum_{j \in I(t)} \Delta_2(t; A_2, \phi_2, \lambda_2, \tau_j) \right] \\
\end{align*}
\] (3.4.4)

Here, \( \theta = (A_2, \phi_2, \lambda_2) \) is the parameter vector to be estimated.

---

**Model 5. Constant Infectivity and Geometric Kernel Model (CG)**

In this model, the infectivity is assumed to not vary over time; that is, a constant infectivity is allowed over the infectious period; The geometric kernel is used, giving the model:

\[
\begin{align*}
    p_{it}^{(CG)} &= 1 - \exp \left[ -\alpha \sum_{j \in I(t)} d^{-\beta_1}_{ij} \right] \\
\end{align*}
\] (3.4.5)

where, \( \alpha \) is a susceptibility parameter. Here, the parameter vector \( \theta = (\alpha, \beta_1) \).

---

**Model 6. Constant Infectivity and Homogeneous Model (CH)**

Here, a model with constant infectivity and no spatial element (i.e. \( k(d_{ij}) = 1 \)) is considered. Thus, the infectivity is only based on the number of infectious individuals within \([t, t + 1)\) and, therefore, each susceptible individual in the population has an equal probability of being infected at any given time. Specifically the CH model is given by:

\[
\begin{align*}
    p_{it}^{(CH)} &= 1 - \exp \left[ -\alpha \sum_{j \in I(t)} 1 \right] \\
\end{align*}
\] (3.4.6)
where the parameter vector is univariate with $\theta = \alpha$.

**Model 7. Constant Infectivity and Exponential Kernel Model with Misspecified Infectious Period (CEMIP)**

Here, a model with constant infectivity, exponential distance kernel, and a misspecified infectious period, is introduced. The probability of a susceptible individual $i$ becoming infected within $[t, t + 1)$ is as follows:

$$
    p_{it}^{(CEMIP)} = 1 - \exp \left[ -\alpha \sum_{j \in I(t)} \exp\{-\beta_2 d_{ij}\} \right]
$$

(3.4.7)

The parameter vector is given by $\theta = (\alpha, \beta_2)$.

**Model 8. Weibull Infectivity and Exponential Kernel Model with Misspecified Infectious Period (WEMIP).**

Finally, a discrete Weibull curve for the time-varying infectivity is used. Here also, the infectious period in the model is assumed to be misspecified. Thus, the model is as follows:

$$
    p_{it}^{(WEMIP)} = 1 - \exp \left[ - \sum_{j \in I(t)} \Delta_3(t; A_3, \phi_3, \lambda_3, \tau_j) \exp\{-\beta_2 d_{ij}\} \right]
$$

(3.4.8)
where,

\[
\Omega_T(j, t) = \Delta_3(t; A_3, \phi_3, \lambda_3, \tau_j) = A_3 \left[ \frac{t + 1 - \tau_j}{\lambda_3} \right]^{\phi_3 - 1} \exp \left[ - \left( \frac{t + 1 - \tau_j}{\lambda_3} \right)^{\phi_3} \right]
\]

\(A_3\) is also a scale parameter, \(\phi_3\) is a shape parameter and \(\lambda_3\) is a scale parameter.

Here, the parameter vector is \(\theta = (A_3, \phi_3, \lambda_3, \beta_2)\).

### 3.4.2 Epidemic Simulation

Assuming there are 4 spatial clusters in the population, the random spatial locations of 200 individuals are generated from truncated bivariate Gaussian clusters on a 10×10 grid. The Gaussian mixture model used is:

\[
f(x|\theta) = \sum_{i=1}^{4} \frac{1}{4} g(x|\mu_i, \Sigma_i)
\]

where, \(x = \{(x_1, x_2)^T : x_1 \in [0, 10], x_2 \in [0, 10]\}\), \(\mu_1 = (2.5, 2.5)^T\), \(\Sigma_1 = \begin{pmatrix} 3.5 & 0 \\ 0 & 4.5 \end{pmatrix}\), \(\mu_2 = (3.0, 7.5)^T\), \(\Sigma_2 = \begin{pmatrix} 4.0 & 0 \\ 0 & 3.5 \end{pmatrix}\), \(\mu_3 = (7.5, 4.5)^T\), \(\Sigma_3 = \begin{pmatrix} 3.5 & 0 \\ 0 & 4.0 \end{pmatrix}\), \(\mu_4 = (7.0, 7.0)^T\), \(\Sigma_4 = \begin{pmatrix} 3.5 & 0 \\ 0 & 4.5 \end{pmatrix}\); and \(g(x|\mu_i, \Sigma_i)\) is a bivariate Gaussian density with mean \(\mu_i\) and covariance matrix \(\Sigma_i\). A typical simulated population is shown in Figure 3.1. As can be seen, the 4 clusters heavily overlap.

Second, an epidemic through this population is simulated from the EG model of equation (3.4.1) with parameters values \(A_1 = 2.5\), \(\lambda_1 = 1.2\) and \(\beta_1 = 6.0\). The latent and the infectious period for each individual are assumed to be constant, and are set
Figure 3.1: A typical simulated heterogeneous population on a 10×10 grid.

to be 2 and 5 days, respectively.

Further, the EG model (3.4.1) (the true model that generated the observed data) and the models of equations (3.4.2) to (3.4.8) which are misspecified models are fitted, to the simulated epidemic. Thus, the effect of a misspecified infectivity profile (e.g. gamma infectivity, Weibull infectivity, and time-invariant infectivity), misspecified distance kernel (e.g. exponential distance kernel) and/or misspecified infectious period, are being considered. Here, the misspecified infectious period is assumed to be 20 days. Ten epidemics are generated from the exponential infectivity model (EG), all based on the spatial locations of 200 individuals given in Figure 3.1.

3.4.3 Modeling Procedure

Vague prior distributions for all parameters are chosen: independent positive half-normal prior distributions with mode 0 and a large variance of 10,000. Note that, the parameters are restricted to be non-negative real numbers.
A random walk Metropolis-Hastings (MH) Markov Chain Monte Carlo (MCMC) algorithm is employed for simulating from the posterior distribution of the parameters. For the gamma infectivity model with non-spatial information (GNSI), in order to allow for more efficient MCMC convergence, the scale parameter is exponentially transferred in all the 10 simulated epidemics (i.e. $A_2 = \exp(A'_2)$). Then, correlated parameters are updated in blocks, in which the proposal densities are tuned with correlation structure that approximately matches the underlying posterior distributions so that the efficiency of the MCMC chain can be improved. For example, in the WEMIP model, parameters $A'_3, \phi_3, \lambda_3, \beta_2$ are found to be highly correlated, so these parameters are updated in one block. For the model with constant infectivity, uniform proposals are used. For the models with time-varying infectivity, multivariate normal distributions are used as proposal distributions for all blocks of correlated parameters. For each epidemic data set, the random walk MH algorithm is run for 100,000 iterations after a burn-in period of 2,000 iterations for all the models. Convergence is verified visually.

### 3.5 Simulation Results

Typical MCMC trace plots after burn-in and marginal posterior distributions for the Exponential Infectivity and Geometric Kernel Model (EG) are illustrated in Figure 3.2 as an example. The estimated posterior means and the associated 95% percentile intervals for the fitted models are reported in Table 3.2.
Table 3.2: Estimated posterior means and associated 95% percentile interval

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Posterior Mean</th>
<th>95% PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>EG</td>
<td>$A_1$</td>
<td>2.600</td>
<td>(1.830, 3.561)</td>
</tr>
<tr>
<td></td>
<td>$\lambda_1$</td>
<td>2.575</td>
<td>(1.294, 4.489)</td>
</tr>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>6.053</td>
<td>(5.374, 6.745)</td>
</tr>
<tr>
<td>EE</td>
<td>$A_1$</td>
<td>81.198</td>
<td>(38.806, 147.060)</td>
</tr>
<tr>
<td></td>
<td>$\lambda_1$</td>
<td>2.552</td>
<td>(1.292, 4.437)</td>
</tr>
<tr>
<td></td>
<td>$\beta_2$</td>
<td>3.724</td>
<td>(3.270, 4.177)</td>
</tr>
<tr>
<td>GG</td>
<td>$A_2$</td>
<td>96.208</td>
<td>(21.108, 232.008)</td>
</tr>
<tr>
<td></td>
<td>$\phi_2$</td>
<td>2.480</td>
<td>(0.257, 4.858)</td>
</tr>
<tr>
<td></td>
<td>$\lambda_2$</td>
<td>3.476</td>
<td>(2.144, 4.563)</td>
</tr>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>6.011</td>
<td>(5.374, 6.700)</td>
</tr>
<tr>
<td>GNSI</td>
<td>$A_2$</td>
<td>84.276</td>
<td>(9.114, 228.957)</td>
</tr>
<tr>
<td></td>
<td>$\phi_2$</td>
<td>9.617</td>
<td>(5.974, 12.519)</td>
</tr>
<tr>
<td></td>
<td>$\lambda_2$</td>
<td>8.292</td>
<td>(6.340, 9.604)</td>
</tr>
<tr>
<td>CG</td>
<td>$\alpha$</td>
<td>1.945</td>
<td>(1.401, 2.623)</td>
</tr>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>6.417</td>
<td>(5.783, 7.094)</td>
</tr>
<tr>
<td>CH</td>
<td>$\alpha$</td>
<td>$5.174 \times 10^{-3}$</td>
<td>(4.479 $\times 10^{-3}$, 5.918 $\times 10^{-3}$)</td>
</tr>
<tr>
<td>CEMIP</td>
<td>$\alpha$</td>
<td>78.130</td>
<td>(35.946, 148.237)</td>
</tr>
<tr>
<td></td>
<td>$\beta_2$</td>
<td>4.004</td>
<td>(3.526, 4.491)</td>
</tr>
<tr>
<td>WEMIP</td>
<td>$A_3$</td>
<td>170.838</td>
<td>(91.078, 281.975)</td>
</tr>
<tr>
<td></td>
<td>$\phi_3$</td>
<td>2.559</td>
<td>(1.099, 4.610)</td>
</tr>
<tr>
<td></td>
<td>$\lambda_3$</td>
<td>1.140</td>
<td>(0.710, 1.435)</td>
</tr>
<tr>
<td></td>
<td>$\beta_2$</td>
<td>3.545</td>
<td>(3.143, 3.944)</td>
</tr>
</tbody>
</table>
3.5.1 Model Comparison

Table 3.3 shows the results of quadratic score (QS), spherical score (SS) and logarithmic score (LS) averaged over 10 trials based on the overall data sets. Once again, all the probability scores in this Chapter are positively oriented; which means the model with higher score is favored. These results are compared to those from the Deviance Information Criterion (DIC) suggested by Spiegelhalter et al. (2002). Here, the model with the smallest DIC is preferred according to this criterion.

Comparing the results from these three probability scores to the DIC values, the highest values of the scores are all obtained for the true fitted model (EG); however, the lowest value of DIC is assigned to the misspecified gamma infectivity and geometric kernel model (GG). In addition, for the DIC result, the rank of the WEMIP model
is slightly higher than the EE model, with values of 363.9281 and 365.5282, respectively. Second, among these three strictly proper scores, the ranks of the misspecified ILMs for the averaged QS and the averaged SS are in the same order, but a slight different result is observed for the averaged LS where the scores (rank) of the GG, EE and WEMIP models are -0.1021 (2), -0.1023 (3) and -0.1026 (4), respectively.

