Spring 2012

A path-specific approach to SEIR modeling

Aaron Thomas Porter
University of Iowa

A PATH-SPECIFIC APPROACH TO SEIR MODELING

by

Aaron Thomas Porter

An Abstract

Of a thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Biostatistics in the
Graduate College of The
University of Iowa

May 2012

Thesis Supervisor: Assistant Professor Jacob J. Oleson

ABSTRACT

Despite being developed in the late 1920s, compartmental epidemic modeling is still a rich and fruitful area of research. The original compartmental epidemic models were SIR (Susceptible, Infectious, Removed) models, which assume permanent immunity after recovery. SIR models, along with the more recent SEIR (Susceptible, Exposed, Infectious, Removed) models are still the gold standard in modeling pathogens that confer permanent immunity. This dissertation expands the SEIR structure to include a new class of spatial SEIR models.

The exponential assumption of these models states that the latent and infectious times of the pathogen are exponentially distributed. Work that relaxes this assumption and still allows for mixing to occur at the population level is limited, thereby making strong assumptions about these times. We relax this assumption in a flexible way, by considering a hybrid approach that contains characteristics of both population level and individual level approaches.

Next, we expand the Conditional Autoregressive (CAR) class of spatial models. This is to account for the Mumps data set we have procured, which contains mismatched lattice structures that cannot be handled by traditional CAR models. The use of CAR models is desirable here, as these models are known to produce spatial smoothing on lattices, and are a natural way to draw strength spatially in estimating spatial effects.

Finally, we develop a pair of spatial SEIR models utilizing our CAR structure. The first utilizes the exponential assumption, which is very robust. The second develops a highly flexible spatial SEIR model by embedding the CAR structure into the SEIR structure. This allows for a realistic analysis of epidemic data occurring on a lattice.

These models are applied to the Iowa Mumps epidemic of 2006. There are three questions of interest. First, what improvement do the methods proposed here provide over the current models in the literature? Second, did spring break, which occurred approximately 40 days into the epidemic, have an effect on the overall number of new infections? Thirdly, did the public's awareness of the epidemic change the rate at which mixing occurred over time? The spatial models in this dissertation are adequately constructed to answer these questions, and the results are provided.

Abstract Approved: _____
Thesis Supervisor

_____
Title and Department

_____
Date

A PATH-SPECIFIC APPROACH TO SEIR MODELING

by

Aaron Thomas Porter

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Biostatistics in the
Graduate College of The
University of Iowa

May 2012

Thesis Supervisor: Assistant Professor Jacob J. Oleson

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Aaron Thomas Porter

has been approved by the Examining Committee
for the thesis requirement for the Doctor of
Philosophy degree in Biostatistics at the May 2012
graduation.

Thesis Committee: _____
Jacob Oleson, Thesis Supervisor

_____
Kathryn Chaloner

_____
Joseph Cavanaugh

_____
Brian Smith

_____
Philip Polgreen

To Mom, Dad, Ashley

ACKNOWLEDGMENTS

I would like to acknowledge the help and support of my advisor, Dr. Jacob Oleson, who helped me decide on epidemic modeling as a research area, and helped me develop the necessary skills in both spatial statistics and stochastic epidemic modeling to perform this research.

I would also like to acknowledge the help and support of Dr. Kathryn Chaloner, Dr. Joseph Cavanaugh, Dr. Brian Smith, and Dr. Philip Polgreen, who sat on my committee.

Finally I would like to specifically acknowledge Philip Polgreen, who provided the data and helped me to develop an understanding of it.

ABSTRACT

Despite being developed in the late 1920s, compartmental epidemic modeling is still a rich and fruitful area of research. The original compartmental epidemic models were SIR (Susceptible, Infectious, Removed) models, which assume permanent immunity after recovery. SIR models, along with the more recent SEIR (Susceptible, Exposed, Infectious, Removed) models are still the gold standard in modeling pathogens that confer permanent immunity. This dissertation expands the SEIR structure to include a new class of spatial SEIR models.

The exponential assumption of these models states that the latent and infectious times of the pathogen are exponentially distributed. Work that relaxes this assumption and still allows for mixing to occur at the population level is limited, thereby making strong assumptions about these times. We relax this assumption in a flexible way, by considering a hybrid approach that contains characteristics of both population level and individual level approaches.

Next, we expand the Conditional Autoregressive (CAR) class of spatial models. This is to account for the Mumps data set we have procured, which contains mismatched lattice structures that cannot be handled by traditional CAR models. The use of CAR models is desirable here, as these models are known to produce spatial smoothing on lattices, and are a natural way to draw strength spatially in estimating spatial effects.

Finally, we develop a pair of spatial SEIR models utilizing our CAR structure. The first utilizes the exponential assumption, which is very robust. The second develops a highly flexible spatial SEIR model by embedding the CAR structure into the SEIR structure. This allows for a realistic analysis of epidemic data occurring on a lattice.

These models are applied to the Iowa Mumps epidemic of 2006. There are three questions of interest. First, what improvement do the methods proposed here provide over the current models in the literature? Second, did spring break, which occurred approximately 40 days into the epidemic, have an effect on the overall number of new infections? Thirdly, did the public's awareness of the epidemic change the rate at which mixing occurred over time? The spatial models in this dissertation are adequately constructed to answer these questions, and the results are provided.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 The Importance of SIR and SEIR Models

Compartmental epidemic modeling began with a seminal paper by Kermack and McKendrick, published in 1927 [50]. This paper developed a nonlinear system of ordinary differential equations that laid the foundation for SIR (Susceptible, Infectious, Removed) and SEIR (Susceptible, Exposed, Infectious, Removed) models. Stochastic epidemic models followed shortly after, with the development of the Reed-Frost model. Though this model was first published in 1952, it was developed in the 1920's [1]. In the time since these developments, both the deterministic and stochastic frameworks for SIR and SEIR modeling have shown fruitful development, and SIR and SEIR models are still the models of choice when modeling infectious agents which yield permanent immunity upon recovery by the individual.

The general framework of the stochastic SEIR model is as follows:

1. There is a group of individuals in the Susceptible class and at least one individual in the Infectious class in a population at the start of the epidemic.

2. The infectious individuals mix with susceptible individuals. Any susceptible individual contacted moves to the Exposed or latent class based on a probabilistic process.

3. Once in the Exposed class, the individual spends a number of days without spreading the infection.

4. Based on some probabilistic process, the individual moves from the Exposed class to the Infectious class.

5. The newly infectious individual may contact susceptible individuals and spread the disease.

6. Based on some probabilistic process, the infectious individual recovers and moves

to the Removed class.

7. Once in the Removed class, the individual may no longer be infected.

Since the inception of these models, research in the field has been rich and fruitful, with many frameworks being developed. The deterministic tradition continues to be a rich and active area of research. Much of the original theory can be found in Anderson and May's text [2]. New research continues to be performed as well. Deterministic spatial models were developed as early as 1991 [4] [62].Additionally, deterministic methods for considering non-exponential generations times have recently been considered [59].

Stochastic modeling has recently become a popular method for accounting for the variability intrinsic in the epidemic process and will be considered heavily throughout this dissertation. Early Bayesian work in stochastic compartmental epidemic modeling was being done in 1998 [30], and early Bayesian work in stochastic SEIR models was underway by 1999 [65].

Contact networks and agent based modeling represent a third method of analyzing more granular data. These models have been researched intensively in the past fifteen years due to increased computing power. Early work was underway as early as 1999 [53]. Complex networks appropriate for diseases with highly variable transmission contact rates have been developed, as these types of pathogens are not well modeled by standard stochastic approaches [6] [46] [45]. Recent work has developed time-varying contact network methodology to further account for contact heterogeneity [35]. Contact networks which emphasize physical distance have also been developed [67]. A recent development in these models are gravity models, which allow for contact probabilities to depend on a power relationship to distance [57].

The contact network approach and stochastic approaches are particularly active today, with most current development occurring in those frameworks.

1.2 Assumptions Typically Made in SIR and SEIR Models

Many of the assumptions in the Kermack and McKendrick framework are still present in the majority of the models in the literature today, though most of these assumptions have been relaxed in at least one model or framework. The main assumptions are:

1. The Exponential Assumption. Exponentially distributed latent and infectious times (relaxed in [9] [40] [48] [82] [84]).

2. Homogeneous mixing. Any individual is equally likely to contact any other individual in the population on a given day (relaxed in contact network and agent based models).

3. Identical disease processes for every individual (this assumption is typically considered valid at the population level).

4. Equal susceptibility for every individual (relaxed in [66], agent based models).

5. Linearity in the number of new infectious with respect to the number of infectious individuals (relaxed in contact network models).

The sheer number of assumptions relaxed in agent based and contact network models seems to argue for their usage, and is part of the reason that these models have become so popular. However, these models require a great amount of additional information and require granular data. Oftentimes, constructing a contact network for a disease outbreak in the general population is impossible, and so there is still validity in other approaches. The current popularity of the stochastic approach is due, in part, to the strong data requirements of the contact network approach.

1.3 Outline of the Dissertation

The goal of this thesis is to create a single, flexible model which can simultaneously relax the first two assumptions in the spatial lattice framework. This model would be useful for data sets collected at some level of aggregation that does not have an easily accessible contact network to use (such as the Mumps data we analyze here, or the Google Influenza data). We also wish to develop a model which can smooth the data predictions. Due to the stochastic nature of the epidemic process, variations in disease counts oftentimes result from the contraction mechanism, rather than some spatial consideration. It is possible that, over the course of an epidemic, this process can lead to widely varying disease counts in neighboring locations. Smoothing in the prediction process would allow more accurate prediction of new epidemics and yield a better understanding of the epidemic process.

Most population level stochastic models utilize the exponential assumption, because it is difficult to relax at the population level. Such models typically only allow a single other distribution to be used (e.g. [9]). Chapter Two contains the necessary background with regards to the assumptions to be relaxed in this dissertation, as well as background on the Mumps epidemic. In Chapter Three of this thesis, we will develop a flexible framework that allows the exponential assumption to be relaxed and any discretized distribution to be utilized for the latent and infectious times. The key difference between this model and others is that mixing will occur at the population level, and it will allow more than one latent and infectious time distribution to be used. This makes it ideal for embedding into a spatial model.

We will demonstrate in this thesis that the Mumps epidemic spread through the Iowa county lattice along two major spatial conduits: the highway system and though the crossing or borders into adjacent counties. An intuitive way of allowing this sort of spread and still performing spatial bleeding is the use of a conditional

autoregressive (CAR) model. These are known in the spatial literature to induce spatial smoothing. However, not every county in Iowa contains a major highway, so a model which can incorporate the use of mismatched lattices is desirable. A literature search will reveal that a CAR model which can handle multiple correlated outcomes on mismatched lattices does not exist. In Chapter Four we develop such a model, for use in embedding into the SEIR structure.

In Chapter Five, we will develop a spatial SEIR model which embeds the CAR model from Chapter Four into the SEIR model from Chapter Three. This will allow for a single, flexible model which allows epidemic modeling on lattices to be modeled in a new and more realistic way. The model will not only relax the exponential assumption and homogeneous mixing assumptions on a lattice, it will allow for spatial smoothing of epidemic predictions to occur, which allows for more realistic prediction of future epidemics. This model represents the culmination of the thesis.

We will end with methodological conclusions and conclusions with regards to the Mumps data in Chapter Six.

CHAPTER 2
BACKGROUND

2.1 Chapter Goals

This chapter is designed to give the required background regarding the motivating data for this dissertation work. The Mumps Epidemic will be described as well as a literature review of work that has been done to relax the exponential assumption and the homogenous mixing assumption in SEIR models.

2.2 The Mumps Data

In 2006, the state of Iowa was the center of a large Mumps epidemic. The data set we procured contained 214 individuals in Iowa and the surrounding states whose Mumps infection was confirmed by swab culture. Of these 214 individuals, 205 were diagnosed in Iowa, and their county of residence was recorded, along with the date of diagnosis. In Chapter Three, the 214 individuals will be considered as a single population. In Chapters Four and Five, we will consider only the 205 individuals, and additionally consider the pattern of spatial spread using the individuals' counties of residence.

There are considerations which must be addressed before considering a SEIR model for this epidemic. The first is that an exponential distribution will fit the latent time of Mumps quite poorly. Mumps is known to be latent for a typical period of 16-18 days, and almost never less than 12 days or more than 25 days [12][13]. The memoryless exponential distribution is not adequate to account for this type of distribution. A Gamma or Weibull distribution may be flexible enough to fit this process, however. These are typical time to event distributions in the parametric survival literature, and so they are desirable distributions to fit a latent or infectious time distribution.

Another issue that must be accounted for are time dependent interventions. Two events occurred during the epidemic which may have changed the mixing process. The first was spring break, which occurred approximately 40 days into the epidemic. Polgreen et. al. have shown that this affected the age distribution of new infections, shifting it from being primarily college aged students to an increasing number of individuals outside of the typical college ages [69]. We do note that their analysis considered all probable cases, whereas our analysis considered only confirmed cases. The epidemic curve gives reason to believe that this may have lead to an increase in infections. The second is what we will term the "public health awareness intervention". The CDC declared an epidemic of Mumps in Iowa March 30, 2006 [11], and made suggestions to reduce contacts. Recent work has been done suggesting that epidemic models must account for public awareness in modern epidemics, because contemporary medicine has given individuals knowledge of how to prevent infections [26][43][56][70]. We will also model this process for the Mumps epidemic. It is important to note that we consider this the likely cause of the change in mixing, but that other explanations may exist for this change. Natural changes in the Mumps transfer process due to weather, or even differences in the criteria for swab culture may also have contributed to the time dependent nature of the mixing in this model.

Because the state of Iowa is a large enough geographic area to consider as having a spatial pattern, the homogeneous mixing assumption will need to be addressed. While our analyses in Chapter Three will consider the state as a single unit, homogeneous mixing is certainly violated in these analyses. Chapter Five will utilize homogeneous mixing at the county level in order to alleviate these violations. This will require us to account for the spatio-temporal pattern in the infections. In Chapter Five, we will build a flexible spatial SEIR model to handle this pattern.

As was previously mentioned, graphing the Mumps data over time shows two main conduits for epidemic spread. The first being county to county transmission along borders and the second being spreading along highways. In a rural area such as Iowa, highways may serve as an important measure of proximity in terms of epidemic spread. Methods utilizing air travel as a proxy for country to country spread are known in the literature (e.g., [3][19][39]). In fact, simulations based on prior information regarding the latent and infectious periods of Mumps yields an average of only 6 to 10 cases of Mumps that cannot be explained by one of these two methods of spread. This argues strongly for these being the primary proximity measures in this epidemic.

2.3 Developments with Regard to the Exponential Assumption

One of the salient features of the original models is that the latent and infectious times of the infectious disease under consideration are exponentially distributed, known as the exponential assumption [2]. The exponential assumption was one of many assumptions from the early deterministic models that carried into the stochastic methodology.

The exponential assumption tends to be a convenient assumption in modeling, largely due to the simple form of the models and simpler imputation in the stochastic framework. Unfortunately, this assumption is unrealistic for many infectious diseases. The memoryless property of the exponential distribution requires a constant probability of moving to the infectious compartment on day $j+1$ after j days of a latent infection, regardless of j.

Kenah showed that the infectious time distribution has a marked effect on the probability of a major epidemic [48], and Wearing et. al. demonstrated bias in the basic reproductive number of the microorganism when the time to infection

is incorrectly assumed to be exponentially distributed [82]. The work has strong implications for the accuracy of predictions of future epidemics in the stochastic framework. The basic reproductive number is defined as the average number of secondary infections caused by an infectious individual in a fully susceptible population. If the estimate of this quantity is biased, it may result in inaccurate predictions of the final sizes of future epidemics and inaccuracies in the analysis of the efficacy of public health interventions. To our knowledge, no one has investigated the variance of the parameters when the exponential distributions on the latent and infectious times are wrongly assumed. Because the parameter variances in SEIR models are already large [24], we feel this is worth investigating.

The idea of relaxing the exponential assumption is not new, with work done in the stochastic framework as early as 1948 [49]. Recently, work has been done to relax the exponential assumption in both the Bayesian and frequentist frameworks. However, one still tends to see two types of models: (1) models that utilize the exponential framework (e.g., [56][65]), or (2) models which make very strong assumptions in the exposure times (e.g. where it is assumed the initial exposure times are known (e.g., [78]), or where the latent period is assumed to be fixed (e.g.,[66])). A notable exception is found in Boys' and Giles' paper, which provides a model which can use gamma distributions for the latent and infectious classes and does not require initial exposure times [9]. While their method only handles gamma distributions, this approach may work well for many infectious diseases. It has been suggested that a gamma distribution may fit many infectious diseases [59][82]. Additionally, Jewell et. al. propose an SIR model allowing general infectious periods, but do not extend this model to the SEIR structure and propose a very different method than we propose here [40].

At the individual level, work has been done to relax the exponential assumption (e.g,[66][84]). These models work quite well when individual level information is available for small populations. However, they are limited in the regard that many interventions will be applied at the population level, and the indices of the individuals receiving the interventions may need to be imputed. Additionally, many epidemics occur in large populations, and computation may be slow with individual level models.

In order to embed a spatial pattern into a SEIR model for an area as large as Iowa, a SEIR model which allows for mixing at the population level and still allows for non-exponential latent and infectious times is desirable computationally. Such a model will be developed in Chapter Three.

2.4 Developments with Regard to the Homogeneous Mixing Assumption and Spatial Epidemic Modeling

There are at least two reasons for relaxing the homogeneous mixing assumption. One is that the contact distribution is variable, meaning that the number of contacts made by individuals is highly dispersed. A second is that the spatial area affected by the epidemic is too large for the assumption to adequately hold. In the case of the Mumps epidemic, the latter reason cited is the primary concern. Due to the contact mechanism of Mumps, it is likely that the contact distribution for individuals with Mumps is not too dispersed to be modeled by a Poisson contact distribution, which is the typical way of handling diseases such as Mumps, Measles, or Influenza [61]. For diseases with highly variable contact rates, such as sexually transmitted diseases, graph methods may work better [23].

Traditionally, the vast majority of epidemic models found in the literature have considered only a single location, though work was being done in as early as 1990

to consider the spatial epidemics (e.g., [4]). This appears to be due in large part to the ubiquitous nature of the homogeneous mixing assumption. As mentioned, the homogeneous mixing assumption dates back to the original models by Kermack and McKendrick, and states that any Susceptible individual in the population is equally likely to contact any Infectious individual in the population within any fixed time frame. For populations large enough to consider spatially, this assumption is almost certainly violated. Additionally, it is known that correctly considering the spatial framework within the mixing has implications for vaccine efficacy analysis [74]. Other key parameters and interventions may be similarly affected.

Recent work has relaxed this assumption or removed it altogether. Typically, one of three types of models are applied to create a spatial SEIR model. The first are SEIR models which used a point referenced spatial approach to model the distances between individuals or locations (e.g., [22][38][54]). The second are person to person contact networks. These have been well developed in the literature and work quite well for modeling populations where individual level information is known (e.g., [6][16] [17] [31][52] to name a few). These can also be used to completely relax the homogeneous mixing assumption within a single location as well. However, these tend to work best when the contact network is known based on *a priori* information or a previous study, and can be computationally limiting. To construct a person to person contact network for the state of Iowa would require its own study, and would potentially make as many assumptions as a population based approach.

The preferred types of spatial models for lattice data like the Mumps data are location to location contact networks (eg, [3][39][73]). These assume homogeneous mixing within a given location, but often develop contact probabilities which allow nonhomogeneous mixing between different locations. Weights are often defined *a priori* [39] or based on prior information [33] (the latter being a relatively recent

development in the epidemic modeling literature). These models work well and are intuitive, but there is little flexibility in methods for weighting contacts and additionally in drawing strength in estimation. Contact graph methods which allow weights to be estimated based on the idea of drawing strength from related edges do not appear in the literature. For a data set that is sparse, such as the Mumps data, flexible ways to draw strength in estimating spatial mixing are a key component to accurately modeling the disease process. A new method beyond the standard contact graph analysis may allow the flexibility to accurately model sparse epidemic data. Gravity models offer an obvious alternative to the traditional contact method approach. However, an approach which would allow spatial smoothing to occur is desirable, because low counts and sparse counts are likely due to the probabilistic nature of the infection process, rather than the properties of the units themselves.

The Conditional Autoregressive (CAR) family of models is commonly employed to induce spatial correlation on a lattice. Lawson has employed a CAR model in the analysis of spatial epidemics [55]. The technique is designed to smooth the diseases counts of Influenza data, however, and does not consider the spatial bleeding mechanism of the virus. This is reasonable for Influenza data due to the incomplete nature of the infection counts, but the Mumps data are likely to be more complete. A CAR approach to directly model spatial bleeding may be more appropriate for the Mumps data, and many other data sets where counts can be assumed to be relatively accurate. This technique could be embedded into the graph technique to create a more flexible graph structure than currently exists in the literature. This is the approach that will be developed in this thesis.

CHAPTER 3
A PATH-SPECIFIC SEIR MODEL FOR USE WITH GENERAL LATENT AND INFECTIOUS TIME DISTRIBUTIONS

3.1 Chapter Goal

In this chapter of the thesis, we develop a model where mixing occurs at the population level but does not require exponentially distributed latent and infectious times. This model is helpful in and of itself as a computationally efficient model relaxing the exponential assumption, but its mixing properties also allow it to be utilized in a spatial framework for a location to location graph approach. This property will allow it to be used to develop a spatial model in Chapter Four. We call this model the Path-Specific SEIR (PS SEIR model), because it mixes properties of population level models and individual models, and operates at what we will call the "path level", a term defined later.

3.2 Methods

In this section, we first propose the PS SEIR model and derive it as the analog to a class of deterministic SEIR models.

3.2.1 Proposed Model

The main goal of this section is to demonstrate how the exponential assumption can be relaxed, and how discretized distributions can be implemented for the latent and infectious periods. We utilize the more realistic assumption that there is a maximum time that an individual may sustain a latent infection before becoming actively infectious. We assume that all individuals in the exposed category will eventually move to the infectious category, as is done in Lekone and Finkenstädt [56] and Anderson and May [2], and we do not consider cases of exposure without

latent infection at this juncture.

The population averaged SEIR model of interest is found in Lekone and Finkenstädt, which is itself the generalization from the SIR model found in Mode and Sleeman[61]:

$$\begin{aligned} S_i \to E_{i+1} &= binomial(S_i, 1 - exp(-f(\underline{\psi}, i)h\tfrac{I_i}{N})); \\ E_i \to I_{i+1} &= binomial(E_i, 1 - exp(-h/\rho)); \\ I_i \to R_{i+1} &= binomial(I_i, 1 - exp(-h/\gamma)). \end{aligned} \tag{3.1}$$

Define i=1,...,T as a subscript for discrete time and S_i, E_i, I_i, and R_i represent the counts of individuals in the susceptible, exposed, infectious, and removed compartments at time i, respectively. The notation $S_i \to E_{i+1}$ denotes a change of category. Let $f(\underline{\psi}, i)$ represent the mixing and possible intervention functions controlling the number of new exposures at time $i+1$ and is constrained to be nonnegative, and let h represent the number of days between time points in the data collection partition. The total number of individuals in the population is denoted by N.

For models utilizing the exponential assumption, the exposure data are typically arranged as a T-dimensional vector of counts, $E = (E_1, ..., E_T)'$. Note that, in these models, the only necessary information for the evaluation of the likelihood is the total count in the exposed category at each time point, E_i. We relax this assumption by not only counting the number of exposed individuals at each time point, but also by utilizing the length of time each individual has been in the exposed compartment. Consider collecting the exposure counts in a TxM_1 matrix $\mathbf{E}$, where M_1 is the maximum amount of time the infectious agent can remain latent. Cell (i,j) then contains a count of the number of individuals who are at time point j of the latent infection process on time point i of the epidemic. In other words, i represents objective, calendar time since the start of the epidemic, and j denotes the subjective, individual time in the diseases process. In practice, i and j will typically

be measured in days, although this is certainly not required. The TxM_2 infectious matrix $\mathbf{I}$ is defined analogously to $\mathbf{E}$, with the rows representing the number of time points elapsed since the start of the epidemic, and the columns representing the number of time points an individual has remained in the infectious compartment.

When an individual is newly exposed and contracts a latent infection at time i, the individual moves from the susceptible class into row i, column 1, of the exposed matrix $\mathbf{E}$. The individual then takes a diagonal path, moving one column to the right, j+1, and one row down, i+1, for every time unit in which the individual does not become infectious. When the individual becomes infectious at time i', the individual moves to row i', column 1, of $\mathbf{I}$, and repeats the process until removed. This process allows the length of time each individual is in the exposed category to be imputed, and allows for many latent time and infectious time distributions to be discretized and utilized. Specifying a maximum length of time which an individual may have a latent infection, or be infectious, allows the number of columns of the matrix to be defined *a priori*, and removes the need to adaptively choose the size of the matrix as the analysis is running. While an adaptive scheme may be possible, it is not necessary to do so, since the maximum amount of time an infectious agent may remain in a latent state is often known. Additionally, an adaptive scheme may not be computationally efficient.

