Distribution Agreement

In presenting this dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I agree that the Library of the University shall make it available for inspection and circulation in accordance with its regulations, governing materials of this type. I agree that permission to copy from, or to publish, this dissertation may be granted by the professor under whose direction it was written, or, in his absence, by the Dean of the Graduate School when such copying or publication is solely for scholarly purposes and does not involve potential financial gain. It is understood that any copying from, or publication of, this dissertation which involves potential financial gain will not be allowed without written permission.

Signature:

_____ _____

Jeffrey M. Switchenko Date

Estimation of Epidemic Model Parameters:

A Spatial Analysis using Bayesian Techniques

By

Jeffrey M. Switchenko

Doctor of Philosophy

Biostatistics

Lance A. Waller, Ph.D.
Advisor

Michael J. Haber, Ph.D.
Committee Member

Andrew N. Hill, Ph.D.
Committee Member

Leslie A. Real, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Estimation of Epidemic Model Parameters: A Spatial Analysis using Bayesian Techniques

By

Jeffrey M. Switchenko

M.S., Emory University, 2010

B.A., Bowdoin College, 2006

Advisor: Lance A. Waller, Ph.D.

An abstract of

A dissertation submitted to the Faculty of the

James T. Laney School of Graduate Studies of Emory University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in Biostatistics

2011

Abstract:

Infectious disease models attempt to evaluate the effects on the spread and transmission of disease. One particular model, the susceptible-infected-recovered (SIR) model, places individuals into classes of disease progression, where a series of differential equations tracks the rates of transmission and recovery for a given disease through a susceptible population. Two parameters, the transmission parameter and the recovery parameter, drive the dynamics of the model, and their ratio, R_0, is the average number of cases caused by one infectious individual within a completely susceptible population. R_0 is seen as one of the most important quantities in the study of epidemics, and signals how quickly a particular disease can spread amongst a susceptible population. Previous analyses have focused primarily on tracking these epidemic disease parameters over time, and classifying individuals due to baseline differences which reflect heterogeneity within the population. For example, these differences can be based on age, gender, vaccination status, or behavior.

However, we choose to quantify the spatial heterogeneity that exists in spatially-referenced data in an effort to define core areas of disease rates and transmission. We first consider geographically weighted regression (GWR) models in an effort to assess the spatial variability that exists between disease rates and baseline tract-level characteristics which can define core disease areas. Next, we build hierarchical Bayesian models which incorporate random effects structures, inducing correlation in local estimates of disease transmission with exchangeable random effects, which smooth local estimates based on global averages, and conditionally autoregressive (CAR) random effects, which smooth local estimates based on neighboring estimates. We extend a chain binomial model to predict the spread of disease, while considering two different parameterizations of the chain binomial model, and simulate outbreaks to assess model performance. In addition, we extend a general epidemic model, which incorporates aspects of frailty models in assessing heterogeneity within the population. Through our modeling approaches, we are able to identify cores areas for the transmission of sexually transmitted infections (STIs) in Baltimore, Maryland from 2002-05.

Estimation of Epidemic Model Parameters: A Spatial Analysis using Bayesian Techniques

By

Jeffrey M. Switchenko

M.S., Emory University, 2010

B.A., Bowdoin College, 2006

Advisor: Lance A. Waller, Ph.D.

A dissertation submitted to the Faculty of the

James T. Laney School of Graduate Studies of Emory University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in Biostatistics

2011

Acknowledgements

I would like to thank my advisor, Lance Waller, for his incredible support and advice over these last five years. The challenges I have faced along the way would have been impossible to tackle without his help and advice.

I would also like to thank my committee members, Andrew Hill, Michael Haber, and Leslie Real, for their help and suggestions on my dissertation. Their guidance has been invaluable in my progress towards finishing my dissertation. For her help in accessing the data, I would like to thank Jacky Jennings from Johns Hopkins University who made this research possible.

In my time at Emory University, I have had the opportunity to consult on a number of projects, which would not have been possible without the assistance and opportunities provided by Kirk Easley and Michael Kutner. In addition, I would like to thank Patrick Kilgo for providing me with teaching opportunities in a problem-based learning environment as well as the opportunity to work on creative projects in statistics in baseball research. Thanks to all the various students and faculty members in my department for their help and encouragement.

Finally, I would like to thank my parents for their constant support, guidance, and love through every step of my life. My completion of the degree is a reflection of their dedication and sacrifice to make my goals possible.

Contents

List of Figures

List of Tables

1 Introduction

Since epidemiologic methodology has arisen in the last century, researchers attempted to quantify, measure, and track infectious diseases over time and space. The goals primarily involved the containment, control, and prevention of infectious disease spread, and the attempt to define the most vulnerable populations at risk for obtaining disease. Measures such as calculating disease rates over time or mapping the locations of disease helped researchers identify disease outbreaks, as with John Snow plotting cholera cases on a Broad Street map in London in an effort to determine the source of an outbreak and eliminate it. Attempts to model disease spread involve the basic premise of controlling infectious diseases and reducing disease incidence. Any model in infectious diseases is only as effective as the quality and reliability of the available data, and each model brings a specific insight into the dynamics of a particular infectious disease. For example, some models are effective in accounting for vaccination rates and others address the inherent heterogeneity of those at risk.

Our research focuses on quantifying the spatial heterogeneity that exists in infectious disease data. By allowing model parameters to vary according to a position in space, we can identify core areas of disease transmission, i.e. areas of higher disease rates which are defined geographically and typically can be characterized by socioeconomic factors such as poverty and poor health care access. The objective of our model choice is to estimate parameters which effectively describe the level of infectiousness and contagiousness of a given disease, and we hope to identify areas where future outbreaks are likely. We test our models on sexually transmitted infection (STI) data collected in Baltimore City County, Maryland from 2002 through 2005.

1.1 SIR Disease Modeling

Infectious disease systems can be thought of as a complex network of interactions, representing processes at different scales, and among individuals in a population. In order to describe some of the underlying mechanisms of disease dynamics, we can use mathematical and statistical models with sets of assumptions that generate predictions about a specific system. The process of modeling infectious diseases can be broken down into a series of objectives where we set up simple models, account for contact heterogeneities, account for seasonal transmission dynamics, add stochastic extensions, and ultimately estimate parameters of interest [36]. At the outset, we can start by assigning individuals within a population to specific categories of disease transmission, and tracking the progress of the disease using a series of basic differential equations. These equations account for changes in the numbers of those who have contracted the disease and those who have recovered from the disease.

The Kermack-McKendrick model is an epidemiological model that computes the theoretical number of people infected with a contagious illness in a closed population over time [37]. At any time t, individuals in that population can be classified as susceptible $X(t)$, infectious $Y(t)$, or recovered $Z(t)$. At the outset of an epidemic, we assume that the number of people in the susceptible class is approximately equal to the total population and that the number of infectives is relatively small (close to zero). We also generally assume that zero individuals are in the recovered stage at the outset.

1.1.1 Parameters of interest

Under an assumption of random mixing, i.e. that every individual is equally likely to interact with any other individual, we define κ as the contact rate between individuals per unit of time, and c as the probability of transmission given contact. For a susceptible, a fraction Y/N of contacts will be with those who are infected, and in a small time interval δt, the number of contacts with infectives is $\kappa(Y/N)\delta t$. Thus, the probability of escaping infection is equal to $(1-c)^{\kappa(Y/N)\delta t}$ or $1-\delta q$. If we define the transmission parameter β as $-\kappa\log(1-c)$, then:

- $\delta q = 1 - \exp(-\beta Y \delta t/N)$
- $\delta q = 1 - \left(1 - \beta Y \delta t/N + \frac{(\beta Y \delta t/N)^2}{2!} - \frac{(\beta Y \delta t/N)^3}{3!} + ...\right)$
- $\delta q/\delta t \approx \beta Y/N \rightarrow dq/dt = \beta Y/N$

The transmission rate per susceptible is equal to $\beta Y/N$, and the resulting transmission rate for the entire susceptible class is $-\beta XY/N$. If we assume the recovery rate is constant, then γ is equivalent to the infectious period or the inverse of the length of time spent infectious. The transmission rate is assumed to depend on the frequency of contacts between susceptibles and infecteds in the population Y/N. Although we assumed κ to be constant, we easily could assume that the contact rate is proportional to the population size N, i.e. the contact rate is equal to κN, resulting in density-dependent transmission. Frequency dependence is generally more appropriate in large populations with heterogeneous mixing, sexually transmitted infections, or vector-borne pathogens, while density-dependence is more likely for wildlife diseases with smaller population sizes [36].

1.1.2 Differential equations

Differential equations form the foundation of the SIR (susceptible-infected-recovered) model, as rates of change for each class are tracked over time. The dynamics of the outbreak are governed by two parameters in this basic model: β, the transmission parameter, and γ, the recovery parameter. The SIR differential equations are defined as follows:

$X(t)$ = susceptible population size at time t.

$Y(t)$ = infected population size at time t.

$Z(t)$ = removed/recovered population size at time t.

$N = X + Y + Z$ = Total population size, assumed to be constant.

The population dynamics follow:

$$\frac{dX}{dt} = \frac{-\beta XY}{N},$$

$$\frac{dY}{dt} = \frac{\beta XY}{N} - \gamma Y, \text{ and}$$

$$\frac{dZ}{dt} = \gamma Y,$$

where $X(0) = X_0 \approx N$, $Y(0) = N - X_0 \approx 0$, $Z(0) = 0$.

The SIR model can also be rewritten to display proportions, as opposed to counts. We denote $S = \frac{X}{N}, I = \frac{Y}{N}, R = \frac{Z}{N}$, where S, I, R are now dimensionless proportions and the dynamics follow:

$$\begin{aligned}
\frac{dS}{dt} &= \left(\frac{1}{N}\right)\frac{dX}{dt} = -\frac{\beta XY}{N}, \\
&= -\beta\left(\frac{X}{N}\right)\left(\frac{Y}{N}\right), \\
&= -\beta SI, \\
\frac{dI}{dt} &= \beta SI - \gamma I, \text{ and} \\
\frac{dR}{dt} &= \gamma I,
\end{aligned}$$

where $S(0) = S_0 \approx 1$, $I(0) = 1 - S_0 \approx 0$, $R(0) = 0$.

The basic SIR model offers many opportunities for customization. For instance, seasonal changes in disease dynamics and births/deaths (demography) can be taken into account. An exposed class can be added for diseases with a latency period, and the model can be adapted to an SIS system, where no individuals develop immunity, i.e., individuals move directly from the infective class back to the susceptible class with no recovery time. Finally, we can account for stochasticity, which involves fluctuations in population processes that arise from the random nature of events at the level of the individual, where the baseline probability associated with each event is fixed at the proportions above, but individuals can experience different outcomes.

1.2 R_0

For frequency-dependent transmission, R_0 is defined as the ratio of the transmission parameter to the recovery parameter, $\frac{\beta}{\gamma}$, and can be thought of as the maximum reproductive potential of a pathogen - also referred to as the basic reproduction number. If the SIR model is established with density-dependent transmission in a closed population of N individuals, R_0 is equivalent to $\frac{\beta N}{\gamma}$, with homogeneous mixing where each individual produces on average βN offspring per unit of time for an average of $1/\gamma$ time-units.

Conceptually, R_0 is the average number of susceptible individuals infected by one infectious individual in a completely susceptible population, and can also be thought of as the product of the average transmission rate per contact, the average number of new contacts per time unit, and the average length of infectious period or βcD [3]. As noted by Halloran [28], the value of R_0 is not specific to a parasite or pathogen, but to a population within a particular host population at a particular time. We further extend this idea by considering R_0 as a function of location; that is R_0 is also a function of a *particular place*. For example, the contact rates in rural areas should be lower than contact rates in urban areas, so we would expect R_0 of measles, for example, to be lower in rural areas than in urban areas. R_0 values for a variety of well-known infectious diseases are listed below (Table 1).

Disease	**Transmission**	R_0
Measles	Airborne	12-18
Pertussis	Airborne droplet	12-17
Diphtheria	Saliva	6-7
Smallpox	Social contact	5-7
Polio	Fecal-oral route	5-7
Rubella	Airborne droplet	5-7
Mumps	Airborne droplet	4-7
HIV/AIDS	Sexual contact	2-5
SARS	Airborne droplet	2-5
Influenza (1918)	Airborne droplet	2-3

Table 1: List of R_0 values for well-known infectious diseases. Note that each disease has a range of R_0 estimates, not one specific value.

According to Heesterbeek, R_0 is the most important quantity in the study of epidemics and notably in comparing population dynamical effects of control strategies [31]. If everyone is initially susceptible, a sustained outbreak requires $X/N = 1 > 1/R_0$ or $R_0 > 1$, i.e., each infection must do more than replace itself, on average, for an epidemic to occur. If everyone is not susceptible, then a successful pathogen invasion requires $X/N = 1/R_0$ so the fraction susceptible must be above $1/R_0$. In a spatial setting, this reproduction number is very useful for determining core areas of disease as it signals how quickly the disease can spread amongst a susceptible population and is directly related to the amount of control effort needed to eliminate an infection from a population.

It is possible to reduce the parameters β and γ to a single parameter, by rescaling the SIR model. With a rescaled time variable $\tau = \gamma t$, the mean duration of infection is $1/\gamma$, where a unit increase in τ corresponds to the real elapsed time equal to the mean duration of infection [22]:

$$\frac{dS}{d\tau} = \left(\frac{1}{\gamma}\right)\frac{dS}{dt} = -\left(\frac{\beta}{\gamma}\right)SI,$$

$$= -R_0 SI,$$

$$\frac{dI}{d\tau} = R_0 SI - I, \text{ and}$$

$$\frac{dR}{d\tau} = I,$$

$$\text{where } R_0 = \frac{\beta}{\gamma}.$$

At time 0, $\frac{dY}{d\tau} = Y(R_0 X_0 - 1) \approx Y(R_0 - 1)$ and the relationship between R_0 and R_∞ (the fraction of all individuals who contract the disease before it dies out) is: $R_0 = -\frac{1}{R_\infty} \ln(1 - R_\infty)$. Thus, the SIR model parameters can be reduced to one parameter, R_0, again demonstrating the central role of R_0 in determining the dynamics of an infectious disease.

1.3 Analysis and Data

We attempt to extend these infectious disease models to a specific population, and account for differences in transmission based on location, or spatial heterogeneity. As a result, we hope to estimate infectious disease parameters spatially in order to address public health problems in Baltimore, Maryland. Our goal is to provide quantitative evidence of core areas of disease transmission.

Sexually transmitted infections (STIs) represent a challenging public health dilemma. In the community of Baltimore, Maryland, researchers at the Baltimore City Health Department (BCHD) collected information on several STIs over the course of four years, 2002-05. The data for this analysis include a set of coordinates for each STI case location in the greater Baltimore area, as well as the date (year, month, day) of diagnosis. The BCHD recorded a total of 12,556 cases of gonorrhea (GC) over the study period. For the purposes of our study, we will be grouping the cases by census tract location in order to perform areal data analysis [6]. Specifically, we consider the 200 census tracts which make up Baltimore City County. According to the 2000 US Census, 651,154 people live in Baltimore City County, and consistent with the STI literature, we consider the 336,551 individuals aged 15-49 in the population to be "at-risk" for GC. This results in a median tract-level incidence rate of 40 GC cases per 1000 individuals at risk from 2002-05 [51].

We will consider four years of data with a total of 12,556 cases of GC contracted in the Baltimore area. The data will be aggregated by census tract as well as month of infection, and we will assume that the aggregated cases within each tract at each time point represent the number of newly infectious individuals. However, in order to estimate our primary function of parameters, R_0, we need both infection times and removal times. We will assume a fixed infectious period of one month for each infectious individual, which will fix γ at 1.

GC is a sexually transmitted infection caused by the bacterium *Neisseria gon-*

orrhoeae, which can grow and easily multiply in the reproductive tract. It can be spread through sexual contact, or from mother to child during childbirth. Although some individuals will remain asymptomatic, symptoms can occur less than 5 days after infection. Like syphilis, GC can be treated with antibiotics, but the disease can be reacquired through sexual contact with another individual with the disease. According to the CDC, gonorrhea is very common, with approximately 700,000 new cases in the United States each year, although only half are usually reported [25]. From 1975 to 1997, rates declined, although the national rate has steadily climbed since 2000. Previous spatial analyses of sexually transmitted infections have focused on gonorrhea [27], chosen because the geography of the disease has been widely studied, incidence is relatively high, and the disease is easily and readily diagnosed. It has been noted that because of the disease's short incubation and absence of acquired immunity, GC incidence responds rapidly to changes and does not exhibit the wider oscillations characteristic of other sexual transmitted infections such as syphilis [7].

Hethcote and Yorke developed mathematical models for gonorrhea incidence in the 1970s and 1980s [56, 32]. Their modeling procedures address the concept of a saturation factor which limits the prevalence of disease, and they note that acquired immunity cannot be a saturation factor for GC. A saturation factor occurs when infectious individuals contact individuals who are already infected from different sources - also known as the preemption effect. Since some individuals may have many more sex partners than others, the population is not uniform and homogeneously mixing. Thus, the preemption that limits gonorrhea occurs primarily in a subset of the at-risk population. As a result, Hethcote and Yorke propose separating the population at risk into many subgroups according to demographic characteristics and other relevant factors such as the number of sex partners. Hethcote and Yorke describe the population within groups with significant preemption effects as the "core" driving the disease dynamics.

Other epidemiologic studies of STIs have noted substantial associations between demographic variables and high incidence of STIs. In a number of US cities, African-Americans typically live in highly segregated neighborhoods, and sexually transmitted infections, as a result, tend to remain confined to those neighborhoods because of racially and geographically segregated sexual networks [35]. The analyses below build on earlier studies which have focused primarily on the presence of "core areas" of STI transmission in Baltimore, Maryland in the mid-1990s [7, 35, 57]. Core areas primarily are defined geographically and can be characterized by socioeconomic factors such poverty and poor health care access [7]. Other study areas have also been considered in STI core area analysis [1, 16, 39, 47, 50]. Hethcote and Yorke note that if the core infections are removed so that there is no saturation in the remaining groups at risk, the disease will eventually die out [56]. The primary focus of this research will be to build on the concept of core area identification over space and time through quantitative methods.

2 Mathematical vs. Statistical Modeling

In describing the spread of infectious diseases, it may be necessary to quantify the results using particular mathematical functions to describe the deterministic patterns in the data. It is possible to use mechanistic descriptions with meaningful parameters, derived from a theoretical model describing the underlying process driving the pattern [13]. There are a variety of analytical and computational methods available to explain ecological phenomena, and the differential equations from an SIR model are an example of a more complex system of functions used to determine the deterministic flow of disease spread in a population. Many of these theoretical models can provide general insight on an ecological question, and although quantitatively precise, they result in qualitative conclusions. In 1966, Levins addressed the nature of theoretical model building as a trade-off between generality, realism, and precision, and the truths about these ecological questions could be revealed by "robust theorems" [41].