In terms of the misspecified infectivity profile, since the true infectivity profile is assumed to vary over time and follow a discrete exponential curve in the simulated epidemic, it is expected that the fitted time-varying infectivity model should perform better than the time-invariant infectivity model. It is indeed observed that the time-varying EG and GG models have higher scores than the associated constant infectivity CG model. Also, the time-varying EE and WEMIP models are assigned higher scores compared to the CEMIP model. For the non-spatial models, the time-varying GNSI model performs better than the constant infectivity CH model. Additionally, all the probability scores favor the true exponential infectivity EG and EE models over the misspecified time-varying infectivity GG and WEMIP models, respectively.

With respect to the misspecified distance kernel, it appears that the misspecified exponential distance kernel in model with correctly specified infectivity (EE) does indeed have a slightly lower score than the true EG model. Then, for models with the same infectivity, the GG and CG models outperform the respective GNSI and CH models, respectively.

Considering the course of an epidemic over time, Figure 3.3 shows the averaged
Table 3.3: The averaged DIC values and averaged quadratic score (QS), spherical score (SS) and logarithmic score (LS) over 10 simulated epidemics for the true fitted model EG and the seven misspecified models, with performance rankings in parentheses.

<table>
<thead>
<tr>
<th>Model</th>
<th>Averaged DIC</th>
<th>Averaged QS</th>
<th>Averaged SS</th>
<th>Averaged LS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential Inf. &amp; Geometric Kernel (EG)</td>
<td>363.5283 (2)</td>
<td>0.9699 (1)</td>
<td>0.9671 (1)</td>
<td>-0.1014 (1)</td>
</tr>
<tr>
<td>Exponential Inf. &amp; Exponential Kernel (EE)</td>
<td>365.4486 (4)</td>
<td>0.9698 (2)</td>
<td>0.9670 (2)</td>
<td>-0.1023 (3)</td>
</tr>
<tr>
<td>Gamma Inf. &amp; Geometric Kernel (GG)</td>
<td>360.0046 (1)</td>
<td>0.9697 (3)</td>
<td>0.9668 (3)</td>
<td>-0.1021 (2)</td>
</tr>
<tr>
<td>Gamma Inf. &amp; Non-spatial Info. (GNSI)</td>
<td>1025.7550 (7)</td>
<td>0.9137 (7)</td>
<td>0.9055 (7)</td>
<td>-0.2902 (7)</td>
</tr>
<tr>
<td>Constant Inf. &amp; Geometric Kernel (CG)</td>
<td>382.4056 (5)</td>
<td>0.9677 (5)</td>
<td>0.9644 (5)</td>
<td>-0.1072 (5)</td>
</tr>
<tr>
<td>Constant Inf. &amp; Homogeneous (CH)</td>
<td>1121.7950 (8)</td>
<td>0.9094 (8)</td>
<td>0.9020 (8)</td>
<td>-0.3174 (8)</td>
</tr>
<tr>
<td>Constant Inf. &amp; Exponential Kernel with MIP (CEMIP)</td>
<td>401.1674 (6)</td>
<td>0.9676 (6)</td>
<td>0.9643 (6)</td>
<td>-0.1091 (6)</td>
</tr>
<tr>
<td>Weibull Inf. &amp; Exponential Kernel with MIP (WEMIP)</td>
<td>363.9281 (3)</td>
<td>0.9697 (3)</td>
<td>0.9668 (3)</td>
<td>-0.1026 (4)</td>
</tr>
</tbody>
</table>

probability score lines of the quadratic score (QS), spherical score (SS) and logarithmic score (LS) with posterior means under the one-step ahead predictive distribution over time for the true fitted EG (black line), misspecified GG (green dotted line) and CG (blue dotted line), respectively. As shown, these score lines tend to decrease and diverge as the epidemic progresses as larger number of "events" need to be predicted. It is obvious that both time-varying infectivity models (EG and GG) have higher scores than the constant infectivity model (CG) over time. Although only a tiny difference is observed between the scores of the EG model and the misspecified GG model as time progresses, the score line of the true EG model (black line) is slightly higher than that of the GG model (green dotted line). The exponential curve is a special case of the gamma curve and so this tiny difference is perhaps not surprising. More alarming is that the DIC implies a slight preference for the misspecified GG model rather than the EG model. Similar results are observed for all three strictly proper scoring rules in Figures 3.3(a) to 3.3(c).
Figure 3.4(a) shows the average quadratic score lines with posterior means under the one-step ahead predictive distribution over time for the true fitted model EG (black line), the misspecified distance kernel model EE (orange line), the misspecified model WEMIP (blue line) and the CEMIP model (green line). With the same exponential infectivity, the misspecified EE model gives lower probability scores than the true EG model over time. Figure 3.4(b) shows the average quadratic score lines with posterior means under the one-step ahead predictive distribution over time for the true EG, and the misspecified GNSI and CH models, respectively. These two non-spatial models appear to have a much worse fit than the true model during the entire epidemic progress.

3.5.2 Model Diagnostic

3.5.2.1 Posterior Predictive Distribution of the cumulative number of cases

Figure 3.5 presents a typical example of the posterior predictive distribution of the cumulative number of cases under the various fitted models together with the observed cumulative number of cases over time. It seems that the performances of the true EG model and the GG model are better than the others. The figures show that the observed cumulative numbers of cases for both the GNSI and CH models generally fall out of the 95% credible bands of the posterior predictions, respectively.

Here, the predictive model choice criterion (PMCC) proposed by Gelfand and
Figure 3.3: The quadratic score (QS), spherical score (SS), and logarithmic score (LS) with posterior means under the one-step ahead predictive distribution over time for the true EG model (black line), the misspecified GG model (green dotted line), the misspecified CG model (blue dotted line).
Figure 3.4: (a) The quadratic score lines with posterior means from the EG model (black line), the misspecified EE model (orange line), the misspecified WEMIP model (blue line) and the misspecified CEMIP model (green line). (b) The quadratic score lines with posterior means from the EG model (black line), the misspecified GNSI model (green line) and the CH model (blue line) over time, respectively.

Ghosh (1998), the continuous ranked probability score (CRPS), the quantile score ($\alpha.QS$) and the interval score ($\alpha.IS$) are applied to reward numerical values for each model in the same context. Here, the PMCC is treated as a positively oriented score. The results averaged over the 10 trials are shown in Table 3.4. As the distribution of the cumulative number of cases is not explicitly available, the PMCC, CRPS, the quantile score ($\alpha.IS_{\alpha=0.1,0.5}$) and the interval score ($\alpha.QS_{\alpha=0.75,0.95}$) are each approximated using the obtained MCMC samples. These approximations are based on 2,000 posterior predictive simulations.

First, the ranks from the PMCC, CRPS, $\alpha.IS_{\alpha=0.5}$ and $\alpha.QS_{\alpha=0.75}$ are generally in accord with each other. Second, comparing the PMCC to the CRPS, the PMCC does not agree that the rank of the EE model is higher than the WEMIP model.
Table 3.4: The averaged values of PMCC, CRPS, α.IS(α = 0.1, 0.5), α.QS(α = 0.75, 0.95) over the 10 trials for the true fitted model EG and the 7 misspecified models, with performance, rankings in parentheses.

as indicated in the CRPS. Simultaneously, comparing the α.IS(0.5) and α.QS(0.75) scores to the CRPS, both of them assign the highest values to the misspecified GG model rather than the true EG model, which is ranked second by both.

It appears that when using the PMCC, CRPS, α.IS(0.5) and α.QS(0.75) misspecification of the infectivity profile and kernel can generally be detected. With respect to the misspecified infectivity profile, first of all, the results indicate that the time-varying infectivity models outperform the constant infectivity model when considering the same distance kernel. In addition: the time-varying EG and GG models obtain higher scores than the constant CG model; the time-varying EE and WEMIP models have larger scores than the CEMIP model; and the score of the GNSI model is greater than the CH model. Further, the true exponential infectivity model (EG) is assigned higher scores than the misspecified time-varying infectivity profile model (GG) when using the PMCC and CRPS. Additionally, in terms of the misspecified distance kernel, the EE model with true exponential infectivity but misspecified exponential distance kernel appears to have lower scores than the true fitted EG model.
Meanwhile, the GNSI and CH models obviously perform worse than the respective GG and CG models.

However, although the $\alpha.IS_{\alpha=0.1}$ and $\alpha.QS_{\alpha=0.95}$ scores also suggest the same three misspecified models as being poorest, both of them fail in terms of rewarding the true model the highest score, and award the misspecified EE, WEMIP and CG models higher scores than the true EG model. This may be attributed to the fact that the $\alpha.IS$ and $\alpha.QS$ award a penalty if the observed cumulative number of cases are not covered within the corresponding interval or beyond the quantile at the level of $\alpha$.

### 3.5.2.2 Posterior Predictive Distribution of the probability scores

Here, the use of probability scoring rules is also proposed and examined as a model fit diagnostic. Epidemics are simulated repeatedly from the model under parameters drawn from the posterior distribution, and the probability scores as time progresses is calculated for each of the simulated epidemics. This results in a Monte Carlo estimate of the posterior predictive distributions of scoring rules over time. Further, the posterior predictive distribution is simulated in two different ways: one is to generate the whole epidemic progress; the second is using one-step-ahead simulation at each time point, given the current and past observed data. Three examples of strictly proper scoring rules—quadratic score (QS), spherical score (SS) and logarithmic score (LS) are considered in this section.

An example of such a model fit diagnostic is shown in Figure 3.6, in which the
The 95% posterior predictive interval under the whole epidemic simulation. That is, the score lines for the observed epidemic apparently fall below the model fit is insufficient at certain times of the epidemic. The first row is for the true EG model and the 7 misspecified models, with performance, rankings in parentheses, based on the posterior predictive distribution of quadratic score over time.

Table 3.5: The averaged values of PMCC, CRPS, α.IS(α = 0.1, 0.5), α.QS(α = 0.75, 0.95) over the 10 trials for the true EG model and the 7 misspecified models, with performance, rankings in parentheses, based on the posterior predictive distribution of quadratic score over time.