Because the exposure data and infectious data are being collected in matrices, the probability of compartmental change can vary with the amount of time an individual has stayed in the compartment. This allows the exponential assumption to be relaxed, and any distribution can be discretized and used to approximate the true, underlying latent and infectious time distributions. As noted in the introduction, this allows more realistic distributions to be used for infectious diseases.

With this structure in place, the investigator is able to use strong prior knowledge of the length of time that individuals spend in the exposed and infectious categories. Typically, this information is available and multiple distributions may be fit and compared. It is unlikely that there will be strong prior information for the mixing and intervention parameters, so relatively weak priors can be used for these parameters.

The proposed PS SEIR model follows: Let i denote discrete calendar time since the beginning of the epidemic, and j denote discrete time that an individual has spent with a latent infection or in an infectious state. Then,

$$\begin{aligned}
&S_i \to E_{i+1,1} = binomial(S_i, 1 - exp(-f(\underline{\psi}, i)h\tfrac{I_{i+}}{N})) \equiv W_i;\\
&E_{ij} \to I_{i+1,1} = binomial(E_{ij}, P(Z_1 \leq j + h | Z_1 > j)) \equiv X_{ij};\\
&E_{ij} \to E_{i+1,j+1} = E_{ij} - X_{ij};\\
&I_{ij} \to R_{i+1} = binomial(I_{ij}, P(Z_2 \leq j + h | Z_2 > j)) \equiv Y_{ij};\\
&I_{ij} \to I_{i+1,j+1} = I_{ij} - Y_{ij}.
\end{aligned} \tag{3.2}$$

These definitions follow from Equation 1, where X_{ij}, Y_{ij}, W_i, and E_{ij} are all unobserved, while $\sum_j X_{ij}$ and $\sum_j Y_{ij}$ are known. Z_1 is a random variable defined by the exposure distribution and Z_2 is defined by the infectious distribution. Let $f(\underline{\psi}, i)$ represent the mixing and possible intervention functions controlling the number of new exposures at time $i + 1$, and is constrained to be nonnegative, with $\underline{\psi}$ representing the vector of parameters controlling mixing and interventions. Let I_{i+} represent the total number of infectious individuals at time i, h represent the number of days between data collection times, and N represent the total number of individuals in the population.

The compartments are the Susceptible, Exposed, Infectious, and Removed classes, respectively. Define a bin as the amount of time between data collection times. In our discretization scheme, a bin will be h time units (often measured in

days). Bins are used within the Exposed and Infectious compartments as the basic time unit for the discretizations. Most data sets will use $h = 1$, but all that is required is $0 < h < \infty$. This style of discretization allows for the analysis of large data sets, while still providing the flexibility to use a time-dependent conditional probability of changing compartments. By defining bins within the Exposed and Infectious compartments, it is possible to vary the conditional probability of a compartment change depending on the length of time an individual has spent in the compartment, which, in turn, allows for distributions other than the exponential distribution to be used for the latent and infectious times.

3.2.2 Derivation of the PS SEIR Model

The PS SEIR model can be derived as a stochastic analog to the following nonlinear system of ordinary differential equations:

$$\begin{aligned} \tfrac{dS}{dt} &= -f(\underline{\psi}, t)S\tfrac{I}{N}; \\ \tfrac{dE}{dt} &= f(\underline{\psi}, t)S\tfrac{I}{N} - g(\underline{\alpha}, E); \\ \tfrac{dI}{dt} &= g(\underline{\alpha}, E) - h(\underline{\gamma}, I); \\ \tfrac{dR}{dt} &= h(\underline{\gamma}, I). \end{aligned} \tag{3.3}$$

Several assumptions are made in this process, and we outline the core assumptions here.

1. Assume a homogeneous population with regards to susceptibility. This is commonly assumed in population averaged models.

2. Assume a Poisson contact rate for infectious individuals. This works well for diseases such as Mumps or Measles, but works poorly in models for sexually transmitted diseases, such as Gonorrhea or Chlamydia. 3. Define the Exposed compartment as only containing those who will eventually become infectious, and do not consider the possibility of a return to the Susceptible class.

4. Assume constant infectivity throughout the course of the infectious process.

5. Assume independent probabilities of moving from the Exposed Compartment to the Infectious Compartment (as well as from the Infectious Compartment to the Removed Compartment). Individuals are treated as having identical latent and infectious time distributions.

6. Homogeneous Individuals in terms of the disease process.

Suppose $g(\underline{\alpha}, E) \geq 0$ and the equation $\frac{dE}{dt} = -g(\underline{\alpha}, E)$ has the solution $E = G(\underline{\alpha}, C, t)$ where C is the constant of integration. If there exists a C^* such that $\frac{E}{E_0} = F(\underline{\alpha}, C^*, t)$ where E_0 is the total number of individuals who will become exposed throughout the epidemic, and the conditions $F(\underline{\alpha}, C^*, 0) = 1$ and $F(\underline{\alpha}, C^*, \infty) = 0$ are met, then $F(\underline{\alpha}, C^*, t)$ is a survival function, and a path-specific analog can be found to the deterministic system under consideration.

Given that an individual is in the Exposed compartment at time t, the probability of a compartmental shift to I before time $t + h$ is $\frac{F(\underline{\alpha},C^*,t)-F(\underline{\alpha},C^*,t+h)}{F(\underline{\alpha},C^*,t)}$. To discretize the above nonlinear system, partition the time over which the epidemic occurs into regular blocks of length h. Then $F(\underline{\alpha}, C^*, t)$ can be approximated by $\prod_{k=0}^{j} \frac{F(\underline{\alpha},C^*,kh)-F(\underline{\alpha},C^*,(k+1)h)}{F(\underline{\alpha},C^*,kh)}$ for $t\epsilon((jh, (j+1)h)$. This places a point mass for the discretization on the left endpoint of the interval, which is consistent with the discretization of the binomial process leading to new exposures below.

Note that $\frac{F(\underline{\alpha},C^*,j)-F(\underline{\alpha},C^*,j+h)}{F(\underline{\alpha},C^*,j)} = P(Z_1 \leq j + h | Z_1 > j)$. Consider breaking the Exposed compartment into M_1 distinct bins. In bin j, there is a constant probability, $P(Z_1 \leq j + h | Z_1 > j)$, of a compartment change. Using i.i.d. Bernoulli random variables to accommodate the process of a random compartmental change for each individual in bin i, we see that the number of individuals moving from $E_{i,j}$ into $I_{i+1,1}$ is distributed as $binomial(E_{ij}, P(Z_1 \leq j+h | Z_1 > j))$. Assuming that the number of individuals experiencing compartmental change is independent of the bin

after conditioning, we retrieve the path-specific process for the latent times found in the PS SEIR formulation. Note that the same process will yield the path-specific process for infectious times.

Next, we derive the binomial process leading to new exposures. The derivation is very similar to the one found in Mode and Sleeman [61], with the difference being that we use discretized Poisson contact rates and probabilities of contracting infections from infectious individuals. This does not change the overall flow of the derivation found in their book, but our additions are informative to the properties of our model, and so we include the derivation here.

Assume contacts are made at a rate $P(C(t) = c) = x(c, t)$, where $C(t)$ is a random variable indicating the number of contacts an individual makes at time t, and c is a realization of this random variable. Assuming independent contacts, the probability of escaping infection between time t and time $t+h$ is $Q(t) = \sum_{c=0}^{\infty} x(c, t)q^c(t)$, where $q(t)$ is the discretized probability of not developing an infection based on a contact with a single individual in this interval. Assume contacts are made according to a process that can be approximated discretely by a set of T Poisson random variables with rates $\lambda(t)$, which may vary over the T time points. This will allow for differences in contact rates due to interventions. Additionally, denote the probability of contracting a latent infection based on contacting a random member of the population as $q(t)$, where $q(t)$ has been discretized into T values. Then, for a fixed time point t, we have $Q(t) = \sum_{c=0}^{\infty} \frac{exp(-\lambda(t)h)(\lambda(t)h)^c}{c!} q^c(t)$, implying $Q(t) = exp(h\lambda(t)(q(t) - 1)) = exp(-h\lambda(t)p(t))$.

In order to include public health style interventions into our model, assume $p(t) = p^*(t)I(t)/N$, where $p^*(t)$ is the time dependent probability of contracting an infection based on contacting a single infectious individual from time t to time $t + h$, and is discretized into T values. Note that we have parameterized $\lambda(t)p^*(t)$

as $f(\underline{\psi}, t)$ because there is only data to estimate one parameter in the case where there are no interventions. Therefore a single mixing parameter is typically used in these cases, rather than the product $\lambda(t)p(t)$.

Finally, approximate the mixing process $f(\underline{\psi}, t)SI/N$ by discretizing it and placing a pointmass at the left endpoint of the interval $(ih, (i+1)h)$. This yields the probability distribution of the number of new latent infections at time $i+1$, given the state of the system at time i, as $binomial(S_i, 1 - exp(-f(\underline{\psi}, i)h\frac{I_{i+}}{N}))$. This completes the derivation.

As an example of this process of determining the conditional probabilities for the path-specific process through the exposure matrix, consider the standard case of exponentially distributed latent times. Anderson and May [2] show that the form of $\frac{dE}{dt}$ is:

$$\frac{dE}{dt} = f(\underline{\psi}, t) - \alpha E.$$

In deriving the survival function, one only needs to consider $\frac{dE}{dt} = -\alpha E$. Solving this equation yields $E(t) = Cexp(-\alpha t)$, where C is determined by the initial conditions of the system. In order to determine the survival function, we consider $E(t = 0) = E_0$, where E_0 is the total count of individuals who will move into the Exposed compartment over the course of the epidemic. The survival function is then $\frac{E}{E_0} = exp(-\alpha t)$, which can be identified as the survival function of an exponential random variable. If we call this random variable Z_1, the discretization we consider is $P(Z_1 < t + h | Z_1 > t)$. This yields $exp(-\alpha h)$, which is constant in t, and is of the form commonly seen in the SEIR model literature.

3.2.3 Equivalency of the PS SEIR and Population Averaged SEIR Models Under the Exponential Assumption

Given the assumption that we do not place a maximum limit on the amount of time that a patient spends in the latent category, we are able to demonstrate the equivalency of our model, as defined in Equation 2, with the population averaged model as defined in Equation 1. The following assumptions are made:

1) Once in the exposed class, a patient must wait at least one time unit before moving to the infectious class.

2) Assume that the total number of patients entering and exiting the infectious class is known and fixed at every time point. This causes cancelation in the ratio of likelihoods for the intervention and mixing parameters when applying MCMC sampling. The results that hold for the exposed class also hold for the infectious class, with the only modification being that the number entering the class at any given time point is known.

3) Homogeneous mixing.

4) We consider a homogeneous population in terms of susceptibility.

5) There is only one individual in the infectious compartment, no individuals in the exposed compartment, and a fixed and known number in the removed compartment at the start of the epidemic.

6) WLOG, h=1 day.

The following notation will be used: P_M will represent the full probability distribution of the population averaged model, and P_I for the PS SEIR model. In the population averaged model, E_i will represent the total number of exposed individuals at time i, and E_i' will represent the number of individuals who are newly exposed on the i^{th} day of the epidemic. In the PS SEIR model, the notation $\mathbf{E}_{i,j}$ will

be used, where each element will represent the number of individuals who have been exposed for j days on the i^{th} day of the epidemic. For example, $\mathbf{E}_{i,1}$ represents the number of new exposures on day i. The notation $\mathbf{E}_{i+}$ will represent the marginal total of exposed individuals on day i. In both models, the same notation will be used for the infectious category, using I_i, I'_i, $\mathbf{I}_{i,j}$, $\mathbf{I}_{i+}$.

The full distribution for the population averaged model can be written as:

$$P_M(\underline{\psi}, \rho | E_i \forall i, I_i \forall i) = \prod_i L_{E'_i, S_i, \underline{\psi}} \binom{E_{i-1}}{I'_i} (1 - g(\rho))^{I'_i} g(\rho)^{(E_{i-1} - I'_i)} p(\underline{\psi}, \rho), \tag{3.4}$$

where i subscripts time, $\underline{\psi}$ represents all mixing and intervention parameters, and ρ represents the parameter of the exponential distribution for the exposed class, with $g(\rho) = exp(-1/\rho)$. $L_{E'_i, S_i, \underline{\psi}}$ is the likelihood of the process producing new exposures. This process is assumed to have the same parameterization for both models.

In what follows, we show that an MCMC sampler will draw from the same posteriors for the parameters whether the specification is the population averaged model or the PS SEIR model.

By assumption, we know the number of new infectious individuals at day i. We assumed there was only one infectious individual at day 1, so S_i is fixed for all i. Note that with E_{i+} fixed and S_i known, we have $E_{i,1}$ fixed and known. Consider the PS SEIR model. For a given realization of $\mathbf{E}$, $E_{i+} = E_i$, where i subscripts the day of the epidemic, and j subscripts the bin of the exposed category.

The PS SEIR model can be written as:

$$P_I(\underline{\psi}, \rho, E_{i,j}) = \prod_i (L_{E_{i,1}, S_i, \underline{\psi}} \prod_j \binom{E_{i-1,j}}{E_{i,j+1}} (1 - g(\rho))^{I_{i1}} g(\rho)^{(E_{i-1,+} - I_{i,1})}) p(\underline{\psi}, \rho). \tag{3.5}$$

If we assume that $E_{i,1} = E'_i$ in the population averaged specification (and therefore E_{i+}=E_i) we have

$$P_M(\underline{\psi}, \rho | E_{i,1} \forall i, I_{i,1} \forall i) \propto P_I(\underline{\psi}, \rho | E'_i \forall i, I'_i \forall i)$$

. Thus, for a given realization of the $\mathbf{E}$ in the PS SEIR model, the full distribution of

the parameters in the PS SEIR model can be written in a population averaged form. Because the kernel of this full distribution of the PS SEIR model is proportional to that of the original population averaged model, it follows that the parameter realizations of the PS SEIR model come from the same posterior distribution as those from the original population averaged model whenever E_{i+}=E_i for all i.

To show that MCMC chains will draw from the same posterior distributions after burn in, it suffices to show that, given a realization of the parameters in the model, the distribution of E_{i1} and the distribution of E'_i are the same distribution. This follows because the full distributions for both the population averaged model and the PS SEIR model can be written in terms of $\underline{\psi}, \rho, E_i, E'_i$ under the assumptions in the population averaged model.

First, note that knowledge of the full set of infectious times $\{I'_i \forall i\}$ and initial conditions is sufficient to determine E_i from the set $\{E'_k, k \leq i\}$. Define Υ_i=$P(E'_i = E_i^{**} | E'_k = E_k^{**}\ \forall k < i, I'_k\ \forall k \leq i)$, under the population averaged model. Note that, given the starting conditions, the number of new infections, $I'_k\ \forall k \leq i$, and the number of new exposures, $E'_k\ \forall k \leq i$, the total number exposed at time i, E_i is known. We denote this by E_i^* for ease of notation. Now given a realization of $\underline{\psi}, \rho$ and assuming that the total number of patients in the infective category is known at every time point, consider time $i|i-1$:

$$\prod_i \Upsilon_i = \prod_i (L_{E_i^{**}, S_i, \underline{\psi}} \binom{E_i^*}{I'_{i+1}} g(\rho)^{E_i^* - I'_{i+1}} (1 - g(\rho))^{I'_{i+1}}) p(\underline{\psi}, \rho). \tag{3.6}$$

Now, we turn to the PS SEIR model to show that $P(E_{i,1} = E_i^{**} | E_{k,1} = E_{k,1}^{**} \forall k < i, I_{k,1} \forall k \leq i) = \prod_i \Upsilon_i$. To begin, assume that all of the patients l=1,...,n in the PS SEIR model are distinguishable once exposed. The full posterior for the PS SEIR model can be written as the likelihood of $E_{i,1}$ newly exposed individuals, multiplied by a product of Bernoulli likelihoods, representing the other exposed

bins. That is,

$$P_I(\underline{\psi}, \rho, E_{i,+,1}, E_{i,l,j\neq 1}) = \\ \prod_i (L_{E_{i,+,1},S_i,\underline{\psi}} \prod_l \prod_{j\neq 1} (1-g(\rho))^{I_{i,1,l} * E_{i-1,l,j}} g(\rho)^{(E_{i-1,l,j} - I_{i,l,1} * E_{i-1,l,j})}) p(\underline{\psi}, \rho) \tag{3.7}$$

where E_{ilj} is the indicator that person l is in exposed category j at time i, and $E_{i,+,1}$ is the number of new exposures at time i. For clarity on the terms $I_{i,l,1}E_{i-1,l,j}$, consider the following: if person l is not in exposed category j at time $i-1$,then both $E_{i-1,l,j}$=0 and $I_{i,l,1}E_{i-1,l,j}$=0 in the Bernoulli likelihood, and these terms do not change the value of full likelihood. If person l is in exposed category j at time $i-1$, then there are two options: 1) person l can become infectious, indicating the contribution to the likelihood is $(1-g(\rho))$ or 2) person l can remain exposed, with a contribution of $g(\rho)$, which are the conditional probabilities in the exponential assumption framework of 1) a person becoming infectious given they were exposed the previous day, or 2) staying exposed given they were exposed the previous day, respectively.

Note that knowledge of the set $\{E_{k,+,1}, k \leq i\}$ and the set $\{I_i \ \forall i\}$, along with the initial conditions (one individual in the infectious compartment, none in the exposed compartment, a fixed and known number in the removed compartment), fully determines E_{i++}, where $E_{k,+,1}$ is the number of individuals who entered the exposed class at time k, and E_{i++} is the population averaged total number of individuals in the exposed class at time i in the distinguishable framework.

Given a single path from $i-1$ to i, we have the following posterior:

$$P(E_{i,l,j\neq 1}, E_{i,+,1} | \underline{\psi}, \rho, E_{k,+,1} \forall k < i, E_{i-1,l,j}, I_{k,+,1} \forall k \leq i) = \\ L_{E_{i,+,1},S_i,\underline{\psi}} \prod_{j,l} (g(\rho)^{E_{ilj} - I_{i+1,l,1}E_{ilj}} (1-g(\rho))^{I_{i+1,l,1}E_{ilj}} * 1_{(E_{i-1,l,j-1} \geq E_{ilj})}) p(\underline{\psi}, \rho) \tag{3.8}$$

where we have assumed there is no maximum limit on the length of time a patient may stay in the exposed category. The indicator variables $1_{(E_{i-1,l,j-1} \geq E_{ilj})}$ are in place to indicate that paths through the exposed matrix are diagonally non-increasing.

For example, it is impossible for there to be an individual in exposed category j at time i, given that there was not an individual in exposed category $j-1$ at time $i-1$.

Next, we derive $P(E_{i,+,1} = E_i^{**} | I'_k \ \forall k \leq i, E_{k,+,1} = E_k^{**} \ \forall k < i)$ in the PS SEIR model. This will demonstrate that the distribution of the marginal sums for the PS SEIR model is equal to the probability distribution of the missing data for the population averaged model. To demonstrate this more clearly, we first consider a single path,

$$P(E_{i,+,1} = E_i^{**} | I_{k,+,1} \forall k \leq i, E_{k,+,1} = E_k^{**} \forall k < i) = \\ L_{E_{i,+,1}^{**}, S_i, \underline{\psi}} \prod_{j \neq 1} \prod_l (g(\rho)^{E_{ilj} - I_{i+1,l,1} E_{ilj}} (1 - g(\rho))^{I_{i+1,l,1} E_{ilj}} 1_{(E_{i-1,l,j-1} \geq E_{ilj})}) p(\underline{\psi}, \rho), \tag{3.9}$$

where we have evaluated the first product over l and considered only one particular path. The second product remains unevaluated because it is a fixed quantity for that given path.

To evaluate the second product, recall $E_{i++} = E_i^*$, which is fixed where the number of new exposures for $E_k, k \leq i$ and $I_k, k \leq i$, and the starting conditions of the system are known. Also note that $\sum_l \sum_j I_{i+1,l1} E_{ilj} = I_{i+1,+,1}$, which denotes all patients who were in an exposed class at time i and moved to the infectious class at time $i+1$. Thus, for a single path we write

$$P(E_{i,+,1} = E_i^{**} | E_{k,+,l} = E_k^{**} \forall k < i, I_{k,+,1} \forall k \leq i) = \\ \prod_i L_{E_{i,+,1}^{**}, S_i, \underline{\psi}} g(\rho)^{E_i^* - I_{i+1,+,1}} (1 - g(\rho))^{I_{i+1,+,1}}. \tag{3.10}$$

Now, consider the probability over all possible paths. We see the probability becomes

$$P(E_{i,+,1} = E_i^{**} | E_{k,+,l} = E_k^{**} \forall k < i, I_{k,+,1} \forall k \leq i) = \\ \prod_i L_{E_{i,1}^{**}, S_i, \underline{\psi}} \binom{E_i^*}{I_{i+1,+,1}} g(\rho)^{E_i^* - I_{i+1,+,1}} (1 - g(\rho))^{I_{i+1,+,1}} = \prod_i \Upsilon_i. \tag{3.11}$$

Summing across all valid paths yields the combinatorics. Considering only valid paths allows us to drop the indicator variables, as the cases relating to paths which

are not valid yields a probability of zero. There are $\binom{E_i^*}{I_{i+1,+,1}}$ valid possible paths that yield the combinatoric, because, given an exposed vector at time i, we will have $I_{i+1,+,1}$ patients leaving the exposed class. Once these $I_{i+1,+,1}$ patients move out of the exposed class, the remaining patients will be forced to move diagonally through the exposure matrix. This demonstrates that the distribution of the margins of the missing data in the PS SEIR model with distinguishable patients is the same as the distribution of the missing data in the population averaged model.

The previous result is for a given realization of patients. Now consider the patients to be indistinguishable. Note that when there are multiple patients in bin $E_{i,j}$ but fewer patients in bin $E_{i+1,j+1}$, this represents multiple potential paths. The number of paths represented is $\binom{E_{i,j}}{E_{i+1,j+1}}$. If we fix the values of the two rows of the exposure matrix (i.e. $E_{i-1,j}$ and $E_{i,j}$ are known for all j and a fixed i), then we see that the full posterior for $E_{i,1}$ given this particular arrangement of the exposure matrix is

$$\begin{aligned} &P(E_{i,1}|S_{k,+}\forall k \leq i, E_{k,1}\forall k \leq i, I_{k,+,1}\forall k \leq i) = \\ &\textstyle\prod_i L_{E_{i,1},S_i,\underline{\psi}} \prod_{j\neq 1} (\binom{E_{i-1,j-1}}{E_{i,j}} g(\rho)^{E_{i,j}-I_{i+1,1}E_{i,j}}. \\ &(1-g(\rho))^{I_{i+1,1}E_{i,j}} 1_{(E_{i-1,j-1}\geq E_{i,+,j})}) p(\underline{\psi},\rho). \end{aligned} \tag{3.12}$$

This accounts for certain paths being more likely than others, but only accounts for the current arrangement of the exposure matrix. There will be many arrangements of the matrix that results from the same E_{i+1} totals.

Due to the relationship between the Bernoulli distribution and the binomial distribution, we know that summing across all possible arrangements with patients being indistinguishable will also yield a full probability $\prod_i \Upsilon_i$. This can be seen by a simple argument. If we consider all possible paths as arising from a set of Bernoulli distributions, and we know that there are $\binom{E_i^*}{I_{i+1,+,1}}$ possible combinations that will yield $I_{i+1,1}$ new exposures, we can view this set as summing to a binomial distribution. Alternatively, we can partition the Bernoulli random variables into

groups, each as a binomial random variable with $E_{i-1,j-1}$ sample size and $E_{i,j}$ as the value taken by the random variable, as we have in the path-specific case. Now, the sum of all the Bernoulli random variables was a binomial as in the population averaged case (i.e. with $\binom{E_i^*}{I'_{i+1}}$ possible combinations), so we know that summing the binomial random variables arising from grouping the Bernoulli random variables will yield the same binomial random variable. This relies on the exponential assumption, because, for this proof to hold, the probabilities of success must be equal for all the bins, which is a consequence of the exponential assumption. So, for the PS SEIR model with indistinguishable individuals, we have

$$\begin{aligned} &P(E_{i,1} = E_i^{**} | E_{k,1} = E_k^{**} \ \forall k < i, newI_k \ \forall k \leq i) = \\ &L_{E_{i,1}^{**}, S_i, \underline{\psi}} \binom{E_i^*}{newI_{i+1}} g(\rho)^{E_i^* - newI_{i+1}} (1 - g(\rho))^{newI_{i+1}} = \prod_i \Upsilon_i. \end{aligned} \tag{3.13}$$

This shows the distribution of E_{i1} in the PS SEIR model is equivalent to the distribution of E'_i in the population averaged model.