However, for many ecologists and statisticians, it is apparent that true systems have a level of variability in the system. Purely deterministic differential equations can describe nicely the general spread of disease from susceptibles to infectives to recovered individuals, but do not account for "noise" or variability in the estimates of the parameters driving the dynamics of the model. Thus, the introduction of stochasticity, or randomness, to a deterministic model provides a more accurate description of how the disease is actually behaving and spreading in a given population. More specific insight is available than a purely mathematical model, and these applied systems can capture the real complexity and quirkiness of the behavior of the disease through added variability in the model. Bolker details this discrepancy as a trade-off between process, the theoretical set of functions describing expectation, and pattern, the phenomenological application to real-world dilemmas [13].

Overall, the type of model one chooses and implements is the result of determin-

ing the ecological questions one wants to answer and the ecological questions one actually can answer. The data available will help to determine the direction of the modeling process, and hopefully yield reliable estimates for the questions to be answered. We propose development of mathematical and statistical analyses of R_0 for sexually transmitted infections (STIs) within a geographically defined population.

In the following, we distinguish between mathematical calculations of R_0 and statistical estimation of R_0, and we briefly review the literature and common approaches for both here. Anderson and May proposed the most well-known equation for calculating R_0, where R_0 is the product of the transmission probability, the contact rate, and the duration of the disease (βcD) [3]. However, more advanced methods are available for determining the theoretical value of R_0 for a given disease in a given population.

2.1 Calculation of R_0

2.1.1 Basic calculation

Two indirect, simple approaches for calculating R_0 can be used when the transmission system is assumed to be in dynamic equilibrium [28]. If one assumes that the average incidence rate and prevalence of disease are not changing, then the infectious case will produce one other infectious case, on average, so $R = 1$. Also, from the relation $R = R_0 x = 1$, the proportion susceptible at equilibrium would be equal to $x = 1/R_0$. If one assumes random mixing, then R_0 is calculated simply by the reciprocal of the proportion susceptible. In the second method, derived by Dietz, the incidence rate is assumed to be independent of age, and the average age of infection A is equal to the inverse of the incidence rate I. If one knows the average life expectancy L of a population, then R_0 is calculated by dividing the life expectancy L by the average age of infection A [21].

2.1.2 Survival function

Dietz [21] moved from mathematical calculation toward statistical estimation by considering an underlying survival model. As proposed by Dietz, consider a large population and let:

- $p(\alpha)$ = the probability of surviving to age α, or the probability a newly infected individual remains infectious for at least time α

- $\beta(\alpha)$ = the rate of giving birth for an individual of age α, or the average number of newly infected individuals that an infectious individual will produce per unit time when infected for total time α

Then:

$$R_0 = \int_0^\infty p(\alpha)\beta(\alpha)\, d\alpha$$

It is straightforward to handle situations in which infectivity depends on time since infection and other transmission probabilities between states vary with time [48]. If we assume:

- X = number of susceptibles,
- Y = number of infectives,
- h = proportion of contacted persons who are infected given a contact,
- κ = number of persons contacted per unit of time by one infectious individual,
- γ = rate of transfer to a non-infectious state (either susceptible or immune),
- N_0 = initial size of population,
- 1 = initial number of infectious individuals,
- $X = N_0 - 1$ initially (t=0),

then:

$$\frac{dY}{dt} = \kappa h \frac{X}{N_0 - 1} Y - \gamma Y, \text{ and}$$

Y will increase only if $R_0 = \dfrac{\kappa h}{\gamma} > 1$ where $D = \dfrac{1}{\gamma}$ = the duration of the infectious period, and κD equals the number of persons contacted during this period. As noted by Dietz, R_0 is only meaningful for diseases where contacts are clearly defined such that they can be counted, and his methods can be difficult to apply in practice [48].

2.1.3 Multitype model

In the multitype model, Britton [15] and Ball and Clancy [5] assume a closed population of size N, consisting of k different types of individuals, $i = 1, ..., k$, from a heterogeneous population, and define a stochastic SIR epidemic model where:

- n_i = the number of i-individuals (individuals in type i), and
- $\pi_i = \frac{n_i}{n}$, the corresponding population proportion.

If an i-individual becomes infected, he/she becomes infectious, possibly after a latency period. An i-individual has a "close contact" with any given j-individual at rate $\frac{\beta_{ij}}{n}$, where a "close contact" is defined as contact which could result in infection if the other individual is susceptible, else the contact has no effect. The matrix $\{\beta_{ij}\}$ of contact intensities is assumed to be irreducible, omitting the possibility of a major outbreak for some but not all types of individuals. The infectious period I_i has distribution F_i, mean μ_i, and standard deviation σ_i. The parameter $\lambda_{ij} = \mu_i \beta_{ij}$ is also of interest, where $\lambda_{ij}\pi_j$ is the expected number of close contacts which an i-individual has with j individuals during the infectious period. When the infectious period is over, the individual recovers and becomes immune, and the individual is removed. The epidemic evolves until there are no more infectious individuals in the population. All contact processes and infectious periods are defined to be mutually independent.

If we allow for time-varying infectivity, including an initial latency period, with $\{I_i(t); t \geq 0\}$, then $I_i(t)$ is the infectivity t time units after infection of an i-individual and F_i denotes the distribution of $\int_0^\infty I_i(t)\, dt$. Thus, a branching process determines how the infectious individuals infect new individuals independently, and R_0 is defined as largest positive eigenvalue of the matrix $\lambda_{ij}\pi_j$. In the branching process, $\lambda_{ij}\pi_j$ corresponds to the matrix of the mean offspring distribution, where the proportion s_j are susceptible.

Similar to Britton, Höhle defines R_0 in terms of an eigenvalue of a matrix within a branching process [33]. R_0 is then equivalent to the mean number of offspring in a limiting multitype branching process, which is the Perron-Frobenius eigenvalue of the matrix: $nE(T_1)B\Pi$ where $E(T_1) = \frac{\gamma_I}{\delta_I}$ = the mean infectious period, B = the contact rate matrix, and $\Pi = \text{diag}\left(\frac{n_1}{n}, ..., \frac{n_k}{n}\right)$ which is the proportion of the initial susceptible individuals in the individual units.

According to Grassly, one can also define R_0 based on the idea of an offspring distribution [26]. If Y is the number of secondary infections that are generated by a single infected individual, then the probability function that describes the distribution of Y is referred to as the offspring distribution. Thus, the number of secondary infections would depend on the infectiousness of the index case over time τ since the index case became infectious, which is equivalent to $\beta(t)$. This quantity is the product of biological infectiousness and contact rates or a product of biological, behaviorial, and environmental infectiousness. R_0 can be calculated through the integral:

$$E(Y) = c\int_0^\infty \beta(\tau)\, d\tau = R_0.$$

The SIR model typically assumes a constant infectivity β while an individual remains infected and a constant rate of recovery from infection γ such that the time spent infectious is exponentially distributed. Under an assumption of random infectious contacts among individuals in a population, the reproduction number can be thought of as a mixture of Poisson distributions with exponentially distributed means. In reality, the infectiousness and susceptibility of an individual are influenced by many different factors, where individuals can be categorized by any factors that are considered to be important for infectious disease transmission, denoted as h-states by Diekmann and Heesterbeek [20]. The h-state variables are characteristics which describe an individual or group of individuals, and are defined by varying levels of susceptibility to disease, where h denotes heterogeneity in a population. Similar methods for determining R_0 assuming an age-stratified heterogeneous population are

discussed by Farrington [24]. The h-state concept also motivates our consideration of spatial heterogeneity in the model parameters.

2.1.4 Next-generation operator

Heesterbeek and Diekmann provide another approach to calculating R_0 using the next-generation method [20, 31]. To define R_0, one must first specify a linear positive operator - "the next-generation operator" - which maps generations of infected individuals into each other, as distributed over the possible individual characteristics. Susceptibles are assumed to be in a steady state in the absence of an infectious agent, and one only regards the initial stage of an invasion by an infectious agent into a fixed population of susceptibles.

If we express the next-generation operator as a matrix, then the dominant eigenvalue is equal to R_0. Hence, to determine R_0, one must identify relevant heterogeneous characteristics, construct elements of the next-generation operator in terms of the basic parameters and ingredients, and compute the dominant eigenvalue of the operator. The "type-at-birth" causes individuals to have different susceptibility - similar to heterogeneities within host populations with complex risk structures [36]. The next-generation operator is an $n \times n$ matrix where n equals the number of types-at-birth. $K(\epsilon, \eta)$ is equal to the expected number of new cases with type-at-birth ϵ caused by a single infected individual with type-at-birth η. Infectivity, contacts, and disease status depend on the time elapsed since becoming infected τ. Requirements for calculation include infectivity as a function of τ and η, survival as a function of τ and η, and total contacts towards susceptibles that could be born with type ϵ. Next, define operator K as:

$$(K\psi)(\epsilon) = \int_\Omega K(\epsilon, \eta)\psi(\eta)\, d\psi$$

where ψ is equal to the generation of infecteds as distributed over the types-at-birth, and R_0 is equal to the dominant eigenvalue of K.

2.2 Statistical Estimation of R_0

A number of methods exist for statistically estimating R_0 from observed data, ranging from simpler techniques such as measuring the slope of an epidemic curve, to more complex model-based methods involving Bayesian sampling techniques such as Markov chain Monte Carlo (MCMC) methods. Each method provides an approximation to the reproduction number by using different sets of data available for estimation; thus, more complex methods require more data of more different types. Direct estimation, according to Dietz, is not easy [21].

2.2.1 Epidemic curve estimation

An epidemic curve is a plot of the number of infections over time, where the frequency of infectious individuals is plotted across the y-axis and time (usually in days, weeks, or months) is plotted over the x-axis. During the early stages of an outbreak, the number of infected individuals is approximately $I(t) \approx I_0 \exp[(R_0 - 1)(\beta + \gamma)t]$. After taking the logarithm of both sides, one can show that the log of the number of infected individuals is approximately linear in time with a slope which can be corrected to be roughly equivalent to R_0 [36]. As a result, simple linear regression fit to the first several data points on a log-scale, corrected to account for β and γ, provides a rough estimate for R_0.

2.2.2 Final outbreak size

An additional method for estimating R_0 comes from information contained in the final size of an epidemic. Although not helpful at the early stages of an epidemic, this method can be a useful tool for later analysis. Keeling and Rohani propose that we assume the epidemic is started by a single infectious individual in a completely susceptible population [36]. As mentioned previously, this individual would infect R_0 others, on average. The probability a particular individual escaped infection is $\exp(-R_0/N)$. If Z individuals have been infected, then the probability of an individual escaping infection from all potential sources is $\exp(-ZR_0/N)$. At the end of an epidemic, a proportion $R(\infty) = Z/N$ have been infected and the fraction remaining susceptible is $S(\infty) = \exp(-R(\infty)R_0)$, which is equal to $1 - R(\infty)$. Thus, we can obtain the equation $1 - R(\infty) - \exp(-R(\infty)R_0) = 0$, and numerical methods are then required to find the value of R_0 solving the equation.

2.2.3 Least squares estimation

Of course, more complicated methods exist for estimating R_0 more efficiently and exactly. The next methods for estimation involve estimating the parameters of the SIR model and then calculating R_0 from the parameter estimates. One technique involves least squares fitting or trajectory matching, where we find the values of the model parameters which minimize the squared differences between model predictions and the observed data [13]. **R** functions [45], such as `optim`, enable estimation by finding parameter values minimizing this sum of squared errors.

2.2.4 Chain binomial models

Next, we move toward parameter estimation based on maximum likelihood - a very general method for model parameterization, estimation, and importantly inference [13]. After specifying a stochastic model presumed to have generated the observed data, we next determine the probability that the observed data would be generated by the model. Maximum likelihood estimation (MLE) attempts to identify which parameterizations make this probability the greatest.

As an example, consider a chain binomial epidemic. Chain binomial models are dynamic models developed from the simple binomial model by assuming that infection spreads from individual to individual in populations in discrete units of time, producing chains of infection governed by the binomial probability distribution. In the Reed-Frost model [8], we assume that people pass through three states: susceptible, infectious, and recovered. If we assume that p equals the probability of infection for one susceptible individual when contacting one infectious individual at a given time point, and $q = 1 - p$ equals the probability of avoiding infection from that contact, then q^{i_t} is the probability that a susceptible individual avoids infection from i_t infectious individuals. The transition probability of getting $I_{t+1} = i_{t+1}$ new infectives at time $t + 1$, given $S_t = s_t$ and $I_t = i_t$ susceptibles and infectives [30] is:

$$P(I_{t+1} = i_{t+1} | S_t = s_t, I_t = i_t) = \binom{s_t}{i_{t+1}} (1 - q^{i_t})^{i_{t+1}} q^{i_t(s_t - i_{t+1})}, s_t \geq i_{t+1}$$

The number of new infectives depends on the number of old infectives in the Reed-Frost model. In the Greenwood model, the number of new infectives does not depend on the number of old infectives, but on the presence of one or more infectives, such that:

$$P(I_{t+1} = i_{t+1} | S_t = s_t, I_t = i_t) = \left\{ \begin{array}{cc} \binom{s_t}{i_{t+1}} (1 - q^{i_t})^{i_{t+1}} q^{i_t(s_t - i_{t+1})}, s_t \geq i_{t+1}, i_t > 0 \\ \\ 0 \qquad \text{otherwise} \end{array} \right\}$$

The binomial probabilities for each of the chain binomial models define their respective likelihood functions. Then, we maximize the likelihood, or log-likelihood, functions with respect to the model parameters. With the likelihood function in place, we can extend the models into a Bayesian framework as shown in the next section.

3 Bayesian Inference for Epidemic Modeling

The mathematical and statistical models represent a progression from deterministic solutions to stochastic estimation. Likelihood-based methods provide adequate modeling techniques in the field of epidemic modeling; however, we can define a full probability model and estimate parameters using the framework of Bayesian inference. In classical inference, the parameters from our epidemic models are regarded as fixed quantities, where the values of the parameters are estimated from data using estimators that are random variables and whose distributional properties may be known [43]. However, in Bayesian inference, the parameters are assumed to be random variables, and the posterior distribution is the desired density for estimation purposes. The posterior density, or the distribution of the parameters given the data, is defined using Bayes' Theorem as the normalized product of the likelihood and the prior density. We can choose a prior density specific to our modeling needs, such that uncertainty can be represented by uninformed prior distributions. Likewise, in the case of stronger epidemiological prior evidence, we can use informative priors.

Bayesian inference has a number of advantages over the techniques used in classical inference. Using a Bayesian framework enables us to make inferences on the probability of a given parameter or model - an issue for classical inference. Additionally, in sparse data cases, we can introduce prior information on the parameters of interest. Finally, problems involving random effects, process and measurement error, and unobserved states are typically more difficult to solve using classical techniques [13]. Oftentimes, it can be straightforward to obtain credible intervals under Bayesian inference for parameters, whereas classical confidence intervals may require the development of appropriate theoretical results [43].

Another advantage of Bayesian inference involves the imputation of missing data. We often encounter missing data in epidemic modeling due to missing infection or removal times, for example, and classical inferential techniques along with the stan-

dard likelihood may be difficult to evaluate. This becomes a substantial problem with temporal data, as the likelihood usually involves integration over all possible infection times. With Bayesian inference, we can assume that missing infection times are random quantities or extra model parameters to be estimated. Although classical techniques such as the EM algorithm have been considered for analysis of epidemic data, the expectation step can become more complicated [43]. Techniques available in Bayesian inference are more straightforward.

For the purposes of parameter estimation in epidemic modeling, Bayesian inference assumes the following:

- $y = (y_1, ..., y_n)$: the observed data, such as counts of cases or infection/removal times,
- $f(y|\theta)$: a stochastic model for the observed data, usually a probability distribution, defining the likelihood,
- θ: a vector of unknown parameters, assumed a random quantity, and
- $\pi(\theta)$: the prior distribution of θ.

Therefore, the posterior distribution for inference concerning θ is $p(\theta|y) = \dfrac{f(y|\theta)\pi(\theta)}{\int f(y|u)\pi(u)du}$. We can avoid evaluating the denominator integral by the use of Markov chain Monte Carlo (MCMC) methods, wherein we focus on the unnormalized posterior density: $p(\theta|y) \propto f(y|\theta)\pi(\theta)$ or $p(\theta|y) \propto L(\theta)\pi(\theta)$. By sampling from the unnormalized posterior density, we hope to build the posterior density of the parameters under the knowledge that the denominator of our posterior distribution is independent of the parameters and will cancel out in the algorithm.

MCMC methods are an established suite of methodologies, and the literature on MCMC methods is extensive. In Bayesian inference, the target density is the joint posterior density of the model parameters, and these methods work by defining

a Markov chain whose stationary distribution is equal to the normalized target density [43]. We simulate the chain until we establish adequate convergence, and samples from the target density of interest can be drawn. MCMC sampling techniques have the advantage of eliminating the summation and high-dimensional integration that can make classical inferential techniques numerically complicated or intractable for models involving integration over infectious periods. These techniques also enable us to analyze any model parameters or a specific function of them. In our case, if we attempt to estimate both transmission and recovery parameters, we still have the ability to analyze the ratio of the two - R_0.

Two algorithms are typically associated with MCMC methods: the Gibbs sampler and the Metropolis-Hastings algorithm. The Gibbs sampler is an algorithm used to generate a sequence of samples from the joint posterior density of two or more random variables. The goal of the Gibbs sampler is to approximate the joint distribution or the marginal distributions of one or more of the variables. In the algorithm, the next value of a particular parameter is sampled from a distribution conditioned on the previous values of the other parameters. The Gibbs sampler is a special case of the more general Metropolis-Hastings algorithm.

With the Metropolis-Hastings (MH) algorithm, we start with a target distribution π for a parameter θ, and construct a Markov chain according to the following rules:

- Start with some initial value $X_0 = x_0$.
- For $n = 0$ to N do:
  - Simulate a candidate value $Y \sim q(j|X_n = i)$. Suppose $Y = j$.
  - Compute the MH acceptance probability: $a_{ij} = min\left\{\frac{\pi_j q(i|j)}{\pi_i q(j|i)}, 1\right\}$.
  - Generate $U \sim$ Unif[0,1].
  - Accept the candidate $Y = j$ if $U \leq a_{ij}$, otherwise $X_{n+1} = X_n$.