<table>
<thead>
<tr>
<th>Model</th>
<th>Averaged PMCC</th>
<th>Averaged CRPS</th>
<th>Averaged α.IS(_{0.1})</th>
<th>Averaged α.IS(_{0.5})</th>
<th>Averaged α.QS(_{0.75})</th>
<th>Averaged α.QS(_{0.95})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(EG)</td>
<td>-4.986 \times 10^{-4} (5)</td>
<td>-1.620 \times 10^{-4} (5)</td>
<td>-2.554 \times 10^{-4} (5)</td>
<td>-5.944 \times 10^{-4} (4)</td>
<td>-9.688 \times 10^{-4} (5)</td>
<td>-2.480 \times 10^{-4} (5)</td>
</tr>
<tr>
<td>(EE)</td>
<td>-4.622 \times 10^{-4} (6)</td>
<td>-1.915 \times 10^{-4} (6)</td>
<td>-2.925 \times 10^{-4} (4)</td>
<td>-7.926 \times 10^{-5} (6)</td>
<td>-10.145 \times 10^{-5} (6)</td>
<td>-3.571 \times 10^{-5} (6)</td>
</tr>
<tr>
<td>(GG)</td>
<td>-4.658 \times 10^{-4} (4)</td>
<td>-1.518 \times 10^{-4} (4)</td>
<td>-6.141 \times 10^{-5} (6)</td>
<td>-5.879 \times 10^{-4} (3)</td>
<td>-9.315 \times 10^{-4} (4)</td>
<td>-2.27 \times 10^{-4} (4)</td>
</tr>
<tr>
<td>(GNSI)</td>
<td>-2.937 \times 10^{-4} (7)</td>
<td>-29.049 \times 10^{-4} (7)</td>
<td>-74.95 \times 10^{-4} (7)</td>
<td>-119.58 \times 10^{-4} (7)</td>
<td>-194.12 \times 10^{-4} (7)</td>
<td>-30.358 \times 10^{-4} (7)</td>
</tr>
<tr>
<td>(CG)</td>
<td>9.226 \times 10^{-4} (1)</td>
<td>-1.375 \times 10^{-4} (1)</td>
<td>-2.739 \times 10^{-4} (1)</td>
<td>-5.126 \times 10^{-4} (1)</td>
<td>-7.298 \times 10^{-4} (1)</td>
<td>-2.768 \times 10^{-4} (1)</td>
</tr>
<tr>
<td>(CH)</td>
<td>-471.215 \times 10^{-4} (8)</td>
<td>-81.255 \times 10^{-4} (8)</td>
<td>-132.063 \times 10^{-4} (8)</td>
<td>-172.525 \times 10^{-4} (8)</td>
<td>-128.614 \times 10^{-4} (8)</td>
<td>-31.025 \times 10^{-4} (8)</td>
</tr>
<tr>
<td>(CEMIP)</td>
<td>-3.241 \times 10^{-4} (2)</td>
<td>-1.401 \times 10^{-4} (2)</td>
<td>-2.461 \times 10^{-4} (2)</td>
<td>-3.541 \times 10^{-4} (2)</td>
<td>-7.401 \times 10^{-4} (2)</td>
<td>-3.305 \times 10^{-4} (2)</td>
</tr>
<tr>
<td>(WEMIP)</td>
<td>-4.95 \times 10^{-4} (3)</td>
<td>-1.573 \times 10^{-4} (3)</td>
<td>-2.479 \times 10^{-4} (3)</td>
<td>-6.081 \times 10^{-4} (3)</td>
<td>-8.293 \times 10^{-4} (3)</td>
<td>-2.975 \times 10^{-4} (2)</td>
</tr>
</tbody>
</table>

posterior predictive mean, and the associated 95% credible bands, of the probability score over time are compared to the score line approximated from the observed epidemic under that fitted model. Once again, deviations here may provide evidence that the model fit is insufficient at certain times of the epidemic. The first row is for the true fitted EG model. Here, it seems that the score lines approximated from the observed epidemic under the first 5 models all fall within the respective 95% posterior predictive interval for all the three different scores, and only slight discrepancies among these 5 models for each score can be observed. However, for the misspecified GNSI and CH models, the score lines for the observed epidemic apparently fall below the 95% posterior predictive interval under the whole epidemic simulation. That means the method of using the probability scores as a model fit diagnostic is able to distinguish between the true and two of the least well-fitting models. It also appears that the one-step ahead prediction tends to reduce the variation of the posterior predictive distribution of the scoring rules.
However, the PMCC, CRPS, the quantile score ($\alpha.IS_{\alpha=0.1,0.5}$) and the interval score ($\alpha.QS_{\alpha=0.75,0.95}$) are also calculated to provide a numerical measure for each model based on the posterior predictive distribution of the quadratic, spherical, and logarithmic score, respectively. Only the results averaged over the 10 trials for the posterior predictive distribution of the quadratic score (QS) over time are reported in Table 3.5, since similar results hold for the posterior predictive distribution of the spherical and logarithmic scores. For the whole epidemic, it is observed that the highest values of the scores are all given to the constant infectivity and geometric kernel model (CG). Also, except for the $\alpha.IS_{\alpha=0.5}$ score, all the scores generally result in the same rank of the fitted models. More worryingly, the rank shows that the scores for the CEMIP and WEMIP models are slightly higher than those for the corresponding true EG model. This may be explained by the fact that these scores tend to reward narrow posterior predictive distributions (see Figure 3.6). Meanwhile, the models under the one-step ahead prediction generally obtain lower absolute numerical scores comparing to the associated score under the whole epidemic simulation.

In general, it seems that the scores obtained from the posterior predictive distribution of the scoring rules fail to rank correctly the true EG model relative to the misspecified EE, GG, CG, CEMIP and WEMIP models. It therefore appears that the method of using the posterior predictive distribution of the scoring rules might not be as useful as the posterior predictive distribution of the cumulative number of cases for assessing poor model fit. However, it does seem that using the posterior
predictive distribution of the scoring rules still flags up problems with the GNSI and CH models quite starkly.

3.6 Case Study

The objective in this case study is to illustrate the use of probability scoring rules for assessing competing ILMs which are fitted to data from the 2009 H1N1 epidemic observed in the Greater Toronto Area (GTA). In contrast to the simulation study, the "individual-level" at which the epidemic transmission is modeled refers to an aggregated set of individual units; here, census regions over the GTA. In such situations, it is reasonable to assume that the infectivity of the census regions varies over time as the infectious status of the individuals within the regions changes. Therefore, two time-varying infectivity ILMs are fitted, with comparison to two time-invariant infectivity ILMs. It may expected that such TV-ILMs can better fit the data than their time-invariant equivalences.

3.6.1 Influenza Data

*Individual-level Data.* The individual-level data are consist of hospitalization cases recorded in the GTA from April to June 2009, and on individuals' postcodes, onset-date (recorded or estimated by clinicians), data-collecting date and lab-testing date (the data are available at Dr. Dongmei Chen’s lab in Queen’s University). A total of 1435 individual cases are identified. In this study, the lab-testing date is used to
approximate the date of individuals being exposed or infectious. No information in
the data set allows for the identification of individual cases.

Regional-level Data. The individual-level data are combined with 2006 census popu-
lation data by matching the postcode coordinates of infected individuals to the census
region in which there was an infection. A total of 1003 regions are contained in the
study region and 642 regions had infections recorded within them. Information, in-
cluding number of infectious cases reported each day, population size, coordinates of
centroid of the regions and region size, are recorded. These regional-level data are
provided by Dr. Dongmei Chen of Queen’s University.

3.6.2 Model Descriptions

Here, the latent period and the infectious period for each person in a region
are estimated to be approximately 2 and 8 days, respectively (WHO, 2009a). The
individual-level data has been aggregated within particular regions in which there are
infection events so that the number of cases for those regions is recorded at each time
point. For the time-invariant infectivity models, the ILMs are placed within an SEIS
framework, which means, regions go through infection states from susceptible (S),
exposed (E), infectious (I), and then back to susceptible (S) again, as soon as there
are no exposed or infectious individuals reported in the region. Additionally, the
number of observed infectious individuals in each region is incorporated to indicate
the infectivity for each particular regions. For these models, the latent and infectious
periods are assumed to be two and eight days, in line with the individual WHO assumptions mentioned previously.

For the time-varying infectivity models, the ILMs are placed within an SEIR framework. The date on which the first individual becomes infected indicates the beginning of infection for a region, and the date on which the last individual observed is removed from the disease indicate the end of the "infectious period" for this region.

The time-invariant ILMs are placed in an SEIS framework in order to allow for regions to become reinfected after the infectious period for a previously observed infection has passed. For the time-varying ILMs it is assumed no longer necessary and that re-infections are accounted for in the time-varying infectivity function. An alternative to placing the time-invariant ILMs in an SEIS framework would, of course, be to place them in an SEIR framework and allow an estimated infectious period to be much longer than eight days.

The following models are fitted to the regional-level data.

Model 1. Constant Infectivity & Exponential Kernel Model (CE). This model is similar to equation (3.4.7).

\[
p_{it}^{(CE)} = 1 - \exp \left[ -\alpha \sum_{j \in I(t)} n_{jt} \exp\{-\beta_2 d_{ij}\} \right]
\]

where, \( n_{jt} \) is the number of infectious cases for infectious regions \( j \) at time \( t \), and \( d_{ij} \) is the Euclidean distance between susceptible region \( i \) and infectious region \( j \).

### Table 3.6: The DIC values and quadratic score (QS), spherical score (SS) and logarithmic score (LS) for the time-invariant infectivity CE and PNSI models and the time-varying infectivity EG and GS models, with performance, rankings in parentheses.

<table>
<thead>
<tr>
<th>Model</th>
<th>DIC</th>
<th>QS</th>
<th>SS</th>
<th>LS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant Inf. &amp; Exponential Kernel (CE)</td>
<td>7579.840 (4)</td>
<td>9.825 × 10^{-1} (4)</td>
<td>9.823 × 10^{-1} (4)</td>
<td>−8.8876 × 10^{-1} (4)</td>
</tr>
<tr>
<td>Population-size Model with Non-spatial Info (PNSI)</td>
<td>7480.403 (3)</td>
<td>9.826 × 10^{-1} (3)</td>
<td>9.824 × 10^{-1} (3)</td>
<td>−8.6848 × 10^{-1} (3)</td>
</tr>
<tr>
<td>Exponential Inf. &amp; Geometric Kernel (EG)</td>
<td>6146.612 (2)</td>
<td>9.827 × 10^{-1} (2)</td>
<td>9.825 × 10^{-1} (1)</td>
<td>−8.5035 × 10^{-1} (2)</td>
</tr>
<tr>
<td>Gamma Inf. &amp; Spark term (GS)</td>
<td>5948.149 (1)</td>
<td>9.828 × 10^{-1} (1)</td>
<td>9.825 × 10^{-1} (1)</td>
<td>−8.2455 × 10^{-1} (1)</td>
</tr>
</tbody>
</table>

\[ p_{it}^{(PNSI)} = 1 - \exp \left[ -\alpha m_i \sum_{j \in I(t)} n_{jt} \right] \]

where, \( m_i \) is the number of population recorded in the census for susceptible region \( i \).

**Model 3. Exponential Infectivity & Geometric Kernel Model (EG).** This model is the same as described in equation (3.4.1).

**Model 4. Gamma Infectivity & Spark Term Model (GS)**

Once again \( k(d_{ij}) = 1 \), and here a sparks term \( \varepsilon(i, t) = \varepsilon \) is included. This model with gamma infectivity (see equation 3.4.3) is given by:

\[ p_{it}^{(GS)} = 1 - \exp \left[ -\sum_{j \in I(t)} \Delta_2(t; A_2, \phi_2, \lambda_2, \tau_j) - \varepsilon \right] \]

### 3.6.3 Results

Regarding model comparison, the DIC, quadratic score (QS), spherical score (SS) and logarithmic score (LS) for the four fitted models are reported in Table 3.6. First of all, the results from the probability scores are compared to those from DIC. It is
Table 3.7: The values of PMCC, CRPS, $\alpha_{IS}(\alpha = 0.1, 0.5)$, $\alpha_{QS}(\alpha = 0.75, 0.95)$ for the time-varying infectivity EG and GS model based on the posterior predictive distribution of cumulative number of infected regions over time.

noticed that the ranks under the probability scores are essentially in agreement with the ranks of the DIC. Secondly, these results indicate that both time-varying infectivity models (EG and GS) perform better than the time-invariant infectivity models (CE and PNSI). Of the time-varying infectivity models, the GS model has higher QS and LS scores compared to the EG model. Of the time-invariant infectivity models, the scores for the PNSI model are higher than the models with the exponential kernel, which implies the inclusion of a spatial effect does not appear to significantly improve model fit. Here, the logarithmic score (LS) seems to be more sensitive to fitted ILM chosen.