Therefore, we have proven that, given a realization of the missing data, the posterior distributions of the parameters are equivalent between the population averaged and path-specific models. We have also shown that, given a set of parameter realizations, the marginal distributions of the missing data are equivalent between the two models. Since the posterior distributions are known to be equivalent, the MCMC chains for both models will draw from the same posterior distributions after burn-in when the assumptions stated at the beginning of the proof hold.

This proof demonstrates that the PS SEIR model proposed in Equation 2 is a generalization of the population averaged model found in Equation 1, and contains it as a special case.

3.3 Computing

In this section we propose an efficient sampling scheme for the exposure matrix, as we have defined it in Section 2. We then provide simulation results demonstrating the improvement of the PS SEIR model over the population averaged approach.

3.3.1 The Sampling Scheme

We recommend the following update scheme for the exposure matrix: 1) Select a time at which an individual moved to the infectious category. 2) Select, at random, a path that corresponds to this removal time. 3) Remove this path. 4) Select a new starting time for the exposure path, with equal probability placed on every day between one day before the transfer to the infectious compartment, and M_1 days before the transfer. 5) Add this path to the exposure matrix, and keep this update if $\frac{\Pi(\mathbf{E'})}{\Pi(\mathbf{E})}$ is greater than a randomly generated uniform random variable where $\Pi(\mathbf{E'})$ is the likelihood corresponding to the new exposure matrix, and $\Pi(\mathbf{E})$ corresponds to the likelihood of previous exposure matrix. 6) Repeat this until there have been a number of updates equivalent to 10% of the final epidemic size. We note that Lekone and Finkenstädt [56] first utilized the 10% value and indicated that it balanced mixing and speed for their update algorithm. For our algorithm, we note that updating 10% of the exposure yields very similar chains to a system where 50% of the exposures are updated, but runs faster.

This process varies the possible exposure times of each individual. Of course, one must find an appropriate place to start the MCMC chain, as it is not permissible for any individual to stay exposed for more than the maximum time, even in the starting condition of the chain. The prior information regarding the infectious disease is helpful here. We propose several chains be started at different possible

values of the mixing and intervention parameters, and **E** generated via a simulation that makes use of the best guess for the latent period distribution. In our experience, this technique allows for the most variability in the starting conditions of the chains, so convergence is easy to assess.

3.3.2 Properties of the Sampling Scheme

The aforementioned algorithm gives good convergence for reasonable exposure distributions, such as gamma or Weibull distributions which are not exponentially distributed. It is important to note that without an intervention or with a constant intervention (outlined in the simulation results), convergence was attained on every attempt, but convergence failed with one particular data set using a small epidemic generated from exponential distribution times and an exponential decay intervention. This was because the algorithm would occasionally generate an exposure matrix where none of the individuals contracted a latent infection after the intervention began. The intervention parameter realizations would increase, and it became very difficult for the algorithm to return to realistic values. We do caution researchers when using this approach for exponentially or nearly exponentially distributed latent times to check convergence. If the exponential assumption holds, we recommend the Lekone and Finkenstädt model. When the exponential assumption is violated, our PS SEIR model has good convergence properties, and offers advantages over other models in the literature.

Additionally, with weak prior information and large data sets, the sampling algorithm tends to move towards an exponential scheme. This is acceptable, as the simulation results demonstrate that the algorithm is most helpful for small epidemics. In these cases, weak prior information easily contains this process. With large data sets, the algorithm can be still be used. The strong prior information

typically available for the latent period can easily control this flattening of the conditional probabilities. If weak prior information is used, the algorithm will typically select a latent period where the conditional probabilities of compartmental change from the Exposed to Infectious compartments are more uniform than the prior information. This still yields advantages over the exponential assumption case, though these advantages are not as marked. When sampling unknown infectious paths via this scheme, where the total number moving into and out of the Infectious compartment are known at each time point, or if these data are available for the Exposed class, there is no flattening of the conditional probabilities of transferring from the Infectious to the Removed compartment, even with weak prior information and large data sets.

Additionally, when sampling from exponentially distributed latent and infectious time distributions, the MCMC chains of all the model parameters for the PS SEIR model draw from equivalent distributions to those of the population-averaged approach we outlined above. Table 3.1 demonstrates the equivalence between the MCMC chains generated by the population averaged and path-specific models. An epidemic consisting of 30 cases was simulated using exponential latent and infectious times. No intervention was used, and all infectious and removal times were assumed known. The quantiles provided were based on 7,000 iterations after burn-in. The mixing parameter posterior quantiles are almost exactly equal between the population averaged and path-specific model, but it may be noted that all the quantiles of the mean exposure time are slightly higher for the population averaged model than for the path-specific model. This is due to a practical consideration when using the PS SEIR model. The maximum time for the latent period was set to be 100 days in the path-specific approach, removing approximately the longest 1% of latent times from consideration. Thus, the differences are due to the coding

of the models, rather than a theoretical concern.

3.3.3 Simulation Results

For each simulated data set, there were 20,000 total individuals, with one member in the infectious category and all the other individuals susceptible at the start of the epidemic. The mixing parameter chosen to simulate the data was 0.25, in order to give a large degree of variability to the epidemic sizes. Let ψ_1 be the mixing parameter and ψ_2 be the intervention parameter. The exponential decay intervention we consider has the form $f(\underline{\psi}, i) = \psi_1 exp(-\psi_2 1_{(i \geq i_0)})$, where i_0 is the time that the intervention began. This represents an exponential decay in the probability of moving from the susceptible class to the exposed class. For this form, ψ_2 was selected to be 0.1, and i_0 to be 100 days. Additionally, we have considered a second intervention parameterization. The form of the intervention was $f(\underline{\psi}, i) = \psi_1(1_{(i<i_0)} + (1 - \psi_2)1_{(i \geq i_0)})$. This represents a constant intervention, where the probability of moving from the susceptible to exposed class is decreased instantaneously at the time of the intervention, then held constant over the course of the intervention. For the simulations employing this intervention parameterization, ψ_2 was fixed at 0.7.

The parameterization selected for the exponential distributions was chosen as $f(x) = \frac{1}{\lambda} exp(-\frac{x}{\lambda})$. For the gamma distributions, the parameterization was chosen as $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} exp(-\beta x)$. For the Weibull distributions, the parameterization was selected as $f(x) = \frac{\alpha}{\beta}(\frac{x}{\beta})^{\alpha-1} exp(-(\frac{x}{\beta})^\alpha)$. The true values for α and β. The parameter values were chosen to approximate a disease such as Mumps. Mumps has a very well known latent period of 16-18 days, although this period can last as few as 12 or as many as 25 days [13]. The infectious period is less well known, but shedding typically lasts five days or less, though there is the possibility of shedding for much longer, and therefore having a longer infectious time [68]. A throat swab

can isolate the viruses from 40% of individuals infected with Mumps 2-3 days prior to the onset of parotitis, and individuals are typically infectious prior to displaying symptoms [12, 13].

For the simulations, the true value for λ in the exponential distribution was chosen to be 18.71 for the exposed mean time, and 8.62 for the infectious mean time. The true α and β related to the gamma distributions, were chosen to be 30 and 1.603 for the exposure distribution, and 100 and 11.6 were chosen for the infectious time. In the Weibull distributions α and β were chosen to be 7 and 20 for the exposed time, and 12 and 9 for the infectious time.

Four data sets were simulated with the final epidemic sizes being within two individuals of 35, 65, 125, and 225. These final sizes were chosen to loosely correspond to a small, medium, large and very large epidemic sizes, based on a simulation study in which 3,000 epidemics were simulated.

Tables 3.2 and 3.3 present the simulation results obtained for the exponential decay intervention with Weibull and gamma distributed latent and infectious times, and all the removal times known. The PS SEIR model typically performs as well as, and often better than, the population averaged model in terms of the size of the credible interval and median parameter values, with the exception being the very large Weibull data set. This is not surprising, as large epidemic sizes will lead to biased results in the mixing and intervention parameters. The PS SEIR model is more sensitive to epidemic size than the population averaged approach, because the hidden assumption in SEIR modeling is that the epidemic being modeled is a "typical" major epidemic. While the improvement offered by the PS SEIR model often appears to be small, it is important to note that small changes in parameter values can have a large effect on the distribution of final epidemic size in stochastic models. The priors used for the mixing and intervention parameters were Gamma(0.25

,1) and Gamma(0.1,1) for all models. For the exponential analysis of each data set, λ was assigned a Gamma(187.09,10) prior. For the Weibull analyses, α was assigned a prior of Gamma(70,10) and β was assigned a Gamma(200,10) prior. For the Gamma analyses, α was assigned a prior of Gamma(300,10) and β was assigned a Gamma(100,100/18.709).

Table 3.4 presents the simulation results obtained for the constant intervention with Weibully distributed latent and infectious times, and all the removal times known. The PS SEIR model typically performs as well as, and often better than, the population averaged model for these simulation as well. The priors used for the mixing and intervention parameters were Gamma(0.25,1) and Gamma(0.7,1) for both models. For the exponential analysis of each data set, λ was assigned a Gamma(187.09,10) prior. For the Weibull analyses, α was assigned a prior of Gamma(70,10) and β was assigned a Gamma(200,10) prior.

P-values greater than 0.05 for the Geweke diagnostic were used to indicate convergence for all model parameters [29]. Note that the PS SEIR model typically offers some improvement, with the most improvement coming when the epidemic sizes are small to moderately sized, due to the greater effect each individual path has on the mixing and intervention parameters in these cases. There is less of an effect in the larger epidemics, and we see that the population averaged model begins to fit the data in the very large epidemics. Beyond just the decrease in variance that one would expect from fitting the true model, we hypothesize that the main reason for this phenomenon is that the gamma distribution model supplies more information about the latent process, which allows for much more accurate parameter realizations in situations with small epidemic sizes.

3.4 Data Analysis

The motivating data set consisted of the onset times for the 214 cases of Mumps confirmed via swab culture during the 2006 Iowa Mumps epidemic, which lasted from January 29, 2006 to June 25, 2006. We note that cases were less likely to be confirmed via swab culture early in the epidemic, which may lead to longer estimates of the infectious time distribution and lower estimates of the mixing parameter. In fact, all of the models we fit tended to underestimate the final size of the epidemic, and the low values of the mixing parameters are likely part of the reason. However, it is quite common in epidemic research that not every infectious individual is diagnosed, and our main goal is to demonstrate the improvement of the PS SEIR structure over similar population averaged approaches.

There are two goals for the following analysis. The first is to obtain a more realistic analysis of the Iowa Mumps epidemic than can be obtained by utilizing the exponential distribution alone, and to demonstrate the improvement of the PS SEIR formulation over the population averaged formulation. The second is to decide on a reasonable parametric form for the public's awareness of the epidemic, which acts as an intervention in the data. This will demonstrate the importance of recognizing changes in behavior resulting from public awareness in modern epidemics, as well as the importance of quantifying these changes.

Polgreen et. al. analyzed the Iowa Mumps epidemic using a Generalized Linear Mixed Model (GLMM) approach to map the data, and employed a test of proportions to analyze the effect of spring break. They found there to be a spring break effect in the age composition of Mumps cases after spring break [69]. We again note that their analysis considered all probable cases, whereas our analysis considered only confirmed cases. Considering a more granular treatment of time in a SEIR structure may allow us to expand upon the results yielded by their research

and identify more temporal structures in the data set.

Simply modeling the data set with no intervention accounting for public awareness was not successful. Epidemics rarely occurred based on the estimated parameter posterior values obtained from models not accounting for public awareness, and epidemics that did occur severely underestimated the final epidemic size, with almost no simulated epidemics reaching half the true epidemic size. Previous literature has used public awareness as an intervention successfully [26][43][56]. Note that Lekone and Finkenstädt use an exponential decay intervention to model public awareness due to a government awareness campaign with promising results [56].

In modeling public awareness, we will use two parameterizations. The first will be the same style of exponential decay intervention found, which will begin on March 30, the day that the CDC posted a dispatch to the MMWR website [11]. The second will be a logistic intervention, which will have the form $\frac{exp(\phi_0-\phi_1*day)}{1+exp(\phi_0-\phi_1*day)}$, where $\phi_1 > 0$, $\frac{exp(\phi_0)}{1+exp(\phi_0)} > 0.99$, and day is the number of days since the initial case. This function will be able to reduce the mixing parameter from its initial value over the course the epidemic. In an attempt to accommodate the effect of spring break on mixing, we use a three week constant effect intervention, beginning on March 6, 2006.

One of the aspects that must be accounted for in working with Mumps is the presence of the MMR vaccine. The CDC states that the 2004-2005 MMR vaccination rate for kindergartners in Iowa was 97% [12]. We therefore use that as our best estimate of the vaccination rate in the state of Iowa. A study performed in 1985 suggests that the efficacy of the MMR vaccine in preventing Mumps is 85% [25]. One simplifying assumption in the model is that the vaccine is an all or nothing vaccine yielding permanent immunity. According to our vaccination estimates, this yields 523,000 individuals susceptible to Mumps in Iowa. 2,570,000

individuals will start in the Removed category, accounting for their immune status. This is an important consideration, as the high rate of immunity plays a role in determining the magnitude of the mixing and intervention parameters. We do note a potential caveat to our vaccination assumption. Individuals who were naturally infected with Mumps have much better immunity to the virus than those who are vaccinated. Because there are individual in the state of Iowa who were still alive before vaccination was common, older individuals may have a much higher rate of immunity than younger individuals in our model. We have not accounted for this phenomenon.

Because the prior information available for the infectious time of Mumps is not as strong as the prior information available for the latent time, we will use two lengths for the infectious period for each distribution. The first set will be short infectious times. These correspond to Exponential(8.6), Gamma(100,11.6), and Weibull(12,9). The second set will be long infectious times. These correspond to Exponential(11), Gamma(25,2.27), and Weibull(6,12).

Models were constructed, and their fit assessed using the posterior predictive p-value approach [28]. At each iteration of the MCMC chain, a single epidemic was generated. The model fit statistic used was an indicator that the final epidemic size was between 107 and 428 (half to twice as large as the epidemic) and the day the simulated epidemic ended was between 117 and 197 (within 40 days of the length of the actual epidemic). These values were chosen based on the variability seen in simulated epidemics.

Table 3.5 shows the posterior predictive p-values for all the models we ran. The path-specific approach yields the highest posterior p-values for model fit. The best path-specific model (Weibully distributed with an exponential form of the intervention and long infectious times) generated over six times as many accepted epidemics

as the best fitting population averaged model (Exponentially distributed with the logisitic form of the intervention and long infectious times). The path-specific approach also yields some of the lowest p-values, indicating that it is sensitive to the form of the public-health intervention chosen. Table 3.6 gives descriptive statistics for the best fitting PS SEIR model, as well as its corresponding exponential model. We see that, unlike the results from the simulated data sets, the estimated parameter posteriors are centered at different values. This is likely due to the fact that the exact parametric forms of the interventions are not known, and were modeled using an approximation. The Weibull model has a higher estimate of the mixing parameter, a quantity we expected to be underestimated in our analysis. Additionally, it is known that Mumps is typically latent 16 to 18 days. The credible interval for the mean latent time in the Weibull model fits this *a priori* known information better than the exponential model. Together, these are leading to more accurate predicted epidemic sizes.

The autocorrelations for the posterior draws are also of interest. Geweke's criterion was used to assess burn in, as was done in the simulations. After burn in, we see lower autocorrelation in PS SEIR mixing parameters found in Table 3.6 as compared to their corresponding population averaged parameters. The mixing parameter ϕ_1 has a lag 10 autocorrelation of 0.10 for the PS approach, but is 0.65 in the population averaged approach. The spring break intervention has a lag 10 autocorrelation of 0.12 for the PS approach versus 0.55 for the population approach. The public health intervention has a lag 10 autocorrelation of 0.07 versus 0.81 for the PS approach versus the population averaged approach. However, the mean of the latent distribution has an autocorrelation of 0.91 in the PS SEIR model versus 0.66 in the population averaged model. The cross correlations between the chains are similar for both models.

Figure 3.1 graphs the epidemic curves that fell in the reasonable range for the best fitting PS SEIR model as well as those for the corresponding population averaged model. One can see that the PS SEIR model provides far more predicted epidemics in the reasonable range (21.84% versus 2.84%). Additionally, the shape of the epidemic is fit more accurately with the PS SEIR approach. Those epidemics that fall into the reasonable range in the population approach tend to overestimate the epidemic size early on, whereas the PS SEIR approach provides a more accurate envelope of the epidemic up until day 50. Around day 50, there is a sharp increase in the number of epidemics, a feature neither model captures well. However, the PS SEIR model captures the shape of the remainder of the epidemic as well, with many more curves above the epidemic. Recall that it was expected that the final epidemic size would be underestimated because only the confirmed cases were used. For this data set, the PS SEIR model captured the true form of the epidemic better than the population averaged approach.

3.5 Discussion

The path-specific formulation typically offers improvement over the population averaged approach to modeling epidemics. In simulation studies, where all the parametric forms of the mixing and intervention processes were known, the path-specific approach typically performed as well as the population averaged approach in terms of the median values for these parameters and almost always gave narrower credible intervals for them. Very small differences in parameter realizations can be important in SEIR modeling, especially in the mixing parameter.

We also note that, in the real data analysis performed on the Iowa Mumps epidemic, the PS SEIR model yielded substantially more reasonably sized predicted epidemics than the population-averaged approach. In fact, the best PS SEIR model yielded over six times as many reasonable epidemic size predictions as the best

population averaged model, as based on the model fit statistic we defined in the data analysis section. For this reason, we recommend the path-specific approach be used when analyzing epidemics related to infectious diseases with latent and infectious periods that are not exponentially distributed.

Additionally, there were two unrelated initial cases. The population averaged approach does not handle such a structure well, while the PS SEIR model can handle it quite easily. For the second initial case, which occurred in Dubuque County, we set the individual as having a latent infection for fourteen days previous to the start of the epidemic, which yields a total latent period of seventeen days for that individual. This minimizes the effect on the latent distribution posterior, and is supported by prior knowledge regarding the length of the latent period of Mumps. The population averaged model will analyze this as a three day latent period, which may have some effect on the posteriors of the parameters.

There are limitations for our Iowa Mumps epidemic analysis. First, the infectiousness of individuals is constant throughout their infectious periods. This is unrealistic, but required by the current form of the PS SEIR model. Secondly, homogeneous mixing is violated in this analysis. Thirdly, vaccinations were handled in a rather naive way, which may affect the accuracy of the mixing and intervention parameters.

Despite these limitations, we have demonstrated that the path-specific approach can yield much more accurate SEIR models for epidemics than the population averaged approach, and avoids many of the weaknesses of the current SEIR and SIR models allowing for general latent and infectious time distributions.

Table 3.1: Equivalency of the Population Averaged Model and the Path-Specific Model.

Model	Parameter	2.5%	25%	50%	75%	97.5%
Population Averaged	Mixing	0.088	0.113	0.132	0.151	0.185
Path-Specific	Mixing	0.088	0.114	0.131	0.149	0.185
Population Averaged	Mean Exposure Time	15.14	17.14	18.48	19.94	22.99
Path-Specific	Mean Exposure Time	14.98	17.01	18.36	19.81	22.73

Table 3.2: Parameter medians and 95% central credible intervals for the analysis of the Gamma data sets featuring an exponential decay intervention. Based on 10,000 realizations after burn-in each.

Analysis	Data Set	Mixing (0.25)	Intervention (0.1)	Alpha (30)	Beta (1.604)
Exponential Analysis	Small Gamma	0.23 (0.14, 0.32)	0.62 (0.07, 2.96)	Not Applicable	Not Applicable
Gamma Analysis	Small Gamma	0.20 (0.13, 0.28)	0.12 (0.05, 0.22)	29.78 (26.45, 33.17)	1.54 (1.21, 1.91)
Exponential Analysis	Medium Gamma	0.23 (0.16, 0.31)	0.12 (0.04, 0.35)	Not Applicable	Not Applicable
Gamma Analysis	Medium Gamma	0.21 (0.15, 0.28)	0.07 (0.03, 0.11)	29.73 (26.54, 33.23)	1.64 (1.30, 1.97)
Exponential Analysis	Large Gamma	0.26 (0.20, 0.34)	0.12 (0.07, 0.22)	Not Applicable	Not Applicable
Gamma Analysis	Large Gamma	0.26 (0.20, 0.32)	0.10 (0.07,0.14)	29.13 (25.93, 32.50)	1.66 (1.39, 1.97)
Exponential Analysis	Very Large Gamma	0.27 (0.22, 0.32)	0.10 (0.07,0.15)	Not Applicable	Not Applicable
Gamma Analysis	Very Large Gamma	0.27 (0.22, 0.31)	0.10 (0.08, 0.13)	28.81 (25.56, 32.16)	1.57 (1.34, 1.82)

Table 3.3: Parameter medians and 95% central credible intervals for the analysis of the Gamma data sets featuring an exponential decay intervention. Based on 10,000 realizations after burn-in each.

Analysis	Data Set	Mixing (0.25)	Intervention (0.1)	Alpha (7)	Beta (20)
Exponential Analysis	Small Weibull	0.22 (0.13, 0.27)	0.20 (0.03, 0.89)	Not Applicable	Not Applicable
Weibull Analysis	Small Weibull	0.20 (0.13, 0.27)	0.08 (0.03, 0.14)	6.87 (5.30, 8.60)	20.82 (18.84, 22.90)
Exponential Analysis	Medium Weibull	0.24 (0.17, 0.33)	0.22 (0.07, 0.72)	Not Applicable	Not Applicable
Weibull Analysis	Medium Weibull	0.22 (0.16, 0.28)	0.11 (0.06, 0.16)	6.33 (4.87, 7.95)	20.01 (18.23, 22.08)
Exponential Analysis	Large Weibull	0.24 (0.19, 0.30)	0.11 (0.06, 0.19)	Not Applicable	Not Applicable
Weibull Analysis	Large Weibull	0.24 (0.19, 0.29)	0.11 (0.07,0.15)	5.76 (4.19, 7.55)	20.83 (18.72, 23.60)
Exponential Analysis	Very Large Weibull	0.26 (0.22, 0.31)	0.12 (0.08,0.17)	Not Applicable	Not Applicable
Weibull Analysis	Very Large Weibull	0.28 (0.24, 0.33)	0.15 (0.12, 0.20)	5.30 (4.21, 6.59)	21.41 (20.14, 22.80)

Table 3.4: Parameter medians and 95% central credible intervals for the analysis of the Weibull data sets featuring a constant intervention. Based on 7,000 realizations after burn-in each.

Analysis	Data Set	Mixing (0.25)	Intervention (0.7)	Alpha (7)	Beta (20)
Exponential Analysis	Small Weibull	0.23 (0.11, 0.38)	0.61 (0.14, 0.89)	Not Applicable	Not Applicable
Weibull Analysis	Small Weibull	0.25 (0.13, 0.39)	0.68 (0.25, 0.89)	6.86 (5.37, 8.51)	20.56 (18.08, 22.78)
Exponential Analysis	Medium Weibull	0.25 (0.16, 0.34)	0.70 (0.46, 0.86)	Not Applicable	Not Applicable
Weibull Analysis	Medium Weibull	0.26 (0.19, 0.34)	0.73 (0.59, 0.84)	6.39 (4.94, 8.04)	20.81 (18.23, 23.19)
Exponential Analysis	Large Weibull	0.26 (0.20, 0.33)	0.73 (0.59, 0.84)	Not Applicable	Not Applicable
Weibull Analysis	Large Weibull	0.24 (0.19, 0.30)	0.69 (0.56,0.79)	6.41 (5.03, 7.97)	19.06 (17.64, 20.61)
Exponential Analysis	Very Large Weibull	0.28 (0.22, 0.34)	0.79 (0.68,0.87)	Not Applicable	Not Applicable
Weibull Analysis	Very Large Weibull	0.26 (0.22, 0.31)	0.75 (0.67, 0.81)	6.33 (5.16, 7.83)	20.39 (19.27, 21.37)

Table 3.5: Posterior predictive p-values for the models run for the real data analysis. Each posterior predictive p-value is based on 7,000 realizations.

Distribution	Intervention	Infectious Period	P-value
Exponential	Exponential	Short	0.0203
Exponential	Exponential	Long	0.0284
Exponential	Logistic	Short	0.0245
Exponential	Logistic	Long	0.0305
Gamma	Exponential	Short	0.0881
Gamma	Exponential	Long	0.1061
Gamma	Logistic	Short	0.0028
Gamma	Logistic	Long	0.0033
Weibull	Exponential	Short	0.1668
Weibull	Exponential	Long	0.2184
Weibull	Logistic	Short	0.0004
Weibull	Logistic	Long	0.0335

Table 3.6: Descriptive statistics for the posteriors of key parameters or parametric forms for Model 1, which has exponentially distributed exposure times, an exponential decay intervention, and long infectious times, as well as for Model 2, which has Weibully distributed exposure times, an exponential decay intervention, and long infectious times. Based on 7,000 realization each.