The post-convergence values of the Markov chain X_i for i=1 to N iterations represent a sample from the posterior density of θ. Note that a symmetric proposal density $q(i|j) = q(j|i)$ yields the original Metropolis algorithm. The acceptance probability simplifies to $a_{ij} = \min\{\frac{\pi_j}{\pi_i}, 1\}$. For the Gibbs Sampling algorithm, we consider the target distribution to be $f(\mathbf{x})$ and we assume that we can sample from full conditional distributions $x_i|\mathbf{x}_{-i}$, where the notation $\mathbf{x}_{-i}$ indicates all elements of $\mathbf{x}$ except the ith component. We construct the Gibbs Sampler as follows:

- Start with some initial value $X_0 = x_0$.
- For $n = 0$ to N do:
  - Sample $x_1^{n+1} \sim f_1(x_1|\mathbf{x}_{-1}^{(t)})$.
  - Sample $x_2^{n+1} \sim f_2(x_2|x_1^{(t+1)}, x_3^{(t)}, ..., x_p^{(t)})$.
  - Sample $x_p^{n+1} \sim f_p(x_p|\mathbf{x}_{-p}^{(t+1)})$.

As with the Metropolis-Hastings algorithm, the values of the Markov chain $x_1^{(i)}$ from i=1 to N iterations represents the posterior density of our parameter of interest [46, 14]. After establishing a burn-in period, we estimate the parameters of interest by either taking the mean or median of the posterior samples in the generated chain.

3.1 Chain binomial models

We next move to more specific models for estimating the parameters of our epidemic models using the MCMC techniques listed above. Although a wide variety of epidemic models can work, the processes of the Metropolis-Hastings algorithm and Gibbs Sampling algorithm remain consistent. We choose the stochastic chain binomial models as a way of defining the relationship between our parameters and the data.

Chain binomial models are stochastic, dynamic models, which assume that the numbers of individuals who are susceptible or infected are binomially distributed and that infection spreads from individual to individual in discrete time units. The numbers of individuals who are susceptible and infectious are assumed to be known at each time point. In the Reed-Frost model, individuals start out susceptible to infection S, then can be become infected/infectious I, and eventually recover from the disease R. Thus, the SIR model assumes no latency period, and can be used to track the progression of person-to-person infectious diseases [30].

3.1.1 Transitional approach

We extend the basic chain-binomial model, which assumes that the infection time series is a chain of binomially distributed random variables. In previous work done by Lekone and Finkenstädt [40], the numbers of newly infectious individuals and newly recovered individuals (transition compartments) are binomially distributed. The transitions of individuals from one stage of disease to the next are stochastic movements between the corresponding model compartments. In each time period, an individual can either stay or move on to the next compartment. If we assume an exponentially distributed length of time that an individual spends in each compartment with compartment specific rate $\lambda(j)$, then the probability of staying within that compartment for an additional length of one time unit (months, for example)

is $\exp(-\lambda(j))$. Thus, the probability of leaving is $1 - \exp(-\lambda(j))$. Let NI_j denote the number of susceptible individuals who become infectious at time j. Likewise, let NR_j denote the number of cases who are removed from the infectious class at time j. We use a discrete-time approximation to a continuous SIR model, and we also define individuals as susceptible, infectious, or removed from the population as follows for the transition approach to the chain binomial model:

S_j = those susceptible at time point j,

I_j = those infectious at time point j,

R_j = those removed at time point j,

$S_{j+1} = S_j - NI_j$,

$I_{j+1} = I_j + NI_j - NR_j$, and

$R_{j+1} = R_j + NR_j$.

NI_j and NR_j are random variables with binomial distributions:

$$NI_j \sim \text{Bin}(S_j, p_j),$$
$$NR_j \sim \text{Bin}(I_j, p_R),$$

$$\text{where } p_j = 1 - \exp\left(\frac{-\beta}{N} I_j\right),$$

$$\text{and } p_R = 1 - \exp(-\gamma)$$

where p_j represents the probability of a susceptible individual becoming infectious at the given timepoint j. p_R represents the probability that an infectious individual is removed from the study population. These probabilities of staying in a compartment are the result of the compartment-specific exponential rates. We can establish a Bayesian model as follows:

$$\begin{aligned} NI_j &\overset{ind}{\sim} \text{Bin}(S_j, p_j), \\ p_j &= 1 - \exp\left(\frac{-\beta_j}{N} NI_j\right), \text{ and} \\ \beta_j &= \beta_0. \end{aligned}$$

We next assign a vague gamma prior to the fixed effect β_0, since β_0 can take values of 0 or greater. As we are assuming the infectious period is equal to one time unit (1 month), R_0 is equivalent to β_0.

3.1.2 Reed-Frost approach

In the Reed-Frost model of disease transmission, the assumption is made that individuals pass through three states - susceptible, infective, and recovered. The model assumes a fixed population size N, and each person is in one of the three states where S_j, I_j, and R_j are defined in the previous chain binomial model. Binomial models are often used to estimate the transmission probability, and the exposure to infection can occur in discrete time contacts, which can also be discrete time units of exposure [30]. It is typically assumed that each contact is independent of other contacts. We can define p_j as the transmission probability during a contact between a susceptible individual and an infectious individual at time point j, and we can further define q_j as the escape probability which is equivalent to $1 - p_j$. The probability that a susceptible individual escapes infection from all infectious individuals at time point j is $q_j^{I_j}$. Thus, the probability of not escaping infection is $1 - q_j^{I_j}$ or h_j.

In this case, R_0 is a function of the number of initial susceptibles S_0 and the transmission probability p_j, such that $R_0 = S_0 * p_j$. No assumptions are made concerning the recovery parameter γ, as well as the transmission parameter β. The Reed-Frost chain binomial model is set up as follows:

$$\begin{aligned} I_{j+1} &\overset{ind}{\sim} \text{Bin}(S_j, h_j), \\ h_j &= 1 - q_j^{I_j}, \\ q_j &= 1 - p_j, \\ \text{Logit}(p_j) &= \alpha_0, \text{ and} \\ \alpha_0 &\sim N(0, \tau_z). \end{aligned}$$

We can assign a vague normal prior to the fixed effect α_0, as well as a vague gamma hyperprior distribution to τ_z.

3.2 General epidemic model

We can define a general epidemic as a counting process for infections and removals [2, 8, 43, 44]. A counting process is a stochastic process $N(t), t \geq 0$ that contains the following properties:

- $N(t) \geq 0$,
- $N(t)$ is an integer, and
- If $s \leq t$, then $N(s) \leq N(t)$.

If $s < t$ then $N(t) - N(s)$ is the number of events that occurred during the interval $(s, t]$. In addition, we consider a closed population of M individuals, and we assume that multiple cases introduce the infection into a population of initially susceptible individuals, starting an outbreak. Once the outbreak starts, the hazard of infection depends only on the presence or number of infectives in the population. If an individual becomes infected, he or she is infective for an exponentially distributed period of time, after which he/she becomes removed when developing symptoms. Those individuals who are removed will no longer contribute to the outbreak, and there is no latency period.

For an individual i, the events of infection and removal could be described in terms of two counting processes, where $N_I^{(i)}(t)$ jumps by one at the time of infection, and $N_R^{(i)}(t)$ jumps by one at the time of removal. We assume $N_I^{(i)}(0)$ and $N_R^{(i)}(0) = 0$, except for the initial cases where $N_I^{(i)}(0) = 1$. Additionally, we can denote $H_t^{(i)}$ as the history of the two processes up to time t: $H_t^{(i)} = \{N_I^{(i)}(s), N_R^{(i)}(s); 0 \leq s \leq t\}$. The two counting processes are then specified in terms of their stochastic intensities:

$$\begin{aligned} P(dN_I^{(i)}(t) = 1 | H_{t-}^{(i)}) &= \frac{\beta}{M} I(t) S^{(i)}(t) dt, \text{and} \\ P(dN_R^{(i)}(t) = 1 | H_{t-}^{(i)}) &= \gamma I^{(i)}(t) dt \end{aligned}$$

where $S^{(i)}(t) = 1 - N_I^{(i)}(t-)$, $I^{(i)}(t) = N_I^{(i)}(t-) - N_R^{(i)}(t-)$, and $I(t) = \sum_{i=1}^{M} I^{(i)}(t)$. Next, we count the total numbers of infections and removals that occur up to time t:

$$N_I(t) = \sum_{i=1}^{M} N_I^{(i)}(t) \text{ and } N_R(t) = \sum_{i=1}^{M} N_R^{(i)}(t)$$

Thus, the numbers of susceptibles and infectives at time t are:

$$S(t) = \sum_{i=1}^{M} S^{(i)}(t) = M - N_I(t-) \text{ and } I(t) = \sum_{i=1}^{M} I^{(i)}(t) = (N_I(t-) - N_R(t-))$$

We then define the general epidemic as a counting process for infections and removals:

$$\begin{aligned} P(dN_I(t) = 1|H_{t-}) &= \frac{\beta}{M} I(t)S(t)dt \\ P(dN_R(t) = 1|H_{t-}) &= \gamma I(t)dt \end{aligned}$$

where H_t is the history of the aggregated processes. The infection times are: $\{0 = t_1 < t_2 < ... < t_n\}$ and removal times: $\{\tau_1 < \tau_{n-1} < \tau_n = T\}$. The likelihood of parameters β and γ based on the complete data is as follows:

$$\begin{aligned} L(\beta, \gamma; y_{complete}) &= \prod_{i=1}^{n} \left\{ \gamma I(\tau_i) \exp\left(- \int_{\tau_i - 1}^{\tau_i} \gamma I(u) du \right) \right\} \\ &\quad \times \prod_{j=2}^{n} \left\{ \frac{\beta}{M} I(t_j) S(t_j) \exp\left(- \int_{t_j - 1}^{t_j} \frac{\beta}{M} I(u) S(u) du \right) \right\} \\ &\quad \times \exp\left(- \int_{t_n}^{T} \frac{\beta}{M} I(u) S(u) du \right) \\ &= \prod_{i=1}^{n} \{\gamma I(\tau_i)\} \prod_{j=2}^{n} \{\frac{\beta}{M} I(t_j) S(t_j)\} \\ &\quad \times \exp\left(- \int_{0}^{T} (\gamma I(u) + \frac{\beta}{M} I(u) S(u)) du \right) \end{aligned}$$

The likelihood aggregates the amount of time spent in the infectious class by all infectives over the time period, the amount of time susceptibles spend interacting with infectives, as well as the hazard of infection and hazard of removal. The survival component (the integral) aggregates the cumulative hazard over time. We assume independent gamma priors for β and γ as follows:

$$\begin{aligned} f(\beta) &\propto \beta^{\nu_\beta - 1} \exp(-\lambda_\beta \beta), \text{ and} \\ f(\gamma) &\propto \gamma^{\nu_\gamma - 1} \exp(-\lambda_\gamma \gamma). \end{aligned}$$

The following Metropolis-Hastings algorithm can be used in the estimation of β and γ:

- Initialize chain for β_0 and γ_0.
- Draw candidate β_1 from proposal distribution $q(.|\beta_0)$.
- Set numerator equal to $\dfrac{f(t,\tau|\beta_1,\gamma_0)f(\beta_1)}{q(\beta_1|\beta_0)}$.
- Set denominator equal to $\dfrac{f(t,\tau|\beta_0,\gamma_0)f(\beta_0)}{q(\beta_0|\beta_1)}$.
- Generate U from U(0,1).
- If $U <$ numer/denom, then $\beta_1 = \beta_1$ else $\beta_1 = \beta_0$.
- Draw candidate γ_1 from proposal distribution $z(.|\gamma_0)$.
- Set numerator equal to $\dfrac{f(t,\tau|\beta_1,\gamma_1)f(\gamma_1)}{z(\gamma_1|\gamma_0)}$.
- Set denominator equal to $\dfrac{f(t,\tau|\beta_1,\gamma_0)f(\gamma_0)}{z(\gamma_0|\gamma_1)}$.

- Generate U from U(0,1).
- If $U <$ numer/denom, then $\gamma_1 = \gamma_1$ else $\gamma_1 = \gamma_0$.
- Run for N iterations, until convergence then draw samples from posterior density.

After establishing a burn-in period, we can estimate the parameters of interest by either taking the mean or median of the generated chain representing the posterior density of a given parameter. From these methods, we can estimate R_0 as the ratio of the transmission parameter β to the recovery parameter γ. By running this process over the entire study space as well as each census tract, we can estimate an overall R_0 as well as a series of tract-specific values of R_0. We use the software package **R** to implement the MCMC sampling techniques for the general epidemic model.

4 Introduction to Conceptual Epidemic Models

4.1 Initial Spatial Analysis: Tracking Spatial Patterns in Prevalence

Our data include 12,556 cases of gonorrhea (GC) in the years 2002-05 within Baltimore City County - the specific area of interest for our analysis. A total of 651,154 people live in the 200 census tracts of Baltimore City County, and we consider the 336,551 individuals aged 15-49 in the population to be "at-risk". The GC data were collected by the Baltimore City Health Department, and accessed by researchers at the The Johns Hopkins University. The raw data were geocoded in ArcMap [23] using the TIGER/Line files for the street map and census block group boundary file. Addresses outside Baltimore City County were excluded based on street name and zip code, and all years had similar geocoding rates (between 93-96%). In addition, addresses were automatically geocoded with a 10-meter offset, a minimum 70% match score, and allowed for ties.

A total of 651,154 people live in the 200 census tracts of Baltimore City County, yielding a median tract-level incidence rate of 40 cases per 1000 people at risk over the four years. At the census tract level, the average percent of the population with a college degree is 17.9%, the average percent with a high school degree is 66.2%, and the average median home value is $71,514 (Table 2). Also at the census tract level, the average median age is 35 years, the average per capita income is $16,872, and the average percent below the poverty line is 24.6% (Table 2). The average percent black is 63.5%, the average percent white is 31.7%, and the average percent foreign born is 4.3% (Table 2). Figure 2 illustrates some of the discrepancies between census tract level total cases and case rate per 1,000 people over four years in each census tract, although generally a higher number of cases yields higher rates of cases. Demographic choropleth maps are also shown in Figure 2, and similar spatial patterns

exist in nearly all maps. Areas of lower educational qualifications are typically areas of lower income, higher percent in poverty, lower home values, and also high GC rate. It appears that GC rates are also highly associated with race, as evidenced by particularly high numbers and rates within the African-American communities in Baltimore.

The data for analysis include a set of coordinates for each GC case location in the greater Baltimore area, as well as the date (year, month, day) of diagnosis. These observations were taken over the course of four years from 2002 through 2005, and are aggregated over the time period to produce one set of total cases. The U.S. Census Bureau provides the rest of the data used for analysis [51]. These variables include percent with a college degree, percent with a high school degree, median home value, median age, per capita income, percent black, percent white, percent below the poverty line, and the percent foreign born. Demographic variables such as race and age are Census 2000 100-percent Data, while socioeconomic variables such as education level and housing data are Census 2000 Sample Data. Census tract level population numbers are used to establish rates of GC infection per 1,000 people at risk in each census tract. We use Hawth's Tool "Count Points in Polygons" within ArcMap [12] to attribute a GC count number for each census tract in the Baltimore area, which was then standardized by the population and multiplied by 1000 to establish a GC rate of infection per 1,000 persons per census tract [23]. Choropleth maps were generated for GC count, GC rate, and several of the socioeconomic census variables (Figures 1, 2). Although cases exist throughout the greater Baltimore area, only census tracts within Baltimore City County are analyzed in this study.

Descriptive Statistics of Census Tracts (n=200)

Effect	Mean	St. Dev.	Median
% w/ College Degree	17.9	17.2	11.1
% w/ High School Deg.	66.2	13.3	67.8
Home Value	71,514.21	44,400.91	62,600
Median Age	35.0	5.0	35.3
Per Capita Income	16,872.34	9,007.96	14,419
% Black	63.5	35.9	79.7
% Below Poverty Line	24.6	13.7	22.0
% White	31.7	33.3	16.3
% Foreign Born	4.3	4.4	3.3

Table 2: Data and summary statistics taken from American Fact Finder - US Census Bureau.

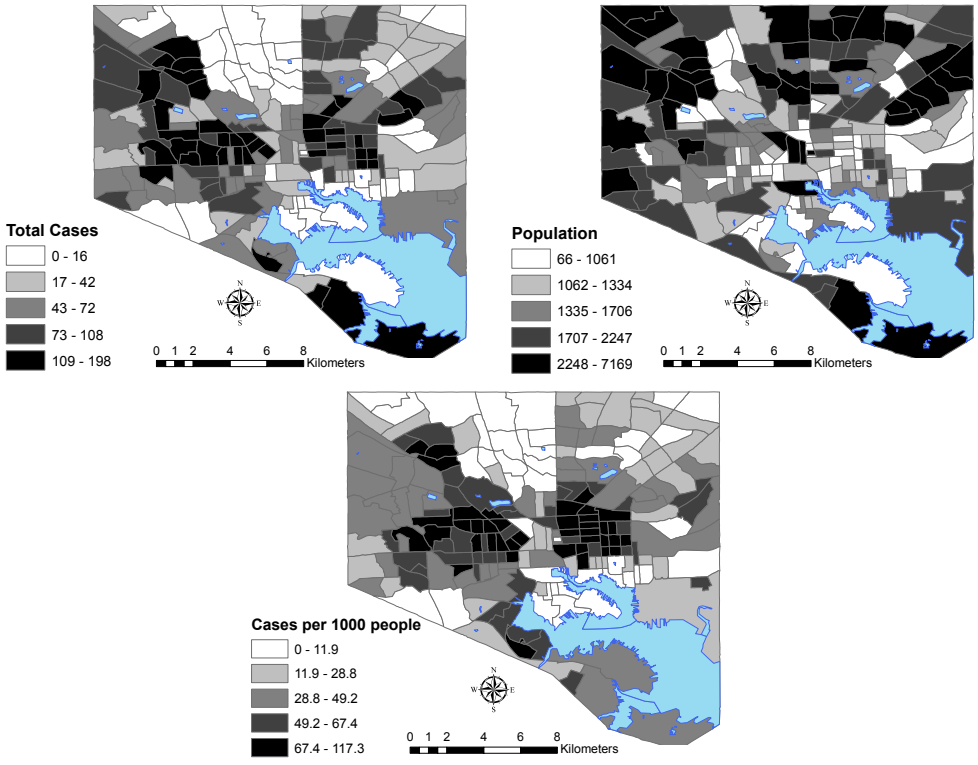

Figure 1: Maps of total GC cases, total population at risk, and GC case rate per 1,000 individuals at risk from 2002-05.