In terms of model diagnostic, the posterior predictive distribution of the cumulative number of infected regions is first generated by using the method of one-step ahead prediction under the fitted EG and GS models, respectively. Plots of the posterior predictive mean of the cumulative number of infected regions, along with the associated 95% confidence bands, are illustrated in Figure 3.7(a) and 3.7(b). Under the one-step ahead prediction, the observed cumulative number of infected regions obviously deviate from the corresponding posterior predictive prediction for both the EG and GS models after about day 35. However, the posterior predictive perfor-
mance under the GS model seems to outperform the posterior prediction under the EG model for almost all the first 35 days.

Based on these posterior predictive simulations and the observed cumulative number of infected regions, the results of the aforementioned PMCC, CRPS, $\alpha.IS_{\alpha=0.1,0.5}$ and $\alpha.QS_{\alpha=0.75,0.95}$ scores are shown in Table 3.7. Although the GS model results in lower values of PMCC, all the other five scores reward the GS model a higher score in comparison to the EG model. In other words, all these five scores favor the time-varying gamma infectivity model over the exponential infectivity model, which is consistent with results in Table 3.6.

Probability scoring rules are also applied as a model fit diagnostic in this case study (see Section 3.5.2.2). The plots for the quadratic score (QS) are shown in Figure 3.7(c) and 3.7(d). Plots for the spherical score (SS) and logarithmic score (LS) are quite similar to those of the QS (results not shown). In general, these two plots are consistent with the posterior predictive plots of the cumulative number of infected regions shown in Figure 3.7(a) and 3.7(b), respectively. Over about the first 35 days, the performance of the QS under the GS model is generally better than those under the EG model. Especially before day 20, the score line approximated from the observed regional-level data under the EG model tends to fall below the 95% posterior predictive credible bands of the QS. At around day 40, both score lines for the observed regional-level epidemic begin to increase. It is also evident in Figure 3.7(a) and 3.7(b) that the posterior predictive distribution of the cumulative
Table 3.8: The values of PMCC, CRPS, $\alpha$.IS($\alpha = 0.1, 0.5$), $\alpha$.QS($\alpha = 0.75, 0.95$) for the time-varying infectivity EG and GS model based on the posterior predictive distribution of quadratic score over time.

<table>
<thead>
<tr>
<th>Model</th>
<th>PMCC</th>
<th>CRPS</th>
<th>$\alpha$.IS(0.1)</th>
<th>$\alpha$.IS(0.5)</th>
<th>$\alpha$.QS(0.75)</th>
<th>$\alpha$.QS(0.95)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(EG)</td>
<td>$-22.895 \times 10^{-5}$</td>
<td>$-16.716 \times 10^{-4}$</td>
<td>$-41.551 \times 10^{-4}$</td>
<td>$-6.956 \times 10^{-3}$</td>
<td>$-6.048 \times 10^{-3}$</td>
<td>$-1.495 \times 10^{-4}$</td>
</tr>
<tr>
<td>(GS)</td>
<td>$-9.439 \times 10^{-5}$</td>
<td>$-7.526 \times 10^{-4}$</td>
<td>$-9.858 \times 10^{-4}$</td>
<td>$-2.967 \times 10^{-3}$</td>
<td>$-3.657 \times 10^{-3}$</td>
<td>$-1.192 \times 10^{-4}$</td>
</tr>
</tbody>
</table>

number of infected regions appears to deviate from the observed cumulative number of infected regions after around day 40. This implies none of the models fit the tail of the epidemic particularly well.

Table 3.8 shows the results of the PMCC, CRPS, $\alpha$.IS($\alpha = 0.1, 0.5$), $\alpha$.QS($\alpha = 0.75, 0.95$) based on the posterior predictive distribution of the quadratic score (QS) over time and the score line approximated from the observed regional-level epidemic and the fitted EG and GS models, respectively. It is obvious that all the scores for the GS model are higher compared to those for the EG model, in accord with the results in Table 3.7. The results based on the posterior predictive distribution of the spherical and logarithmic score are not reported, but they all agree with the results in Table 3.8.

Note the posterior predictive distribution of both the cumulative number of infected regions and the probability scores under the time-invariant ILMs are not considered since it is difficult to simulate re-infections under these models.
3.7 Discussion

In this Chapter the tool of probability scoring rules was developed as a method for comparing and validating different forms of competing ILMs under a Bayesian statistical framework. The scoring rule assigns a numerical value to a model according to the posterior predictive distribution and the data observed. Essentially, each model is treated as an expert, forecasting the probability of an infection, or non-infection, event occurring for each susceptible individual of the population at each time point. Here, three strictly proper scoring rules – quadratic, spherical and logarithmic – were examined. In terms of model diagnostic, the use of the PMCC, CRPS, $\alpha.IS_{\alpha=0.1,0.5}$ and $\alpha.QS_{\alpha=0.75,0.95}$ scores for each model were examined.

Such probability scoring rules could be successfully used as a tool for model comparison when fitting various individual level models of infectious disease spread. This can be done both globally, obtaining scores averaged over the entire epidemic, and also locally, over time as the epidemic unfolds. The latter might prove useful in situations where parameters vary over the course of an epidemic, helping to pinpoint such heterogeneity. This is an area that may be of interest for future work. The use of probability scoring rules for model comparison may have an advantage over tools such as the DIC, in that they may be more intuitive for the non-statistician. They also appear to have the advantage that, unlike the DIC for example, they are not dependent upon underlying assumptions about the posterior distribution (e.g. symmetry). They certainly share the advantage, along with criterion such as the DIC,
of being computationally cheap compared to the "gold standard" method of using reversible-jump MCMC (Richardson and Green, 1997) to calculate posterior model probabilities.

Attempts to use the posterior predictive distribution of probabilities scoring rules to ascertain model adequacy did not appear to work in as satisfactory a manner. Although, what might be termed extremely misspecified models fell foul of such an analysis, comparing the posterior predictive distribution over various misspecified models often lead to the conclusion that a misspecified model was more likely to have generated the observed data than the true model. This did not appear to be the case using the more conventional method of examining the posterior predictive distribution of the epidemic curves.

Here are some future research of interest that might be explored. First, the tool of probability scores might be further applied to compare and assess different fitted ILMs when accounting for heterogeneity over space as well as time. For example, transmission parameters may change over space if control policies vary over regions. Second, some other time-varying ILMs could be developed in which time-varying susceptibility function and/or time-varying infection kernel are incorporated to account for the varying infection levels and/or population movement between regions. Moreover, the use of time-varying infectivity models in which the parameters of the time-varying infectivity curves depend on some other covariates (e.g. population size within a region) could be examined.
Figure 3.5: Predictive cumulative number of cases with posterior means (dotted line) and 95% percentile intervals (blue lines) and the observed cumulative number of cases (black line) for all the fitted models over time.
Figure 3.6: Posterior prediction of the quadratic score (QS), spherical score (SS) and logarithmic score (LS) with posterior means and 95% percentile intervals (black lines) over time and the observed score line (dotted line) approximated from the observed epidemic under the corresponding fitted model.
Figure 3.7: (a) and (b) The predictive cumulative number of cases under the fitted EG and GS models with posterior means (dotted line) and 95% percentile intervals (blue lines) and the observed cumulative number of cases (black line) over time, respectively. (c) and (d) The posterior prediction of the quadratic score (QS) with posterior means (dotted line) and 95% percentile intervals (blue lines) over time and the observed score line (black line) approximated from the observed epidemic under the EG and GS models, respectively.
4.1 Introduction

Statistical infectious disease modeling is essential for providing insights into the dynamics of epidemic spread. To design efficient strategies for outbreak control it is critical to have the risk factors governing the transition dynamics identified and evaluated. Recently, statistical models have been widely applied in various scenarios of infectious disease epidemiology. For example, Numminen et al. (2013) developed a variant of the so-called approximate Bayesian computation approach for parameterizing household models to investigate different strains of *S.pneumoniae* in a study conducted in day care units in Oslo, Norway. Lawson (2006) applied a space-time interaction model based on individual-level data in an analysis of a historic measles epidemic in Germany. Cook et al. (2007) investigated the spread of *Heracleum man-
tegazzianum (Giant Hogweed) in Britain, identifying several risk factors via a stochastic spatial-temporal model.

Individual-level models (ILMs), as discussed by Deardon et al. (2010), are a framework of discrete-time stochastic models which allow for the spatial-temporal modeling of epidemic dynamics. These models can incorporate individual-level risk factors on which the risk of disease transmission depends, for example, the distance between individuals. Such models were applied by Vrbik et al. (2012) to study the spread of an experimental fire, where the "individuals" were a set of cells constituting a piece of burning wax paper. These ILMs are commonly fitted within a Bayesian framework and the posterior distributions of parameters of interest can be estimated via Markov chain Monte Carlo (MCMC) techniques.

However, the modeling of epidemic spread is usually complicated by the challenges of collecting complete epidemiological outbreak data. For example, the date of symptom onset for an infectious individual may be observed, but the date of this individual being infected is rarely known. Several of the associated risk factors may be unobserved not only for the infectious individuals, but also for individuals remaining uninfected. Such problems can be tackled through the use of a data-augmented MCMC approach (see, e.g., Jewell et al., 2009a; Cauchemez and Ferguson, 2011; Neal and Roberts, 2004).

There are also some situations in which disease transmission is actually modeled at a level that consists of an aggregated set of units (e.g. farms, census regions, or
health units). For example, for the 2001 foot-and-mouth (FMD) disease in the UK, several studies were carried out at the individual farm level rather than animal level, often including the number and type of animals (e.g. sheep and cattle) on each farm as covariates in the models (see, e.g., Keeling et al., 2001; Jewell et al., 2009b; Deardon et al., 2010). In such cases, some other regional-level risk factors may include: geographic distance between farms; trading relationships; environmental factors; time, species, and number, of recent animal acquisitions; biosecurity measures imposed.

In the sphere of public health, regional-level data can be obtained by aggregating human-level disease prevalence data so that "individual-level" modeling can be implemented at the regional level, with regional-level prevalence and demographic covariates being utilized to inform the model. In such situations, since the total number of infectious individuals within that aggregated unit changes over time, a reasonable assumption is that the infectivity/transmissibility, of aggregated units varies during the course of regional-level infectiousness.