Model	Parameter	Median	95% Central Credible Interval
Model 1	Mixing	0.91	(0.43, 1.72)
Model 2	Mixing	1.23	(0.89, 1.66)
Model 1	Spring Break	-0.59	(-2.21, 0.22)
Model 2	Spring Break	-0.21	(-0.82, 0.21)
Model 1	Public Awareness	0.08	(0.03, 0.13)
Model 2	Public Awareness	0.04	(0.03, 0.06)
Model 1	Mean Exposure	18.88	(11.60, 25.23)
Model 2	Mean Exposure	17.25	(16.93, 18.07)

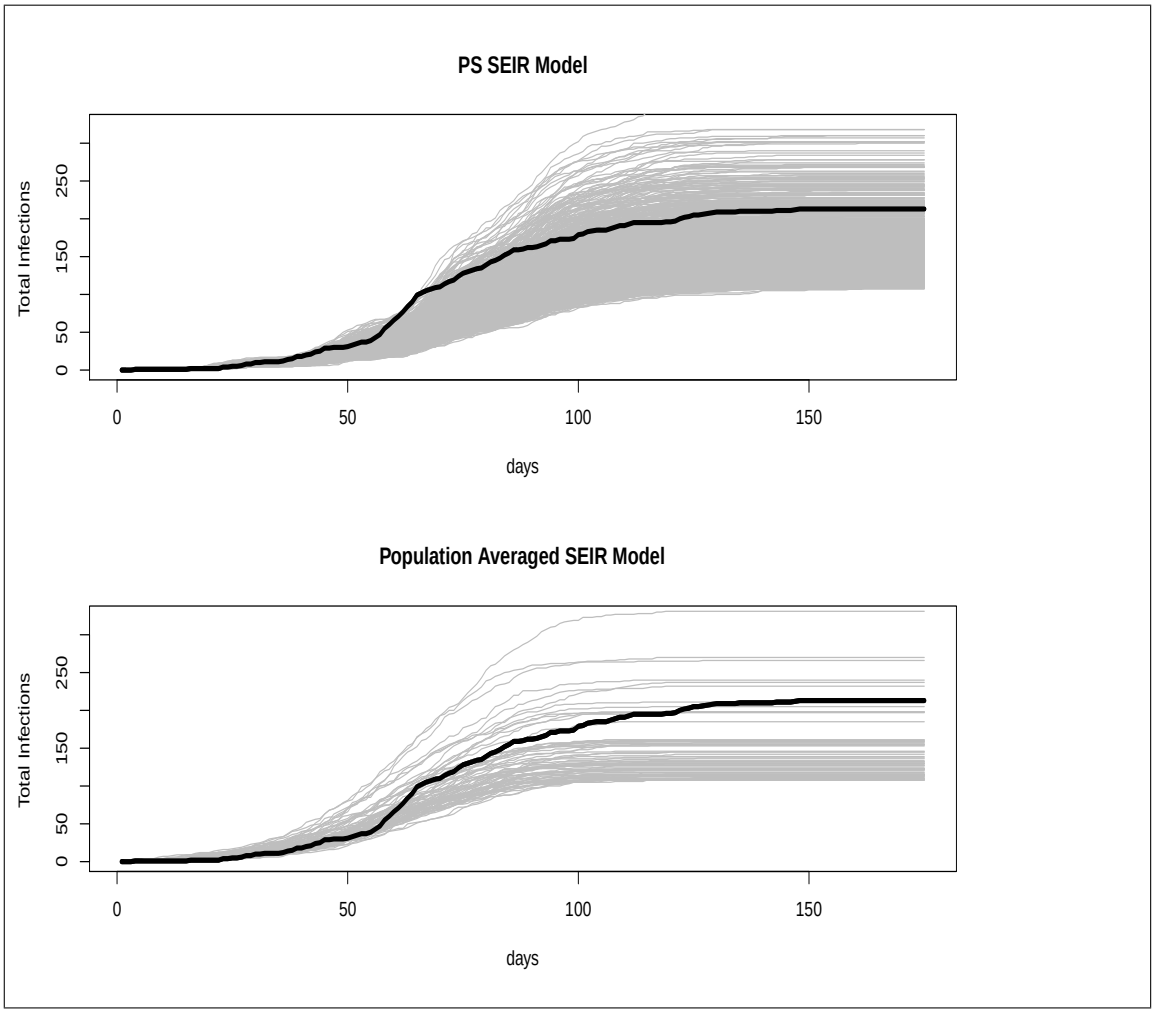

Figure 3.1: Accepted epidemic curves for the PS SEIR model with Weibull latent and infectious times, exponential public health intervention and long infectious distributions versus Accepted epidemic curves for the same population averaged model. Gray curves are model predictions while the black curve is the actual epidemic.

CHAPTER 4
A CAR MODEL FOR MULTIPLE OUTCOMES ON MISMATCHED LATTICES

4.1 Chapter Goal

The goal of this chapter is to develop a spatial model which can be embedded into the SEIR structure in order to create a flexible Spatial SEIR model. The model will be able to draw strength in estimating spatial mixing within spatial structures and between spatial structures. This will be done in enough generality that more than two spatial bleeding conduits can be considered.

4.2 Introduction

Conditional Autoregressive (CAR) models are common in the literature, as these models represent an intuitive way of incorporating spatial autocorrelation on lattices. The univariate CAR structure has been well developed, with major contributions in the field coming as early as 1974 [7]. Excellent references on the topic include Cressie's 1993 text for the frequentist formulation of CAR models [21], and the 2004 text by Banerjee, Carlin and Gelfand for the Bayesian formulation [5]. These models are extremely flexible and have been used in a variety of fields, including noninfectious disease mapping [18][51][79], ecology [58], and image processing [15], to name just a few.

Since multivariate outcomes are often of interest, multivariate CAR models have also been developed. Early multivariate work was done by Mardia [60], who proposed a very general multivariate CAR structure. Kim et. al. proposed a two-fold model for mapping two correlated mortality rates [51]. Additional work has been done utilizing the same neighborhood structure for each of the multivariate outcomes [27][41][71]. Recent work by Sain and Cressie has extended the univariate

CAR structure to include different neighborhood structures for each variable, but does so assuming the same lattice for every neighborhood structure [72]. Additionally, there is the restrictive requirement that the variance for each outcome differ by a known constant within each lattice point.

While noninfectious disease mapping with CAR models is common,there is far less literature using CAR models for infectious disease mapping. One of the reasons is likely that many infectious diseases are spread via human to human transmission, and the geographical proximity of two locations may not be an accurate assessment of the overall proximity of the locations in terms of disease spread. This makes a reasonable neighborhood structure difficult to construct. In the modern world, proximity is often defined by travel, rather than geographical measures. Hufnagel et. al. showed that airline travel served as an accurate measure of proximity in the severe acute respiratory syndrome (SARS) pandemic of 2003 [39]. In general, it is likely that modern epidemics are spread via more than one conduit. For many epidemics, a realistic model may have to account for multiple modes of travel.

In this chapter, we develop a type of multivariate CAR model in keeping with Sain and Cressie's concept of differing neighborhood structures. We utilize a similar mean structure, but construct our model in such a way that different lattice structures may be used for each outcome of interest. While a change of basis is often possible for mismatched lattices, this technique is not always desirable, and is occasionally impossible.

Allowing mismatched lattices means that the requirement that the variances differ by a known proportionality constant for each outcome at a given lattice point must be relaxed. The model we propose allows correlated, multivariate outcomes that occur on different lattices to be analyzed, which is helpful when lattice structures are developed in such a way that a change of basis is not desirable. Possible

applications include infectious disease mapping and environmental data.

Specific examples of applications that would benefit from mismatched lattices are easy to construct. When mapping soil pollutants, one might consider adjacency of soil plots to be one reasonable neighborhood structure, but also want to consider that two soil samples of the same land-use type may be related, even if the two sites are not geographically close. This lattice structure may not be readily accounted for in a change of basis. Our motivating example considers infectious disease modeling. In 2006, there was a large Mumps epidemic in Iowa (described in detail in Section 5). As Iowa is predominantly rural, a standard CAR model which utilizes geographic borders to determine adjacency could be used to relate neighboring counties in the infectious disease process. However, it is clear from the data that cases of Mumps tended to spread along the highway system. This is an example of how multiple modes of travel and contact can spread an infectious agent. In a relatively rural state like Iowa, highway systems may serve as an important measure of adjacency for infectious disease spread, just as airline travel may be an important measure of adjacency in pandemic modeling. More examples can be readily constructed involving travel as a measure of proximity between human populations.

We have already mentioned that the Mumps epidemic spread across borders and along highways. Separating the spatial dependence in the disease process using two distinct structures certainly allows both measures of proximity to be modeled, but does not consider that the two latent effects are potentially correlated. A given county not only has correlation with other elements of its neighborhood structure, but across neighborhood structures as well. *A priori*, one would expect positive correlation between the modes of travel, as counties that have large numbers of people traveling along highways almost certainly have large numbers of people crossing county borders unassociated with highway travel. A model allowing for multiple

outcomes that have different spatial correlation parameters between multiple outcomes is desirable.

In this chapter, we propose a multivariate CAR model that allows mismatched lattices to be utilized, and give conditions for consistency and propriety. We also discuss model properties, as well as giving a recommended parameterization guaranteeing consistency. We proceed to show model adequacy when analyzing model-based data through simulation results. Next, we perform a disease mapping example considering the Mumps epidemic final counts in order to demonstrate model properties. We close with a discussion.

4.3 Mismatched Conditional Autoregressive Model

4.3.1 Model Definitions

Consider m outcomes, denoted by $\underline{X} = (\underline{X}_1', ...\underline{X}_m')'$, where $\underline{X}_k' = (X_{k1}, ..., X_{kn_k})', k = 1, ..., m$ and every X_{ki} coming from a univariate normal distribution. Each $\underline{X}_k$ is collected on potentially different lattices and neighborhood structures, $C_1, ..., C_m$. Within each neighborhood structure, we require positive correlation. Between neighborhood structures, we allow either positive or negative correlation. Consider potential spatial relationships between two neighborhood structures C_k and C_l, $l \neq k$. The MMCAR model can be derived as follows:

Suppose

$$X_{ki}|\{X_{kj}, j \neq i, X_{lj}, l \neq k\} \sim N\left(\mu_k + \sum_{kj \sim ki} \beta_{kij}(x_{kj} - \mu_k) + \sum_{l} \sum_{lj \simeq ki} \delta_{klij}(x_{lj} - \mu_l), \sigma_{ki}^2\right), \tag{4.1}$$

where β_{kij} is the weight of the mean for X_{ki} associated with X_{kj}, δ_{klij} is the weight of the mean for X_{ki} associated with X_{lj}, μ_k and μ_l are the overall means of $\underline{X}_k$ and $\underline{X}_l$, respectively, and σ_{ki}^2 is the conditional variance of X_{ki}. For convenience, $\beta_{kii} \equiv 0$ for every i and k. We use $ki \sim kj$ to denote that elements i and j are

both units within neighborhood structure C_k and are neighbors. We use $ki \simeq lj$ to denote that element i of C_k is a neighbor of element j of C_l.

In this model, we have $\underline{\beta}_{ki} = (\beta_{ki1}, ..., \beta_{k,1,n_k})'$ representing a vector of weights that define the contribution of members of C_k to the conditional mean of X_{ki}, and $\underline{\delta}_{kli} = (\delta_{kli1}, ..., \delta_{k,l,i,n_l})'$ representing a vector of weights that define the contribution of members of C_l to the mean of X_{ki}.

We now derive necessary and sufficient conditions for which $\underline{\beta}$ and $\underline{\delta}$ create a consistent Markov Random Field. In order to derive the conditions under which the conditionals are self-consistent, we begin by using Brook's Lemma. WLOG, assume $\underline{\mu}_k = \underline{0}$ for all k. Henceforth, we denote by $f(\underline{z})$ the value of the probability density function of $\underline{z}$ given the model parameters. To begin note that:

$$\begin{gathered} 2 * log(f(x_{ki}|x_{kj}, j \neq i, x_{lj}, l \neq k)) \propto \\ \frac{x_{ki}^2}{\sigma_{ki}^2} - 2\frac{\sum_{kj \sim ki} \beta_{kij} x_{kj} x_{ki}}{\sigma_{ki}^2} - 2\frac{\sum_l \sum_{lj \simeq ki} \delta_{klij} x_{lj} x_{ki}}{\sigma_{ki}^2} \end{gathered} \tag{4.2}$$

Brook's Lemma states

$$\frac{f(x)}{f(y)} = \prod_{i=1}^{N} \frac{f(x_i|x_1, ..., x_{i-1}, y_{i+1}, ..., y_N)}{f(y_i|x_1, ..., x_{i-1}, y_{i+1}, ..., y_N)}. \tag{4.3}$$

Plugging (1) into (2), with $y_i = 0$ for all i yields

$$\begin{gathered} -2 * log(f(x)/f(0)) \propto \sum_{k=1}^{m} \sum_{i=1}^{n_k} \frac{x_{ki}^2}{\sigma_{ki}^2} - 2\sum_{k=2}^{m} \sum_{i=2}^{n_k} \frac{\sum_{j<i} \beta_{kij} x_{kj} x_{ki}}{\sigma_{ki}^2} \\ -2\sum_{k=2}^{m} \sum_{i=2}^{n_k} \frac{\sum_{l<k} \sum_{j=2}^{n_l} \sum_{h<j} \delta_{klih} x_{lh} x_{ki}}{\sigma_{ki}^2}. \end{gathered} \tag{4.4}$$

Now, for the conditionals to be self-consistent, any permutation in the variable order when applying Brook's Lemma must equal (3). A particularly informative permutation is the "forward" version of Brook's Lemma:

$$\frac{f(x)}{f(y)} = \prod_{i=1}^{N} \frac{f(x_i|y_1, ..., y_{i-1}, x_{i+1}, ..., x_N)}{f(y_i|y_1, ..., y_{i-1}, x_{i+1}, ..., x_N)}. \tag{4.5}$$

Plugging (1) into (4), and once again using $y_i = 0$ for all i yields

$$-2 * log(f(x)/f(0)) \propto \sum_{k=1}^{m} \sum_{i=1}^{n_k} \frac{x_{ki}^2}{\sigma_{ki}^2} - 2 \sum_{k=1}^{m-1} \sum_{i=1}^{n_k-1} \frac{\sum_{j>i} \beta_{kij} x_{kj} x_{ki}}{\sigma_{ki}^2} \\ -2 \sum_{k=1}^{m-1} \sum_{i=1}^{n_k-1} \frac{\sum_{l>k} \sum_{j=1}^{n_l-1} \sum_{h>j} \delta_{klih} x_{lh} x_{ki}}{\sigma_{ki}^2}. \tag{4.6}$$

In general, one sees that the likelihood will contain all of the terms $\frac{x_{ki}}{\sigma_{ki}^2}$, and all of the cross products between x_{ki} and x_{kj}, $j \sim i$, as well as all the cross products between x_{ki} and x_{lj}, regardless of the permutation of variables selected when applying Brook's Lemma.

Because expressions (3) and (5) must be equal for the conditionals to be self-consistent, we must have $\frac{\beta_{kij}}{\sigma_{ki}^2} = \frac{\beta_{kji}}{\sigma_{kj}^2}$ and $\frac{\delta_{klij}}{\sigma_{ki}^2} = \frac{\delta_{lkji}}{\sigma_{lj}^2}$. Additionally, it is desirable for $\sum_{kj \sim ki} \beta_{kij} + \sum_{l \neq k} \sum_{lj \simeq ki} \delta_{klij} = 1$, as this indicates the weights in determining that the mean of X_{ki} sum to one. Given these weights, one can write the pdf of $\underline{X}$:

$$f(\underline{x}) \propto |\Sigma^{-1}(I - \Omega)|^{-\frac{1}{2}} exp(-\frac{1}{2}(\underline{x} - \underline{\mu})' \Sigma^{-1}(I - \Omega)(\underline{x} - \underline{\mu})), \tag{4.7}$$

where $\underline{\mu}$ is the vector of overall means, Σ is a diagonal matrix with elements σ_{ki}^2, and

$$\Omega = \begin{bmatrix} \beta_1 & \delta_{12} & \delta_{13} & \cdots \\ \delta_{21} & \beta_2 & \delta_{23} & \cdots \\ \delta_{31} & \delta_{32} & \beta_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Note that β_k is defined as the weight matrix for neighborhood structure k, and that δ_{kl} is an n_k by n_l matrix with entries of δ_{klij} as element (i, j), and zeroes elsewhere.

Algebra will confirm that the pdf in (6) will yield the same expression as equations (3) and (5). We also see that post-multiplying $(I - \Omega)$ by the one vector yields the zero vector, indicating that the covariance matrix in (6) is singular. Because $\underline{\delta}$ is used to construct the contingency matrix, the model can only be used as an improper prior if these values are known *a priori*. In practice, these values are almost

certainly not known by investigators, so we must derive a condition for which this distribution can be made proper.

In order to create a proper distribution, one can multiply Ω by a propriety parameter α. If $\alpha\epsilon\left(\frac{1}{\lambda_{(1)}}, \frac{1}{\lambda_{(n)}}\right)$, where $\lambda_{(1)}$ and $\lambda_{(n)}$ are the minimum and maximum eigenvalues of Ω, then $\Sigma^{-1}(I - \Omega)$ will be positive definite, and $\underline{X} \sim N(\underline{\mu}, (I - \alpha\Omega)^{-1}\Sigma)$.

This result follows because Σ^{-1} is a positive definite diagonal matrix, implying its determinant is positive. Letting $P\Lambda P'$ be the spectral decomposition of Ω, we have $(I - \alpha\Omega) = P(I - \alpha\Lambda)P'$. If $\alpha\epsilon(1/\lambda_{(1)}, 1/\lambda_{(n)})$, then $(I - \alpha\Lambda)$ is a diagonal matrix with every diagonal entry greater than zero. Therefore, $\Sigma^{-1}(I - \alpha\Omega)$ is positive definite, and can be inverted to yield a valid covariance matrix.

Of course, with $\alpha \neq 1$, the mean structure now serves as a reaction function, where the deviation from the mean at location i is updated by some proportion of the deviation from the mean of its neighbors. As α approaches zero, this removes spatial correlation. When the δ parameters are all assumed positive, one should set α as high as possible if one wants to induce spatial correlation.

4.3.2 Model Properties

A set of natural parametric forms satisfying the conditions on $\underline{\delta}$, $\underline{\beta}$ and $\underline{\sigma}^2$ would be $\sigma^2_{ki} = \frac{\sigma^2_k}{(\frac{w_{ki+}}{1-\delta_{k+i+}})}$, $\beta_{kij} = \frac{w_{kij}}{(\frac{w_{ki+}}{1-\delta_{k+i+}})}$ and $\delta_{klij} = \frac{\sigma^2_{ki}}{\sigma^2_{lj}}\delta_{lkji}$, where w_{kij} is an indicator that elements i and j are neighbors in neighborhood structure k, and the "+" indicates summing over the subscript. Additionally, this induces the constraint $\sum_l \sum_{lj \simeq ki} \delta_{klij} < 1$ for all i and k, in order to ensure that the conditional variances are all positive.

Of course, these are not the only constraints that work for this model, but they do provide a natural interpretation and make sampling more efficient. It is also important to note that $\underline{\delta}$ allows for information to be drawn asymmetrically

across the different neighborhood structures. The constraints yield the desirable result that outcomes with higher variances will draw less from their neighboring outcomes and more from neighborhood structures with lower variances. Using $\beta_{kij} = \frac{w_{kij}}{(\frac{w_{ki+}}{1-\delta_{k+i+}})}$ keeps the form of the univariate structure for the CAR model presented in [5]. In the case where δ_{k+i+} is positive, the division by $1-\delta_{k+i+}$ implies that a fraction of the outcome's own neighborhood structure will be used in determining the outcome's mean. Therefore, the smoothing performed within the neighborhood structure is reduced, and information is drawn across neighborhood structures to perform additional smoothing. This is anticipated to be the typical case, but the positivity of δ_{k+i+} is not required by the model. If δ_{k+i+} is negative, the division by $1-\delta_{k+i+}$ increases the smoothing performed within the neighborhood structure. The form $\sigma^2_{ki} = \frac{\sigma^2_k}{(\frac{w_{ki+}}{1-\delta_{k+i+}})}$ is a natural way to reduce the number of parameters in the model while still satisfying model constraints. From this point forward, we use this particular parametrization scheme.

Additionally, if this particular parameterization scheme is used, and one assumes $min(\underline{\delta}) \geq 0$, the Ω matrix is row stochastic. Because the maximum and minimum eigenvalues of row stochastic matrices are greater than -1 and less than 1, respectively, any $\alpha\epsilon(-1,1)$ is sufficient to guarantee the existence of the inverse of $\Sigma^{-1}(I-\alpha\Omega)$. One can then fix α *a priori* rather than estimating this parameter. If $min(\underline{\delta}) < 0$, this result does not necessarily hold, and α must be estimated.

The MMCAR structure is very flexible in modeling the spatial patterns across multiple outcomes. Consider the simple case of two outcomes and a single δ parameter. There are seven cases of interest.

1. With α positive and δ near one, we have little spatial correlation within outcomes and strong positive correlation within each areal unit.

2. With α positive and δ zero, we have independent CAR models.

3. With α positive and δ negative, we have negative correlation within areal units and spatial over-smoothing.

4. With α zero, we have purely independent latent spatial effects.

5. With α negative and δ near one, we have little spatial correlation within outcomes and strong negative correlation within each areal unit.

6. With α negative and δ zero, we have independent CAR models that induce negative spatial correlation.

7. With α negative and δ negative, we have positive correlation within areal units and negative spatial over-smoothing.

We note that one of the assumptions of Brook's Lemma is that $\underline{\beta}$ and $\underline{\delta}$ are known and fixed. In the frequentist framework, these weights will have to be fixed *a priori*, and this model reduces to the same assumptions as Sain et. al.'s model with the ability to handle mismatched lattices. However, the Bayesian framework offers an additional option. If the spatial effects can be thought of as a latent process, a Bayesian update scheme can be utilized. Of course, $\underline{\mu}$, $\underline{\sigma}^2$, and α can be drawn regardless of whether the weighting scheme is assumed known. The overall MCMC chain will not draw from a CAR model directly, but rather reduces to a variance estimation problem for a normal prior distribution. However, the conditional CAR structure still allows the asymmetric drawing of weights in determining the latent effects.

4.4 Computation of the Weights

When considering a latent spatial process, Brook's Lemma can be used conditionally. At each MCMC iteration, $\underline{\beta}$ and $\underline{\delta}$ can be drawn. Then, considering these as known and fixed weights, the latent spatial effects can be drawn. Now, using these data points as a new data set, the posteriors of $\underline{\beta}$ and $\underline{\delta}$ can be sampled. The assumed underlying spatial process can then be sampled without making

assumptions on the relationship between the multiple outcomes. The series of simulations below will demonstrate evidence that this process will approximate the true underlying spatial process.

Consideration must be given to the computation of $\underline{\delta}$ and $\underline{\beta}$, due to the restriction that $\sum_{kj\sim ki}\beta_{kij}+\sum_{l\neq k}\sum_{lj\simeq ki}\delta_{klij}=1$. In the case of only two outcomes, the following scheme is recommended:

1. Propose a set of values for $\underline{\delta_{12}}$.
2. Compute $\underline{\delta_{21}}$ as $\delta_{21ji}=\frac{1-\delta_{12ij}}{\frac{\sigma_1^2}{\sigma_2^2}\frac{w_{2i+}}{w_{1i+}}(\frac{1-\delta_{12ij}}{\delta_{12ij}}-1)+1}$.
3. Compute $\underline{\beta}$
4. Accept or Reject $\underline{\delta}$ and $\underline{\beta}$ given other model parameters.
5. Slice sample all variance components.
6. If utilizing a latent spatial effect, use Metropolis Hastings Sampling to sample the latent spatial effect.

In the case of three or more outcomes, we amend the scheme:

1. Propose a set of values for $\underline{\delta_{kl}}_{k=1,\ldots,M,l=1,\ldots,M-1}$.
2. Compute $\underline{\delta_{lk}}$ as $\delta_{lkji}=\frac{1-\sum_k\delta_{klij}}{\frac{\sigma_k^2}{\sigma_l^2}\frac{w_{li+}}{w_{ki+}}(\frac{1-\sum_k\delta_{klij}}{\sum_k\delta_{klij}}-1)+1}$. This constitutes M equations with M unknowns per location. If a solution exists, it is unique. If a solution does not exist, return to step 1.
3. Compute $\underline{\beta}$
4. Accept or Reject $\underline{\delta}$ and $\underline{\beta}$ given other model parameters.
5. Slice sample all variance components.
6. If utilizing a latent spatial effect, use Metropolis Hastings Sampling to sample the latent spatial effect.