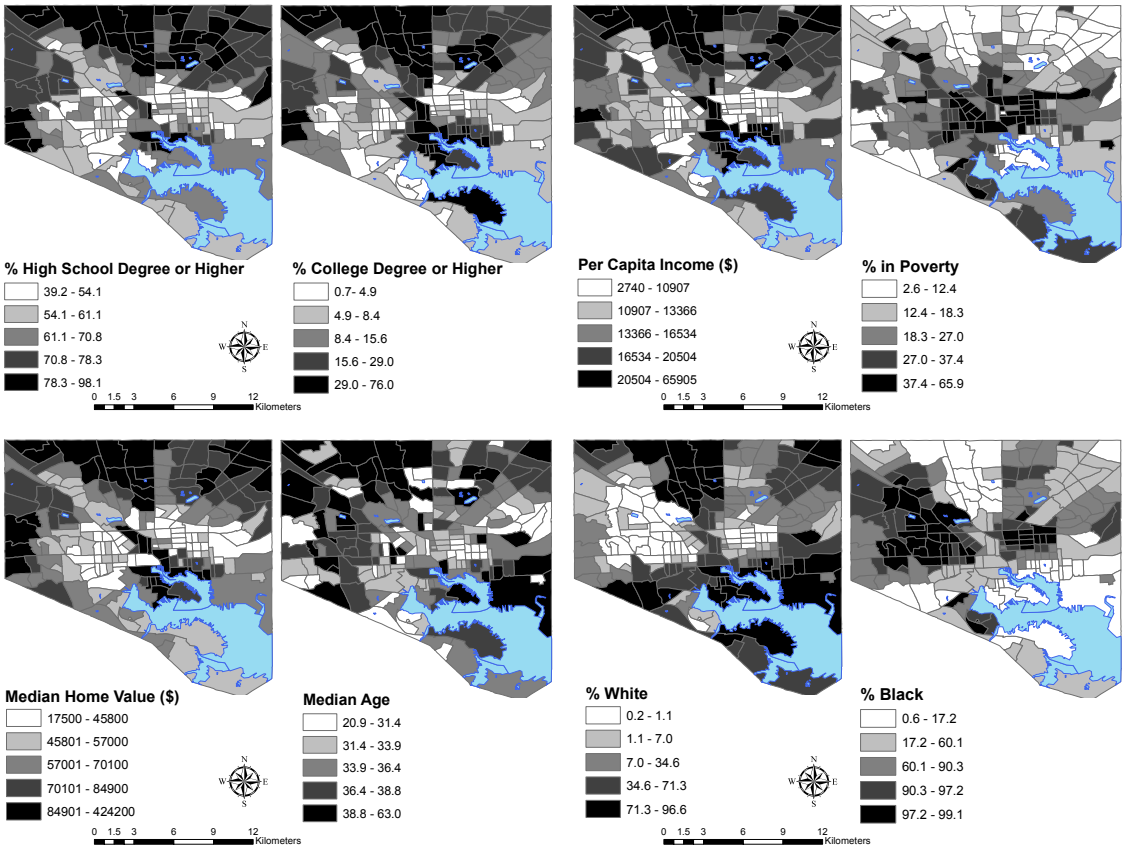

Figure 2: Maps of Baltimore demographics generated in ArcMap using data from US Census Bureau.

We first analyzed the data for spatial associations, and to measure global and local spatial autocorrelation in an effort to find core areas of disease transmission. Using ClusterSeer [49], it was possible to obtain a Global Moran's I statistic for GC case rates in Baltimore. Moran's I coefficient of autocorrelation quantifies the similarity of an outcome variable observed in one area with values observed in neighboring areas. It is defined as follows:

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2}$$

where N is the number of spatial units indexed by i and j, X is the variable of interest (disease rates, for example), $\bar{X}$ is the mean of X, and w_{ij} is an element of a matrix of spatial weights. The spatial weight for a given element w_{ij} is given a value of 1 if

two regions are neighbors, and a value of 0 if not neighbors. Under a null hypothesis of independent observations, the expected value of Moran's I is $\frac{-1}{N-1}$. A Moran's I value of zero indicates no clustering, while a positive Moran's I indicates a positive spatial autocorrelation or a clustering of areas of similar attribute values. Thus, values far from 0 indicate disease rates are not spatially independent. It is then possible to test the significance of the Moran's I value using Monte Carlo randomization where $p < 0.05$ yields a significant Moran's I value.

Additionally, using GeoDa [4], Local Moran's I values were calculated for GC case rates, where neighbors are once again defined by contiguity, and the local values were then used to generate maps showing local clustering. The Local Moran test detects local spatial autocorrelation in group-level data, termed Local Indicators of Spatial Association or LISAs. LISAs sum to the global indices of spatial association, which yield one statistic over the entire study space. LISA is defined as follows:

$$I_{i,std} = \frac{Y_i - \bar{Y}}{s} \sum_{j=1}^{N} w_{ij} \frac{Y_j - \bar{Y}}{s}.$$

Sub i denotes the estimation for each i region, s is the standard deviation, $\bar{Y}$ is the average value for the study area, and w_{ij} is equal to 1 if regions i and j share a boundary and 0 otherwise. The global Moran's I statistic is equal to $\sum_i \frac{I_i}{N}$, where N is the number of observations. High-high areas indicate high rate areas surrounded by other high rate areas, while low-low areas indicate low rate areas surrounded by other low rate areas.

A Global Moran's I value of 0.588 (p=0.002) was obtained for the GC rate data. As a result of the p-value from the Monte Carlo simulations in ClusterSeer [49], we conclude significant positive global spatial autocorrelation in the GC rate spatial data. From the maps showing local clustering mapped across Baltimore City County, it is apparent that there are clear high-high clustering areas near downtown Baltimore and substantial low-low rate clustering areas around the outskirts of the city (Figure 3). This seems to indicate clear distinctions in local clustering, indicative of

tracts of similar disease rates around most observations. Both the Global Moran's I generated from ClusterSeer and the Local Moran test from GeoDa indicate high spatial autocorrelation at both the global and local level.

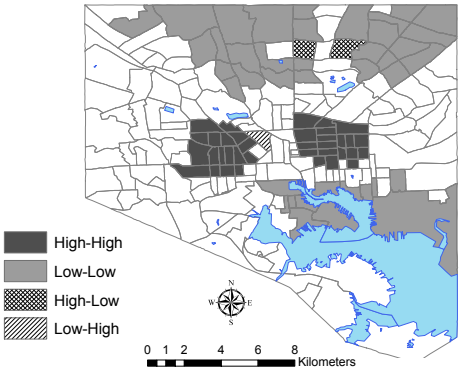

Figure 3: The Local Test for spatial autocorrelation is shown above. The black color indicates significant high-high areas (high rate areas surrounded by other high rate areas), light gray indicates low-low areas, crosshatch indicates high-low areas, and solid stripes indicate low-high areas.

Additionally, we applied spatial scan statistics via SaTScan to detect significant clusters of disease transmission in Baltimore, using a discrete Poisson model with varying cluster sizes [38]. SaTScan provides an objective approach in defining high and low prevalence cutoff points, and allows for adjustment for binary covariates. SaTScan identifies the most likely clusters from a set of circular potential clusters centered at each tract centroid. Under the null hypothesis of the Poisson model, the expected number of cases in each region is proportional to the population size. The likelihood under the assumption of the Poisson model is as follows:

$$\left(\frac{c}{E[c]}\right)^{c}\left(\frac{C-c}{C-E[c]}\right)^{C-c} I(),$$

where C is the total number of cases, c is the observed number of cases within the window, $E[c]$ is the expected number of cases within the window under the null hypothesis, and $I()$ is the indicator function which is equal to 1 if $c > E[c]$ or 0 otherwise.

As we are only interested in finding clusters of high rates, $I()$ should be set equal to 1 [54]. We first applied SaTScan using standard default settings of the largest potential clusters comprised of 50% of the population at risk. In order to assess any variation in the local likelihood ratio within this cluster, we consider limiting the maximum spatial cluster size to a smaller fraction of the study population: 25% of the population at risk. We also adjusted the results using a binary race variable which was dichotomized based on high and low rates of white residents (above and below the median value of 16.3% white). This work was done in part to compare results to previous studies analyzing relationships between demographic characteristics and sexually transmitted disease infection rates using SaTScan, which adjust for race as well [35].

The results shown in Figure 4 tend to mirror clusters found by the Local Moran test (Figure 3) and past studies of GC transmission in Baltimore [7, 35, 57]. When allowing the cluster size to approach 50% of the population at risk, two large clusters appear, which also occur when refining the maximum allowable cluster size to 25% of the population at risk (i.e. focusing on geographically smaller clusters). One can distinguish two very significant clusters in each situation, as is the case in the Jennings study [35]. Nevertheless, an adjustment for race at the smaller allowable cluster size provides little change in cluster location or size in contrast to the formerly mentioned GC study.

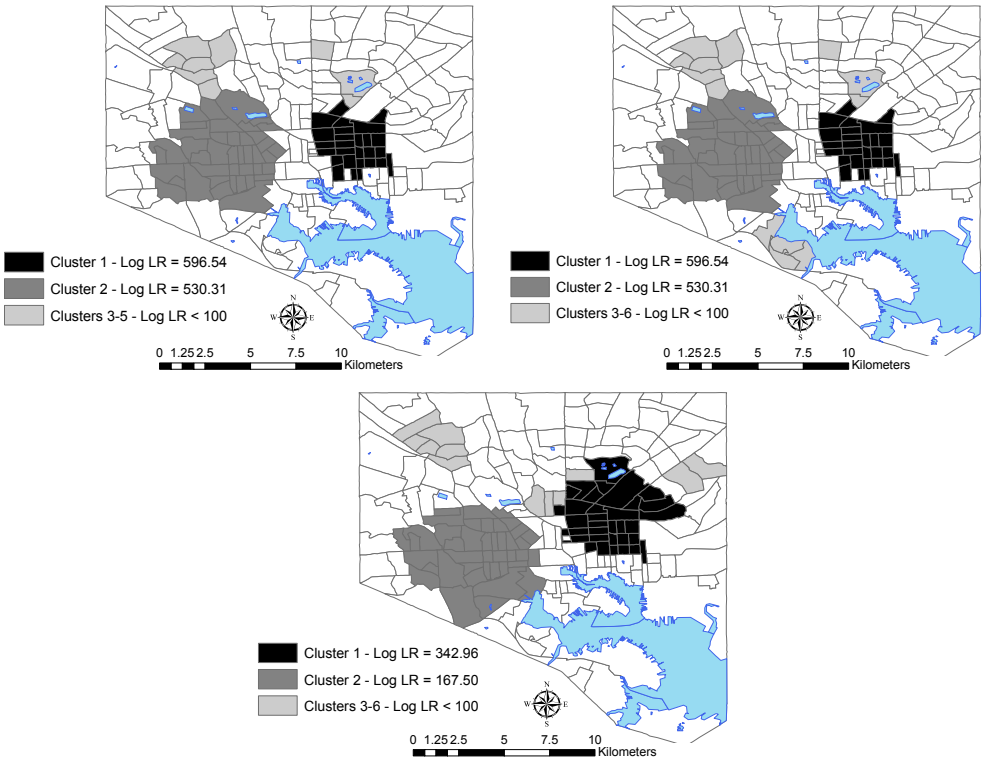

Figure 4: Statistically significant most likely clusters based on a spatial scan statistic for circular clusters ranging up to 50% of the population at risk (upper left), 25% of the population (upper right), and 25% of the population at risk adjusting for race (lower left).

4.2 Methods and Model Descriptions: A Geographically Weighted Regression Approach

Our next step assessed associations with tract-level demographics via a linear regression model. In this case, the response variable is the GC rate per 1,000 individuals at risk in the population, and we considered a linear model of the following form:

$$y_i = \beta_0 + \beta_1 * X_{1i} + \beta_2 * X_{2i} + ... + \epsilon_i \qquad \text{where } i = 1, ..., 200$$

where the coefficients β assess the magnitude of association between disease rates and demographic variables over the 200 census tracts. A univariate analysis in SAS [34] determined significant risk factors at the $\alpha = 0.05$ level of significance, and a final

model was defined using a Wald-type test for adding and removing variables using a 0.05 cut-off point. R^2 values assess the overall goodness-of-fit for the model. In linear regression (and other forms of regression), observations are assumed to be independent, and associations between observations and covariates are assumed to be constant. These assumptions could potentially be violated with spatially referenced data [53].

Next, we consider spatial variation in the observed associations with significant predictors. In this situation, geographically weighted regression analyzes the spatial effects of significant factors in predicting GC rates. Geographically weighted regression (GWR) is a technique for exploratory data analysis, which allows the relationships of interest to vary over space, i.e. the parameter coefficients need not be the same everywhere. With GWR, instead of assuming fixed global parameter estimates, estimates can now vary according to a position in space, characterized by latitudinal and longitudinal coordinates.

Under the typical linear regression framework, maximum likelihood methods are used to estimate model parameters using the standard linear model:

$$y_i = \beta_0 + \beta_1 * X_{1i} + \beta_2 * X_{2i} + ... + \epsilon_i \qquad \text{where } i = 1, ..., 200$$

In contrast, GWR involves estimating regression coefficients locally by weighted least squares where we weight data based on distance from each of a given number of estimation locations [53]. The weighting function used by GWR typically takes the form: $W_j = \exp\left(\frac{-d_j}{b^2}\right)$, where d_j is the distance from the point i at which the regression model is being fitted, and b is the bandwidth defining the smoothness of the parameter surfaces. The bandwidth is calibrated by a cross-validation technique to minimize the score $\frac{y_i - y_i(b)}{2}$ where $y_i(b)$ is the fitted value of y_i using the bandwidth b and the weighted regression model centered at the point i. What follows is a regression model where the coefficients are specific to a location i:

$$y_i = \beta_{i0} + \beta_{i1} * X_{1i} + \beta_{i2} * X_{2i} + ... + \epsilon_i \qquad \text{where } i = 1, ..., 200$$

We applied GWR in **R** [45] using the spgwr package, and the kernel bandwidth was fixed at values of 0.01 and 0.02. The bandwidth will determine the level of smoothing in the surface of estimates, where a larger bandwidth indicates more smoothing or weighting the local estimates closer to the overall global mean. The predictors were the same as those of the final global multivariate linear regression model. Maps of predictor variables with local β estimates, adjusting for other variables in the model, were reported for each bandwidth type.

As counts and rates are often modeled via Poisson regression, we also fit a GWR Poisson model using the spgwr package in **R** [45]. We let y_i denote the number of cases in each census tract where $i = 1, ..., 200$, and we assume the cases follow independent Poisson distributions with the tract-specific mean $E_i \exp(\mu_i)$. E_i represents the number of cases expected under a null model where every individual is equally likely to become infected, and it is fixed and proportional to the population size in tract i, denoted n_i. The observed GC rate over the study population, R, is equal to $\frac{E_i}{n_i}$, with $\exp(\mu_i)$ representing the local relative risk due to local covariates [53]. Using the same local covariates as the linear GWR model, we create the following model with a population offset:

$$\mathrm{E}(y_i) = \exp[\ln(n_i) + \beta_{i0} + \beta_{i1} * X_{1i} + \beta_{i2} * X_{2i} + \beta_{i3} * X_{3i}] \qquad \text{where } i = 1, ..., 200$$

Offsets typically appear in Poisson models since the outcome is assumed to occur at an underlying fixed rate per individual. Thus, by including an offset, the covariate terms are defined based on the impact of an underlying rate as opposed to the underlying counts. As with the GWR linear model, the coefficients are allowed to vary spatially. However, unlike the linear framework, estimation requires more complex computation using Taylor series and iteratively reweighted least squares due to the non-linear nature of the Poisson model [53]. For data which is Poisson-distributed, the variance will depend on the mean, and as a result, the weighted least squares equations implemented in GWR will be weighted by a diagonal variance matrix based

on the variance function for the data [53]. Several fixed bandwidths were chosen for analysis in the Poisson model in order to display the effect of bandwidth on spatial variability. Maps generated through the linear GWR and Poisson GWR techniques provide quick, descriptive results relaying the smoothed general differences in association.

In summary, we present three types of models for analysis, a non-spatial linear model for determining socio-demographic characteristics most highly associated with high/low GC rates, a spatial linear GWR model assessing the additive effect of each characteristic on GC rate, and a spatial Poisson model assessing the multiplicative effect of each characteristic on infection count. Each provides a unique interpretation on how case rates vary over population characteristics, and the latter two on how those associations may vary spatially.

4.3 Application to Baltimore STI Data

In the univariate linear regression analysis, every census tract socioeconomic demographic variable considered was found to be strongly associated with GC rate. The percent with a college degree ($p < 0.0001$), percent with a high school degree ($p < 0.0001$), median home value ($p < 0.0001$), median age ($p < 0.0001$), per capita income ($p < 0.0001$), percent black ($p < 0.0001$), percent white ($p < 0.0001$), percent below the poverty line ($p < 0.0001$), and the percent foreign born ($p < 0.0001$) were all strong individual predictors of STI risk (Table 3).

Effect	Estimate	95% CI	p-value
% w/ College Degree	-1.043	(-1.219, -0.869)	<0.0001
% w/ High School Deg.	-1.330	(-1.559, -1.100)	<0.0001
Home Value (per $1000)	-0.335	(-0.410, -0.261)	<0.0001
Median Age	-1.561	(-2.317, -0.804)	<0.0001
Per Capita Income (per $1000)	-1.956	(-2.294, -1.619)	<0.0001
% Black	0.602	(0.533, 0.672)	<0.0001
% Below Poverty Line	1.374	(1.163, 1.585)	<0.0001
% White	-0.650	(-0.725, -0.575)	<0.0001
% Foreign Born	-3.025	(-3.812, -2.239)	<0.0001

Table 3: The linear regression univariate analysis of demographic characteristics in relation to GC rate in Baltimore. Significance assessed at 0.05 level.

In order to determine the factors most highly correlated with GC rate, one would need to observe the magnitude of the parameter estimates and p-values generated for the corresponding demographic data. For example, the percent below the poverty line is influential in predicting higher GC rates in the multivariate model (Table 4). A 1% increase in the percent of the population which is below the poverty line results in a 0.388 increase in GC rate per 1,000 individuals at risk in the population. Other factors in the multivariate model, such as percent black and percent with a high school degree or higher, are also significant.

Effect	Estimate	95% CI	p-value
% Black	0.460	(0.403, 0.516)	<0.0001
% Below Poverty Line	0.388	(0.198, 0.578)	<0.0001
% w/ High School Degree	-0.655	(0.020, -0.471)	<0.0001

Table 4: The multivariate linear regression analysis of demographic characteristics in relation to infections in Baltimore. Note, % black, % below the poverty line, and % with a high school degree or higher represent the three most significant covariates in the multivariate setting, controlling for the other variables. Significance assessed at 0.05 level

Next, in an effort to assess spatial variation in the parameter estimates of the multivariate model, we perform a geographically weighted regression. It appears that the associations between GC rates and education level are strongest in the outskirts

of the city, where there exists a stronger negative association between disease rate and education level in western and eastern Baltimore, and a weaker effect towards the central parts of the city (Figure 5). However, the local estimates for the percent in poverty display stronger associations in the central parts of the city. Higher levels of poverty result in higher GC rates in central Baltimore. The percent black variable seems to have a stronger effect in the central and western parts of the city.