The general ILMs of Deardon et al. (2010) can be extended to account for the modeling of time-varying susceptibility, infectivity, and/or contact functions, leading to so-called "time-varying individual-level models" (TV-ILMs). For example, we might want to allow for time-varying susceptibility when the control measures (e.g. changing vaccination status) in a region are introduced, time-varying infectivity when the infectious status change over time within a region, and/or time-varying infection kernel to account for meteorological factors or population movement restrictions.
Here, the use of time-varying infectivity focusing on regional-level modeling is specifically explored after data has been spatially aggregated. An assumption is made that the regional-level infectivity varies during the course of infectiousness of a region, and follows some discrete curve.

In this Chapter a simulation study, in which the individual-level data is recorded under various underreporting scenarios, is conducted. Then the data are spatially aggregated to some "regional" level. Two global time-varying infectivity ILMs (TVI-ILMs) (discrete exponential/gamma infectivity) that assume homogeneous infectivity curves for regions are explored. Then, two TVI-ILMs in which parameters of the discrete exponential/gamma infectivity curves are dependent upon some regional-level risk factor are introduced; thus, the time-varying infectivity curves can be different for each infectious region.

One objective of this study is to show that these four TVI-ILMs can perform better than a time-invariant infectivity ILM (TII-ILM) when fitted to the regional-level data while underreporting occurs. A second objective is to show that the performance of time-varying infectivity models with parameters depending upon regional-level covariates is better than both time-invariant infectivity models and the global time-varying infectivity models in which the regional-level covariates are not considered. A third objective is to explore whether, and how much, spatial-temporal information can be detected with these models after data aggregation and underreporting.

First, in Section 4.2, the individual-level model (ILM) framework of Deardon et al.
(2010) is introduced and then generalized to allow for time-varying ILMs. In Section 4.3, five specific ILMs are presented, and then the simulation study is described and inference for ILMs is addressed. The results of the simulation study are shown in Section 4.4, and then some potential future research of interest is discussed in Section 4.5.

4.2 Individual-Level Model

Deardon et al. (2010) proposed a class of individual-level models (ILMs) for modeling infectious disease spread over time and space, incorporating various individual-level risk factors. The compartmental framework within which the ILMs of Deardon et al. (2010) are formulated is known as a stochastic susceptible-exposed-infectious-removed (SEIR) framework (Anderson and May, 1991), commonly applied in infectious disease epidemiology. In an SEIR framework, the status of an individual $i$ within the population is divided into four categories: susceptible (S), exposed (E), infectious (I) and removed (R), during the observed epidemic period. If individual $i$ is susceptible, that means they remain uninfected to the disease. Being in the exposed state means that individual $i$ is infected but can not spread the pathogen. Once individual $i$ has become infectious and can transmit the pathogen they are in the infectious state. If in the removed state, individual $i$ has been removed from the infectious class, either through recovery, acquired immunity, isolation, or death, etc. It is assumed that individuals transit in the specific order: $S \rightarrow E \rightarrow I \rightarrow R$ or remain
in the susceptible state.

Here, time $t$ is discretized to represent an interval $[t, t + d)$ in continuous time, where $d$ represents the time lag between the time points at which data are collected; for example, in this Chapter $d = 1$ day. Let $N$ and $M$ denote the total number of individuals in the population and the final size of individuals that are infected as part of the epidemic, respectively. Let $T$ denote the last day on which data is recorded.

The knowledge of the event history of individuals (actual or estimated) is essential as to parameterise the ILM (Deardon et al., 2010); that is, for $t = 1, \ldots, T$, if individual $i$ is in state S, E, I or R. Let $S(t)$, $E(t)$, $I(t)$, $R(t)$ be the sets of individuals becoming susceptible, exposed, infectious and removed within continuous time interval $[t, t + 1)$, respectively.

### 4.2.1 The General ILM

$P_{it}$ represents the probability that a transition from $S \rightarrow E$ for a susceptible individual $i$ occurs at discrete time point $t$ (i.e. the probability that $i$ is infected within the time interval $[t, t + 1)$ in continuous time). A general form of this probability is given by Deardon et al. (2010) as:

$$P_{it} = 1 - \exp \left\{ -\Omega_S(i) \sum_{j \in I(t)} \Omega_T(j) k(i, j) \right\} - \varepsilon(i, t) \quad (4.2.1)$$

where a susceptibility component containing risk covariates related to susceptible individual $i$ contracting the disease is represented by $\Omega_S(i)$; a transmissibility component
containing risk covariates related to infectious individual $j$ passing on the disease is represented by $\Omega_T(j)$ (for example, the population size in different regions/farms if modeling at a regional/farm level, age, gender, and environmental risk factors, might be such risk covariates incorporated); $k(i, j)$ is the infection kernel, which denotes risk factors shared between the susceptible and infectious individuals $i$ and $j$ and is often a function of the separation distance (e.g. Euclidean distance); and $\varepsilon(i, t)$ is a "spark" term, which is included to describe some other random infections that are insufficiently explained by the former terms in the model (e.g. cases from outside the study population).

Let $\mathcal{S} = \{S(t)\}_{t=1}^{T}$ denote the list of $S(t)$ sets during the period of study $t = 1, \ldots, T$; and define $\mathcal{E} = \{E(t)\}_{t=1}^{T}$, $\mathcal{I} = \{I(t)\}_{t=1}^{T}$, $\mathcal{R} = \{R(t)\}_{t=1}^{T}$ similarly. $\theta$ is vector of unknown parameters. Here, $f_t(\mathcal{S}, \mathcal{E}, \mathcal{I}, \mathcal{R}|\theta)$ is the product of the probability of all new infection event and of all non-infection events occurring in time interval $[t, t+1)$:

$$f_t(\mathcal{S}, \mathcal{E}, \mathcal{I}, \mathcal{R}|\theta) = \left[ \prod_{i \in \mathcal{E}(t+1)\setminus \mathcal{E}(t)} P_{it} \right] \left[ \prod_{i \in \mathcal{S}(t+1)} \left(1 - P_{it}\right) \right].$$

The likelihood of the general ILM is a product of $f_t(\mathcal{S}, \mathcal{E}, \mathcal{I}, \mathcal{R}|\theta)$ during the observed epidemic process $t = 0, \ldots, T$, which is given by:

$$\pi(\mathcal{S}, \mathcal{E}, \mathcal{I}, \mathcal{R}; \theta) = f(\mathcal{S}, \mathcal{E}, \mathcal{I}, \mathcal{R}|\theta) = \prod_{t=0}^{T} f_t(\mathcal{S}, \mathcal{E}, \mathcal{I}, \mathcal{R}|\theta) \quad (4.2.2)$$
Assume that the next two transitions for individual $i$ from $E \rightarrow I$ and $I \rightarrow R$ occur after a latent period $\tau_i$ and an infectious period $\nu_i$, respectively. Let $B_i$, $C_i$ and $D_i$ denote the day at which the individual $i$ becomes infected, infectious and removed (or recovered), respectively. Here, $\tau_i = C_i - B_i$ and $\nu_i = D_i - C_i$, can be modeling in various ways (see Section 4.3.3 for details).

4.2.2 Time-varying Individual-level Models

A generalized form of the ILMs of Deardon et al. (2010) that allows for time-varying susceptibility, infectivity, and/or infection kernel (as well as sparks terms) is given by:

$$P_{it} = 1 - \exp \left[ -\Omega_S(i,t) \sum_{j \in i(t)} \Omega_T(j,t)k(i,j,t) - \varepsilon(i,t) \right]$$

(4.2.3)

where, $\Omega_S(i,t)$ is a susceptibility component containing potential risk factors related to susceptible individual $i$ contracting the disease during time interval $[t, t+1)$; $\Omega_T(j,t)$ is a transmissibility component containing potential risk factors related to infectious individual $j$ passing on the disease in $[t, t+1)$; $k(i,j,t)$ is a time-varying infection kernel representing shared risk factors often based on the separation distance (e.g. Euclidean distance) between the infectious and susceptible individuals $i$ and $j$ in $[t, t+1)$. The likelihood function for these time-varying individual-level models (TV-ILMs) is defined as in equation (4.2.2).
4.3 Simulation Study

In this study, the use of TV-ILMs in which time-varying transmissibility/infectivity function is incorporated ($\Omega_T(j, t)$ defined in equation 2.3) is examined, when fitted to regional-level data obtained by individual-level data aggregation. The infectivity is then assumed to vary during the course of infectiousness for a region, and to follow some discrete curve. Here, two discrete infectivity curves, the discrete exponential and discrete gamma, are considered in order to capture the time-varying nature of the infectivity within regions.

The first two TVI-ILMs focus on modeling the global infectivity curves for regions; that is, the infectivity curve is assumed to be the same for each region, with infectivity beginning when the first infectious individual is observed within a region, and truncated at the time point at which the last infectious individual recorded in that region enters the removed state.

TVI-ILMs in which some parameters of the infectivity curves are dependent upon some regional-level covariate are further developed. Particularly, population density of the regions is taken into consideration. For example, a region with a larger population density may have a higher chance of passing on the disease than one with a low population density; it may also experience a lower rate of decay of the infectivity curve, so that the region tends to be "more infectious for longer". Thus, in these models the time-varying infectivity curves, $\Omega_T(j, t)$, can vary for different infectious region $j = 1, ..., M$. 
The primary purposes of the simulation study are to investigate whether: 1) time-varying infectivity models fit better to the regional-level data than the time-invariant (constant) infectivity ILMs; 2) the TVI-ILMs with parameters depending upon the regional population density perform better than both constant infectivity ILMs and the TVI-ILMs without this regional-level risk factor incorporated, in terms of model comparison and adequacy. Another concern is the question of how much spatial-temporal information can be detected using the TVI-ILMs after data have been spatially aggregated.

Here, individual-level data with an underreporting mechanism is first generated, and this under-reported individual-level data is aggregated into regional levels. A time-invariant infectivity model, the two global time-varying infectivity models, and the two time-varying regional-level covariate-dependent infectivity models are then fitted. For the purpose of this study, the susceptibility function and infection kernel are limited to be time-invariant; i.e., $\Omega_S(i, t) = \Omega_S(i)$ and $k(i, j, t) = k(i, j)$. Further, the susceptibility is assumed to be constant across all regions, i.e., $\Omega_S(i) = \Omega$.

4.3.1 Models for simulation study

Model 1. Constant Infectivity and Exponential Kernel Model (CE).