This process constitutes an efficient way of sampling all model parameters.

4.5 Simulation Results

The motivating example lends itself to the Bayesian framework, where latent variables are used to model the multiplicative effect on the disease counts in each county. The MMCAR model can then be used as a prior for these effects. Therefore, we perform the simulation studies in the Bayesian framework. We use the MMCAR model as the data model, rather than as a prior, to more directly observe the estimation properties of this novel model. In order to verify the model works under ideal conditions, two simulation studies were performed. In the first simulation study, four data sets were generated and analyzed with centered prior information. In the second, those same data sets were analyzed with flat priors being used for all the model parameters.

The lattice used for the simulation studies was a six by six regular lattice. A reference image can be found in Table 4.1. Neighborhood structure C_1 was created considering any two units sharing a border as neighbors. Neighborhood structure C_2 was created by considering units 1, 2, 3, 7, 8, 9, 13, 14 and 15 as one set of neighbors. The second, disjoint set consisted of units 4, 5, 6, 10, 11, 12, 16, 17 and 18. A third, smaller set was comprised of units 25, 26, 31 and 32. A final set contained units 29, 30, 35 and 36. These were chosen to consider the accuracy of posteriors realizations on different sized and disjoint neighbors.

When performing simulations, we considered four cases. In the first, a single δ parameter was set at 0.25, and a single spatial realization was simulated. The propriety parameter α was fixed at 0.999 (to create a strong reaction function) and assumed known. The true variance of C_1 was set at 0.5 and the true variance of C_2 was set at 2. The second simulation was identical in structure, but contained five independent spatial realizations. The third simulation utilized two δ parameters. δ_1 was set at 0.25 for units 1, 2, 3, 7, 8, 9, 13, 14 and 15 as well as 25, 26, 31 and

32. For the remaining units in neighborhood C_2, δ_2 was set at -0.25. These values were chosen to better understand how the model behaves with positive and negative between lattice correlations, and to yield a model in which the propriety parameter α must be estimated. For this simulation, α was set at 0.5, but not assumed known. The true variance of C_1 was set at 2, as was the true variance of C_2. One spatial realization was simulated. The fourth simulation was identical to the third, but with five independent spatial realizations being simulated. For all four of these simulations, every estimated parameter was given a normal prior with a mean equal to its true value and a standard deviation of one third the magnitude of its true value. Simulations five through eight follow the same structure as simulations one through four, except that flat priors are used for every parameter. This will allow us to gauge the effect of the priors on the parameters in the model.

The results for all eight simulations can be found in Table 4.2. All but one of the true parameter values are located within their 95% credible intervals, with the exception being δ_1 in simulation four. Recall that simulation four utilized five independent spatial realizations and flat prior information. Because only one in thirty-two parameters is not contained in its 95% credible interval, we conclude that the model can accurately estimate parameters in the ideal case. Estimation of $\underline{\delta}$ is much more accurate when there is only one δ parameter present in the model, rather than two. This is evidence that reasonable constraints should be imposed on $\underline{\delta}$ when the conditional application of Brook's Lemma is used in practice. Additionally, we note that the median values of the variance parameters tend to be greater than the true values, but that this effect is less marked with multiple observations or in the presence of prior information. The true variance parameter is contained in the 95% credible interval for every simulation, however.

We conclude that the conditional approach to utilizing Brook's Lemma approximates the true underlying spatial process. There may be some positive bias in the variances in this process, but we leave a rigorous justification to future research. Justification of the coverage probabilities for the generated latent effects in the conditional Brook's Lemma scheme is left to the next chapter.

4.6 Data Analysis

Recall that 205 cases of Mumps were confirmed by swab culture in Iowa counties during the 2006 Mumps outbreak. A map of these data can be found in Figure 4.1. A previous Generalized Linear Mixed Model was constructed for these data, in order to map the data using county centroids to define spatial correlation, along with a concept developed therein defined as "zones of risk" [69]. Again, their analysis considered all probable cases, whereas our analysis considered only confirmed cases. The centroid argument may work well for determining the effect of border contacts, but we can additionally consider the highway structure if we utilize the MMCAR approach. This addition may refine the disease map produced in their paper. We wish to assess the spatial spread of the Mumps epidemic by determining how the spread occurred along neighboring counties, how it spread along highways, and also how the two types of spread are related. Our analysis consists of the total number of cases per county at the end of the epidemic.

Two Poisson regression models were constructed. Model One considers two counties to be neighbors if and only if a border was shared. Because counties with initial cases (Dubuque County and Johnson County) and counties with major universities (Black Hawk County, Johnson County, and Story County) tend to have larger infection counts, these counties are modeled using a different log-mean parameter. The data model we consider is:

$$Y_i = exp(\mu_1 + \omega_i)$$

if county i does not contain an initial infection or a major university, and

$$Y_i = exp(\mu_2 + \omega_i)$$

if county i contains an initial infection or a major university.

Here, Y_i is the total number of confirmed cases, μ_1 and μ_2 are the log-means of the counts for each county type, and ω_i is a latent effect to account for the border structure. In our model, $\underline{\omega}$ is given a $CAR(\sigma^2)$ prior, with σ^2 given a flat prior. The parameter μ_1 is given a N(0,1/3) prior and μ_2 is given a N(3,1) prior. These aided in model convergence, and are justified based on looking at the relative sizes of large and small counts of the disease outbreak.

Model Two is the MMCAR model. Two neighborhood structures are defined. The first is the same used in Model One, defined by shared borders. The second considered the highway structure in the state of Iowa. We use ten of the most traveled highways in Iowa as the generating structure (Interstates 80, 380, 29 and 35, as well as State Routes and US Highways 20, 30, 151, 60, 27, 61, and 34). These highways are only considered as part of the structure when they are multilane divided highways, as those stretches contain the most traffic. This structure then consists of every multilane highway in Iowa that passes through more than two counties, yielding a highway lattice consisting of 58 counties. Two counties are considered neighbors in this structure if they both contained the same highway. The data model we consider is:

$$Y_i = exp(\mu_1 + \omega_{1,i} + \omega_{2,i})$$

if county i does not contain an initial infection or a major university, but contains a major highway,

$$Y_i = exp(\mu_1 + \omega_{1,i})$$

if county i does not contain an initial infection or a major university and does not

contain a major highway, and

$$Y_i = exp(\mu_2 + \omega_{1,i} + \omega_{2,i})$$

if county i contains an initial infection or a major university (all of which contain major highways). Again, Y_i is the total number of confirmed cases, and μ_1 and μ_2 are the log-means of the counts for each county type. In the MMCAR model, $\omega_{1,i}$ corresponds to a latent spatial effect generated by the border structure, and $\omega_{2,i}$ corresponds to a latent spatial effect generated by the highway structure.

In this model $(\underline{\omega}_1', \underline{\omega}_2')'$ is given an $MMCAR(\sigma_1^2, \sigma_2^2, \delta)$ prior, with σ_1^2 corresponding to the border structure and given a flat prior, σ_2^2 corresponding to the highway structure and given a flat prior, and a single δ parameter. This δ parameter is chosen in such a way that the association from the highway structure to the border structure was constant. The fact that the number of neighbors per county in the border structure is more homogeneous than the number of neighbors per county in the highway structure serves as justification for this parameterization. The prior on δ is Uniform(0,.9), where the 0.9 prevents the highway structure from dominating the border structure. This concern is due to the low counts of mumps present in most counties (83 of the 99 Iowa Counties had zero cases or a single case). The mean parameters μ_1 and μ_2 are given a N(0,1/3) and N(3,1) priors respectively. The propriety parameter α is fixed to 0.999.

Three chains were run for each model and 45,000 iterations were selected from each of the three chains after a burn in period of 45,000 iterations. Convergence was assessed via the Gelman and Rubin Diagnostic [28]. At every iteration, a sample epidemic was generated from the model. From this epidemic, the MSE of the predicted epidemic, the Spearman correlation of the predicted data and the true data, and an overall epidemic size were generated. These three measures, along with DIC, were used to assess model fit. We found very little difference in the

fit of the model with the MMCAR prior versus the model with the CAR prior in any of the measures of model fit. In fact, the DIC model selection criteria were within five of one another, indicating no noticeable difference in terms of model fit [77]. This demonstrates the flexibility of the CAR model for this particular data set. However, the δ parameter, though highly left skewed, shows strong evidence of correlation between the two latent spatial effects in the MMCAR, with a median of 0.71. While the 2.5^{th} percentile is only 0.03, we do note that the 10^{th} percentile is 0.20, indicating correlation. We also note that the variances associated with the two neighborhood structures are quite different, indicating that the different neighborhood structures do play different roles in the model. This confirms the hypothesis that the spread of Mumps took place through multiple channels and that the rates of spread through such channels are different but correlated. A plot of the posteriors for δ, σ_1^2, and σ_2^2 can be found in Figure 4.2. The low disease counts are a likely rationale for there being very little difference in model fit.

We have additionally shown that the highway structure does play a role in this process. The high variances in this neighborhood structure indicate that some of the latent random effects are far from zero. In fact, the 2.5^{th} and 97.5^{th} percentiles of the multiplicative effect from the highway system are 0.216 and 3.181, providing support that the highway structure does play a role in the analysis.

4.7 Discussion

In this Chapter, we have developed a Gaussian Conditional Autoregressive model allowing for multiple correlated outcomes on mismatched lattices. Additionally, the MMCAR model proposed relaxes the restrictive assumption that the outcomes at each lattice point have variances that differ by a known proportion when using Brook's Lemma conditionally. The model allows for asymmetric correlation information to be used between lattices, and in such a way that lower variance

outcomes yield more information for the estimation of higher variance outcomes.

In simulation studies, the MMCAR model works quite well. We do note that the δ parameters generating the correlation between outcomes on different lattices should be reasonably constrained. Our model performs best when the number of δ parameters is small, but we feel that very few δ parameters would be required to allow a reasonable correlation structure in many applications.

Finally, in Iowa, disease spread from county to county is a natural way to consider disease propagation. However, we have demonstrated that the highway system is another important structure when considering the proximity of two counties in terms of disease spread. This result may apply to other epidemics in rural areas.

This model will prove useful in modeling infectious disease counts, especially in cases where there are multiple modes of travel or contact. A natural extension warranting further research is the application of the MMCAR model to a spatio-temporal setting for modeling epidemics. This would allow a more accurate representation of the modes of contact to be modeled than the purely spatial model considered here.

Table 4.1: The six by six lattice for the simulation study. Neighborhood structure C_1 was created considering any two units sharing a border as neighbors. Neighborhood structure C_2 was created by considering units 1, 2, 3, 7, 8, 9, 13, 14 and 15 as one set of neighbors. A second, disjoint set consisted of units 4, 5, 6, 10, 11, 12, 16, 17 and 18. A third, smaller set was comprised of units 25, 26, 31 and 32. A final set contained units 29, 30, 35 and 36.

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36

Table 4.2: Simulation Results. Simulations 5,6,7 and 8 used centered prior information. Simulations 1,3,5 and 7 used a single observation, whereas simulations 2,4,6 and 8 used five independent observations.

Sim	α	σ_1^2	σ_2^2	δ_1	δ_2
1	0.999	0.600 (0.5)	3.778 (2)	0.347 (0.25)	N/A
		(0.365, 1.130)	(1.353, 17.181)	(0.211, 0.407)	
5	0.999	0.533 (.5)	2.150 (2)	0.263 (0.25)	N/A
		(0.364, 0.753)	(1.294, 3.283)	(0.140, 0.388)	
2	0.999	0.556 (0.5)	2.453 (2)	0.306 (0.25)	N/A
		(0.440, 0.715)	(1.644, 4.247)	(0.211, 0.407)	
6	0.999	0.541 (0.5)	2.232 (2)	0.285 (0.25)	N/A
		(0.443, 0.669)	(1.610, 3.103)	(0.207, 0.369)	
3	0.271 (0.5)	1.361 (2)	1.049 (2)	0.350 (0.25)	-1.104 (-0.25)
	(0.051, 0.673)	(0.788, 2.649)	(0.606, 2.161)	(-1.756, 0.739)	(-3.341, -0.0532)
7	0.528 (0.5)	1.601 (2)	1.444 (2)	0.290 (0.25)	-0.268 (-0.25)
	(0.269, 0.708)	(1.072, 2.421)	(0.881, 2.356))	(0.124, 0.446)	(-1.756, 0.739))
4	0.497 (0.5)	2.500 (2)	2.221 (2)	0.489 (0.25)	-0.278 (-0.25)
	(0.303, 0.697)	(1.949, 3.368)	(1.704, 3.061)	(0.261,0.652)	(-0.708, -0.006)
8	0.520 (0.5)	2.248 (2)	2.092 (2)	0.338 (0.25)	-0.273 (-0.25)
	(0.369, 0.673)	(1.853, 2.757)	(1.704, 3.061)	(0.196,0.473)	(-0.424, -0.137)

Table 4.3: Parameter medians and 95% credible intervals for Model 1 (CAR model only accounting for a border structure) and Model 2 (MMCAR model accounting for both border and highway structures), as well as model fit statistics.

Parameter	Model 1	Model 2
μ_1	0.04 (-0.63, 0.70)	-0.12 (-0.72, 0.45)
μ_2	3.06 (1.75, 4.40)	3.15 (1.80, 4.70)
σ_1^2	10.11 (5.42, 18.83)	4.965 (1.60, 14.08)
σ_2^2	N/A	30.73 (5.46, 98.98)
δ	N/A	0.71 (0.03, 0.89)
MSE	8.13 (3.44, 19.90)	7.70 (3.33, 18.81)
Correlation	0.58 (0.43, 0.71)	0.56 (0.41, 0.70)
Overall Size	204 (165, 248)	206 (167, 248)
DIC	221.88	223.67

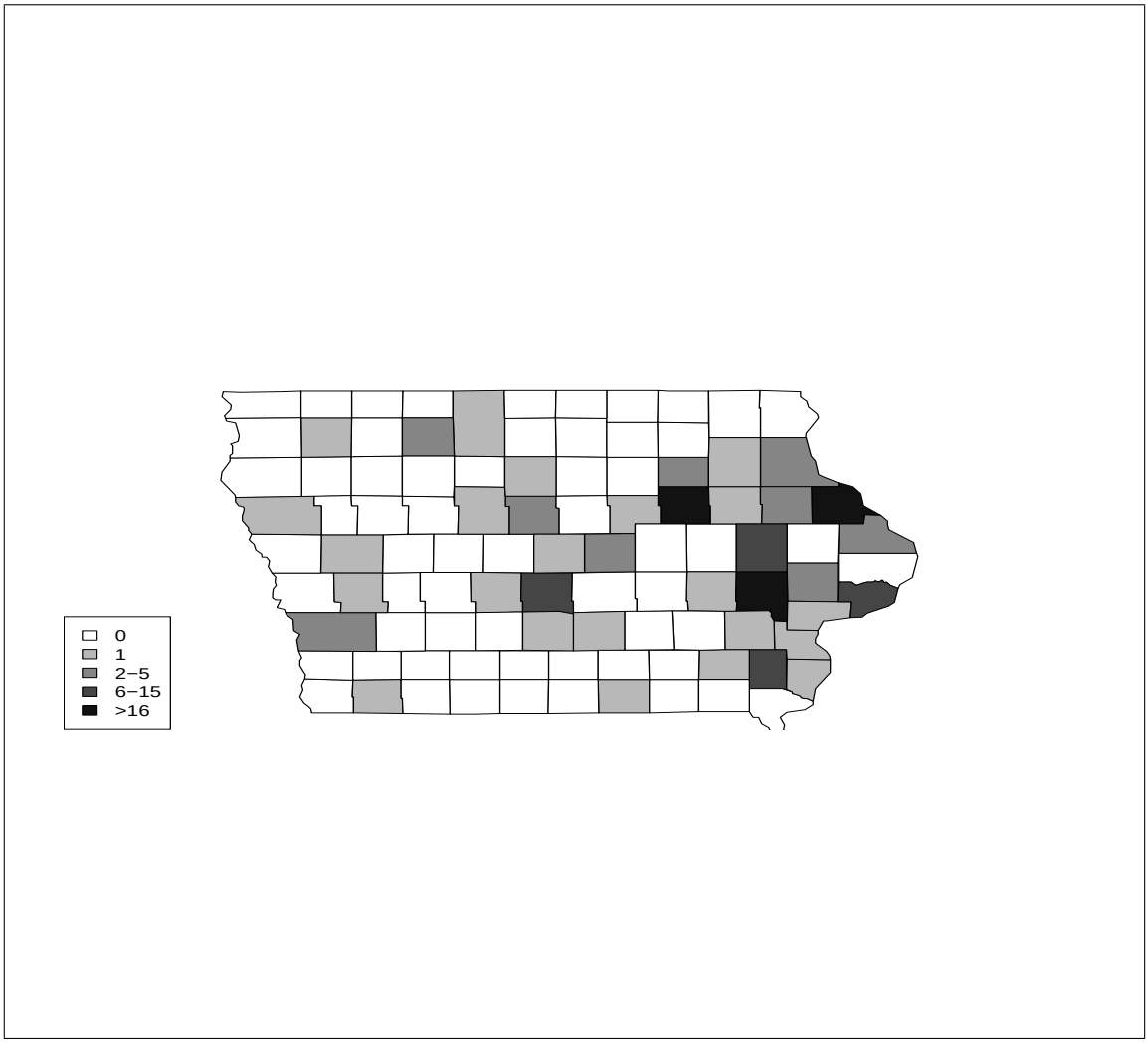

Figure 4.1: A disease map of the total number of confirmed cases of mumps in the 2006 Iowa Mumps epidemic

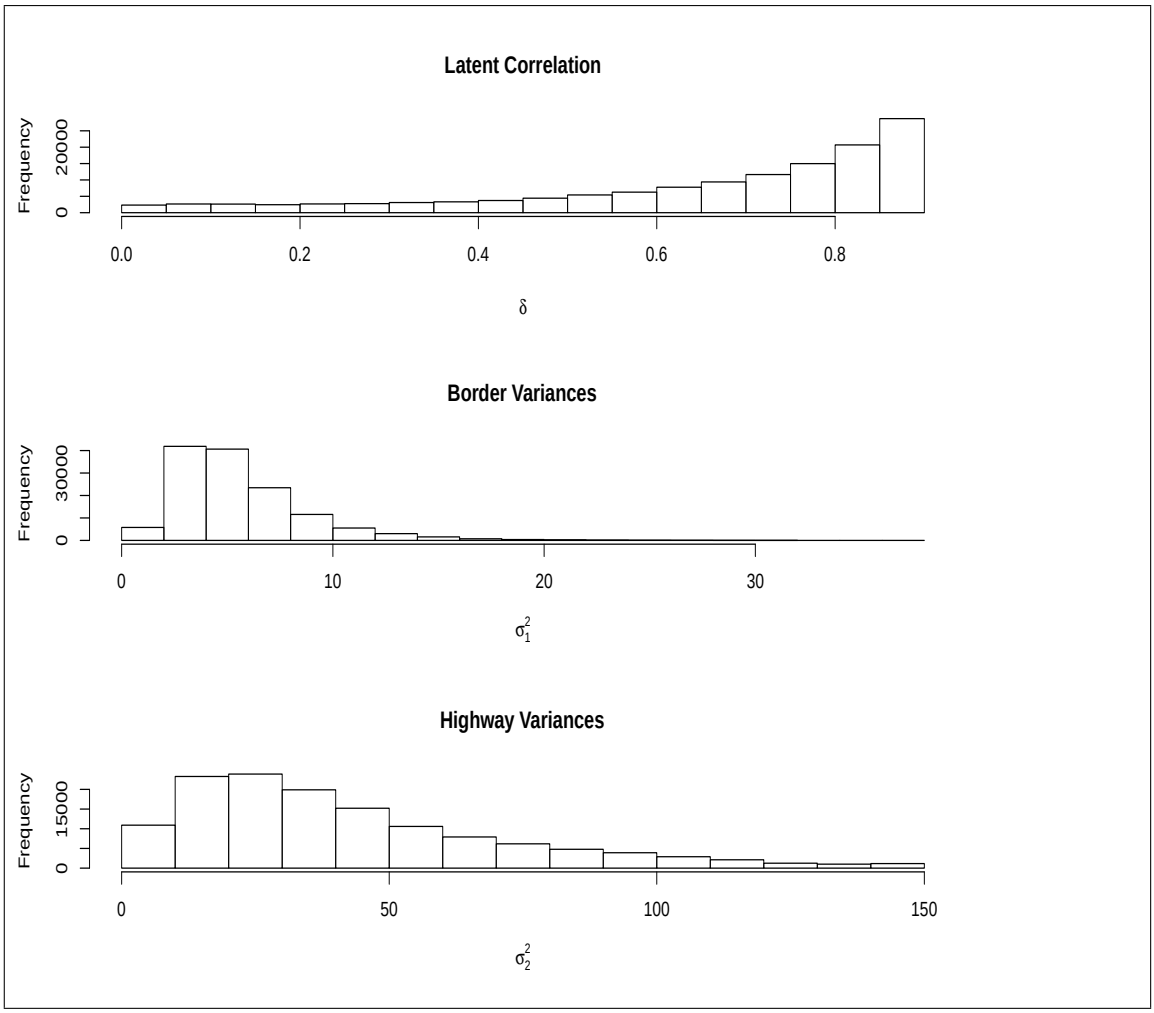

Figure 4.2: Posterior distributions for δ, σ_1^2 and σ_2^2.

CHAPTER 5
A PAIR OF SPATIAL SEIR MODELS UTILIZING THE CAR SPATIAL STRUCTURE

5.1 Chapter Goal

In this chapter, we embed the MMCAR model into the Population Averaged SEIR model from Lekone and Finkenstädt [56] and into the Path-Specific SEIR model from Chapter Three. We derive a spatial mixing function for these models, and discuss how this model relaxes the homogeneous mixing assumption. Simulations are performed to determine model properties. Finally, we utilize both models on the Mumps data set, and compare the spatial performance of both models. We close with a discussion regarding the situations in which each model is preferred.

5.2 Introduction

5.2.1 The Mumps Data and Spatial Spread

As mentioned in the previous chapter, the Mumps data were collected at the county level. This would lend itself to a lattice approach for a spatial analysis. Indeed, it seems that many spatial epidemic data sets will consist of counts of new infections collected on a lattice, with an obvious example being the Google Influenza data. Therefore, a general method for analyzing this type of data would be useful in analyzing transmission and intervention efficacy for many data sets other than just the Mumps data set presented.

It is known that multiple modes of spread need to be accounted for in the human to human transmission model [19]. We have argued that the highway systems were an important conduit for the transfer of Mumps in Iowa. In this chapter, we consider the spatio-temporal ramifications of modeling the Mumps epidemic as transferring along two main pathways: border contacts and highway contacts.

Plotting the data over time reveals a strong highway component. Recall that using prior knowledge of the Mumps latent and infectious time distributions yields an average of only six to ten cases that cannot be explained by one of these two modes of travel. This indicates these two measures as being the primary proxy measures for relative distance in the epidemic process. A graph of this process can be found in Figure 5.1.

5.2.2 The Need for a New Model

The current models in the literature are not ideal for modeling the spatio-temporal spread of the Mumps data. A point referenced approach using county centroids would accurately handle the border contact structure, but may not adequately model the contribution of the highways to the epidemic spread. Mass transportation is known to be an important component in these models (e.g, [22][36]). A person to person contact structure would be difficult to accurately construct for an area as large as the state of Iowa. At the very least, many assumptions would need to be made with regards to such a structure, and just constructing a person to person contact graph for the state of Iowa would likely require its own study.

The most appropriate model in existence would be the location to location contact graph approach (e.g., [3][19][39]). This approach is also not without problems, however. Models are difficult to properly parameterize. Each edge requires a contact probability. Allowing each edge its own contact probability yields a large number of parameters, many of which are nonestimable with the current data. The standard approach is to utilize a small number of probability parameters and weight them according to some proxy measure, as was done in Hufnagel et al [39]. This would be sufficient for the highway structure, where the traffic per day can be utilized as a weighting scheme, but is potentially difficult when considering border contacts. It is difficult to estimate how many individuals are traveling from county

to county each day outside of highway travel. Priors can be used (e.g. [33]), but methods of drawing strength in estimation are not well developed in epidemic modeling. Gravity models may provide a natural solution, and they have been shown to work well with the spatio-temporal analysis of Measles [83], a disease which behaves like Mumps in terms of transmission. Gravity models provide a natural weighting scheme based on distance, but do not smooth in the traditional sense. Smoothing the bleeding parameters based on the values of nearby bleeding parameters in the CAR spatial framework provides an alternative, natural method for smoothing. Additionally, this may have advantages in the border structure, where bleeding may defined not just by the distance between county centroids, but by population factors as well.

Epidemics tend to spread along multiple routes of contact. A general method which allows the data and model to inform the rates of spatial bleeding and additionally draw strength spatially and within locations would be applicable to many data sets beyond the Iowa Mumps data set. This would also allow flexibility in the case where a weighting scheme is not obvious. The data will update any prior information used in the weighting scheme, removing the need to try many weighting schemes in order to determine which fits the model best.