In the results from the Poisson GWR model, there appears to be stronger spatial variability (Figure 6), although the choice of bandwidth is important in determining the level of variation. A lower bandwidth will display areas of stronger parameter estimates in individual census tracts and will account for local variation more accurately, but does not capture a specific spatial trend across Baltimore as well as higher bandwidths. Higher bandwidths show clear, smooth spatial trends, but at the cost of regressing individual tract parameter estimates closer to the non-spatial global values. From the maps in Figure 6, it appears that tract-level education level and percent in poverty have similar patterns to those in the GWR linear regression models. Stronger estimates in the race variable occur in the northern and western regions of Baltimore, in contrast to the stronger effects found in southern and western Baltimore in the linear GWR model. Both models present strong cases for allowing the effects to vary spatially, and the choice of bandwidth will affect the visual interpretation of the spatial effect on the given covariates.

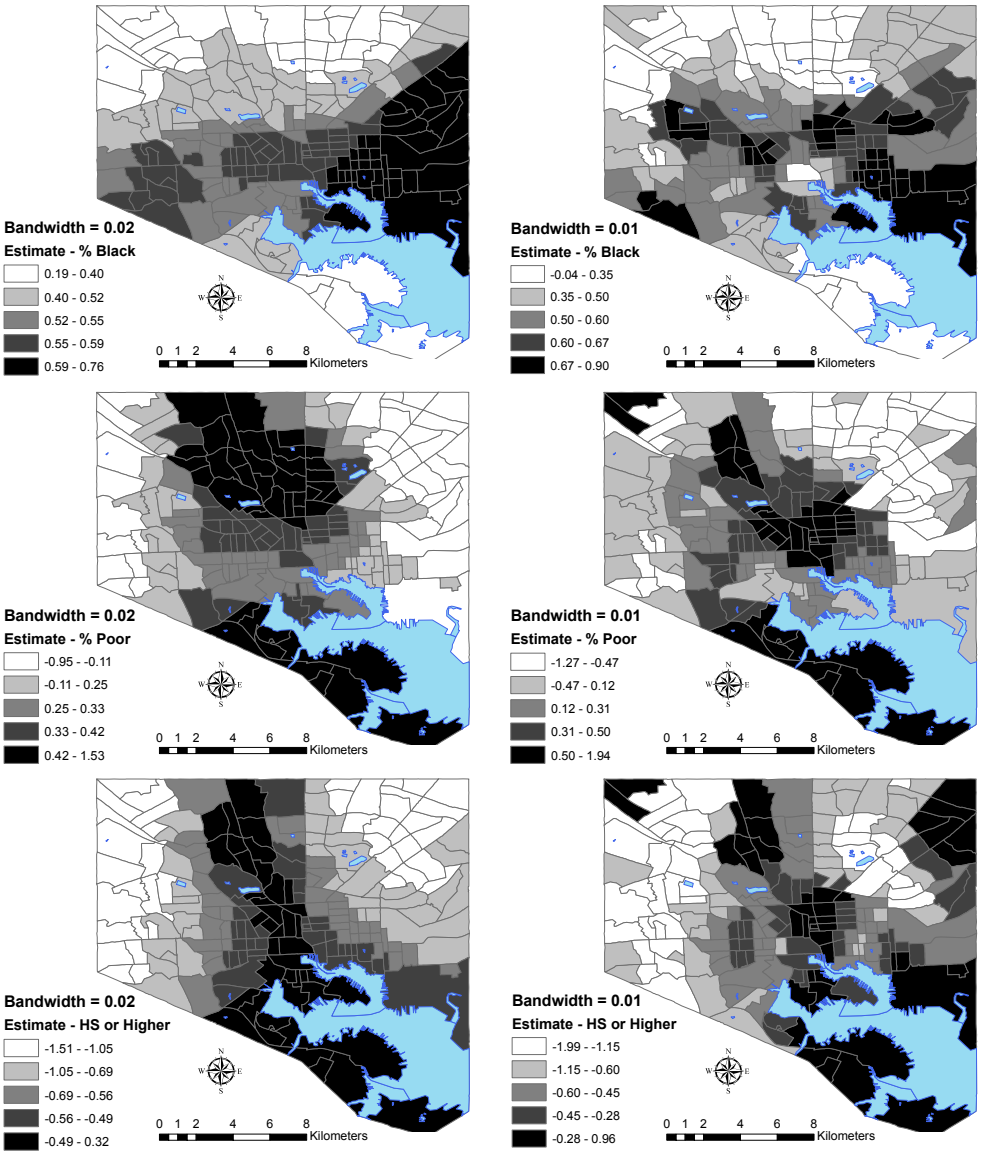

Figure 5: The linear geographically weighted regression maps are shown above. Local estimates for % black with a bandwidth of 0.02 vs. 0.01 (top row), % below the poverty line with a bandwidth of 0.02 vs. 0.01 (middle row), and % with a high school degree or higher with a bandwith of 0.02 vs. 0.01 (bottom row) are calculated with a fixed kernel bandwidth.

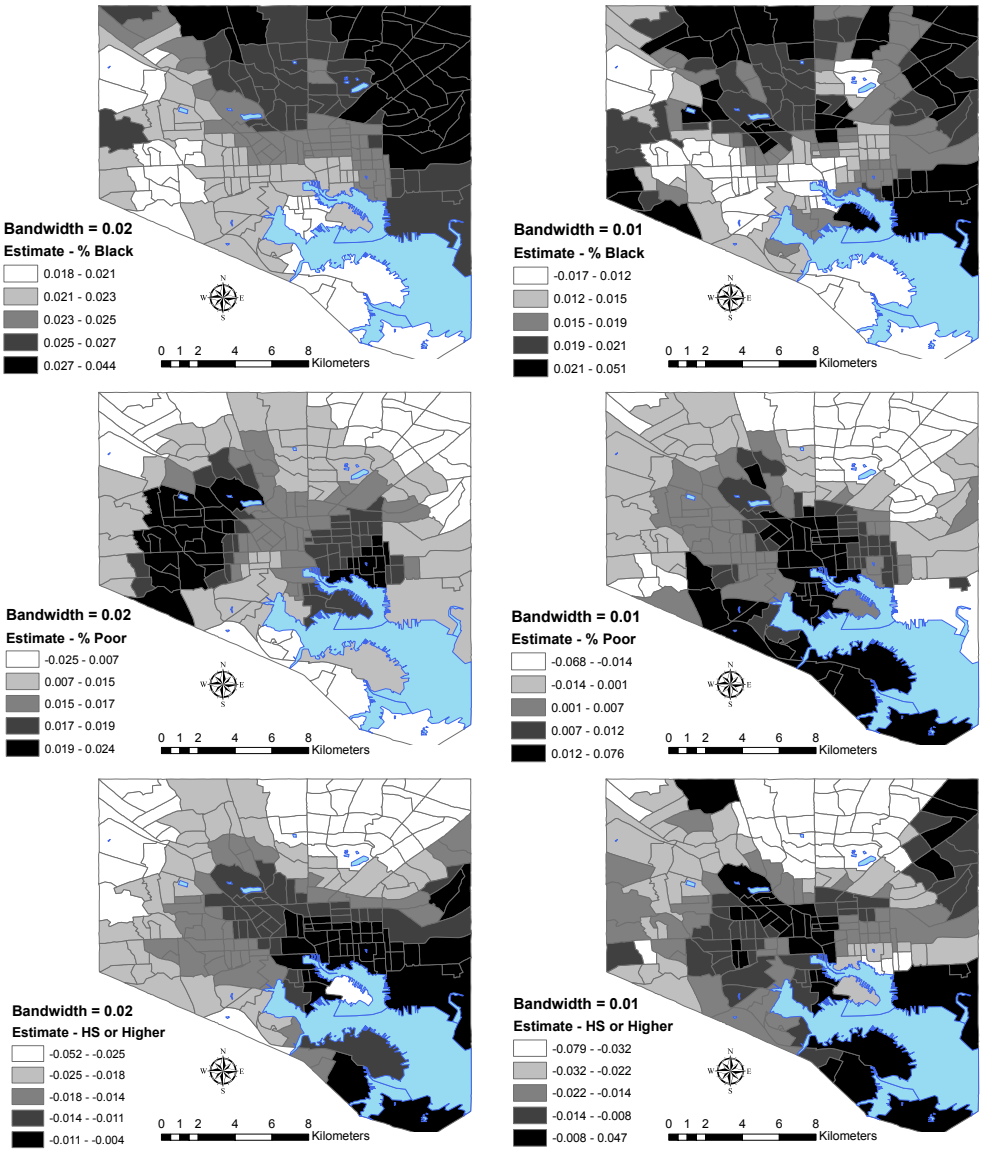

Figure 6: The Poisson geographically weighted regression maps are shown above. Local estimates for % black with a bandwidth of 0.02 vs. 0.01 (top row), % below the poverty line with a bandwidth of 0.02 vs. 0.01 (middle row), and % with a high school degree or higher with a bandwith of 0.02 vs. 0.01 (bottom row) are calculated with a fixed kernel bandwidth.

These results rely on several statistical modeling techniques to assess the demographic variables most highly associated with disease transmission, the spatial variation of GC cases throughout Baltimore, and the spatial association between the significant disease predictors and GC rates. From the linear regression modeling

approach, it is apparent that socioeconomic and demographic factors are highly associated with high GC rates in the Baltimore area from 2002-05. Among the strongest relationships include educational factors such as the percent with a high school or a college degree and economic factors such as median home value, per capita income, and the percent that live below the poverty line. High rate areas for GC exist in central and western Baltimore City County in predominately poorer, African-American communities, with fewer educational credentials.

Two tight spatial clusters in Baltimore were found in both the demographic GC case analysis as well as the SaTScan images, which correspond to similar core areas and clusters from the Jennings GC study [35]. The SaTScan results agree with the Local Moran test in revealing the two clusters of disease risk within central Baltimore. Since local tests seek individual units of high risk surrounded by other high risk units, the smaller, localized clusters are visually displayed. The socioeconomic variables such as the percent in poverty and education level were relatively heterogeneous across space, while the effects of race were also shown to vary spatially, which would perhaps indicate that the geographically weighted regression model describes the data better than a global model. Areas of higher associations with tract-level percent black are at or near boundaries of transition between predominately white and predominately black census tracts in Baltimore, suggesting that there may be an interaction between race and spatial GC rates.

The areas indicated by GWR, SaTScan, and the Local Moran test all focus attention on two areas in central Baltimore, areas previously identified as core areas. However, none of these methods directly assess a hypothesis regarding core areas and none provide local estimates of the risk of transmission, motivating our model development by adding a spatial dimension to the SIR techniques described earlier.

5 Extending the SIR Model to Spatial Analysis

Previous research has addressed the challenges in estimating R_0 for a variety of infectious diseases, and we will build off those methods to estimate R_0 spatially for sexually transmitted infections, specifically. As noted in earlier sections, R_0 is the average number of susceptible individuals infected by one infectious person in a completely susceptible population. Since our data contain geographic locations, it will be possible not only to estimate the reproduction number and its component parameters over time, but also over space.

We begin by building off of the same parameters established in the Kermack-McKendrick model [37]. As noted, the Kermack-McKendrick model is an epidemiological model that computes the theoretical number of people infected with a contagious illness in a closed population over time. Differential equations form the foundation of the SIR (susceptible-infected-recovered) model, as rates of change for each class are tracked over time. Additionally, the dynamics of the outbreak are governed by two parameters in this basic model: β, the transmission parameter, and γ, the recovery parameter. As a reminder, the SIR differential equations are defined as follows:

Let:

X(t) = susceptible population size at time t,

Y(t) = infected population size,

Z(t) = removed/recovered population size, and

N = X + Y + Z = Total population size (fixed).

The dynamics of the system are defined via:

$$\frac{dX}{dt} = \frac{-\beta XY}{N},$$

$$\frac{dY}{dt} = \frac{\beta XY}{N} - \gamma Y,$$

$$\frac{dZ}{dt} = \gamma Y, \text{ and}$$

where $X(0) = X_0 \approx N$, $Y(0) = N - X_0 \approx 0$, $Z(0) = 0$.

Our goal is a spatially-varying estimate of the reproduction number to determine "core areas" of disease since R_0 signals how quickly the disease can spread amongst a susceptible population and is directly related to the amount of control effort needed to eliminate an infection from a population. Our spatial analysis enables further inference addressing the dynamics of outbreaks initiated at different locations, i.e. "How quickly will the outbreak spread if it starts here?" Other methods have also been considered in STI core area analysis [1, 16, 39, 47, 50]. The following Bayesian models represent attempts to visualize the spatial variability of transmission and recovery pertaining to STIs in Baltimore, in addition to quantifying the extent and reliability of core areas.

5.1 Chain Binomial: A Transitional Approach

The Kermack-McKendrick model provides the first look at the types of parameters we wish to estimate spatially. The next step involves developing a statistical model which can give us a likelihood to use for parameter estimation purposes. We focus primarily on the use of chain binomial models.

Chain binomial models are stochastic models, which assume a binomially distributed number of infectious or susceptible individuals. Infection can spread from individual to individual in discrete time units, and the numbers of individuals who are susceptible and infectious are assumed to be known at each time point. An SIR model typically assumes no latency period, and can be used to track the progression of person-to-person infectious diseases [30].

In our analysis, we extend the basic transition chain binomial model described by Lekone and Finkenstädt [40], which assumes that the infection time series is a chain of binomially distributed random variables. In their previous work, the numbers of newly infectious individuals and newly recovered individuals (transition compartments) are binomially distributed. The transitions of individuals from one stage of disease to the next are stochastic movements between the corresponding compartments over time. If we assume an exponentially distributed length of time that an individual spends in each compartment, then the probability of leaving is $1 - \exp(-\lambda(j))$. In our spatial analysis, we let NI_{ij} denote the number of susceptible individuals who become infectious at time j and are located in census tract i. Likewise, let NR_{ij} denote the number of cases who are removed from the infectious class at time j in tract i. Using a discrete-time approximation to a continuous SIR model, we expand upon models defined previously [40] by adding a spatial component to our model. We also define individuals as susceptible, infectious, or removed from the population as follows:

$S_{ij} = i \times j$ matrix of those susceptible in tract i at time point j,

$I_{ij} = i \times j$ matrix of those infectious in tract i at time point j,

$R_{ij} = i \times j$ matrix of those removed in tract i at time point j,

$S_{i,j+1} = S_{ij} - NI_{ij}$,

$I_{i,j+1} = I_{ij} + NI_{ij} - NR_{ij}$, and

$R_{i,j+1} = R_{ij} + NR_{ij}$.

NI_{ij} and NR_{ij} are random variables with binomial distributions:

$$NI_{ij} \sim \text{Bin}(S_{ij}, p_{ij}), \text{ and}$$
$$NR_{ij} \sim \text{Bin}(I_{ij}, p_R),$$

$$\text{where } p_{ij} = 1 - \exp\left(\frac{-\beta_i}{N_i} I_{ij}\right),$$
$$\text{and } p_R = 1 - \exp(-\gamma)$$

As with Lekone's and Finkenstädt's model, we follow the spread of disease over time; however, we also track the location of cases. Here, p_{ij} represents the probability of a susceptible individual becoming infectious at the given timepoint j in tract i. p_R represents the probability that an infectious individual is removed from the study population. We assume this value is independent of time and location.

We begin to explore the idea of extending the epidemic models which cover parameter estimation over time to encompass estimation over space. Given that we possess

coordinates and census tract information for each individual case in the study, it is relevant to investigate spatial correlation between tract-specific estimates of our epidemic parameters. The difficulty in obtaining reliable local estimates of disease transmission has propelled our research towards models which offer both global and local smoothing techniques coupled with adequate geographic resolution for our estimates.

5.1.1 Random effects - Exchangeable, Conditionally Autoregressive, and Convolution Structures

We propose the use of Bayesian methods - a hierarchical approach to induce positive spatial autocorrelation across the estimated local disease transmission parameters, as described by Waller and Carlin [52] through a conditionally autoregressive (CAR) random effects distribution assigned to area-specific intercepts [9]. The CAR model has been extended to a fully Bayesian setting and implemented using MCMC algorithms [11].

The conditionally autoregressive random effects model proposed by Besag, York, and Mollié [11] induces spatial autocorrelation amongst the individual levels of disease risk specified by the observed counts of disease cases and number of individuals in each region. If Y_i is the observed count of disease cases in region $i = 1, ...I$, Besag, York, and Mollié model the counts as Poisson random variables using a log link function. Additional data include the number of individuals at risk in a given region n_i and the local number of "expected" cases under some null model of disease transmission (constant risk for all individuals). We also can assume the n_i values are fixed and known.

In contrast to previous studies regarding the mapping of disease rates using a Poisson model [52], we first implement a Bayesian logistic model to estimate the

probability of becoming infectious. This procedure has been implemented previously in studies mapping the prevalence of schistosomiasis in Tanzania in 2008 [19]. In that study, the number of individuals Y_i having reported experiencing schistosomiasis in ward i was binomially distributed based on n_i the number of individuals questioned/interviewed in ward i and p_i the prevalence of experiencing schistosomiasis in ward i. Both spatial random effects and exchangeable random effects were implemented in their model. Similarly, we adjust this model to complement the transition chain binomial model for disease transmission, and we first consider an exchangeable random effects model where we add the following components:

$$\begin{aligned} NI_{ij}|v_i &\overset{ind}{\sim} \text{Bin}(S_{ij}, p_{ij}), \\ \text{Logit}(p_{ij}) &= \alpha_0 + v_i, \\ \text{where } v_i &\sim N(0, \sigma_v^2), \text{ for } i = 1, \ldots, I. \end{aligned}$$

The structure described above allows us to build the overall distribution of NI_{ij} in two steps. First, the observations NI_{ij} are conditionally independent given the values of the random effects v_i. Second, correlation is induced in the marginal distribution of the NI_{ij}s. The exchangeable random effects are linked to the probability of infection using a logit link, as the newly infectious individuals are binomially distributed. We assign a vague normal prior to the fixed effect α_0, as well as a vague gamma hyper-prior distribution to τ_v (where ${\sigma_v}^2 = \frac{1}{{\tau_v}^2}$). This approach will provide a local estimate defined by the weighted average of the observed data in tract i and the global overall mean.