First, consider a time-invariant ILM. Here, $\Omega_S(i)\Omega_T(j) = \alpha$ (e.g., $\Omega_S(i) = \alpha_S$ and $\Omega_T(j) = 1$, $\forall i, j$) can be viewed as characterizing the global strength of the epidemic. It is assumed that the infection kernel, $k(i, j) = k(d_{ij}) = \exp\{-\beta d_{ij}\}$, is an expo-
nential distance kernel, where $d_{ij}$ denotes the Euclidean distance between susceptible individual $i$ and infectious individual $j$; and $\beta$ is the exponential decay parameter of the infection kernel, governing the extent to which increasing $d_{ij}$ reduces the risk of transmission from infectious $j$ to susceptible $i$. $\varepsilon(i,t) = \varepsilon$, is a sparks term that allows for purely random infections that can not be explained by the previous terms. Then, the exponential kernel model (CE) is given by:

$$P_{it}^{(CE)} = 1 - \exp \left[ \left( -\alpha \sum_{j \in I(t)} \exp \{ -\beta d_{ij} \} \right) - \varepsilon \right], \quad \alpha, \beta, \varepsilon > 0 \quad (4.3.1)$$

**Model 2. Exponential Infectivity and Exponential Kernel Model (EE).**

The second model considered is a TVI-ILM in which the transmissibility is assumed to vary over time and follows a discrete exponential curve. $\Delta_1(t; A_1, \lambda, C_j)$ here denotes the global discrete exponential infectivity curve truncated from the time point $C_j$, at which the last observed infection event is removed from a region $j = 1, \ldots, M$. Additionally, $\Omega_S(i,t) = \Omega_S(i) = 1$. The probability of a susceptible individual $i$ being infected within the time interval $[t, t+1)$ is thus given by:

$$P_{it}^{(EE)} = 1 - \exp \left[ \left( \sum_{j \in I(t)} \Delta_1(t; A_1, \lambda_1, C_j) \exp \{ -\beta d_{ij} \} \right) - \varepsilon \right], \quad \Delta_1, \beta, \varepsilon > 0 \quad (4.3.2)$$
where,

\[ \Omega_T(j, t) = \Delta_1(t; A_1, \lambda_1, C_j) = A_1 \exp\{-\lambda_1(t - C_j)\}, \quad A_1, \lambda_1 > 0; \]

A_1 is a global scale parameter; and \( \lambda_1 \) represents a global decay rate of the discrete infectivity curve.

**Model 3. Gamma Infectivity and Exponential Kernel Model (GE).**

A model with an alternative time-varying transmissibility is considered here; specifically, \( \Omega_T(j, t) \) is modified to follow a global discrete gamma curve, \( \Delta_2(t; A_2, \phi, \lambda_2, C_j) \), truncated from the end of observed infectiousness at time \( C_j \) for \( j = 1, \ldots, M \). The model is therefore:

\[
P_d^{(GE)} = 1 - \exp\left(-\sum_{j \in I(t)} \Delta_2(t; A_2, \phi, \lambda_2, C_j) \exp\{-\beta d_{ij}\} - \varepsilon\right), \quad \Delta_2, \beta, \varepsilon > 0 \quad (4.3.3)
\]

where

\[
\Omega_T(j, t) = \Delta_2(t; A_2, \phi, \lambda_2, C_j) = A_2(t+1-C_j)^{-1} \exp\{-\lambda_2(t+1-C_j)\}, \quad A_2, \phi, \lambda_2 > 0;
\]

\( A_2 \) is a global scale parameter; \( \phi \) is a global shape parameter; and \( \lambda_2 \) denotes a global rate parameter.

Here, a discrete exponential infectivity curve with random scale \((A_{1,j})\), and rate parameters \((\lambda_{1,j})\), is considered for an infectious region \(j = 1, ..., M\). Here, \(\lambda_{1,j} = 1/\mu_{1,j}\), where \(\mu_{1,j}\) is the mean of the discrete exponential distribution defined by the infectivity curve for region \(j\). Both \(A_{1,j}\) and \(\mu_{1,j}\) are assumed to be dependent upon a regional-level covariate (in this study, population density) of region \(j\). Then, suppose that both are positive half-normally distributed. Specifically:

\[
A_{1,j} \sim N^+(\log(\kappa_j)\alpha_{A_1}, \gamma_{A_1}), \quad \gamma_{A_1} > 0, \alpha_{A_1} \in \mathbb{R}
\]

\[
\mu_{1,j} \sim N^+(\log(\kappa_j)\alpha_{\mu_1}, \gamma_{\mu_1}), \quad \gamma_{\mu_1} > 0, \alpha_{\mu_1} \in \mathbb{R}
\]

where, \(\kappa_j\) is the population density for region \(j\); and \(\mathbb{R}\) is the real line, \(\mathbb{R} = (-\infty, +\infty)\).

The probability of a susceptible individual \(i\) becoming infected within the time interval \([t, t+1)\) is thus given by:

\[
P^{(EP)}_{it} = 1 - \exp\left[-\left(\sum_{j \in I(t)} \Omega_T(j,t) \Delta_{1,j}(t; A_{1,j}, \lambda_{1,j}, C_j) \exp\{-\beta d_{ij}\}\right) - \varepsilon\right], \quad \Delta_{1,j}, \beta, \varepsilon > 0 \quad (4.3.4)
\]

where

\[
\Omega_T(j,t) = \Delta_{1,j}(t; A_{1,j}, \lambda_{1,j}, C_j) = A_{1,j} \exp\{-\lambda_{1,j}(t - C_j)\}, \quad A_{1,j}, \lambda_{1,j} > 0;
\]
Model 5. Gamma Infectivity and Population Density Model (GP).

Finally, the discrete gamma infectivity curve is also modified to allow for a random scale ($A_{2,j}$), random rate ($\lambda_{2,j}$), and random shape ($\phi_j$) parameters of the discrete gamma curve for an infectious region $j = 1, ..., M$. Let $\mu_{2,j}$ and $\sigma^2_j$ be the mean and the variance of a discrete gamma distribution defined by the infectivity curve for $j$. Therefore, $\lambda_{2,j} = \mu_{2,j}/\sigma^2_j$ and $\phi_j = \mu^2_{2,j}/\sigma^2_j$. Once again, these parameters are made dependent upon population density of the region. $A_{2,j}$, $\mu_{2,j}$ and $\sigma^2_j$ are also assumed to have positive half-normal distributions. Specifically,

$$
A_{2,j} \sim N^+(log(\kappa_j)\alpha_{A_2}, \gamma_{A_2}), \quad \gamma_{A_2} > 0, \alpha_{A_2} \in \mathbb{R}
$$

$$
\mu_{2,j} \sim N^+(log(\kappa_j)\alpha_{\mu_2}, \gamma_{\mu_2}), \quad \gamma_{\mu_2} > 0, \alpha_{\mu_2} \in \mathbb{R}
$$

$$
\sigma^2_j \sim N^+(log(\kappa_j)\alpha_{\sigma^2}, \gamma_{\sigma^2}), \quad \gamma_{\sigma^2} > 0, \alpha_{\sigma^2} \in \mathbb{R}
$$

The probability of a susceptible individual $i$ becoming infected within the time interval $[t, t+1)$ is thus given by:

$$
P_{it}^{(GP)} = 1 - \exp\left[ -\left( \sum_{j \in I(t)} \Delta_{2,j}(t; A_{2,j}, \phi_j, \lambda_{2,j}, C_j) \exp\{-\beta d_{ij}\} \right) - \varepsilon \right], \quad \Delta_{2,j}, \beta, \varepsilon > 0 \quad (4.3.5)
$$

where,

$$
\Omega_T(j,t) = \Delta_{2,j}(t; A_{2,j}, \phi_j, \lambda_{2,j}, C_j) = A_{2,j}(t + 1 - C_j)^{\phi_j - 1} \exp\{-\lambda_{2,j}(t + 1 - C_j)\}, \quad A_{2,j}, \phi_j, \lambda_{2,j} > 0.
$$
4.3.2 Epidemic Simulation

4.3.2.1 Individual-level epidemic

First, the random spatial locations ($S^{ind}$) of 10,000 individuals ($N^{ind} = 10,000$) are simulated from three bivariate Gaussian distributions such that each has a spatial location truncated within a 40 × 40 unit square area. They are simulated, following the steps:

1. Generate a group membership from a discrete Uniform proposal, $z \sim DU[1, 3]$;

2. A candidate of the spatial location is generated: $s^{ind} = (x^{ind}, y^{ind})^T \sim \mathcal{N}(\mu_z, \Sigma_z)$;
   
   where, $\mu_1 = (29.5, 29.5)^T$, $\Sigma_1 = \begin{pmatrix} 5.5 & 0 \\ 0 & 5.5 \end{pmatrix}$; $\mu_2 = (20.5, 20.0)^T$, $\Sigma_2 = \begin{pmatrix} 8.5 & 0 \\ 0 & 7.5 \end{pmatrix}$;
   $\mu_3 = (12.5, 26.0)^T$, $\Sigma_3 = \begin{pmatrix} 7.0 & 0 \\ 0 & 8.0 \end{pmatrix}$.

3. if $(x^{ind}, y^{ind})^T$ satisfy $s^{ind} = \{(x^{ind}, y^{ind})^T \in \mathbb{R}^2 : 0 \leq x^{ind} \leq 40, 0 \leq y^{ind} \leq 40\}$, then $s^{ind}$ is accepted; otherwise it is rejected, and repeat steps 1 and 2;

4. Repeat until locations of 10,000 individuals are simulated.

Figure 4.1 shows the simulated population on the 40 × 40 unit square area.

Second, constant infectivity and exponential kernel model (CE) is used as the generating model to simulate ten epidemics through this population. In each epidemic simulation, the ILM is formulated within an SEIR framework. It is assumed that the latent period of infected individuals, $\tau^{ind}_i$, follows a truncated discretized exponential distribution with mean 2 days ($1 \leq \tau^{ind}_i \leq 5$ days), and the infectious period, $\nu^{ind}_i$, is...
Figure 4.1: Simulated heterogeneous population with 100 square regions.

assumed to follow a truncated discretized exponential distribution with mean 7 days
\((5 \leq \nu_i^{ind} \leq 10)\). At time \(t = 0\), all individuals are considered to be susceptible,
and two individuals are then randomly selected to be infectious at \(t = 1\). The model
parameters are set to be \(\alpha = 0.1\), \(\beta = 4.5\) and \(\varepsilon = 0\). A typical simulated epidemic
results in 9,950 infections \(M^{ind} = 9,950\) and terminates at \(T^{ind} = 78\) days.

Three different scenarios of underreporting for each of the ten epidemics are in-
vestigated. The situations where 25%, 50% and 75% of the total individual infection
events in the population are randomly recorded for each epidemic, respectively, are
considered. A data set with a 100% reporting rate is also analyzed.

4.3.2.2 Regional-level epidemic

As shown in Figure 4.1, the 40×40 unit square area is evenly divided into 100
square cells each of which is a region considered in this study. For each epidemic
and under-reporting scenario, each individual-level data set is aggregated into these 100 regions. That means, the total number of aggregated "individuals" is \(N = 100\). Let the centre of each region represent its coordinates \((x, y)\) for every combination of \(x, y = 2, 6, \ldots, 38\). For the infection times, once again, \(C_i\) and \(D_i\) denote the beginning and end of observed infectiousness for region \(i\), respectively. Note that \(C_i\) is the time at which the first individual is reported infectious in region \(i\), and \(D_i\) is the time point at which the last infectious individual recorded is removed from that region. However, the infection time \(B = (B_1, \ldots, B_N)\) is assumed to be unobserved and needs to be imputed in the model fitting procedure. The latent period is assumed to be exponentially distributed: \(\tau_i = C_i - B_i \sim \exp(\psi)\) where \(\psi\) is unknown parameter. Therefore, \(B\) is treated as unknown parameters in the model. The parameter vectors for the CE, EE, GE, EP and GP models are given by \(\Theta_1 = (\alpha, \beta, \varepsilon, \psi, B)\), \(\Theta_2 = (A_1, \lambda_1, \beta, \varepsilon, \psi, B)\), \(\Theta_3 = (A_2, \phi, \lambda_2, \beta, \varepsilon, \psi, B)\), and \(\Theta_4 = (\alpha_{A_1}, \gamma_{A_1}, \alpha_{\mu_1}, \gamma_{\mu_1}, \beta, \varepsilon, \psi, A_1, \lambda_1, B)\), \(\Theta_5 = (\alpha_{A_2}, \gamma_{A_2}, \alpha_{\mu_2}, \gamma_{\mu_2}, \alpha_{\sigma^2}, \gamma_{\sigma^2}, \beta, \varepsilon, \psi, A_2, \phi, \lambda_2, B)\), respectively. Each of the five models of Section 4.3.1 are fitted to the regional-level data. That means, treating the regions as "individuals" in the fitted ILMs.