Additionally, spatial bleeding will likely be similar from year to year, though not fixed. Spatial bleeding is also independent of the type of pathogen being transmitted. Therefore, spatial bleeding in a location can be utilized as a prior for the next epidemic of any pathogen in a given location. Then, the model will update the spatial bleeding to the current state.

The flexibility of the MMCAR model in determining spatial dependence seems to be desirable here. Embedding the MMCAR model into the Population Averaged SEIR model proposed by Lekone and Finkenstädt [56] seems to be a reasonable

approach. This is a natural and easily understood model that contains a time-dependent mixing parameter. The flexibility of this parameter could be used to incorporate spatial bleeding on a lattice. A spatial SEIR model constructed in this way would yield an intuitive model that contains interpretable parameters. Therefore, we use the Population Averaged SEIR model as our starting point.

Because the spatial bleeding can be considered a latent underlying process, we propose using the conditional Brook's Lemma approach with the MMCAR model. This will remove the need to set the spatial weights *a priori*. We will demonstrate that using the conditional Brook's Lemma approach in estimating the MMCAR weights will yield reasonable coverage of the underlying latent effects through simulations similar to the Mumps data set. Due to the sparsity of the Mumps data set, it is anticipated that this technique will generalize to other, less sparse, data sets.

5.3 Methods

The goal of this section is to develop a pair of chain binomial SEIR models for use on spatial lattices. The first of these models is a member of the population averaged set of SEIR models. This model utilizes the exponential assumption, which states that the latent and infectious times of individuals are exponentially distributed. While this assumption is often violated by infectious diseases, it tends to be very robust, due in part to the fact that the exponential distribution only requires a single parameter to be estimated.

The population averaged SEIR model of interest is the model due to Lekone and Finkenstädt, which is repeated here for easy reference:

$$\begin{aligned} S_i \to E_{i+1} &= binomial(S_i, 1 - exp(-f(\underline{\psi}, i)h\tfrac{I_i}{N})); \\ E_i \to I_{i+1} &= binomial(E_i, 1 - exp(-h/\rho)); \\ I_i \to R_{i+1} &= binomial(I_i, 1 - exp(-h/\gamma)). \end{aligned} \tag{5.1}$$

Define i=1,...,T as a subscript for discrete time and S_i, E_i, I_i, and R_i represent

the counts of individuals in the Susceptible, Exposed, Infectious, and Removed compartments at time i, respectively. The notation $S_i \to E_{i+1}$ denotes a change of category. Let $f(\underline{\psi}, i)$ represent the mixing and possible intervention functions controlling the number of new exposures at time $i+1$ and is constrained to be nonnegative, and let h represent the number of days between time points in the data collection partition. The total number of individuals in the population is denoted by N.

5.3.1 A Spatial Population Averaged SEIR Model

When considering the spread of infectious diseases on a lattice, it is often important to consider multiple modes of contact. The MMCAR model from the previous section can be employed to model this process. Consider the following spatial SEIR model:

$$\begin{aligned}
&S_{ik} \to E_{i+1,k} = binomial(S_{i,k}, 1 - exp(-f(\underline{\psi}, i, k)h(\tfrac{I_{i,k}}{N_k} + \textstyle\sum_{l=1}^{m} \theta_{k,l}(\sum_{k \sim_l k'} \tfrac{I_{i,k'}}{N_{k'}})))); \\
&E_{ik} \to I_{i+1,k} = binomial(E_{i,k}, 1 - exp(-h/\rho)); \\
&I_{ik} \to R_{i+1,k} = binomial(I_{i,k}, 1 - exp(-h/\gamma)).
\end{aligned} \tag{5.2}$$

Define i=1,...,T as a subscript for discrete time and k=1,...,K as a subscript for location. S_{ik}, E_{ik}, I_{ik}, and R_{ik} represent the counts of individuals in the Susceptible, Exposed, Infectious, and Removed compartments at time i and location k, respectively. The notation $S_{ik} \to E_{i+1,k}$ denotes a change of category in county k. Let $f(\underline{\psi}, i, k)$ represent the mixing and possible intervention functions controlling the number of new exposures at time $i+1$ at location k, which is constrained to be nonnegative. Let h represent the number of days between time points in the data collection partition. The total number of individuals in the population of location k is denoted by N_k. The parameter $\theta_{k,l}$ is a bleeding parameter for location

k under neighborhood structure l, and is constrained to be nonnegative. Finally, the notation $k \sim_l k'$ indicates locations k and k' are neighbors under neighborhood structure l.

The model is intended as a Bayesian model. Flat priors will typically be employed for the parameters relevant to $f(\underline{\psi}, i, k)$, and informative priors can be used for ρ and γ, as there is often good information available for the mean times of the latent and infectious processes of many infectious disease. For $\underline{\theta}$=$(\theta_{1,1}, ..., \theta_{K,M})$, a reasonable prior is $(\frac{log(\theta_{k,l})}{1-log(\theta_{k,l})})_{k,l} \sim MMCAR(\sigma^2_{l=1,...,M}, \underline{\delta})$, where the MMCAR model is defined as in the previous chapter. Of course, the logit link is not required for this prior. Any link which projects the real line into $[0, \infty)$ is mathematically acceptable. Links that project the real line into [0,1] are intuitively preferred.

5.3.2 A Spatial Path-Specific SEIR Model

By combining the PS SEIR model developed in Chapter Three with the MMCAR structure from Chapter Four, we develop a model which can handle population level spatial lattice data and still relax the exponential assumption. The Spatial PS SEIR model is:

$$
\begin{aligned}
&S_{ik} \rightarrow E_{i+1,1,k} = \\
&binomial(S_{ik}, 1 - exp(-f(\underline{\psi}, i, k)h(\tfrac{I_{i,+,k}}{N_k} + \textstyle\sum_{l=1}^{m} \theta_{k,l}(\sum_{k \sim_l k'} \tfrac{I_{i,+,k'}}{N_{k'}})))) \equiv W_{ik} \\
&E_{ijk} \rightarrow I_{i+1,1,k} = binomial(E_{ij}, P(Z_1 \leq j + h | Z_1 > j)) \equiv X_{ijk}; \\
&E_{ijk} \rightarrow E_{i+1,j+1,k} = E_{ijk} - X_{ijk}; \\
&I_{ijk} \rightarrow R_{i+1,k} = binomial(I_{ijk}, P(Z_2 \leq j + h | Z_2 > j)) \equiv Y_{ijk}; \\
&I_{ijk} \rightarrow I_{i+1,j+1,k} = I_{ijk} - Y_{ijk}.
\end{aligned}
\tag{5.3}
$$

The notation in Equation 5.3 is identical to the notation in Equation 5.2, with a few additions. X_{ijk}, Y_{ijk}, W_{ik}, and E_{ijk} are all unobserved, while $\sum_j X_{ijk}$ and $\sum_j Y_{ijk}$ are known. The subscript i denotes calender time since the beginning of the

epidemic, j subscripts subjective time within the latent and infectious classes, and k subscripts location. The random variable Z_1 defines the latent time distribution, and Z_2 defines the infectious time distribution.

This Spatial PS SEIR model is identical in terms of mixing to the Spatial Population Averaged SEIR model, but incorporates the path-specific structure to allow the same flexibility in the latent and infectious time distributions as the traditional PS SEIR model. This allows for very realistic epidemic models to be fit to epidemic data.

5.4 Notes on the Spatial Pattern

The probability of infection in Equations 5.2 and 5.3 is a derivable form under given assumptions. Recall the derivation of the transmission probability in Chapter Three. The model was derived based on a Poisson contact distribution.

For the model at hand, consider the probability of escaping infection at time i by an individual in location k. Assuming independence between conduits and locations given spatial bleeding parameters, we have

$$Q_k(i) = \sum_{c_k=0}^{\infty} x(c_k, i) p_k^c(i) (\prod_l \prod_{k' \sim_l k} \sum_{c_{k'l}=0}^{\infty} x(c_{k',l}, i) p_{k',l}^c(i)), \tag{5.4}$$

where $x(c_k, i)$ is the probability mass function of the contact distribution when c_k contacts are made within the "base" unit at time i and $x(c_{k',l}, i)$ is the probability mass function when $c_{k',l}$ contacts made with neighboring location k' under conduit l. The term $p_k^c(i)$ represents the probability of not contracting an infection upon meeting a random individual within the "base" unit at time i, and $p_{k',l}^c(i)$ represents the probability of not contracting an infection upon meeting a random individual location k' via conduit l at time i. For our model, we assume Poisson contact random variables with rate $\theta_{k,l}\lambda_k$. This last parameterization is convenient for parameter reduction in this sparse data set. The first term of the expression for

$Q_k(i)$) then represents the contacts made within the "base" county, and the second term represents the contacts made with neighboring counties.

The derivation begins by considering a fixed time point i. For every interval $[i, i+h)$, we have discretized the contact rate at location k to the left endpoint, and labeled it as $\lambda_k h$. The probability of contracting an infection given an infectious individual is contacted at location k is also discretized with a pointmass on the left endpoint of the interval $[i, i+h)$, and is labeled as p_k^*.

Consider first the neighboring counties. Then we have

$\prod_l \prod_{k'} \sum_{c_{k',l}}^{\infty} \frac{c_{k',l}^{\theta_{k,l}\lambda_k h} exp(-\theta_{k,l}\lambda_k h)}{c_{k'l}!} p_{k'l}^{c} = \prod_l \prod_{k'} exp(-\theta_{k,l}\lambda_k h(1 - p_{k',l}^{c}))$.

Assuming $p_{k',l} = p_k^* \frac{I_{k',t}}{N_{k'}}$, for all l, we have

$\prod_l \prod_{k'} exp(-\theta_{k,l}\lambda_k h(1 - p_{k',l}^{c})) = exp(-\lambda_k h p_k^* (\sum_{k',l} \theta_{k,l} \frac{I_{i,k'}}{N_{k'}}))$.

The probability of escaping infection from contacts made within the "base " location follows the derivation in Chapter Three, with the expression being:

$\sum_{c_k=0}^{\infty} x(c_k, i) p_k^c(i) = exp(-\lambda_k h p_k^* \frac{I_{i,k}}{N_k})$.

Multiplying the expressions for escaping infection within the "base" location and neighboring locations yields:

$Q_k(i) = exp(-\lambda_k p_k^* h((\frac{I_{i,k}}{N_k}) + \sum_{k',l} \theta_{k,l} \frac{I_{ik'}}{N_{k'}}))$.

Now, if we turn our expression to $\lambda_k p^*$ and connect each discretized value over time as a single mixing function, we have

$Q_k(i) = exp(-f(\underline{\psi}, i, k) h(\frac{I_{i,+,k}}{N_k} + \sum_{l=1}^{m} \theta_{k,l} (\sum_{k \sim_l k'} \frac{I_{i,+,k'}}{N_{k'}})))$.

Taking the complements yields the mixing expression found in the Spatial SEIR model.

When considering the derivation performed in this way, it is clear that these

models are stochastic analogs of the deterministic system

$$\begin{aligned}
\frac{dS_k}{dt} &= -f(\psi,k,t,)S_k(I_k/N_k + \textstyle\sum_l \theta_{k,l}(\sum_{k'\sim_l k} \frac{I_{k'}}{N_{k'}})) \\
\frac{dE_k}{dt} &= f(\psi,k,t,)S_k(I_k/N_k + \textstyle\sum_l \theta_{k,l}(\sum_{k'\sim_l k} \frac{I_{k'}}{N_{k'}})) - g(\underline{\alpha}, E_k) \\
\frac{dI_k}{dt} &= g(\underline{\alpha}, E_k) - h(\underline{\gamma}, I_k) \\
\frac{dR_k}{dt} &= h(\underline{\gamma}, I_k).
\end{aligned} \tag{5.5}$$

This is a reasonable deterministic model to use for such a process. Considering the mixing proposed by such a model, there is an increased rate at which the Susceptible individuals move to the Exposed category as the proportion of Infectious individuals increases in every neighboring location. This is intuitively appealing. We also note that the diseases process is identical across all locations and individuals, which is also intuitive, as the diseases process is more likely influenced by the pathogen itself than the location at which it was contracted. We note the similarity of this model to the deterministic spatial model proposed by Sattenspiel and Dietz [73].

Because the spatial models in Equations 5.2 and 5.3 are stochastic analogs of deterministic systems, we obtain the following set of additional assumptions from Chapter Three:

1. Assume a homogeneous population with regards to susceptibility.

2. Define the Exposed compartment as only containing those who will eventually become infectious, and do not consider the possibility of a return to the Susceptible class.

3. Assume constant infectivity throughout the course of the infectious process.

4. Assume independent probabilities of moving from the Exposed Compartment to the Infectious Compartment (as well as from the Infectious Compartment to the Removed Compartment). Individuals are treated as having identical latent and infectious time distributions.

5. Homogeneous Individuals in terms of the disease process.

This model still utilizes the homogeneous mixing assumption, albeit in a much weaker form. Individuals mix homogeneously within their "base" areal unit. These individuals also mix homogeneously with members of neighboring areal units, but at a lower contact rate. They do not mix at all with areal units which are not considered neighbors in any neighborhood structure. Because we have reasonably constrained $\theta_{k,l}$ to [0,1], we conclude that the Poisson contact rate with neighboring units due to the bleeding structure associated with neighborhood structure C_k is a fraction of the contact rate within the "base" areal unit.

This formulation preserves the structure of the mixing parameter, and yields a physical interpretation of $\underline{\theta}$. The elements of $\underline{\theta}$ represent the reduced portion of contacts made on average with each neighboring county in the neighborhood structure. This allows informative hyperpriors to be utilized when assigning the MMCAR prior on $\underline{\theta}$. It is also consistent with the implicit requirement that $S_{ik} + E_{ik} + I_{ik} + R_{ik} = N_k$, which indicates that members of an areal unit are considered as being located in that unit for the duration of the epidemic.

5.5 Computation

This section is broken into two main subsections. The first subsection contains sampling considerations and simulation results for the Spatial Population Averaged SEIR model. The second contains sampling considerations and simulation results for the Spatial PS SEIR model.

5.5.1 The Spatial SEIR Model

5.5.1.1 Sampling Schemes and Considerations

Much of the sampling in this model has been considered in previous chapters. Metropolis Hastings sampling is appropriate for any mixing and intervention parameters in $f(\underline{\psi}, k, i)$, as well as for the mean latent time parameter.

A Metropolis Hastings algorithm can be employed when sampling $\underline{\theta}$. We propose the bleeding parameters be sampled jointly, using the untransformed bleeding parameters (before the logit link is applied) and independent normal proposal distributions for each untransformed bleeding parameter. The prior density based on the MMCAR can be computed, then these parameters can be re-transformed for use in computing the likelihood portion of the model. This technique gives the best mixing of any of the sampling schemes we attempted.

Sampling the Exposure matrix is done similarly to sampling the Exposure vector in the Population Averaged SEIR model. First, we consider a single location. Each location corresponds to a column of the Exposure matrix. The sampling scheme found in Lekone and Finkenstädt's paper is used to update 10% of the exposure times in each column of the matrix. This is repeated for every location, guaranteeing that at least 10% of the latent times of the epidemic are updated.

Certain considerations must be employed when determining priors for the MMCAR model parameters. In sparse data sets (the Mumps epidemic only affected 38 counties) with few total infections, using flat hyperpriors on the MMCAR model was not successful in simulation. Posteriors draws of the parameters sometimes yielded spatial bleeding parameters which were unreasonable *a priori*. The physical interpretation of $\underline{\theta}$ is helpful in this regard. Because $\underline{\theta}$ represents the percent decrease in the mean contact rate based on a mode of travel, $\underline{\mu}$ and $\underline{\sigma}^2$ may be selected to tune this to a reasonable distribution. No information must be assumed about $\underline{\delta}$.

We suspect that flat priors can be used if the majority of counties have multiple cases.

In the event that an epidemic has already been modeled on a given lattice, the spatial bleeding information may be utilized from that data to inform a prior. Oftentimes this information will not be available, especially in developed countries, but it represents an alternative to the tuning of the parameters required by the researcher.

Typical iterations for the spatial SEIR approach take between 0.75 and 1.5 seconds to complete, with an additional 0.5 to 1 seconds required to generate a sample epidemic when coded in R 2.14.1 and run on a Dell Precision T3500 running Linux Ubuntu.

5.5.1.2 Simulations

We performed two sets of simulations for the Spatial SEIR model. Two data sets were simulated, one slightly smaller and one slightly larger than the actual epidemic in terms of final size. The smaller epidemic had 16 counties effected with 161 total cases and the larger epidemic had 26 counties affected with 258 total cases.

For these simulations, Black Hawk County, Johnson County, Story County and Dubuque county were given a different mixing parameter than the other counties, as these counties contained the initial cases (Dubuque and Johnson Counties) or universities (Black Hawk, Johnson, and Story Counties). This mixing parameter (β_2) was set to 0.86, while the remaining counties were given a single mixing parameter (β_1) which was set to 0.79. These were tuned along with the public health awareness intervention (ψ_2, given a value of 0.09) and the spring break intervention (ψ_1, given a value of -0.59) to give reasonably sized epidemics and still be near the values found in the analysis in Chapter Three. The interventions are utilized in

every county. For the spring break intervention we again used a three week constant effect intervention, beginning on March 6. For the public health awareness intervention, we used the exponential decay intervention, which begins on March 30. Finally, we handle vaccination status equivalently to Chapter Three, but make the additional assumption that the vaccination coverage is 97% in every county. These are designed to be as similar to the true parametric forms and vaccination coverage of the actual epidemic as possible, so that the simulations closely mirror our data analyses.

The mean latent time (ρ) was set to 17.5 days, and the mean infectious time was set to 11 days. These values were based on the model selection from Chapter 2. Bleeding parameters were generated from an MMCAR distribution, with the mean of the border structure (μ_1) set to -2, the mean of the highway structure (μ_2) set to -2.5, variances of each spatial bleeding structure (σ_1^2 and σ_2^2) were set to 4, and a δ parameter of 0.5. These yielded median percent contacts with other locations of approximately 0.18, but give a 2.5^{th} percentile of 0.005, and a 97.5^{th} percentile of 1.28. The medians appear to be a reasonable guess at mixing in Iowa, and the variability within these contact rates is very large and flexible.

For each data set, two analyses were performed. One with relatively strong prior information utilized for the MMCAR parameters, and one with relatively weaker prior information used for these parameters. The strong priors were $\mu_1 \sim N(-2, 0.5^2)$, $\mu_2 \sim N(-2.5, 0.5^2)$, $\sigma_1^2 \sim\ N(4, 1^2)1_{(\sigma_1^2>0)}$, $\sigma_2^2 \sim\ N(4, 1^2)1_{(\sigma_2^2>0)}$, and $\delta \sim Unif(0, 0.9)$. Flat priors were utilized for all mixing and intervention parameters. The prior for ρ was N(17.5, 1), which represents the strong knowledge one would have about the mean latent time in practice. The relatively weaker priors were were $\mu_1 \sim N(-2, 1^2)$, $\mu_2 \sim N(-2.5, 1^2)$, $\sigma_1^2 \sim\ N(4, 2^2)1_{(\sigma_1^2>0)}$, $\sigma_2^2 \sim\ N(4, 2^2)1_{(\sigma_2^2>0)}$, and $\delta \sim Unif(0, 0.9)$. We have not yet considered the issue of non-centered priors.

However, the data analysis will indicate that these priors can be adjusted by the data if they are non-centered.

The resulting simulations can be found in Table 5.1. These show accurate estimation of all parameters, with the exception of ψ_1 in the large data set. This may be an artifact of the data set itself. We see that, among all simulations, 95% of the 95% credible intervals contain the true parameter values, with 100% of the credible intervals for the small data set containing the true parameters and 90% of the credible intervals for the large data set containing the true parameter values. The strong prior information yields much narrower credible intervals for the MMCAR parameters, most notably σ_1^2, σ_2^2, and δ. When weak prior information is used δ takes values that are only slightly influenced by the data themselves, evidenced by the fact that it has a 95% credible interval that is only slightly narrower than the one produced by its Unif(0,0.9) prior. However, this a hyperparameter, so this may not be too detrimental, so long as the spatial bleeding parameters it influences are burned in.

Table 5.3 shows the coverage percentages for the 95% credible intervals for the latent effects. We see that the coverage probabilities tend to fall a little short of the nominal 95% probabilities. However, the coverages are all above 80%, and the model performs better in counties with at least one case. This yields evidence that the conditional Brook's lemma approach is an acceptable way to model latent spatial effects, as well as yielding evidence that the spatial effects are accurately modeled even when the parameters of the MMCAR prior come from diffuse posteriors.

5.5.2 The Spatial PS SEIR model

5.5.2.1 Sampling Schemes and Considerations

In simulation studies with data sets similar to the Mumps data, the Spatial PS SEIR was not very robust. With fixed and known infectious data, the model parameters were modeled relatively well, but there was a distinct flattening of the latent period distribution. The mean of this distribution was typically in a reasonable range, but the variance was always much larger than expected. When the removal times were imputed, this effect was intensified, and even strong prior information did not remove the effect.

Therefore, we propose an alternate scheme for utilizing the Spatial PS SEIR model. Assume the parameters of the latent and infectious time distributions can be given distributions *a priori*. Realizations from these prior distributions can then be used to construct a latent and infectious time distribution at each time point. The exposure array and infectious array can then be updated at each iteration. Updating the arrays properly is an important step, as these arrays do act as random quantities in this model. The update scheme is the same as in Chapter Three, where a given path was removed and then a new path generated in the exposure matrix (and infectious matrix). Now we consider each location individually when computing the exposure array, and update each location's matrix by 10% of the possible paths at each time point. We repeat this process for every location at every iteration. The infectious array is done similarly.

All other parameters in this model can be imputed according to the scheme proposed for the Spatial SEIR model.

Typical iterations take between 5 and 11 seconds, with an additional 4 to 5 seconds required to generate a sample epidemic when coded in R 2.14.1 and run on

a Dell Precision T3500 running Linux Ubuntu.

5.5.2.2 Simulations

A single data set was generated for the simulations. For the simulated data set, Black Hawk County, Johnson County, Story County and Dubuque County were given a different mixing parameter than the other counties, as we again note that these counties contained the initial cases (Dubuque and Johnson Counties) or universities (Black Hawk, Johnson, and Story Counties). This mixing parameter (β_2) was set to 1.25, while the remaining counties mixing parameter (β_1) was set to 1.00, which were tuned along with the public health awareness intervention (ψ_2, given a value of 0.04) and the spring break intervention (ψ_1, given a value of -0.21) to give reasonably sized epidemics and still be near the values found in the analysis in Chapter Three. Interventions and vaccinations are identically chosen to those in the Spatial SEIR simulations.

Again, bleeding parameters were generated from an MMCAR distribution, with the mean of the border structure (μ_1) set to -2, the mean of the highway structure (μ_2) set to -2.5, variances of each spatial bleeding structure (σ_1^2 and σ_2^2) were set to 4, and a δ parameter of 0.5. The latent period was given a Gamma(80, 4.6) distribution, and the infectious time was given a Gamma(100, 11.4) distribution. Two aspects of the model are of note here. First, a gamma distribution was used, rather than a Weibull distribution, in order to reduce autocorrelation in the chains. Second, the infectious times here correspond to the short infectious times of Chapter Three. These short infectious times are *a priori* preferred, though they were not selected by the model selection in Chapter Three. However, as this is considered our most realistic model, we have once again considered the short infectious times.

The simulated epidemic consisted of 217 cases spread over 35 counties. This data set is similar to the actual Mumps epidemic data set. Four analyses were

performed. For the first analysis, the latent time parameters were drawn based on centered and precise distributions ($N(350, 10^2)$ for the shape parameter and $N(20, 1^2)$ for the scale parameter). For the second analysis, off centered, precise distributions were used ($N(200, 10^2)$ for the shape parameter and $N(20, 1^2)$. These generate latent times which tend to be shorter than the true latent times. For the third analysis, latent parameters were drawn from centered, imprecise distributions ($N(350, 100^2)$ for the shape parameter and $N(20, 7^2)$. For the fourth and final analysis, off centered, imprecise distributions were used ($N(200, 100^2)$ for the shape parameter and $N(20, 7^2)$.

These four simulations will allow the effect of centering versus systematic off centering to be assessed. They will also allow the effect of higher variance or lower variance priors to be assessed when the data are centered or off centered. They do not systematically allow the assessment the degree of off centering to be assessed (i.e. how much off centering is too much off centering?).

In the analyses, flat priors were used for β_1, β_2, ψ_1, and ψ_2. The weaker versions of the priors were used from the Spatial SEIR model. Recall these were $\mu_1 \sim N(-2, 1^2)$, $\mu_2 \sim N(-2.5, 1^2)$, $\sigma_1^2 \sim N(4, 2^2)1_{(\sigma_1^2>0)}$, $\sigma_2^2 \sim N(4, 2^2)1_{(\sigma_2^2>0)}$, and $\delta \sim Unif(0, 0.9)$.