Although the previous model induces correlation between local estimates based on the overall global mean, we have not yet introduced spatial correlation among the observations. Clayton and Kaldor [17] introduced the concept of replacing exchangeable priors with a spatially structured prior distribution, where local estimates

borrow strength from neighboring regions rather than the entire study space. Using a conditionally autoregressive set of random effects, we introduce a spatially structured prior distribution where local estimates of disease transmission are a weighted average of the local estimates and the neighboring estimates [52] such that:

$$\begin{aligned} NI_{ij}|u_i &\overset{ind}{\sim} \text{Bin}(S_{ij}, p_{ij}), \\ \text{Logit}(p_{ij}) &= \alpha_0 + u_i, \\ \text{where } \mathbf{u} &\sim MVN(\mathbf{0}, \Sigma_u). \end{aligned}$$

In addition to previously defined parameters, we introduce Σ_u, which denotes a spatial covariance matrix, and a vector $\mathbf{u} = (u_1, ...u_I)$ of spatially correlated random effects. The vector $\mathbf{u}$ can be implemented as a joint prior or as conditionally autoregressive, a collection of conditional distributions. The value of a given u_i is conditioned on the random effects of the neighboring regions. We impose the restriction $\sum_{i=0}^{I} u_i = 0$ at each iteration. The spatial effects are linked to the probability of infection using a logit link, as the number of newly infectious individuals follows a binomial distribution. The parameter τ_{CAR} denotes a hyperparameter related to the conditional variance of u_i given the values of the other elements of $\mathbf{u}$ [52]. Once again, a vague normal prior is placed on α_0, and a vague gamma hyperprior is placed on τ_{CAR}. Here, we have a collection of conditional Gaussian priors for each u_i wherein the prior mean is a weighted average of the other u_k, $i \neq k$:

$$u_i|u_{k\neq i} \sim N\left(\frac{\sum_{j\neq i} c_{ik}u_k}{\sum_{k\neq i} c_{ik}},\ \frac{1}{\tau_{CAR}\sum_{k\neq i} c_{ik}}\right),\ i = 1, \ldots, I.$$

This is an implementation of the joint multivariate normal distribution by a collection of conditionals. In order to determine the distributions of the random effects, the set of spatial random effects was modeled using an adjacency matrix, where a weight of one was given to neighboring tracts, and a weight of zero was given to tracts which

did not border one another. According to Besag and Kooperberg [10], to define the connection between the autoregressive spatial dependence parameters c_{ik} and the joint spatial covariance matrix σ_u, if $\mathbf{c}$ follows a multivariate Gaussian distribution with covariance σ_u, then the density $f(\mathbf{u})$, takes the form:

$$f(\mathbf{u}) \propto \exp\left(-\frac{1}{2}\mathbf{u}'\Sigma_u^{-1}\mathbf{u}\right).$$

Standard multivariate Gaussian theory defines the associated conditional distributions as

$$u_i|u_{k\neq i} \sim N\left(\sum_{k\neq i}\left(\frac{-\Sigma_{u,ik}^{-1}}{\Sigma_{u,ii}^{-1}}\right)u_k\ ,\ \frac{1}{\left(\Sigma_{u,ii}^{-1}\right)}\right),$$

where $\Sigma_{u,ik}^{-1}$ denotes the (i,k)th element of the precision matrix Σ_u^{-1}. Note the conditional mean for u_i is a weighted sum of $u_k, k \neq i$, and the conditional variance is inversely proportional to the diagonal of the inverse of Σ_u, just as it is in the CAR specification above.

Besag et al. [11] notes that we could include both global and local borrowing of information within one model using a convolution prior. This would incorporate both exchangeable random effects and conditionally autoregressive random effects for each tract, such that:

$$\begin{aligned}
NI_{ij}|u_i, v_i &\overset{ind}{\sim} \text{Bin}(S_{ij}, p_{ij}),\\
\text{Logit}(p_{ij}) &= \alpha_0 + u_i + v_i,\\
\text{where } u_i|u_{k\neq i} &\sim N\left(\frac{\sum_{j\neq i} c_{ik}u_k}{\sum_{k\neq i} c_{ik}}\ ,\ \frac{1}{\tau_{CAR}\sum_{k\neq i} c_{ik}}\right),\\
\text{and } v_i &\sim N(0, \sigma_v^2),\ \text{for } i = 1, \ldots, I.
\end{aligned}$$

As with the previous models, we assign hyperpriors to the hyperparameters τ_{CAR} and σ_v^2, typically gamma hyperpriors.

5.2 Transmission Estimation in Chain Binomial Models

With the previous model, we have defined a process for estimating the probability of transmission for each susceptible individual at a given time point j and census tract i. The model in 5.1 adds spatial correlation to p_{ij} through the logit link. Here, we prefer to add spatial correlation to β_{ij}. The next step involves converting the estimate for p_{ij} into an estimate for the transmission parameter for each tract β_{ij}. With estimates for β_{ij}, we can then assume removal times for each case occur after one time unit (one month) for a fixed period of recovery. Thus, an estimate for β_{ij} provides the necessary information for an estimate for R_{0i}.

We set the number of susceptible individuals in each tract S_{i0} equal to the number initially at risk in the population in a given tract - i.e. the number of individuals aged 15-49. From the initial chain binomial model, we assume $p_{ij} = 1 - \exp\left(\frac{-\beta_{ij}}{N_i} I_{ij}\right)$; however, we substitute NI_{ij} for I_{ij} since we now assume individuals recover after one time unit. We now estimate β_{ij} using the following exchangeable model, and we note that $\beta_{ij} = \beta_i$ for all j:

$$\begin{aligned} NI_{ij}|v_i &\overset{ind}{\sim} \text{Bin}(S_{ij}, p_{ij}), \\ p_{ij} &= 1 - \exp\left(\frac{-\beta_{ij}}{N_i} NI_{ij}\right), \\ \beta_{ij} &= \beta_0 + v_i, \\ \text{where } v_i &\sim N(0, \sigma_v^2), \text{ for } i = 1, \ldots, I. \end{aligned}$$

We assign a vague gamma prior to the fixed effect β_0, as well as a vague gamma hyperprior distribution to τ_v (where ${\sigma_v}^2 = \frac{1}{{\tau_v}^2}$). Likewise, we can similarly define a

conditionally autoregressive model as follows:

$$\begin{aligned} NI_{ij}|u_i &\overset{ind}{\sim} \text{Bin}(S_{ij}, p_{ij}), \\ p_{ij} &= 1 - \exp\left(\frac{-\beta_{ij}}{N_i} NI_{ij}\right), \\ \beta_{ij} &= \beta_0 + u_i, \\ \text{where } u_i|u_{k\neq i} &\sim N\left(\frac{\sum_{j\neq i} c_{ik}u_k}{\sum_{k\neq i} c_{ik}}, \frac{1}{\tau_{CAR}\sum_{k\neq i} c_{ik}}\right). \end{aligned}$$

Once again, a vague gamma prior is placed on β_0 and a vague gamma hyperprior is place on τ_{CAR}. Finally, we can define a convolution model as follows:

$$\begin{aligned} NI_{ij}|u_i, v_i &\overset{ind}{\sim} \text{Bin}(S_{ij}, p_{ij}), \\ p_{ij} &= 1 - \exp\left(\frac{-\beta_{ij}}{N_i} NI_{ij}\right), \\ \beta_{ij} &= \beta_0 + u_i + v_i, \\ \text{where } u_i|u_{k\neq i} &\sim N\left(\frac{\sum_{j\neq i} c_{ik}u_k}{\sum_{k\neq i} c_{ik}}, \frac{1}{\tau_{CAR}\sum_{k\neq i} c_{ik}}\right), \\ \text{and } v_i &\sim N(0, \sigma_v^2), \text{ for } i = 1, \ldots, I. \end{aligned}$$

As with the previous models, we will assign hyperpriors to the hyperparameters τ_{CAR} and σ_v^2, typically conjugate gamma hyperpriors.

5.3 Reed-Frost Chain Binomial Model

As with the transition chain binomial model, the Reed-Frost model of disease transmission assumes that individuals pass through three states - susceptible, infective, and recovered. Likewise, the model assumes a fixed population size N, and each person is in one of the three states, where S_{ij}, I_{ij}, and R_{ij} are defined in the previous chain binomial model. Binomial models are often used to estimate the transmission probability, and the exposure to infection can occur in discrete time contacts, which can also be discrete time units of exposure [30]. It is typically assumed that each contact is independent of other contacts. We can define p_{ij} as the transmission probability during a contact between a susceptible individual and an infectious individual in tract i at time point j, and we can further define q_{ij} as the escape probability which is equivalent to $1-p_{ij}$. The probability that a susceptible individual escapes infection from all infectious individuals in tract i at time point j is $q_{ij}^{I_{ij}}$. Thus, the probability of not escaping infection is $1 - q_{ij}^{I_{ij}}$ or h_{ij}. This is analogous to the probability of leaving the susceptible compartment within the transition model.

In this case, R_{0i} is a function of the number of initial susceptibles in each tract S_{i0} and the transmission probability p_{ij}, such that $R_{0i} = S_{i0} * p_{ij}$. No assumptions are made concerning the recovery parameter γ, as well as the transmission parameter β. The Reed-Frost chain binomial model is set up as follows using an exchangeable

random effects model initially:

$$\begin{aligned}
I_{i,j+1}|v_i &\overset{ind}{\sim} \text{Bin}(S_{ij}, h_{ij}), \\
h_{ij} &= 1 - q_{ij}^{I_{ij}}, \\
q_{ij} &= 1 - p_{ij}, \\
\text{Logit}(p_{ij}) &= \alpha_0 + v_i, \\
\alpha_0 &\sim N(0, \tau_z), \\
\text{where } v_i &\sim N(0, \sigma_v^2), \text{ for } i = 1, \ldots, I.
\end{aligned}$$

As with the model estimating the transmission probabilities from the transition chain binomial model, we assign a vague normal prior to the fixed effect α_0, as well as a vague gamma hyperprior distribution to τ_v (where ${\sigma_v}^2 = \frac{1}{{\tau_v}^2}$) and τ_z. This approach will provide a local estimate defined by the weighted average of the observed data in tract i and the global overall mean. We next consider a CAR model where:

$$\begin{aligned}
I_{i,j+1}|u_i &\overset{ind}{\sim} \text{Bin}(S_{ij}, h_{ij}), \\
h_{ij} &= 1 - q_{ij}^{I_{ij}}, \\
q_{ij} &= 1 - p_{ij}, \\
\text{Logit}(p_{ij}) &= \alpha_0 + u_i, \\
\alpha_0 &\sim N(0, \tau_z), \\
\text{where } u_i|u_{k\neq i} &\sim N\left(\frac{\sum_{j\neq i} c_{ik} u_k}{\sum_{k\neq i} c_{ik}}, \frac{1}{\tau_{CAR} \sum_{k\neq i} c_{ik}}\right).
\end{aligned}$$

Once again, a vague normal prior is placed on α_0, and vague gamma hyperpriors are placed on τ_{CAR} and τ_z. Finally, the convolution model is defined as follows:

$$\begin{aligned}
I_{i,j+1}|u_i, v_i &\overset{ind}{\sim} \text{Bin}(S_{ij}, h_{ij}), \\
h_{ij} &= 1 - q_{ij}^{I_{ij}}, \\
q_{ij} &= 1 - p_{ij}, \\
\text{Logit}(p_{ij}) &= \alpha_0 + u_i + v_i, \\
\alpha_0 &\sim N(0, \tau_z), \\
\text{where } u_i|u_{k\neq i} &\sim N\left(\frac{\sum_{j\neq i} c_{ik}u_k}{\sum_{k\neq i} c_{ik}}, \frac{1}{\tau_{CAR}\sum_{k\neq i} c_{ik}}\right), \\
\text{and } v_i &\sim N(0, \sigma_v^2), \text{ for } i = 1, \ldots, I.
\end{aligned}$$

We will assign hyperpriors to the hyperparameters τ_{CAR}, σ_v^2, and τ_z, typically conjugate gamma hyperpriors.

5.4 Chain Binomial Model Overview

Below is an overview of the chain binomial models presented above. Each maintains the framework of a stochastic binomial process, tracking groups of individuals in each of the three stages over space and time. Note the unique parameterizations of the transmission parameters and transmission probabilities.

Overall, we present two methods for estimating R_0 using the theory of chain binomial models. We induce correlation in our estimates through distinct techniques: estimating transmission using an identity link, and estimating transmission probability using a logit link. Our results reflect varying levels of success when using these two processes. We fit each model within WinBUGS, which uses a Gibbs sampling and Metropolis step approach to estimate model parameters.

The Reed-Frost model adds spatial correlation to p_{ij} to reflect spatially correlated variations in the transmission probability, while the transition model has a parametric model of p_{ij} and induces spatial correlation on the parameter β_{ij}. The variability in our estimates possibly could be due to spatially correlated variations in unmeasured risk factors associated with disease transmission.

Model Name	**Model Parameters**	R_0
Transition Model	$p_{ij} = 1 - \exp\left(\frac{-\beta_{ij}}{N_i} I_{ij}\right)$ $\beta_{ij} = \beta_0 + u_i$ $u_i \sim \text{CAR}$	$R_{0i} = \beta_{ij}$
Reed-Frost Model	$1 - q_{ij}^{I_{ij}}$=prob. escape infection $p_{ij} = 1 - q_{ij}$= transmission prob. $\text{logit}(p_{ij}) = \alpha_0 + u_i$ $u_i \sim \text{CAR}$	$R_{0i} = p_{ij} * S_{i0}$

Table 5: Summary of our proposed spatial chain binomial models

5.5 A Spatial Approach to the General Epidemic Model

As described in Chapter 3, we can define a general epidemic as a counting process for infections and removals [2, 8, 43, 44], where we consider a closed population of M individuals, and we assume that multiple cases introduce the infection into a population of initially susceptible individuals. The hazard of infection will depend only on the presence or number of infectives in the population, and as with the chain binomial models, if an individual becomes infected, he or she is infective for an exponentially distributed period of time. From these methods, we can estimate R_0 as the ratio of the transmission parameter β to the recovery parameter γ, and by running this process over the entire study space as well as each census tract, we can estimate an overall R_0 as well as tract-specific values R_{0i} for each tract i.

The likelihood and prior information specified by the model can be reformulated as a hierarchical random effects structure. As with the chain binomial models listed earlier, we can fit exchangeable random effects, conditionally autoregressive random effects, or a combination of the two (via a convolution prior) to our general epidemic model. The placement of the exchangeable and spatial random effects in our model is based on similar random effects structures in frailty models [42, 29].

The term "frailty" typically represents the idea that some people are more susceptible than others to experiencing an event. Frailty models are random effects survival models that account for unmeasured heterogeneity between individuals, and can be used to unify biological models of heterogenous distribution of susceptibility.

As an example, in the study of the impact of vaccination on time to infection, we can assume that each person in the population makes contact with others at a rate of c contacts per unit of time [42]. If a susceptible unvaccinated person makes a single contact with an infected person, then that susceptible individual will become infected with probability π, which corresponds to the transmission probability to an unvaccinated person. However, if a susceptible vaccinated person makes a single contact

with an infected person, then that susceptible individual will become infected with rate $\theta\pi$ where θ denotes the multiplicative efficacy of the vaccine against infection.

Let $\phi(t)$ be the infection point prevalence at time t. In order to model individual heterogeneity in susceptibility to infection, we can also define the non-negative missing random variable Z_ν. Thus, the individual level hazard rate to an unvaccinated person at time t would be:

$$\lambda_0(t) = Z_0 c\pi\phi(t)$$

and to a vaccinated person would be:

$$\lambda_1(t) = Z_1\theta c\pi\phi(t)$$

From there, we could then define the survival function $S_\nu(t)$ corresponding to each hazard rate of stratum ν, which would be the fraction of stratum ν considered to be at risk of infection at time t, $t \geq 0$ [42]. In this case, vaccination strata are indexed by $\nu = 0$ for unvaccinated and $\nu = 0$ for vaccinated.

In another example of frailty models, Clayton and Cuzick introduced a further generalization of the survival proportional hazards model which allowed for positive association of survival times [18]. This model is a semiparametric generalization of other work that allowed for a random effect, or frailty, in the hazard mode, and was motivated by epidemiological studies of disease occurrence in families, litter-matched carcinogenesis experiments, and by studies of sojourn times of the same individual in different states in prognostic studies.

Using the typical language of animal experimentation, the hazard function for an animal from litter l with covariate vector $\boldsymbol{z}$ is:

$$\lambda(t|\boldsymbol{z}, l) = \lambda_0(t)\xi_l \exp\left(\beta^T z\right),$$

where ξ_l are random multiplicative effects, or frailties, shared by all members of the same litter. The frailties are assumed to be i.i.d. gamma variates with mean 1 and variance γ, such that:

$$\xi_l \sim G(\gamma^{-1}, \gamma^{-1})$$

Using the concepts of frailty modeling, we extend a similar random effects approach to handle spatial heterogeneity of our study population. We fit a multiplicative random effect within our existing counting process likelihood for the general epidemic model. First, we consider an exchangeable random effect as follows, where the likelihood is:

$$L(\beta, \gamma; t, \tau) =$$
$$\prod_{i=1}^{n} \{\gamma I(\tau_i)\} \prod_{j=2}^{n} \{\frac{\beta}{M} I(t_j) S(t_j)\} \exp(v_i) \exp\left(-\int_0^T (\gamma I(u) + \frac{\beta}{M} I(u) S(u)) du\right)$$

with $v_i \sim N(0, \sigma_v^2)$, for $i = 1, \ldots, I$. A vague gamma hyperprior distribution can be given to τ_v where $\sigma_v^2 = \frac{1}{\tau_v^2}$. As with the chain binomial models, this approach should provide a local estimate defined by the weighted average of the observed data in tract i and the global overall mean outcome. The estimates $\beta_i = \beta \exp(v_i)$ are tract-specific as the infection times and removal times, as well as the hazards of infection and removal, and cumulative hazard are aggregated over each tract, and smoothed using exchangeable random effects. We can write the independent gamma priors for β and γ as:

$$f(\beta) \propto \beta^{\nu_\beta - 1} \exp(-\lambda_\beta \beta)$$

$$f(\gamma) \propto \gamma^{\nu_\gamma - 1} \exp(-\lambda_\gamma \gamma)$$

where vague hyperpriors are assigned to ν_β, λ_β, ν_γ, and λ_γ. We then express the log-likelihood for the purposes of the Metropolis-Hastings algorithm as follows:

$$LogLik(\beta, \gamma) = n \times [\ln(\gamma)] + (n-1) \times [\ln(\beta)] - \int_0^T (\gamma I(u) + \frac{\beta}{M} I(u) S(u)) du + v_i$$

Additionally, we consider a conditionally autoregressive random effect, where we borrow strength in our estimates from neighboring regions rather than the entire study space. As with the exchangeable random effect, we can fit a multiplicative CAR effect in our likelihood as follows:

$$L(\beta,\gamma;t,\tau) =$$
$$\prod_{i=1}^{n}\{\gamma I(\tau_i)\}\prod_{j=2}^{n}\{\frac{\beta}{M}I(t_j)S(t_j)\}\exp(u_i)\exp\left(-\int_0^T(\gamma I(w)+\frac{\beta}{M}I(w)S(w))dw\right)$$
$$\text{where } u_i|u_{k\neq i}\sim N\left(\frac{\sum_{j\neq i}c_{ik}u_k}{\sum_{k\neq i}c_{ik}}\,,\ \frac{1}{\tau_{CAR}\sum_{k\neq i}c_{ik}}\right),\ i=1,\ldots,I$$

A vague gamma hyperprior is also placed on τ_{CAR}. Similar to the exchangeable log-likelihood, the CAR log-likelihood is as follows:

$$LogLik(\beta,\gamma) = n\times[\ln(\gamma)]+(n-1)\times[\ln(\beta)]-\int_0^T(\gamma I(w)+\frac{\beta}{M}I(w)S(w))dw+u_i$$

Convolution models provide a balance between global smoothing techniques implemented with exchangeable random effects, and local smoothing with conditionally autoregressive random effects. Both exchangeable and CAR random effects are placed in the model:

$L(\beta,\gamma;t,\tau) =$

$$\prod_{i=1}^{n}\{\gamma I(\tau_i)\}\prod_{j=2}^{n}\{\frac{\beta}{M}I(t_j)S(t_j)\}\exp(u_i+v_i)\exp\left(-\int_0^T(\gamma I(w)+\frac{\beta}{M}I(w)S(w))dw\right)$$

$$\begin{aligned}\text{where } u_i|u_{k\neq i} &\sim N\left(\frac{\sum_{j\neq i}c_{ik}u_k}{\sum_{k\neq i}c_{ik}}\,,\ \frac{1}{\tau_{CAR}\sum_{k\neq i}c_{ik}}\right),\ i=1,\ldots,I\\ \text{and } v_i &\sim N(0,\sigma_v^2),\ \text{for } i=1,\ldots,I.\end{aligned}$$

The corresponding log-likelihood for the Metropolis-Hastings algorithm is as follows:

$$LogLik(\beta,\gamma) = n\times[\ln(\gamma)]+(n-1)\times[\ln(\beta)]-\int_0^T(\gamma I(w)+\frac{\beta}{M}I(w)S(w))dw+u_i+v_i$$

We define the proposal distribution $q(.|\beta_0)$ as a uniformly distributed step $U(-0.001,0.001)$ around β_0, and the proposal distribution $z(.|\gamma_0)$ is a uniformly distributed step $U(-0.001,0.001)$ around γ_0. We use the software package **R**, to implement the MCMC algorithm over 10,000 iterations per tract.