4.3.3 Bayesian Analysis

Parameterization of the ILMs is performed using a Bayesian Markov chain Monte Carlo (MCMC) approach. To each parameter \(\alpha, A_1, \lambda_1, A_2, \phi, \lambda_2, \gamma_{A_1}, \gamma_{\mu_1}, \gamma_{A_2}, \gamma_{\mu_2}, \gamma_{\sigma^2}, \beta\) and \(\varepsilon\), vague independent positive half normal prior distributions, \(\pi(\cdot) \sim \ldots\)
\(N^+ (0, 10^5)\), are assigned. For parameters \(\alpha_{A_1}\) and \(\alpha_{\mu_1}\) in the EP model, priors \(N(0, 4)\) which are weakly informative priors (see, e.g., Gelman, 2006) are chosen. Also, set \(\pi(\alpha_{\sigma_2}) \sim N(0, 0.5)\) and \(\pi(\zeta) \sim N(0, 0.1)\) for \(\zeta = \alpha_{A_2}, \alpha_{\mu_2}\) in the GP model. In the study, Gamma prior, \(\text{Gam} (\psi_a, \psi_b)\) where \(\psi_a = 10\) and \(\psi_b = 10\), is assigned to parameter \(\psi\) for all the models.

In the Bayesian framework missing or uncertain data, such as \(B\), can be treated as a model parameter to be estimated via data augmented MCMC (Gelman et al., 2004). This results in a joint distributional estimate of both model parameters and the missing data (Casella and Berger, 2002). For example, the posterior distribution for the EP model, \(\pi(\Theta_4|I, R)\), is obtained by using information in the observed data, characterized by the joint likelihood function, to update our prior knowledge about the distribution of \(\Theta_4\). Here, at each iteration of the Markov chain, the infection time \(B\) is updated, and then the series \(S, E\) can be reconstructed.

The random walk Metropolis Hastings (MH) MCMC algorithm is applied to sample realizations from the posterior distributions for each of the parameters \(\alpha, A_1, A_2, \phi, \lambda_2, \alpha_{A_1}, \gamma_{A_1}, \alpha_{\mu_1}, \gamma_{\mu_1}, \alpha_{A_2}, \gamma_{A_2}, \alpha_{\mu_2}, \gamma_{\mu_2}, \alpha_{\sigma_2}, \gamma_{\sigma_2}, \beta\) and \(\varepsilon\).

Efficient MCMC chains can be obtained through various means. First, the parameters \(\alpha, A_1, A_2\) are exponentiated in all the CE, EE and GE models (i.e., \(\alpha = \exp\{\alpha'\}, A_1 = \exp\{A'_1\}\) and \(A_2 = \exp\{A'_2\}\)). These transformations are found to improve the efficiency of MCMC mixing for these models. Additionally, block update is used for correlated parameters and the proposal densities are tuned to have
a correlation structure that approximately matches the underlying posterior distributions for all the fitted models. For example, in the EE model, parameters $A_1'$ and $\beta$ are correlated with each other, but $\lambda_1$ and $\varepsilon$ are not highly correlated with any other. Therefore, the two correlated parameters are updated in one block. Multivariate Gaussian distributions are used as proposal distributions for all blocks of correlated parameters. Here, for each single parameter update in all the models, uniform proposals are used, and are "tuned" in order to have an overall acceptance rate of approximately 0.4 (Gelman et al., 1996).

At each MCMC iteration the parameter $\psi$ and the infection time $B$ are updated. For the EP and GP models, $A_1, \lambda_1, A_2, \phi$ and $\lambda_2$ are also updated at each iteration. Following Neal and Roberts (2004) who found approximately optimal MCMC mixing and computing time when updating around 10% of the infection times at each iteration, 10 of the $B$, $A_1, \lambda_1, A_2, \phi$ and $\lambda_2$ are updated at each iteration and each of them is updated one at a time.

A Gibbs sampler is used for updating the parameter $\psi$. The conditional posterior, given by

$$
\pi(\psi | \tau) \propto \psi^{M+\psi_a-1} \exp \left( - \left( \sum_{i=1}^{M} \tau_i + \psi_b \right) \psi \right)
$$

has a gamma distribution, $Gam(M + \psi_a; \sum_{i=1}^{M} \tau_i + \psi_b)$.

$A_1, \lambda_1, A_2, \phi, \lambda_2, B$, are each updated using a single parameter random walk update. For example, suppose that $A_{1,i}$ is being updated at iteration $k$. A new candidate $A_{1,i}'$ is proposed by simulating from $A_{1,i}' \sim N^+ \left( log(\kappa_i) \sigma_a^{(k+1)} \kappa_1^{(k+1)} \right)$. Then, the ac-
ceptance probability $\delta$ is calculated by:

$$
\delta(A_1', A_1^{(k)}) = \min \left( 1, \frac{\pi(S, E, I, R|A_1', i)}{\pi(S, E, I, R|A_1^{(k)})} \right)
$$

(4.3.6)

After a burn-in period of 2,000 iterations, another 1,000,000 iterations are run to obtain the MCMC posterior sample. In all the cases, convergence of the chain is assessed visually for different starting values.

4.4 Results

The results about posterior means obtained for each individual simulated epidemic are not reported here since they are not of great interest. This is also because, the parameters in the data-generating and data-fitting models are not directly comparable: 1) the true model generates epidemic spread at the individual level; however, all the fitted models describe epidemic spread at the aggregated regional level; 2) the fitted models (excepting the CE) have a different form to the true individual-level model. Instead, the model fit of the various ILMs is compared via the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002), a lower value of DIC indicating a better model fit according to this criterion. Table 4.1 reports the difference of DIC values averaged over the ten simulated epidemics, for each of the four proportions of reporting in the population. In this table, the time-invariant ILM (the CE model) is arbitrarily chosen as the reference, or baseline, level and the difference between
Table 4.1: Averaged Difference of Deviance Information Criterion (DIC) over the ten epidemics

<table>
<thead>
<tr>
<th>Infectivity Profile</th>
<th>Model</th>
<th>100% reporting</th>
<th>75% reporting</th>
<th>50% reporting</th>
<th>25% reporting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time-invariant Inf.</td>
<td>(CE)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Global Time-varying Inf.</td>
<td>(EE)</td>
<td>-49.376</td>
<td>-50.059</td>
<td>-47.942</td>
<td>-33.581</td>
</tr>
<tr>
<td></td>
<td>(GE)</td>
<td>-53.511</td>
<td>-63.365</td>
<td>-55.097</td>
<td>-37.846</td>
</tr>
<tr>
<td>Time-varying Inf. (covariates)</td>
<td>(EP)</td>
<td>-95.025</td>
<td>-82.089</td>
<td>-79.050</td>
<td>-78.559</td>
</tr>
<tr>
<td></td>
<td>(GP)</td>
<td>-123.308</td>
<td>-119.183</td>
<td>-103.735</td>
<td>-97.970</td>
</tr>
</tbody>
</table>

the DICs of each of the four TVI-ILMs with the CE model, averaged over the ten simulated epidemic, is calculated for each reporting scenario.

First of all, the performances of the TVI-ILMs are compared to the CE model. For 100% reporting, the DIC results indicate that the model fit of the time-invariant infectivity model (CE) is worse than that of the TVI-ILMs (EE, GE, EP and GP). For other reporting settings, The DIC differences are also negative, indicating that the TVI-ILMs uniformly outperform the corresponding time-invariant ILM (the CE model). Second, comparing the TVI-ILMs (the EP and GP models) in which some parameters are dependent upon the population density to the other TVI-ILMs (the EE and GE models), it is observed that the average DIC values of the EP and GP models are all lower than the corresponding EE and GE models for each of the four reporting conditions. This suggests that the EE and GE models have a worse fit than the EP and GP models; that is, the TVI-ILMs with random discrete infectivity curves depending upon population density may better improve the model fit when data has been spatially aggregated. Moreover, it is noticed that the GE models appear to perform slightly better than the corresponding EE models for all the reporting rates,
where the absolute averaged differences of DIC values of the GE models are higher. This is also observed when comparing the difference of DIC values of the EP and GP models, where the DIC value for the GP model is noticeably lower than that of the EP model under all four reporting scenarios. This would imply that a tangible benefit is offered by the extra flexibility of the 3-parameter gamma curve.

Model adequacy is also evaluated by looking at the posterior predictive distributions of the epidemic curves for the four different reporting scenarios. Figure 4.2 shows the means and associated 95% percentile intervals of the posterior predictive distribution of the cumulative number of infected regions over time, along with the observed cumulative infected regions for a typical epidemic realization. For the time-invariant ILM (CE), the observed number of cumulative infected regions obviously deviates from the corresponding 95% percentile intervals before around day 20. There is some improvement observed in the posterior predictive ability for the EE and GE models. However, the posterior predictive plots of the EP and GP models show the best fit of the five models. Furthermore, it is evident that the posterior predictive cumulative number of infected regions for the GP model is better than that of the associated EP models after accounting for underreporting mechanism. In conclusion, all the four TVI-ILMs obviously fit better than the associated time-invariant ILM, and the TVI-ILMs with regional-covariate dependent infectivity curves fit the data better than the TVI-ILMs with global parameters.

In each reporting condition, Figure 4.3 illustrates the estimated exponential dis-
Figure 4.2: Predictive cumulative number of cases with posterior means (black line) and 95% percentile intervals (dotted lines) and the observed cumulative number of cases (black line) at the regional level.
Figure 4.3: The estimated exponential distance kernel against $d_{ij}$ under the posterior means for the time-invariant infectivity model (CE) and the time-varying infectivity models (EE, GE, EP and GP).

tance kernel under the posterior means for each fitted model. The figures show that all the five models are able to detect spatial information despite the data aggregation at the level of regions. For 25% reporting, 50% reporting, 75% reporting and 100% reporting, it is observed that the fitted distance kernels decline over relatively short distances, respectively. A slightly lower rate of decay is obtained for the CE model; that is, it allows for more infections over long distances compared to the other four models at each reporting setting. Additionally, the spatial rate of decay estimated at 100% reporting seems lower than in under-reporting scenarios, and the results are actually quite similar for the three under-reporting conditions.
4.5 Discussion

In this Chapter the general ILMs of Deardon et al. (2010) were extended to allow for time-varying covariate-dependent infectivity to account for spatially aggregated data. A simulation study was carried out in which two discrete infectivity curves, discrete exponential and discrete gamma, were tested. These models were considered in two cases: one in which the time-varying infectivity models were homogeneous across regions, and another in which some parameters of the infectivity curves were dependent upon some regional-level covariates.