We see from the simulations that centered, precise priors of the latent effects allow for accurate estimation of the model parameters, while centered, imprecise (higher variance) priors do not do much detriment to parameter estimation. However, off centering causes poor estimation of mixing and intervention parameters. We also note that the degree of precision in the off centered prior specification yields very different estimates of the MMCAR variance components, but has little effect on mixing and intervention parameter estimation.

Table 5.4 shows the coverage percentages for the 95% credible intervals for the

latent spatial effects in the Spatial PS SEIR model. These coverages are all around 95%, indicating that the conditional Brook's Lemma approach is valid here. We also note that latent distribution has little effect on these coverage probabilities.

5.6 Data Analysis

For the comparison of the Spatial SEIR and Spatial PS SEIR models, we perform a final set of analyses on the 205 confirmed cases of Mumps in Iowa. These analyses will give a realistic analysis of the spatial pattern of the Mumps epidemic, as well as a comparison between the two models. For this analysis, interventions and vaccinations are chosen identically to the simulations, which were designed to attempt to capture the true parametric forms of the Mumps epidemic.

The mixing portion of the model has been modified somewhat. As mentioned, an average of six to ten cases of Mumps in the actual epidemic cannot be explained by border contacts or highway contacts. To account for these infections, we modify the Spatial SEIR model infection probability to the following form:

$$1 - exp(-f(\underline{\psi}, i, k)h(\frac{I_{i,+,k}}{N_k} + \sum_{l=1}^{m} \theta_{k,l}(\sum_{k \sim_l k'} \frac{I_{i,+,k'}}{N_{k'}})) + \epsilon_k) \tag{5.6}$$

where $\epsilon_k \sim Gamma(0.5, 10^7)$. This term is tuned to give an average of one to ten infections that will just "appear" via other contact structures per epidemic in the entire state over the course of 175 days. This allows for both contacts that do not occur along borders or highways to be accounted for, and accounts for edge effects. The Spatial PS SEIR infection probability is handled similarly.

For the Spatial SEIR model, three chains were run, and 15,000 iterations discarded as burn in for each chain, with 85,000 iterations being collected from each chain after burn in. Convergence was assessed via the Gelman and Rubin Diagnostic [28]. For the Spatial PS SEIR model, three chains were run, and 30,000 iterations were discarded as burn in for each chain, with 35,000 iterations being

collected from each chain after burn in. Convergence was again assessed via the Gelman and Rubin Diagnostic.

Parameter posterior medians and 95% credible intervals can be found in Table 5.5. We see that many parameter posteriors look quite similar between the two models. We hypothesize that this is due to the large overall size of the epidemic. Even thought the data are sparse, it is possible that this was a large enough epidemic that the advantages of the PS SEIR structure in terms of more accurate parameter estimation are small as compared to the standard population averaged approach.

A few differences are of note in the parameter posteriors. First, the estimate of ψ_2 appears to be somewhat different in the two analyses. This is not surprising, as this parameter is directly related to the infectious time distribution. Because the public health awareness intervention is time varying, the number of infectious individuals in the model at each time point will likely be important. This makes the variance of the infectious time distribution a consideration for the analysis of time dependent interventions.

We also draw the reader's attention to the posterior predictive p-value and Mean Squared Error of the models. Prediction of new epidemics will be greatly affected by differences in the latent and the infectious time distributions. This is particularly true in a spatial model where contacts between spatial locations are somewhat less common. The Mean Squared Error is based on the final size of the epidemic at each spatial location. The Spatial PS SEIR model performs considerably better by this measure. The posterior predictive p-value considers the state of Iowa as a whole, and uses the same criteria as Chapter Three (the predicted epidemic must be between half as large and twice as large as the actual epidemic, and end within 40 days of the actual epidemic). We see that Spatial PS SEIR model predicts epidemics for the state of Iowa in this range almost three times as often as

the Spatial SEIR model, indicating improved model fit.

Also of interest was spatial smoothing of the epidemic. Small scale variations in epidemic modeling are likely the result of the probabilistic infectious process, rather than any true spatial differences. Figure 5.2 shows the spatio-temporal pattern of 3,000 realizations from the Spatial PS SEIR model. Additionally, figures 5.3 and 5.4 show the means and 97.5^{th} percentiles of 3,000 randomly selected epidemics from both the Spatial SEIR model and the Spatial PS SEIR model. We see that both models smooth the data spatially, which was one of the desirable characteristics of embedding a CAR model into the SEIR structure. We also note that the spatial smoothing in the Spatial PS SEIR model yields means that resemble the "zones of risk" formulation in the original analysis by Polgreen et. al. [69]. This lends credence to both models, due to the corroborating evidence the models provide one another.

Finally, we note that the bleeding parameters were able to adjust to the data, despite the use of priors. The priors we provided gave a median percent contacts with other locations of approximately 0.18, but yields a 2.5^{th} percentile of 0.005, and a 97.5^{th} percentile of 1.28. However, the joint posterior distribution of the bleeding parameters yields a median of 0.13, with a a 2.5^{th} percentile of 0.03, and a 97.5^{th} percentile of 0.32, indicating substantial adjustment from the prior information. This is a positive result, because, while informative prior information should be used, it does not fully determine the final results of the model.

5.7 Discussion

In this chapter, we have developed a pair of spatial SEIR models, one which uses the robust exponential assumption, and one which allows for very realistic spatial PS SEIR models to be fit in the event that strong information is known about the latent and infectious time distributions of the disease process. These

models perform spatial smoothing, which is a desirable characteristic in analyzing epidemic data in the spatial framework. This is particularly desirable for sparse spatial epidemic data, but important for all epidemic modeling.

We note that informative prior information must be used for the specification of the MMCAR prior parameters. However, the interpretation of the spatial bleeding parameters is helpful in this regard. Appropriate prior information can be determined for these parameters. We also note the ability of the models to adjust the parameter posteriors to adjust from these priors, and the ability of the latent bleeding parameters to converge to different values than the priors would have generated. This is a favorable result, indicating that the data can adjust even moderately strong prior information for a data set as sparse as the Mumps data set.

To compare the two Spatial SEIR models we proposed here, we computed two posterior p-values. The first posterior p-value is related to model fit, and the statistic is defined as 1 when a predicted epidemic ends naturally within 40 days of the true epidemic with a final size within half as large and twice as large as the actual epidemic, and 0 otherwise. The Spatial PS SEIR yields a p-value of 0.31 and the Spatial SEIR model yields a p-value of 0.11. Both indicate model fit, but a higher proportion of accurate epidemics were produced by the Spatial PS SEIR model. An additional quantity of interest is the peak of the epidemic. To estimate the peak of the epidemic, we utilize the median of the ten day window containing the highest number of new infections. For model generated sample epidemics, we define the peak statistic as 1 if the peak of the sample epidemic is within ten days of the actual peak. Both posterior predictive p-values indicate model fit. However, the Spatial SEIR model yields a higher p-value than the Spatial PS SEIR model. This indicates high accuracy for this model to predict epidemic peak. The summary of the posterior p-value approaches can be found in Table 5.5.

Finally, we computed the model selection criterion DIC [28] for both models, based on 10,000 iterations after burn in. These values can be found in Table 5.5. Because DIC assumes normally distributed parameters, a normal approximation to the binomial distributions in both models was used. Note that these values may be inaccurate due to the extremely low counts of the exposure and infectious arrays. Additionally, n*p for the Spatial SEIR model component that involved transfer from the Susceptible to the Exposed matrix had a maximum of 1.6, which a 97.5^{th} percentile of the order of 0.01, which is far lower than the desired value of at least five required for the normal approximation to hold. We note that penalization term for DIC was actually higher for the Spatial SEIR model, which has fewer parameters. This is likely due to the poor approximation that the normal distribution gives to the binomial distribution under these conditions. This may lead to the selection of the Spatial PS SEIR model even when the Spatial SEIR model may be preferred. In fact, the Spatial SEIR model actually gives larger log-likelihoods than the Spatial PS SEIR model for our data set. This may indicate that the Spatial SEIR model fits the data better, but predicts new epidemics more poorly than the Spatial PS SEIR. The poorer predictions would be due to the high variance of the latent and infectious time distributions, and not due to the actual mixing and intervention parameter estimates themselves.

It is important to note that DIC does have the advantage of being rigorously defined, as opposed to the posterior p-value approaches, which require the modeler to define the fit statistics. This advantage of DIC yields support for future research into possible variants of DIC as a model fit statistic when comparing epidemic models.

Finally, the analysis of the Mumps data has been performed in a new and appropriate way. We once again note that the spatial smoothing yielded by the

realistic Spatial PS SEIR model mirrors the "zones of risk" analysis performed previously [69]. Additionally, the PS SEIR model yields strong evidence of model fit, with our posterior predictive p-value for model fit being 0.31. We have evidence that spring break was not a factor in increasing the overall number of cases of Mumps in this outbreak, even though it is known to have changed the composition of the ages affected during the outbreak [69]. Finally, we note that every analysis has yielded strong evidence of a public health awareness effect in this data. This is arguably an important feature to capture in any modern outbreak, and we have demonstrated this in the Mumps data.

Table 5.1: Medians and 95 % credible intervals for parameter posteriors for the Spatial SEIR simulations. Analysis 1 pertains to the small data set with strong priors, and Analysis 2 to the small data set with weak priors. Analysis 3 pertains to the large data set with strong priors, and Analysis 4 to the large data set with strong priors.

Parameter	Analysis 1	Analysis 2	Analysis 3	Analysis 4
β_1 (0.79)	0.69	0.61	0.87	0.80
	(0.42, 1.02)	(0.37, 0.94)	(0.53, 1.26)	(0.53, 1.16)
β_2 (0.86)	0.81	0.78	1.03	0.98
	(0.47, 1.22)	(0.42, 1.21)	(0.71, 1.42)	(0.69, 1.34)
ψ_1 (-0.59)	-0.59	-0.59	-0.01	0.07
	(-1.48, -0.09)	(-1.33, -0.06)	(-0.58, 0.39)	(-0.54, 0.42)
ψ_2 (0.09)	0.10	0.10	0.09	0.09
	(0.08, 0.13)	(0.08, 0.14)	(0.07, 0.11)	(0.07, 0.11)
ρ (17.5)	17.39	17.43	17.68	17.73
	(15.60, 19.21)	(15.61, 19.25)	(15.94, 19.44)	(15.96, 19.48)
μ_1 (-2)	-1.47	-1.42	-1.90	-1.93
	(-2.32, -0.64)	(-2.79, -0.18)	(-2.68, -1.09)	(-3.21, -0.59)
μ_2 (-2.5)	-3.17	-3.50	-2.83	-3.20
	(-4.00, -2.31)	(-4.91, -1.97)	(-3.58, -2.06)	(-4.44, -1.87)
σ_1^2 (4)	3.88	4.15	3.99	4.24
	(1.79, 5.77)	(0.78, 7.74)	(2.25, 5.76)	(1.52, 8.91)
σ_2^2 (4)	3.94	3.47	4.45	5.38
	(2.15, 5.87)	(0.81, 7.31)	(2.36, 6.29)	(2.30, 8.91)
δ (0.5)	0.48	0.56	0.47	0.38
	(0.18,0.79)	(0.07, 0.95)	(0.18, 0.75)	(0.02, 0.91)

Table 5.2: Medians and 95 % credible intervals for parameter posteriors for the Spatial SEIR simulations. Analysis 1 utilizes precise, centered information, and Analysis 2 utilizes off centered precise information. Analysis 3 utilizes imprecise centered priors, and Analysis 4 utilizes imprecise off centered priors.

Parameter	Analysis 1	Analysis 2	Analysis 3	Analysis 4
β_1 (1.00)	1.04 (0.69, 1.56)	0.58 (0.48, 0.95)	1.05 (0.68, 1.53)	0.76 (0.46, 1.12)
β_2 (1.25)	1.05 (0.67, 1.75)	0.65 (0.55, 1.01)	1.02 (0.63, 1.52)	0.79 (0.48, 1.21)
ψ_1 (-0.21)	0.24 (-0.22, 0.54)	-0.18 (-0.40,0.31)	0.13 (-0.44, 0.48)	0.07 (-0.54, 0.42)
ψ_2 (0.04)	0.04 (0.03, 0.05)	0.02 (0.02, 0.03)	0.04 (0.03, 0.04)	0.03 (0.02, 0.03)
μ_1 (-2)	-1.88 (-3.12, -0.59)	-1.97 (-2.37, -0.77)	-1.93 (-3.28, -0.59)	-2.08 (-3.36, -0.85)
μ_2 (-2.5)	-2.27 (-3.54, -0.98)	-2.17 (-2.61, -1.02)	-2.25 (-3.52, -1.03)	-2.19 (-3.43, -0.99)
σ_1^2 (4)	4.54 (1.18, 7.90)	4.19 (3.14, 8.28)	4.50 (1.68, 8.37)	3.95 (1.00, 7.79)
σ_2^2 (4)	4.37 (1.33, 7.90)	5.21 (4.22, 8.05)	4.89 (1.68, 8.37)	3.95 (1.00, 7.79)
δ (0.5)	0.43 (0.05,0.85)	0.55 (0.37, 0.88)	0.51 (0.03, 0.88)	0.51 (0.07, 0.87)

Table 5.3: Coverage probabilities for the latent spatial bleeding parameters in the Spatial SEIR simulations.

Data Set	Priors	Latent Coverage	Latent Coverage with Data
Small	Strong	91.1 %	95.8 %
Small	Weak	83.5 %	91.7 %
Large	Strong	86.7 %	91.5 %
Large	Weak	88.0 %	89.4 %

Table 5.4: Coverage probabilities for the latent spatial bleeding parameters in the Spatial SEIR simulations.

Priors	Latent Coverage	Latent Coverage with Data
Cen. Prec.	94.3 %	96.4 %
Off Cen. Prec.	93.6 %	94.7 %
Cen. Imp.	97.1 %	100 %
Off Cen. Imp.	96.2 %	100 %

Table 5.5: Medians and 95% credible intervals for the parameter posteriors of the Spatial SEIR and Spatial PS SEIR analysis of the Mumps data.

Parameter	Spatial SEIR	Spatial PS SEIR
β_1	0.56 (0.46, 0.89)	0.57 (0.37,0.82)
β_2	1.14 (0.82, 1.58)	1.17 (0.84, 1.56)
ψ_1	-0.16 (-0.76, 0.31)	0.12 (-0.24, 0.38)
ψ_2	0.07 (0.05, 0.09)	0.05 (0.04, 0.06)
ρ	17.57 (15.78, 19.38)	N/A
μ_1	-2.28 (-3.47, -0.97)	-2.29 (-3.53, -1.02)
μ_2	-3.31 (-4.19, -1.74)	-2.33 (-3.62, -1.12)
σ_1^2	3.63 (0.81, 7.52)	3.42 (1.24, 6.75)
σ_2^2	3.60 (0.80, 7.05)	3.24, (0.86, 7.27)
δ	0.54 (0.11, 0.87)	0.60 (0.04, 0.89)
Spatial MSE	190.67 (131.57, 5471.00)	100.10 (51.73, 1793.01)
Size p-value	0.11	0.31
Peak p-value	0.65	0.12
DIC	130120.4	78402.1

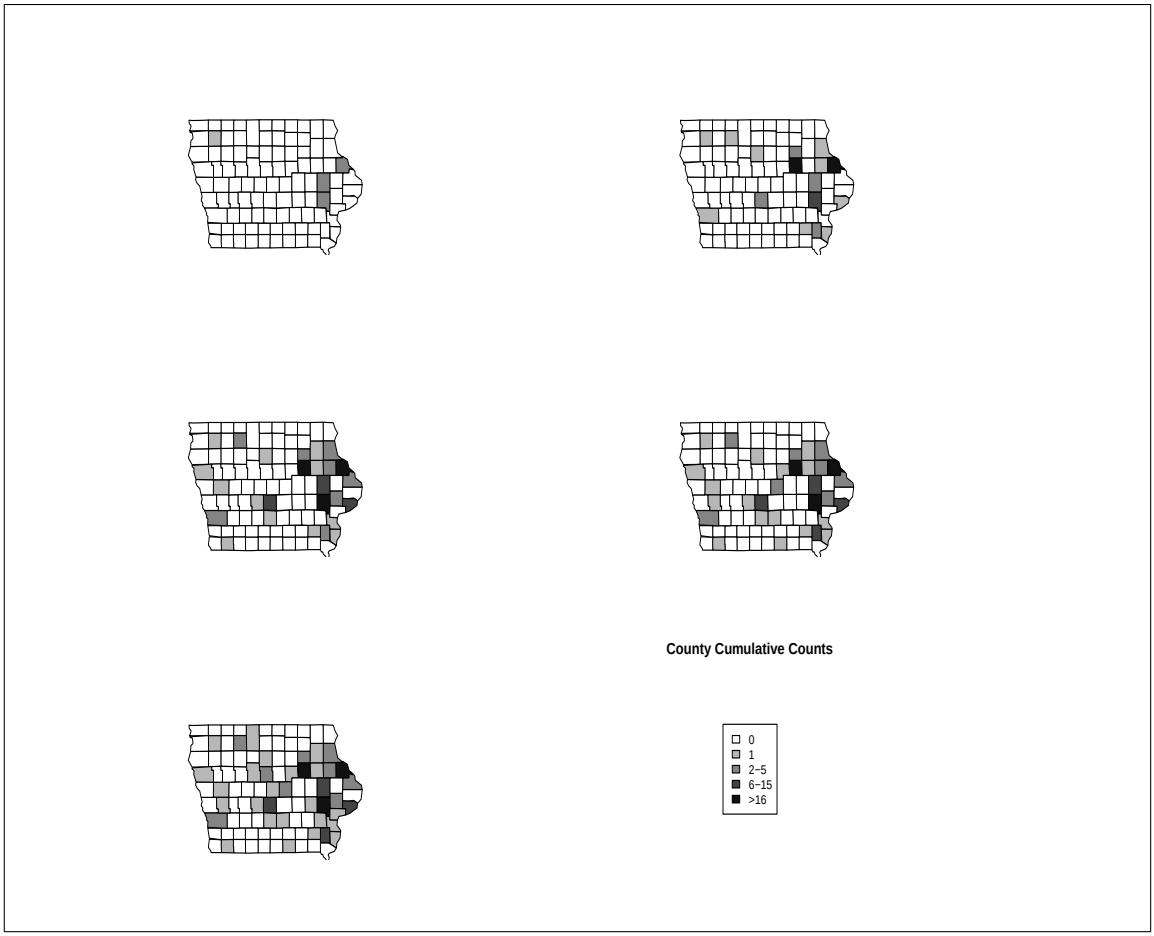

Figure 5.1: The Mumps data graphed over time. Each graph represents 30 additional days since the beginning of the epidemic.

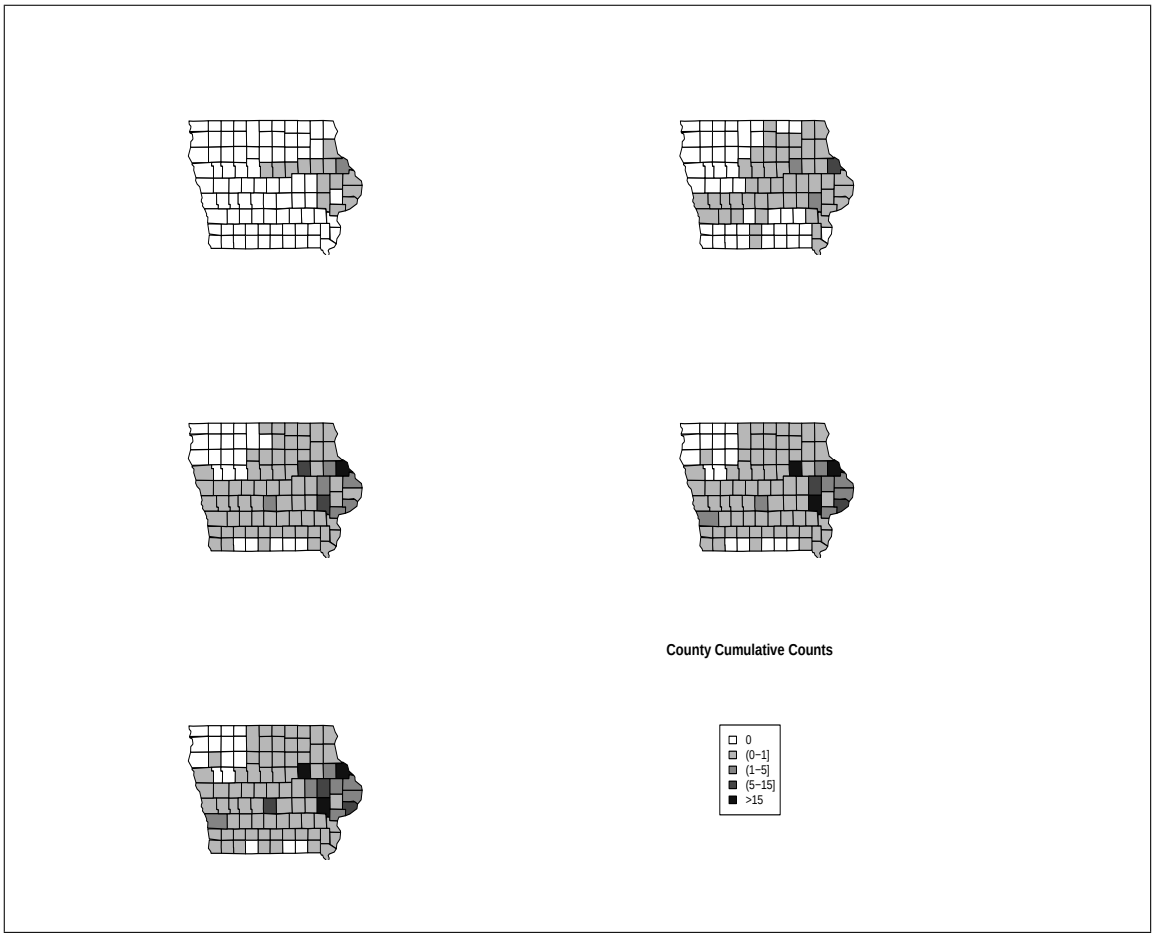

Figure 5.2: The averages counts predicted by the Spatial PS SEIR model graphed over time. Each graph represents 30 additional days since the beginning of the epidemic.

Figure 5.3: The county means of the predicted epidemics generated in the Mumps data analysis.

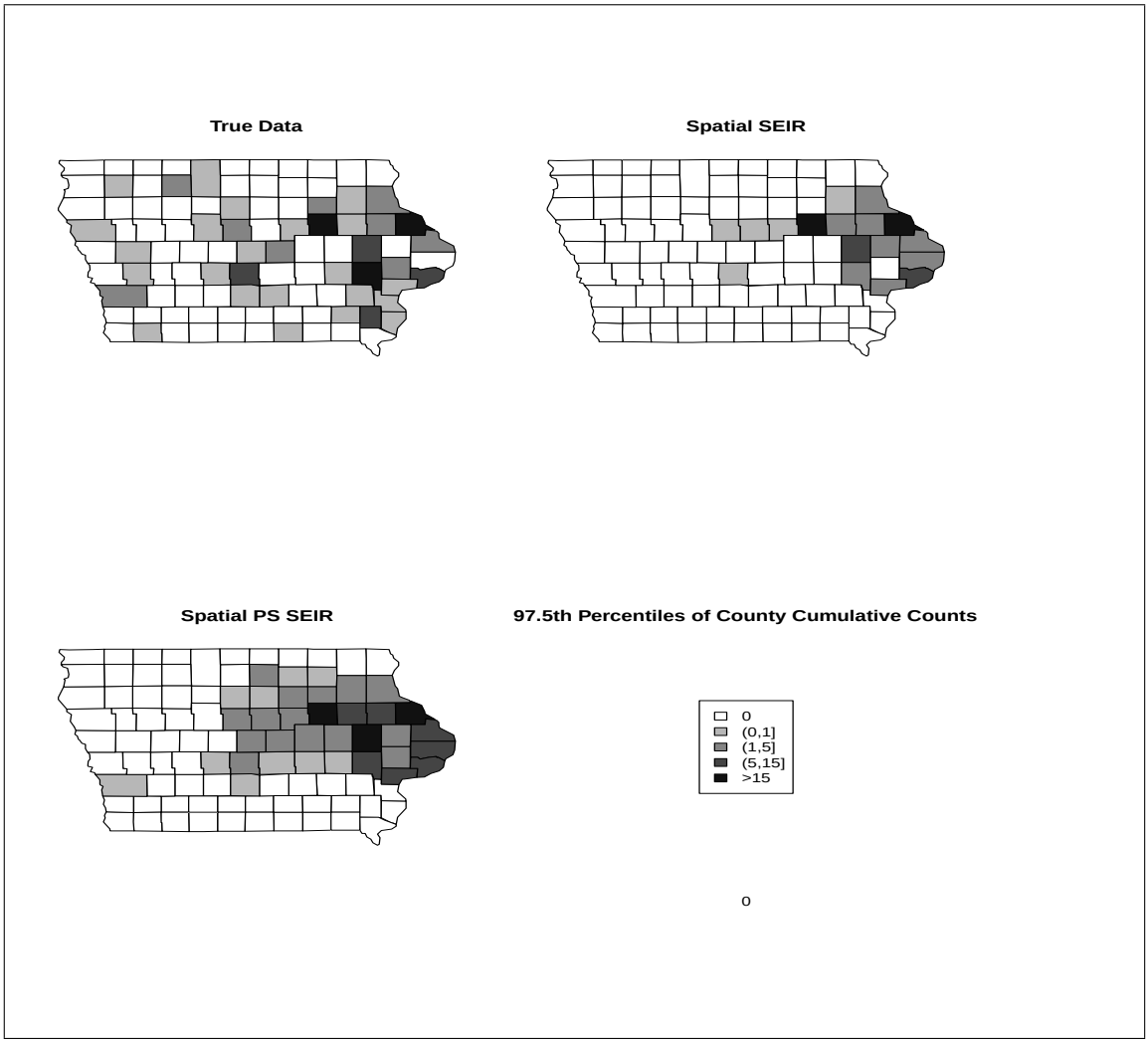

Figure 5.4: The county 97.5th percentiles of the predicted epidemics generated in the Mumps data analysis.