5.6 Results: Chain Binomial - Spatial Model

5.6.1 Estimation of Transmission Probability

We consider the first four years of cases for the transition probability model - an adjustment to the temporal chain binomial model with a specified set of priors. First, we induce global smoothing by using an exchangeable random effects structure built into our spatial hierarchical model. Initially, we borrow strength globally for our local estimates using a set of exchangeable random effects in order to estimate the probability of transmission for each susceptible individual at a given time point in a given tract. Next, we induce local correlation from neighboring census tracts according to the conditionally autoregressive random effects spatial structure. Choropleth maps have been produced in ArcMap, and cut-off intervals of the probability of transmission are fixed across all maps using the quantiles from the CAR model. We run the model in WinBUGS, which uses an assortment of Gibbs and Metropolis steps to estimate the marginal posterior densities of each of our parameters, and we assessed convergence using the Brooks-Gelman-Rubin statistic. If the full conditional is not recognizable in WinBUGS, either adaptive rejection sampling will be used or the Metropolis algorithm. With adaptive rejection sampling, a dynamic envelope function is created to closely mimic the functional form of the full conditional. In the Metropolis algorithm, WinBUGS will sample from a candidate generating normal density, whose variance is ideally close to the true posterior variance. As with the Metropolis-Hastings algorithm described earlier, the candidate is either accepted as the new iterate or rejected and the old value is retained.

Figure 7 shows strong spatial patterns amongst the estimated probabilities of transmission. Higher probabilities are found towards the center of the city - an indication of the location of core areas of disease transmission. For the exchangeable α_0 prior, we chose $N(0, 0.01)$, and for the τ_v hyperprior, we chose $Gamma(0.01, 0.01)$.

Additionally, we addressed spatial correlation by incorporating conditionally autoregressive random effects which allows local estimates to borrow strength from neighboring census tracts. These "neighbors" are defined using an adjacency matrix. As with the map detailing the effect of exchangeable random effects, we find strong spatial correlation between the probabilities of transmission. For the α_0 prior, we again chose $N(0, 0.01)$, and for the τ_{CAR} hyperprior, we chose $Gamma(0.01, 0.01)$. In order to incorporate both global and neighboring weighting techniques, a convolution model was developed. These estimates were mapped, and display strong similarities to both the exchangeable and CAR models (Figure 7). Prior information was the same as the previously described models. A list of parameter estimates for the overall mean effect α_0 as well as for the exchangeable and CAR hyperpriors are listed in Table 6. The numbers suggest strong overlap in the mean effects across the types of random effects structures.

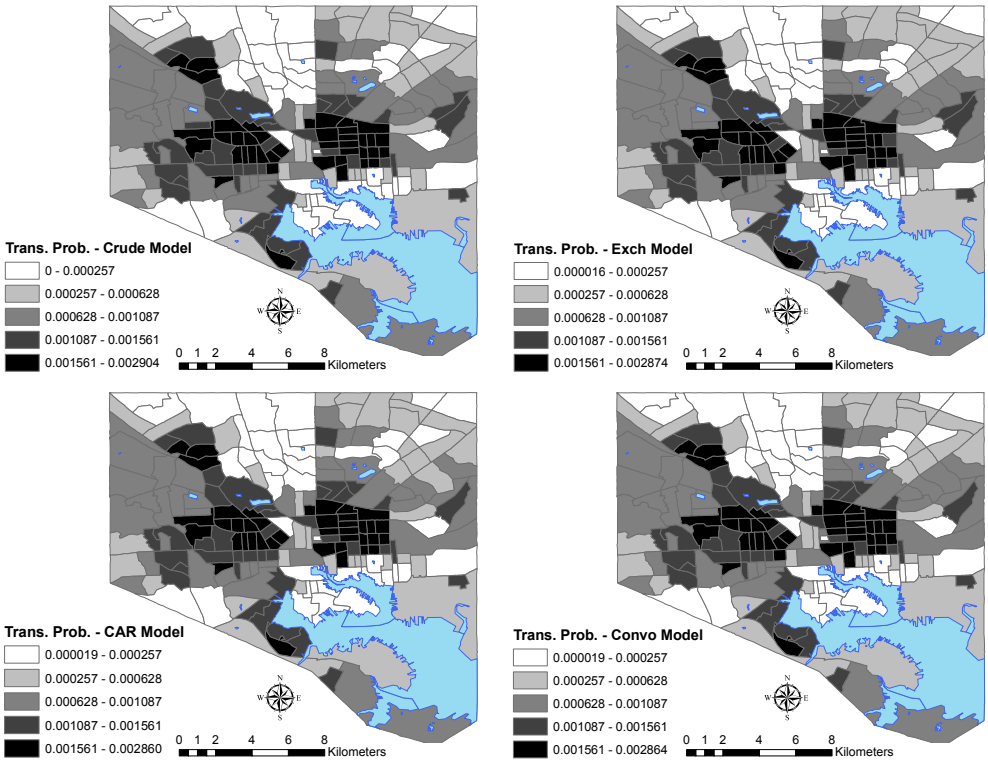

Figure 7: Local median estimates for the transmission probability, i.e. the percent chance of moving from the susceptible class to the infectious class given a contact with an infectious individual based on the transition chain binomial model approach. Maps shown include crude probability estimates, along with exchangeable, CAR, and convolution random effects.

Model	α_0 **[Mean (SE)]**	τ_v **[Mean (SE)]**	τ_{CAR} **[Mean (SE)]**
Exchangeable	-7.340 (0.0687)	0.9767 (0.107)	.
CAR	-7.347 (0.0148)	.	0.3649 (0.0410)
Convolution	-7.346 (0.0170)	0.3322 (0.0480)	64.99 (58.89)

Table 6: List of parameter estimates for α_0 and τ in the transmission probability model.

5.6.2 Estimation of R_0 - Transition Chain Binomial Model

As with our model estimating the probability of transmission, we used the first four years of our data. We apply exchangeable, conditionally autoregressive, and convolution priors within the model, in an effort to assess the level of variation in our

estimates over the study space. In this model, we fix our recovery parameter γ to be one month, so that our estimate for β is equivalent to R_0. In the overall temporal model, we obtain a median R_0 value of 1.019 with a 95% credible set of (1.002, 1.037) (Figure 8) which behaves well over 10,000 iterations.

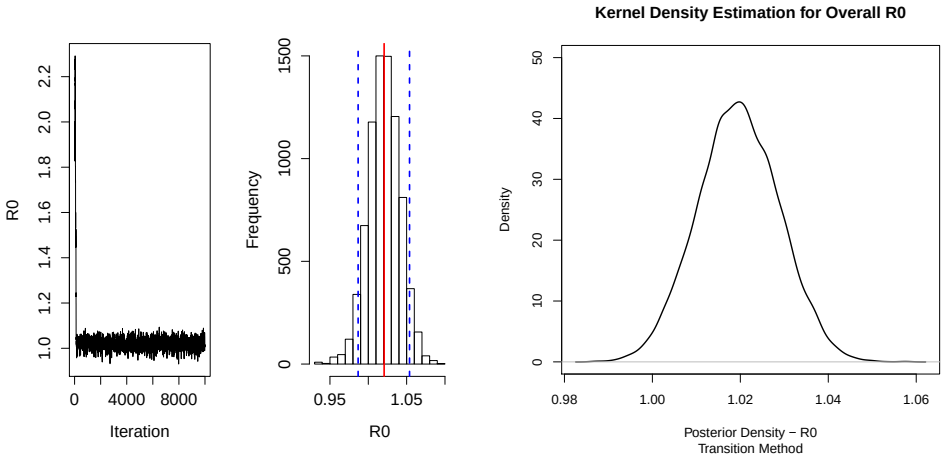

Figure 8: The Markov chain of the temporal R_0 estimate over 10,000 iterations, along with a histogram of the last 8,000 iterations and a kernel density estimate over the last 1,000 iterations.

In our spatial estimation of R_0, we map local estimates of R_0 with no random effects structure, an exchangeable random effects structure, a conditionally autoregressive random effects structure, and a convolution random effects structure. The goal is to visualize and quantify the effect on inducing varying types spatial correlation over a study space. We assign a $Gamma(0.01, 0.01)$ prior to β_0, our overall mean effect, for each model, as well as for τ_{CAR} and τ_v. Estimates were calculated using 10,000 iterations for each model with a 2,000 iteration burn-in, and the number of susceptible individuals per tract at the outset is assumed to be the number of people at risk - i.e., individuals in the population aged 15-49 with 336,551 total susceptibles in the study area. We create choropleth maps in ArcMap, using cut-offs established

by the CAR model quantiles.

All four maps of tract-specific R_0 estimates detail several hot spots for disease transmission, particularly the conditionally autoregressive structure (Figure 9). Core areas of STI transmission can be found towards the center of the city, as was also indicated in our maps from the initial spatial analysis in SaTScan and GeoDa, as well as the spatial estimation of the transmission probability in our previous model.

From Figure 9, we find core areas (dark) of disease spread based on higher median values of R_0. These higher values of R_0 are found primarily in central and western parts of Baltimore, with lower values found primarily in the northern and eastern edges of the city. There are clear spatial patterns and clustering emerging in the CAR model, where the smoother surfaces are due local smoothing techniques. The level of smoothing can be altered based on the strength of the hyperpriors on τ_{CAR}, although we choose vague hyperpriors. In addition, we produce parameter estimates for the overall mean effect β_0 and τ (Table 7). The values of β_0 (overall mean) are equivalent to the overall mean R_0 for each model since the recovery rate γ is fixed at one time unit (month). We find stable estimates of the overall R_0 of about 1.034-1.038 based on three models used, along with stable estimates of τ_v and τ_{CAR}.

In order to compare and evaluate estimates based on the exchangeable, CAR, and convolution random effects structures, we create scatterplots of the median R_0 estimates of exchangeable random effects vs. CAR random effects, CAR vs. convolution, and exchangeable vs. convolution (Figure 10). We find that median estimates across each set of pairs are highly correlated, with stronger correlation for tracts with higher values of R_0, and less correlation for lower values of R_0. The CAR effects and convolution effects produce estimates which are most strongly correlated across all values of R_0. In these plots, we observe no outlying tract estimates away from the diagonal, and we produced these plots using **R** with estimates generated in WinBUGS. Additionally, we track the median estimates of R_0 across census tract for all three

models, broken down into sets of 50 tracts (Figure 11). In most cases, the median tract estimates are very similar between each of the three random effects structures, and we generally find the strongest overlap for higher median R_0 tract estimates, and weaker overlap for lower estimates.

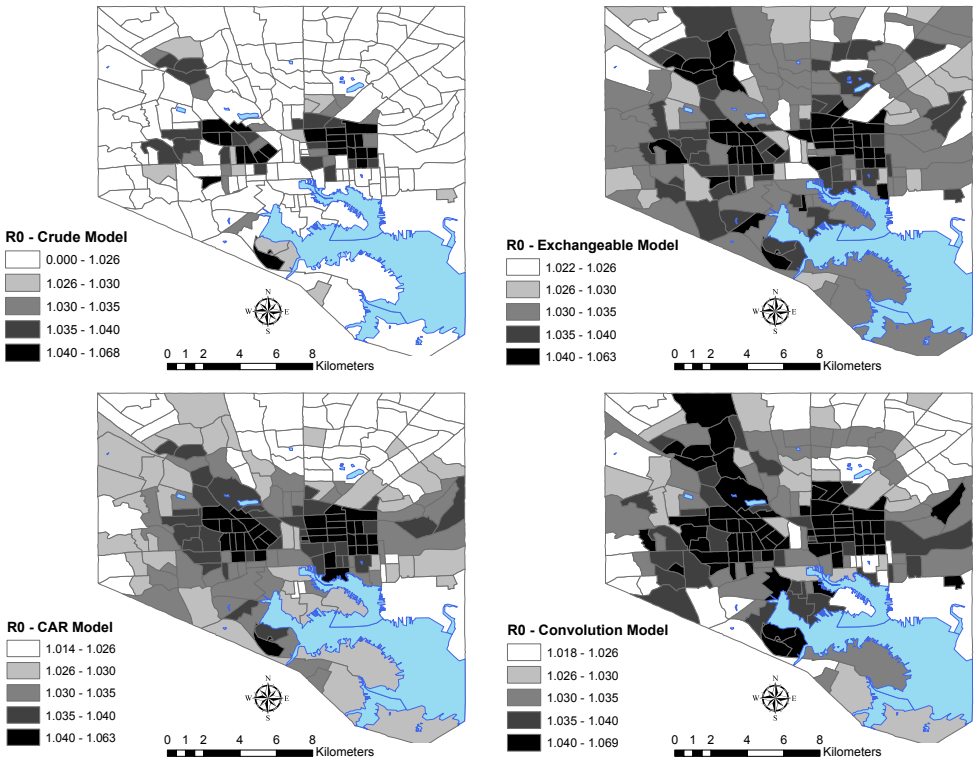

Figure 9: Local median estimates for R_0. Estimates obtained using assumption of a binomially distributed set of newly infected individuals, with exchangeable, CAR, and convolution random effects correlation induced in the transmission parameter. A map of estimates from the crude (non-adjusted) model is also included.

Model	β_0 **[Mean (SE)]**	τ_v **[Mean (SE)]**	τ_{CAR} **[Mean (SE)]**
Exchangeable	1.038 (0.0147)	48.69 (5.797)	.
CAR	1.034 (0.0107)	.	28.89 (4.383)
Convolution	1.038 (0.0153)	41.00 (5.442)	19.69 (3.722)

Table 7: List of parameter estimates for β_0 and τ in the transition chain binomial model for the estimation of R_0.

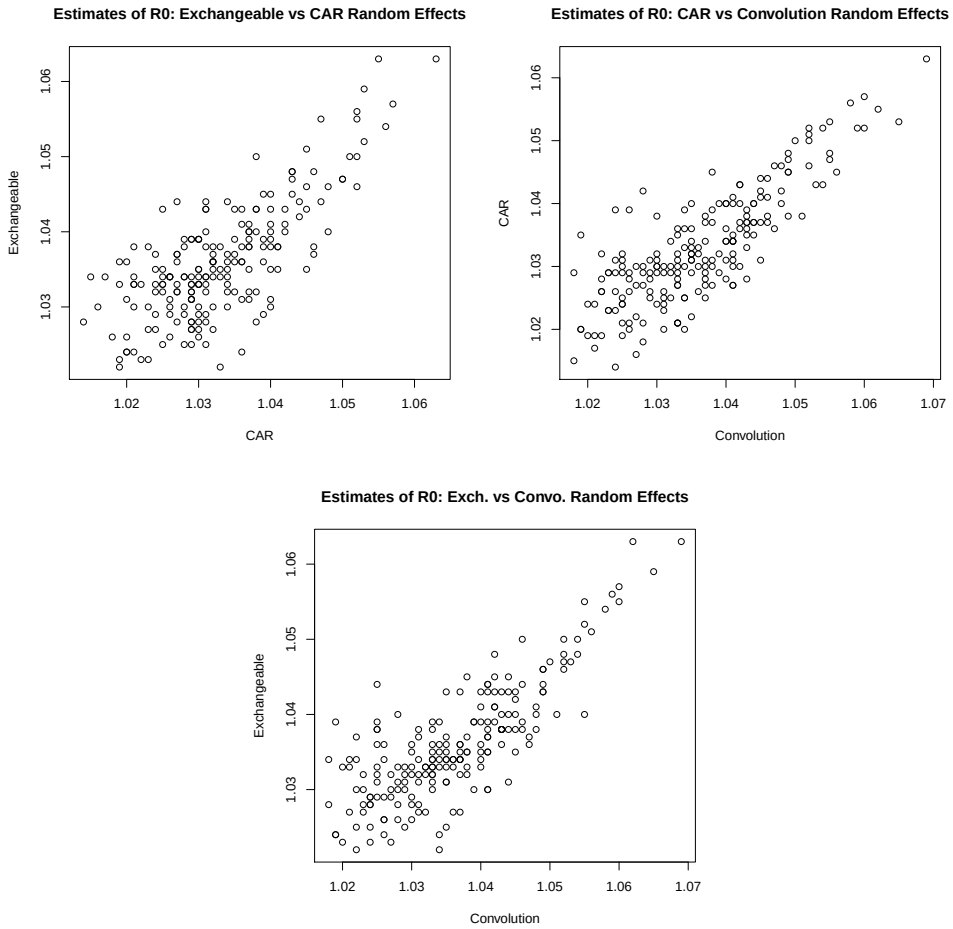

Figure 10: Local median estimates for R_0. Comparing median estimates of exchangeable, conditionally autoregressive, and convolution random effects with the transition chain binomial model.