In the simulation study, for both the 100% reporting and under-reporting scenarios, the results showed that the time-varying infectivity models achieved a better fit to the regional-level data than a time-invariant infectivity model in terms of the DIC. And also, the covariate-dependent infectivity curve models outperformed the corresponding models without the regional-level risk factors incorporated. Furthermore, the TVI-ILMs incorporating gamma infectivity (the GP models) were favored over the associated TVI-ILMs with exponential infectivity (the EP models) under all four reporting settings. This may be explained by the fact that the 3-parameter gamma infectivity is more flexible and can better reflect the underlying infectious status for regions. In fact, plots of the observed number of individuals being infectious per region over time (results not shown) tended to follow gamma-shaped curves across all reporting rates. Additionally, model checking via the posterior predictive distribution of the cumulative number of infected regions showed that the TVI-ILMs with
regional-level covariate-dependent infectivity curves also performed better in terms of accuracy and precision across all four reporting rates.

However, for both time-varying infectivity models and time-invariant infectivity models, the regional-level model fit does appear to be influenced by the issue of underreporting at the individual-level data. First, the posterior predictive performance at 25% reporting seems worse than those of other reporting settings except for the GP model. Second, the averaged differences of DIC values with the time-invariant ILM for each time-varying ILMs appeared to increase with the proportion of reporting at the level of individuals (except the EE and GE models at 100% reporting rate). However, spatial information was found to be captured despite the data aggregation at the regional level by both the time-invariant ILM and time-varying ILMs.

Here some issues and potential future work are now addressed. In this study, the choice of exponential distance kernel, \( \exp(-\beta d_{ij}) \), is somewhat arbitrary. Some other spatial infection kernel, such as a geometric distance kernel (see, for example, Deardon et al. (2012)), is an alternative option for testing the time-varying infectivity models when fitted to a regional-level data.

For TVI-ILMs, infectiousness of a region is assumed to begin from the time point when the first individual recorded becomes infectious, and to end when the last individual recorded is removed from the population within that region. However, since the issue of underreporting is commonly observed in reality, and it is often difficult or impossible for a doctor to infer time of infectiousness beginning from a diagnosis
time, we might want to relax the assumption of knowing this time. We similarly might want to relax the assumption of known removed time. These issues could be tackled by data augmentation (see, e.g., Jewell et al., 2009a; Cauchemez and Ferguson, 2011; Neal and Roberts, 2004), in which time of infectiousness beginning and infectious periods are treated as unknown parameters that need to be imputed. Of course, such techniques can come with added computational costs, however. Additionally, the use of reversible jump MCMC might be considered to address the issue of hidden or unreported infections.

Finally, it would be of interest to apply these time-varying infectivity models to an infectious disease data set (e.g. 2009 H1N1 influenza data or the UK 2001 FMD outbreak). Such analysis could also involve developing TV-ILMs with time-varying susceptibility and/or contact functions incorporated, allowing varying infection status in a region and/or population movement.
CHAPTER 5

Conclusions

This thesis has presented work on three related topics in the area of infectious disease modeling using individual level models (ILMs). Specifically, these three studies show the development of time-varying infectivity ILMs which incorporate time-varying infectivity for regions when data has been spatially aggregated at the regional level, the use of probability scoring rules as a tool for model comparison and validation, and finally the further development of time-varying infectivity ILMs (TVI-ILMs) in which some parameters of the infectivity curve are dependent upon some regional-level covariate.

Here, an overview of each Chapter is presented, along with a discussion of avenues of future research.
5.1 Homogeneous Time-varying Infectivity ILMs

5.1.1 Summary

In Chapter 2, three TVI-ILMs were developed for situations in which individual-level disease data has been regionally aggregated, that allow for the time-varying nature of infectivity within regions. Simulation study results suggested that the performance of the TVI-ILMs is better than that of corresponding time-invariant infectivity ILMs (TII-ILMs) under various individual-level underreporting scenarios. The results also showed that the spatial effect could be detected over reasonable distances by both the TVI-ILMs and TII-ILMs despite the data aggregation and underreporting. Five time-invariant infectivity models and five time-varying infectivity models were also applied to data from 2009 H1N1 influenza outbreak in Southern Ontario. In general, the results from the H1N1 case study were in agreement with those of the simulation study, except that spatial spread was not detectable when modeling was carried out at the level of census regions.

5.1.2 Future Work

In the 2009 H1N1 influenza case study, both the latent period and the infectious period were assumed to be constant at the individual-level data. For modeling at the level of census regions, both the regional-level times of infectiousness beginning and ceasing were assumed to be the time points at which the first case is observed, and the last observed infectious individual is removed from the population, within
each region, respectively. However, these assumptions can be relaxed. Joint posterior
distributional estimates of both the latent and infectious periods of the regions and
model parameters can be obtained using data-augmented MCMC. However, such an
implementation would substantially increases the computational burden.

Potential spatial error in terms of allocating individuals to census regions could
also be modeled. For example, Deardon et al. (2010) introduced Gaussian spatial
random effects into ILMs to deal with the problem of individual-level spatial mea-
surement error. This approach can likely be adapted to model measurement error in
the infectious population in the case study. An an alternative approach, a "nearest-
neighbour" random effect could be incorporated, allowing shifts from each census
tract to neighbouring census tract for a proportion of infectious individuals reported.

Additionally, other census tract-level covariates, such as age distribution and
transportation network, could be incorporated in the case study to account for such
levels of heterogeneity between census tracts.

5.2 Probability Scoring Rules for Model Comparison and Di-
agnostics

5.2.1 Summary

The use of probability scoring rules for model comparison, and as a model diagno-
tic tool for the various competing ILMs, under a Bayesian statistical framework was
investigated. In a simulation study, epidemic data was generated from a geometric
spatial ILM with discrete exponential infectivity and, then, seven misspecified ILMs, as well as the true data-generating model, were fitted to the data. The effectiveness of different probability scoring rules for identifying the correct ILM was explored. Three strictly proper scoring rules, and some other proper probability scores, were examined.

In the simulation study, probability scoring rules successfully demonstrated their ability to identify the true data-generating model. In the H1N1 case study, conclusions drawn from the probability scoring rules were generally in agreement with those discussed in Chapter 2.

Probability scoring rules for ILMs can be carried out at global level, to obtain scores averaged over the entire epidemic, or the scores can also be calculated at each time point over the course of an epidemic. The latter may be useful to investigate heterogeneity when the parameters vary over the epidemic period. Then, the use of probability scoring rules for model comparison may have advantages over the DIC approach, because they may be more intuitive for the non-statistician, and are not dependent upon underlying assumptions about the posterior distribution (e.g. symmetry).

However, using the posterior predictive distribution of probabilities scoring rules to determine model adequacy did not prove to work as well as using the more typical approach of considering the posterior predictive distribution of the epidemic curves.
5.2.2 Future Work

The application of probability scores into situations where parameter-heterogeneity over time and space is present could be explored to compare and assess different fitted ILMs. For example, transmission parameters may change over space if control policies vary over regions. Second, the development of some other TV-ILMs, in which time-varying susceptibility and/or time-varying infection kernel are incorporated to account for, specifically, changing vaccination status and/or restricted population movement, can be explored.

5.3 Covariate-dependent Time-varying Infectivity ILMs

5.3.1 Summary

In this final Chapter, the time-varying infectivity ILMs (TVI-ILMs) of Chapter 2 are extended to allow for time-varying infectivity dependent upon some regional-level covariate. Various models were tested via a simulation study, in terms of model comparison and adequacy.

Once again, better model fit was observed for TVI-ILMs than corresponding time-invariant infectivity ILMs. Moreover, TVI-ILMs with regional covariate-dependent infectivity curves performed better than the associated TVI-ILMs without regional-covariate information considered. Further, models containing discrete gamma infectivity curves appeared to outperform less flexible exponential curve models.
The problem of underreporting tended to have impacts on the regional-level model fit for both TVI-ILMs and TII-ILMs. However, the spatial mechanism behind transmission patterns were detectable by the regional covariate-dependent TVI-ILMs despite data aggregation and underreporting issues.

5.3.2 Future Work

The exponential distance kernel to model the spatial component of disease spread is somewhat arbitrarily chosen in the study. Some other alternative kernels, such as a geometric distance kernel which tends to facilitate a higher proportion of infections over longer distances, or a nearest neighbour kernel, could be considered. The situation where space is not explicitly modeled, but some underlying contact network or contiguity based neighbourhood drives the disease, could also be taken into consideration.

Probability scoring rules (Chapter 3) could be applied to compare the regional covariate-dependent TVI-ILMs with the TVI-ILMs with homogeneous infectivity curves as an alternative model selection tool.

Once again, the assumption of a known time of infectiousness beginning for regions can be relaxed. Since underreporting commonly exists in epidemiological disease data, and there is generally a reporting delay, it is often very difficult to identify the time that infectiousness begins within a region. This problem could be addressed using data-augmented MCMC techniques in which missing data (e.g. infectious times and
periods) can be imputed. Finally, the TVI-ILMs with some parameters depending upon regional covariates could also be applied to the 2009 H1N1 influenza data used in Chapters 2 and 3.
APPENDICES
APPENDIX A

Metropolis-Hastings (MH) Algorithm

A.1 Random Walk Metropolis-Hastings (rw-MH) Algorithm

The random walk Metropolis-Hastings algorithm is a special case of the MH algorithm. At iteration $t$, a proposed parameter value, $\theta'$, is simulated from a proposal density $q(\theta'|\theta^{(t)})$ that is symmetric around $\theta^{(t)}$. Therefore, $q(\theta'|\theta^{(t)}) = q(\theta^{(t)}|\theta') = q(|\theta' - \theta^{(t)}|)$, and the acceptance probability, under the Bayesian framework, becomes (Robert and Casella, 2004):

$$\alpha(\theta'; \theta^{(t)}) = \min\left(1, \frac{\pi(D|\theta') \pi(\theta')}{\pi(D|\theta^{(t)}) \pi(\theta^{(t)})}\right)$$

The candidate $\theta'$, is then accepted ($\theta^{(t+1)} = \theta'$) with probability $\alpha(\theta'; \theta^{(t)})$; otherwise, it is rejected ($\theta^{(t+1)} = \theta^{(t)}$) with probability, $1 - \alpha(\theta'; \theta^{(t)})$. Commonly chosen
proposals are the multivariate uniform, normal, or \( t \)-distributions.

\[ \text{A.2 Independent MH Algorithm} \]

The independent Metropolis-Hastings algorithm is an alternative to the \( \text{rw-MH} \) algorithm. Here, the proposal density \( q(\theta' | \theta^{(t)}) \) is independent of the current position of the chain which means \( q(\theta' | \theta^{(t)}) = q(\theta') \). Therefore, the proposed \( \theta' \) is accepted (i.e. \( \theta^{(t+1)} = \theta' \)) with acceptance probability:

\[
\alpha(\theta' ; \theta^{(t)}) = \min \left( 1, \frac{\pi(D|\theta')\pi(\theta')q(\theta^{(t)})}{\pi(D|\theta^{(t)})\pi(\theta^{(t)})q(\theta')} \right)
\]
BIBLIOGRAPHY


Gschlößl, S., and C. Czado (2005), Spatial modeling of claim frequency and claim size in insurance, *Tech. rep.*, Ludwig-Maximilians-Universität, Munich, Germany.