CHAPTER 6
CONCLUSION

6.1 Conclusions with regards to the Methodology of the Dissertation

The Path-Specific SEIR structure we have developed in Chapter Three is a useful SEIR model in and of itself, even without the spatial consideration. It is computationally more efficient than individual level models, and still has the ability to incorporate any discretizable distribution as a latent or infectious time distribution. It allows mixing and interventions to be considered at the population. Considering interventions at the population level is natural for computations and intuitively reasonable.

The model can reasonably detect the mean of the latent time distribution, but tends to overestimate the variance. This has little effect on mixing for large epidemics, but is important to consider for small epidemics. Prior information is typically available for this process, and readily informs the latent time distribution parameter estimates when the epidemic is small. This is a reasonable use of prior information.

The MMCAR model derived in Chapter Four extends the CAR class of models to include multiple outcomes on mismatched lattices. If the weights are fixed and known, which is the typical requirement for CAR models, an MCMC chain will draw from the CAR class of models. This induces the requirement that variances of each outcome at each lattice point differ by a known proportionality constant. It does not require that this proportionality constant be equal at each location or that every location have the same outcomes, however. In fact, the model does not even require that one lattice be a proper subset of the other.

In order to circumvent this structure on the variances, we have proposed using

the Brook's Lemma formulation of the MMCAR model conditionally. Then, the chains do not draw from a single CAR model, but rather conditionally draw from the CAR structure and still perform spatial smoothing. In the event that the spatial pattern being modeled can be considered a latent process, this may be a reasonable approach. Chapter Five demonstrated reasonable coverage for this approach.

The pair of spatial SEIR models developed in Chapter Five represent an alternative to the standard contact network approach. A graph is still utilized, but rather than a weighting scheme, it indicates where contacts may occur. The actual contact weighting is done by considering each individual as having a Poisson contact distribution with each neighboring location. This is more flexible, because weights do not have to be defined *a priori*, and it allows a CAR structure to be utilized.

The CAR structure approach is important here. These models are known to induce spatial smoothing, a desirable trait for an epidemic model. Heterogeneities in the spatial realizations of the model are often due to the contact and contraction mechanisms, rather than actual spatial heterogeneities. Therefore, it is desirable to smooth estimates of the data.

The spatial SEIR models do require prior information for the variance components when used to model epidemics of the size and sparsity of the Mumps epidemic. Two points are of note here. First, the interpretation of the bleeding parameters will allow informative priors to be reasonably utilized. Second, the real data analysis demonstrated that the bleeding parameters may burn in to reasonable values even when the variance components have diffuse posteriors. We suspect that flat priors can be used for large epidemics.

Finally, the Spatial PS SEIR represents a realistic and reasonable model for modeling epidemic spread. Prior information must be used to adequately incorporate the Path-Specific approach here. If such information is not available, the

Spatial SEIR model should be used. However, centered prior information yields very good results for modeling the Mumps data, and can be expected to perform well with other data sets as well.

6.2 Conclusions with regards to the Mumps Data

We have performed several new analyses of the Mumps data in this dissertation. The most reasonable analyses for drawing conclusions about this data are those from the Spatial PS SEIR model in Chapter Five.

The first question of interest was whether spring break increased or decreased the overall mixing that occurred in the state of Iowa during the Mumps epidemic. Polgreen et. al. have already shown that the age composition was changed [69], but the epidemic curves appeared to indicate an increase in infections over this time frame. It appears that, while spring break did change the age distribution of infectious, it did not cause an overall increase or decrease in the number of infections in the state of Iowa. Once we account for the spatial pattern in the data, the credible intervals for the spring break parameterization do not show any difference in mixing, as all of these intervals contain zero.

The second question was whether the public's awareness of this intervention changed the course of the epidemic. It appears that it did. Without accounting for the public becoming aware of the epidemic and changing its behavior, neither the models considered here nor the traditional models in the literature are able to give reasonable predictions of epidemic size. The fact that medicine has equipped the general population with the ability to protect themselves against infectious diseases is an important consideration when modeling contemporary outbreaks.

The final questions involved how to account for the spatial spread in the Mumps epidemic. The answer seemed to be that the primary measure of proximity

between counties is the sharing of borders and the sharing of a highway. With these two measures in place, only six to ten cases of Mumps cannot be explained in the epidemic from 2006. The use of public transportation proxies is an important consideration when modeling infectious disease spread on a lattice, and the use of highways in a SEIR structure is unique. These models add support to the use of multiple conduits of spread in epidemic modeling.

6.3 Directions for Future Research

There are several future directions for this research. First, the PS SEIR structure determines the means of the latent time distributions quite easily, but overestimates the variances. We would like to test the use of distributions which can be parameterized by only the mean, with a parameter fixed (such as a log-normal distribution with a known variance) or parameterizations of distributions using only a single parameter (such as Gamma(α^2, α)). This may allow flat priors to be adequately used for these models.

The requirement of a parameterization of interventions being decided upon *a priori* may be unnecessarily restrictive. Typically, functions are chosen *a priori*, but a semi-parametric method utilizing splines may be more flexible in this regard. We would like to pursue this research further.

The MMCAR structure is a very flexible spatial model. There may be natural extensions allowing for a spatio-temporal MMCAR model to be utilized. These may be similar to the STAR variety of spatio-temporal models. This structure would be particularly helpful when analyzing the data where the resolution of collection is changing over time.

We would like to apply the Spatial SEIR model to the Google Influenza data. It seems likely that this model could be used for such a data set. It would be helpful to investigate whether flat priors can be used for the variance components for such

a large data set. This also would allow some level of model validation to occur, and would allow for our hypothesis that the spatial bleeding does not change rapidly over time to be investigated.

Finally, the issues of model selection and model validation are almost unmentioned in the SEIR literature containing population level mixing. Model selection criteria should be developed, and model validation methods should be developed. This represents a large open problem in the literature.

REFERENCES

[1] Abbey, H. 1952. "An examination of the Reed Frost theory of epidemics." Human Biology. Volume 24, pages 201–233.

[2] Anderson, R. M. and R. M. May. 1991. *Infectious Diseases of Humans: Dynamics and Control* Oxford: Oxford Science Publications.

[3] Balcan, D., Goncalves, B., Hu, H., Ramasco, J.,Colizza, V. and A. Vespignani. 2010. "Modeling the Spatial Spread of Infectious Diseases: the Global Epidemic and Mobility Computational Model." *Journal of Computational Science* 1:132–145.

[4] Ball, F. G. 1991. "Dynamic Population Epidemic Models." *Mathematical Biosciences* 107:99–324.

[5] Banerjee, S., Carlin, B. P. and A. E. Gelfand. 2004. *Statistical Modeling and Analysis for Spatial Data* Boca Raton:Chapman and Hall/CRC Press.

[6] Barthelemy, M., Barrat, A., Pastor-Satorras, R. and A. Vespignani. 2005. "Dynamical Patterns of Epidemic Outbreaks in Complex Heterogeneous Networks." *Journal of Theoretical Biology* 235:275–288.

[7] Besag, J. 1974. "Spatial Interaction and the Statisitcal Analysis of Lattice Systems" *Journal of the Royal Statistical Society Series B* 36(2):192–236.

[8] Bolker, B.M., and B.T. Grenfell. 1996. "Impact of Vaccination on the Spatial Correlation and Persistance of Measles Dynamics." *Proc. Natl. Acad. Sci. USA.* 93:12648–12653.

[9] Boys, R. J. and P. R. Giles. 2007. "Bayesian inference for SEIR epidemic models with time-inhomogeneous removal rates." *J. Math. Biol.* 55:223–247.

[10] Carley, K. M., Fridsma, D. B., Casman, E.,Yahja, A., Altman, N.,Chen, L. C., Kaminsky, B. and D. Nave. 2006. "BioWar: Scalable Agent-Based Model of Bioattacks." *IEEE Transactions On systems, Man,and Cybernetics Part A: Systems and Humans* 36(2):250-265.

[11] CDC. 2006. "Mumps Epidemic." http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5513a3.htm (accessed October 20, 2010).

[12] CDC. 2009. "Updated Recommendations for Isolation of Persons with Mumps." http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5740a3.htm. (accessed October 20, 2010).

[13] CDC. 2009. "Mumps Clinical Questions and Answers." http://www.cdc.gov/mumps/clinical/qa-disease.html. (accessed October 20, 2010).

[14] Census Bureau. 2006. "Population Estimates." http://factfinder.census.gov. (accessed October 22, 2010).

[15] Chellappa, R. and S. Chatterjee. 1985. "Classification of textures using Gaussian Markov random fields." *IEEE Transactions on Acoustics, Speech and Signal Processing* 33(4) 959–963. doi 0.1109/TASSP.1985.1164641. ISSN 0096-3518.

[16] Chipara, O., Lu, C., Bailey, T. and G. Roman. November 3-5, 2010. "Reliable Clinical Monitoring using Wireless Sensor Networks: Experiences in a Step-down Hospital Unit." SenSys'10. Zurich, Switzerland.

[17] Cho, E., Meyers, S. A. and J. Leskovec. August 21-24, 2011. "Friendship and Mobility: User Movement in Location-Based Social Networks." KDD'11. San Diego, California, USA.

[18] Clayton, D. and J. Kaldor. 1987. "Empirical Bayes Estimates of Age-Standardized Relative Risks for Use in Disease Mapping." *Biometrics* 43(3):671–681.

[19] Colliza, V., Barrat, A., Barthelemy, M. and A. Vespignani. 2006. "The Role of the Airline Transportation Network in the Prediction and Predictability of Global Networks." *PNAS* 103(7): 2015–2020.

[20] Courcoul, A., Vergu, E., Denis,J. and F. Beaudeau. 2010. "Spread of Q Fever Within Dairy Cattle Herds: Key Parameters Inferred Using a Bayesian Approach." *Journal of the Royal Statistical Society Series B* 277: 2857–2865.

[21] Cressie, N. A. C. 1993. *Statistics for Spatial Data (Revised Edition).* New York: Wiley-Interscience.

[22] Diggle, P. J. 2006."Spatio-Temporal Point Processes, Partial Likelihood, Foot and Mouth Disease." *Statistical Methods in Medical Research* 15:325–336.

[23] Earnes, K.T.D. and M. Keeling. 2002. "Modeling Dynamic and Network Heterogeneities in the Spread of Sexually Transmitted Diseases." *PNAS* 99(20): 13330–13335.

[24] Elderd, D.E., Dukic, V.M. and G. Dwyer. (2006). "Uncertainty in Predictions of Diseases Spread." *PNAS* 103:15693–15697.

[25] Farley-Kim, R., Bart, S., Stetler, H., Orenstein, W., Bart, K., Sullivan, K., Halpin, T. and B. Sirotkin. 1985. "Clinical Mumps Vaccine Efficacy." *American Journal of Epidemiology* 121(4): 593–597.

[26] Funk, S., Gilad, E., Watkins, C., and V.A.A. Jansen. 2009. "The Spread of Awareness and its Impact on Epidemic Outbreaks." *PNAS* 106(16):6872–6877.

[27] Gelfand, A.E. and P. Vounatsou. 2003. "Proper Multivariate Conditional Autoregressive Models for Spatial Data Analysis." *Biostatistics* 1(4):11-25.

[28] Gelman, A. and D. B. Rubin. 1992. "Inference from iterative simulation using multiple sequences." *Statistical Science* 7:457–511.

[29] John Geweke, 1991. "Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments," Staff Report 148, Federal Reserve Bank of Minneapolis.

[30] Gibson, G. J. and E. Renshaw. 1998. "Estimating parameters in stochastic compartmental models using Markov chain methods." *IMA Journal of Mathematics Applied in Medicine and Biology.* 15:19–40.

[31] Gray, J. E.,Davis, D. E.,Pursely, D. M., Smallcomb, J. E. and A. Geva. 2011. "Network Analysis of Team Structure in the Neonatal Intensive Care Unit." *Pediatrics* 125(8):1480–1487.

[32] Grassly, N. C.,Fraser, C. and G. P. Garnett. 2005. "Host Immunity and Synchronized Epidemics of Syphilis Across the United States." *Nature* 433(27):417–421.

[33] Groendyke, C., Welch, D. and Hunter, D. R. 2011."Bayesian Inference for Contact Networks Given Epidemic Data." *Scandanavian Journal of Statisics* 38:600-616.

[34] Halloran, M. E., Ferguson, N. M., Eubank, S., Longini, I. M. Jr., Cummings, D. A. T., Lewis, B., Xu, S., Fraser, C., Vullikanti, A., Germann, T. C., Wagener, D.,Beckman, R., Kadau, K., Barrett, C., Macken, C. A. and Burke, D. S. and P. Cooley. 2008. "Modeling Targeted Layered Containment of an Influenze Pandemic in the United States." *PNAS* 105(12) 4639–4644.

[35] Heath, M. F., Vernon, M. C. and C. R. Webb. 2008. "Construction of Networks with Intrinsic Temporal Structure from UK Cattle Movement Data." *BMC Veterinary Research*[Online] 4(11). http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2335102/ (accessed October 22, 2011).

[36] Höhle, M. 2009. "Additive-Multiplicative Regression Models for Spatio-Temporal Epidemics." *Biometrical Journal*. 51: 961–978.

[37] Höhle, M. 2008. "Spatio-temporal Epidemic Modelling Using Additive-Multiplicative Intensity Models." Department of Statistics: Technical Reports, No.41.

[38] Höhle, M., Jorgensen, E., and P.D. O-Neill. 2005. "Inference in Disease Transmission Experiments by using Stochastic Epidemic Models." *Applied Statistics* 54(2):349–366.

[39] Hufnagel, L., Brockmann, D., Geisel, T. and R. May. 2004. "Forecast and Control of Epidemics in a Globalized World." *PNAS* 101(42):15124–15129.

[40] Jewell, C., Kypraiosy, T, Nealz, P. and G. Roberts. 2009. "Bayesian Analysis for Emerging Infectious Diseases." *Bayesian Analysis* 4(4):465–496.

[41] Jin, X., Carlin, B.P. and S.Banerjee. 2005. "Generalized Hierarchical Multivariate CAR Models for Areal Data." *Biometrics* 6:950–961.

[42] Khalil, K.M., Abdel-Aziz, M., Nazmy, T.T. and A.B.M. Salem. 2010. "An agent-based modeling for pandemic influenza in Egypt." 2010 The 7th International Conference on Informatics and Systems (INFOS).

[43] Kamp, C. 2010. "Demographic and Behavioural Change During Epidemics." International Conference on Computational Science, ICCS 2010.

[44] Keeling, M.J. 1999. "The Effects of Local Spatial Structure on Epidemiological Invasions." *Journal of the Royal Statistical Society Series B* 266:859-867.

[45] Keeling, MJ. and K. T. D. Earnes. 2005. "Networks and Epidemic Models." *Journal of the Royal Society Interface* 2:295–307.

[46] Keeling, M. 2005 "The Implications of Network Structure for Epidemic Dynamics." *Theoretical Population Biology* 67:1–8.

[47] Keeling, M. J., Woolhouse, M. E. J., May, R. M., Davies, G. and B. T.Grenfell. 2003. "Modelling Vaccination Strategies Against Foot-and-Mouth Disease." *Nature* 421:136–142.

[48] Kenah, E. and J. C. Miller. 2011. "Epidemic Percolation Networks, Epidemic Outcomes, and Interventions." *Interdisciplinary Perspectives on Infectious Diseases.*[Online] 2011. http://www.hindawi.com/journals/ipid/2011/543520/cta/ (accessed March 20, 2011). doi:10.1155/2011/543520

[49] Kendall, D. G. 1948. "On The Role of Variable Generation Time in the Development of a Stochastic Birth and Death Process." *Biometrika* 35(3/4):316–330.

[50] Kermack, W. O. and A.G. McKendrick. 1927. "A Contribution to the Mathematical Theory of Epidemics." *Journal of the Royal Statistical Society Series A* 115:700–721.

[51] Kim, H., Sun, Dongchu and R. K. Tsutakawa. 2001. "A Bivariate Bayes Method for Improving the Estimates of Mortality Rates with a Twofold Conditional Autoregressive Model." *Journal of the American Statistical Association* 96(456):1506–1522.

[52] Kiss, I. Z.,Green, D. M. and R. R. Kao. 2006. "The Effect of Contact Heterogeneity and Multiple Routes of Transmission on Final Epidemic Size." *Mathematical Biosciences* 203:124–136.

[53] Kleczkowski, A and B. T. Grenfell. 1999. "Mean-field-type Equations for Spread of Epidemics: the "Small World" Model." *Physica A* 274:355–360.

[54] Lawson, A.B. and H. Zhou. 2005. "Spatial Statistical Modeling of Disease Outbreaks with Particular Reference to the UK Foot and Mouth Disease (FMD) Epidemic of 2001." *Preventative Veterinary Medicine.* 71:141–156.

[55] Lawson, A. B. and H. Song. 2010. "Bayesian Hierarchical Modeling of the Dynamics of Spatio-Temporal Influenze Season Outbreaks." *Spatial and Spatio-Temporal Epidemiology* 1:187–195.

[56] Lekone, P. E. and B. Finkenstädt. 2006. "Statistical Inference in a Stochastic Epidemic SEIR Model with Control Intervention: Ebola as a Case Study." *Biometrics* 63:1170–1177.

[57] Lesslet, J., Kaufman, J. H., Ford, D. A. and J. V. Douglas. "The Cose of Simplifying Air Travel When Modeling Disease Spread." *PLoS One* 4(2) [Online] 2009. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2633616/. Accessed April 12, 2012.

[58] Lichstein, J. W., Simons, T. R.,Shriner, S. A. and K. E. Franzreb. 2002. "Spatial Autocorrelation and Autoregressive Models in Ecology." *Ecological Monographs* 72(3):445–463.

[59] Lloyd, A. 2001. "Realistic Distributions of Infectious Periodsin Epidemic Models: Changing Patterns of Persistence and Dynamics." *Theoretical Population Biology* 60:59–71.

[60] Mardia, K.V. 1998. "Multi-dimensional Multivariate Gaussian Markov Random Fields with Application to Image Processing." *Journal of Multivariate Analysis* 24:29-56.

[61] Mode, C.J. and C. K. Sleeman. 2010. "Stochastic Processes in Epidemiology: HIV/AIDS, Other Infectious Diseases, and Computers." Singapore: World Scientific Publishing Co. Pte. Ltd.

[62] Mollison, D. 1991. "Dependence of Epidemic and Population Velocities on Basic Parameters." *Mathematical Biosciences* 107:255–287.

[63] Neal, P. J. and G. O. Roberts. 2004. "Statistical Inference and Model Selection for the 1861 Hagelloch Measles Epidemic." *Biostatistics* 5(2):249–261.

[64] Neri, F. M., Perez-Reche, F. J., Taraskin, S. N. and C. A. Gilligan. 2011. "Heterogeneity in Susceptible-Infected-Removed (SIR) Epidemics On Lattices." *Journal of the Royal Society Interface* 8(55):201–209.

[65] O'Neill, P. D. and G. O. Roberts. 1999. "Bayesian Inference for Partially Observed Stochastic Epidemics." *Journal of the Royal Statistical Society Series B* 162:121–129.

[66] O'Neill, P. D. and N. G. Becker. 2001. "Inference for an Epidemic when Susceptibility Varies." *Biostatistics* 2:99–108.

[67] Parham, P. E. and N. M. Ferguson. 2006. "Space and Contact Networks: Capturing the Locality of Disease Transmission." *Journal of the Royal Society Interface* 3:483–493.

[68] Polgreen, P.M., Bohnett, L.C., Cavanaugh, J.E., Gingerich, S.B., Desjardin, L.E., Harris, M.L., Quinlisk, M.P. and M.A. Pentella. 2008. "The Duration of Mumps Virus Shedding after the Onset of Symptoms." *Clinical Infectious Diseases* 46:1450–1451.

[69] Polgreen, P.M., Bohnett, L.C., Yang, M., Pentella, M.A. and J.E. Cavanaugh. 2010. "A spatial analysis of the spread of mumps: the importance of college students and their spring-break-associated travel." *Epidemiology and Infection* 138:434–441.

[70] Reluga, T. C. 2010. "Game Theory of Social Distancing in Response to an Epidemic." *PLoS Computational Biology* 6(5):1-9.

[71] Sain, S. and N. Cressie. 2002. "Multivariate lattice models for spatial environmental data." 2002 Proceedings of the Joint Statistical Meetings, Section on Statistics and the Environment of the American Statistical Association. Pages 2820–2825

[72] Sain, S. R., Furrer, R. and N. Cressie. 2011. "A Spatial Analysis of Multivariate Output from Regional Climate Models." *The Annals of Applied Statistics* 5(1):150–175.

[73] Sattenspiel, L. and K. Dietz. 1995. "A Structured Epidemic Model Incorporatin Geographic Mobility Among Regions." *Mathematical Biosciences* 128:71–91.

[74] Scheel, I., Magne, A., Frigessi, A. and P.A. Jansen. 2007. "A Stochastic Model for Infectious Salmon Anemia (ISA) in Atlantic Salmon Farming." *Journal of the Royal Society Inferface* 4(15)699–706.

[75] Schimit, P.H.T. and L.H.A. Monteiro. 2009. "On the basic reproduction number and the topological properties of the contact network: An epidemiological study in mainly locally connected cellular automata." *Ecological Modeling* 220:1034–1042.

[76] Schumm, P., Scoglio, C., Gruenbacher, D., and T. Easton. December 10-13, 2007. "Epidemic Spread on Weighted Contact Networks." Bionetics'07, Budapest, Hungary.

[77] Spiegelhalter, D.J., Best, N. G., Carlin, B.P. and A. Van der Linde. 2002. "Bayesian Measures of Model Complexity and Fit (with Discussion)." *Journal of the Royal Statistical Society Series B* 64:583–616.

[78] Streftaris, G. and G. J. Gibson. 2004. "Bayesian analysis of experimental epidemics of foot-and-mouth disease." *Journal of the Royal Statistical Society Series B* 271:1111–1117.

[79] Sun, D., Tsutakawa, R., Kim, H. and Z. He. 2000. "Spatio-Temporal Interaction with Disease Mapping" *Statistics in Medicine* 19:2015–2035.

[80] Svensson, A. 2007. "A Note on Generation Times in Epidemic Models." *Mathematical Biosciences* 208:300–311.

[81] Tolbert, P. E., Mulholland, J. A. Macintosh, D. L., Xu, F., Daniels, D., Devine, O. J., Carlin, B. P., Klein, M., Butler, A. J., Nordenberg, D. F., Frumkin, H., Ryan, P. B. and M. White. 2000. "Air Quality and Pediatric Emergency Room Visits for Asthma and Atlanta, Georgia." *American Journal of Epidemiology* 151(8):798–810.

[82] Wearing, H., Rohani, P. and M. Keeling. 2005. "Appropriate Models for the Management of Infectious Diseases." *PLoS Medicine* 2(7):0621–0627.

[83] Xia, Y., Bjornstad, O. N., and B.T. Grenfell. "Measles Metapopulation Dynamics: A Gravity Model for Epidemiological Coupling and Dynamics." 2004. *The American Naturalist.* 162(2):267-281.

[84] Yang, Y., Longini, I. M. and M. E. Halloran. 2006. "Design and Evaluation of Prophylactic interventions using infectious disease incidence data from close contact groups." *Applied Statistics* 55(3):317–330.