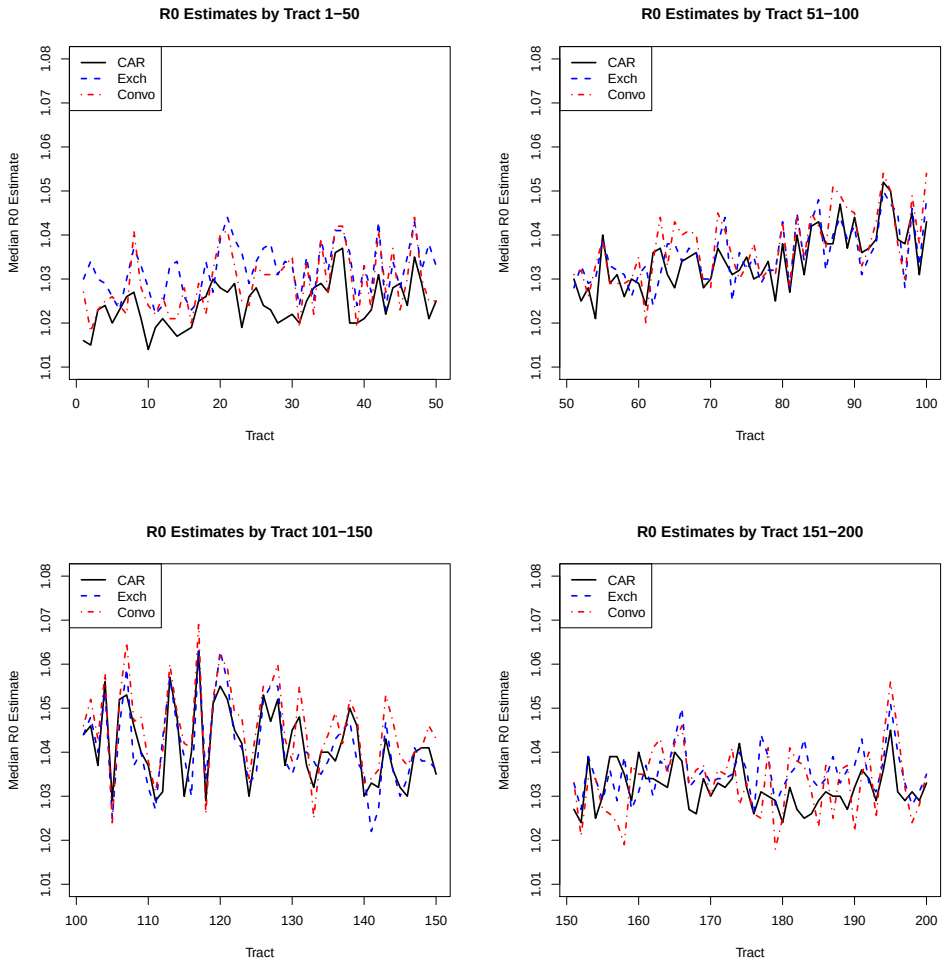

Figure 11: Local estimates for R_0. Comparing median estimates of exchangeable, conditionally autoregressive, and convolution random effects across tract number.

5.6.3 Estimation of R_0 - Reed-Frost Chain Binomial Model

As with our transition chain binomial model, we use four years of data corresponding to 2002-05. We fit exchangeable, conditionally autoregressive, and convolution random effects to the Reed-Frost chain binomial model. In the overall temporal Reed-Frost model, we obtained a median R_0 value of 1.019 with a 95% credible set of (1.002, 1.039) (Figure 12) which behaves well over 10,000 iterations. The Reed-Frost model produces the same median R_0 estimate as the transition chain binomial model.

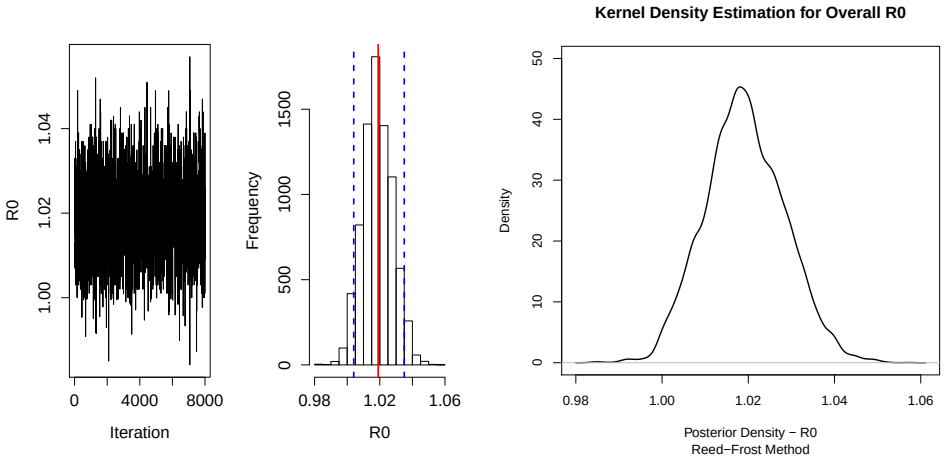

Figure 12: The Markov chain of the temporal R_0 estimate over 10,000 iterations, along with a histogram of the last 8,000 iterations and a kernel density estimate of the last 1,000 iterations.

In our spatial modeling approach, we map local median estimates of R_0 using the assortment of random effects structures listed above. Unlike the transition chain binomial model, we needed to fix the values of τ_{CAR} and τ_v to 0.01 and 0.01 respectively for the algorithms to perform adequately. A number of combinations of fixed values were considered for τ_{CAR} and τ_v, although authors typically suggest values of τ_{CAR} approximately equal to $0.7\tau_v$ so that fair weighting is assigned to both local and global smoothing techniques. Allowing the precision parameters to vary caused a number

of issues in the estimation of R_0 spatially and failed to induce spatial correlation as well as the fixed precision parameters. The overall mean effect, α_0 was given a vague $N(0, 0.01)$ prior for each random effects model.

The maps listed below (Figure 13) display variations in median R_0 values according to the type of random effects structure, and show core areas (dark) of disease spread based on higher values of R_0. Cut-off values for the choropleth intervals are based on CAR model quantiles. As with the transition model, higher values of R_0 are found primarily in central and western parts of Baltimore, and lower values are found primarily in northern and eastern edges of city. From the maps, we also see strong similarities in the patterns within the map, perhaps due to fixing the precision parameters. This level of correlation between the median R_0 estimates is further shown in Figure 14, where the exchangeable, CAR, and convolution random effects produce very similar estimates when fixing τ_{CAR} and τ_v. Aside from a few tracts corresponding to regions with very low case counts, we see almost no variation in the R_0 estimates across random effects structures. CAR effects and convolution effects produce estimates which are most strongly correlated across all values of R_0.

In order to get an idea of the differences in R_0 across the tracts, we also produced line graphs linking the median R_0 values (Figure 15). We compare median exchangeable, CAR, and convolution estimates of R_0 across census tract number, broken down into sets of 50 tracts as with the transition model. In most cases, median tract estimates are very similar between each of the three random effects structures; however, tracts with small numbers of cases, such as Tract 8, Tract 20, Tract 130, and Tract 198 produce median R_0 values which are harder to predict and more dependent on the model.

Again, we find more variability in areas with lower numbers of individuals who contract the disease. It appears that the Reed-Frost model is more sensitive to low counts than the transition chain binomial model. Exchangeable, CAR, and convo-

lution random effects can produce vastly different estimates at these types of small count locations. The overall parameter estimate α_0 (Table 8) varies based on the structure implemented, from -3.79 to -7.32, and the latter estimate most closely follows the α_0 estimates produces by the transmission probability model listed earlier.

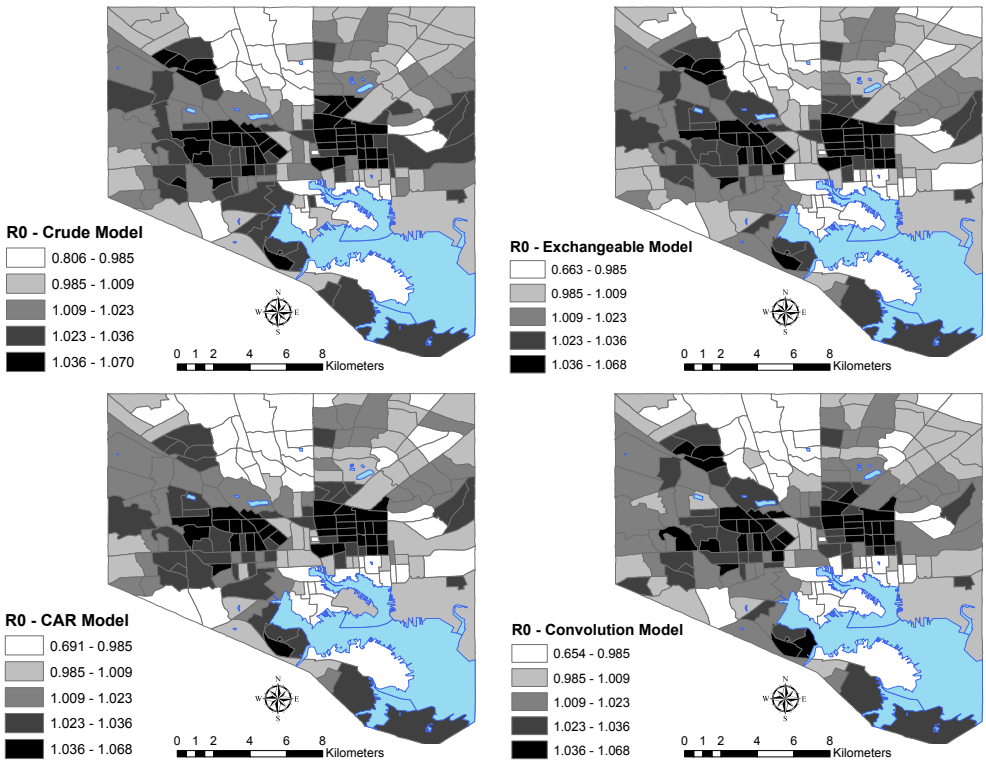

Figure 13: Local median estimates for R_0. Estimates obtained using assumption of a binomially distributed set of infected individuals, with exchangeable, CAR, and convolution random effects correlation induced in the transmission probability within the Reed-Frost model. Crude model map also included.

Model	α_0
Exchangeable	-3.790 (0.0578)
CAR	-7.318 (0.0194)
Convolution	-4.695 (0.0833)

Table 8: List of parameter estimates for α_0 in the Reed-Frost chain binomial model.

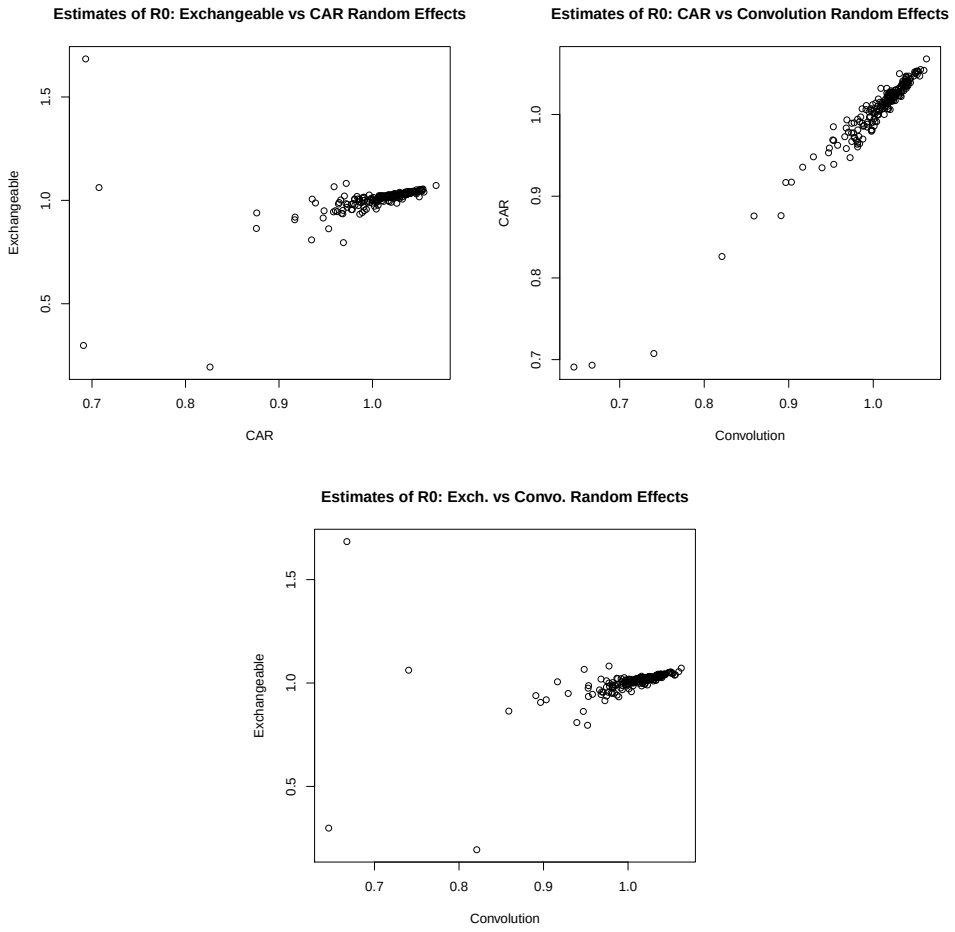

Figure 14: Local median estimates for R_0. Comparing median estimates of exchangeable, conditionally autoregressive, and convolution random effects with the Reed-Frost chain binomial model.

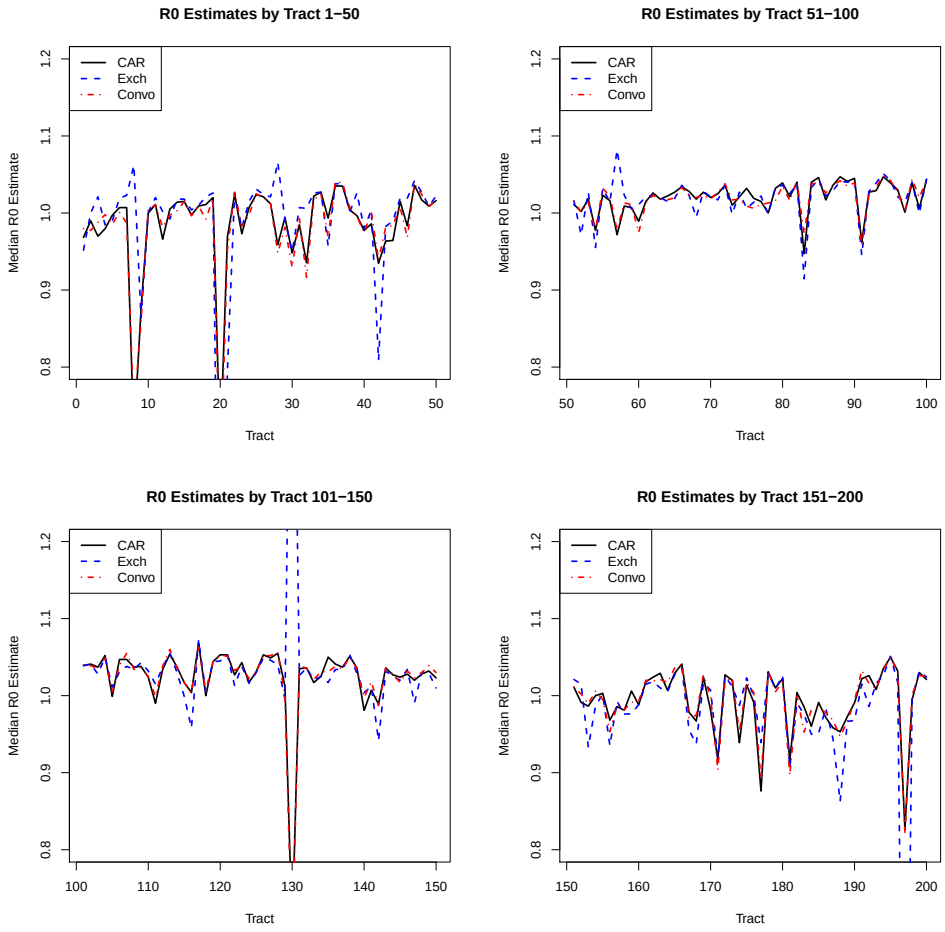

Figure 15: Local median estimates for R_0. Comparing median estimates of exchangeable, conditionally autoregressive, and convolution random effects across tract number using the Reed-Frost chain binomial model.

5.6.4 Chain Binomial Model Comparison

In order to further evaluate and compare models, we recorded information on model fit available in WinBUGS. The Deviance Information Criterion (DIC) is a hierarchical modeling generalization of AIC and BIC, useful in situations with posterior densities obtained using MCMC sampling techniques. We noted for each model - the transmission probability model, transition model, and Reed-Frost model - the three components of the DIC - $\bar{D}$, $\hat{D}$, and pD - as well as the DIC values. The deviance is defined as $D(\theta) = -2\ln(p(y|\theta))$ where y denotes the data, θ are the model parame-

ters, and $p(y|\theta)$ is the likelihood function. The average deviance, or $\bar{D}$, is a measure of how well the model fits the data and is equal to $E[D(\theta)]$. The larger the value the worse the fit. pD denote the effective number of parameters in model, and DIC = $\bar{D}$ + pD. Lower values of DIC indicate better fit. From Table 9, it is apparent that the chain binomial models estimating R_0, as opposed to transmission probability, provide a better fit. Based on the DIC values, the transition chain binomial model with a CAR random effects structure has the lowest value and best fit among the models considered. In general, the transition chain binomial model appears to provide a better overall fit to the data than the Reed-Frost model, as well as providing more stable estimates as noted in the previous section.

In addition, we examined the effect of random effects on median R_0 estimates. Typically, models will have more variability in local estimates which do not induce correlation amongst the other estimates, i.e., each tract is estimated independently. CAR and exchangeable random effects should have a smoothing influence on the parameter estimates and pull observations closer to the neighborhood or global mean, respectively. From Figure 16, we find that the exchangeable and CAR random effects draw extreme values closer to the overall average, and their estimates tend to have less variability than the crude model estimates. Five outlying values with R_0 less than 0.8 in the crude model are pulled back towards the other values closer to the overall mean, as shown in Figure 16. However, in the Reed-Frost model in Figure 17, we also find that the exchangeable and CAR random effects do not have the desired influence on the crude model estimates. In contrast to the transition chain binomial model, inducing correlation can have a detrimental effect on the Reed-Frost model. Small counts for cases in tracts have an undesirable effect when inducing correlation in the Reed-Frost model, and we conclude that the Reed-Frost model as implemented with exchangeable or CAR random effects does not fit our data well.

We also plotted and observed the kernel density estimates of the posterior distribu-

tions of R_0 for the first ten census tracts based on the adjacency matrix identification number, as well as the estimated posterior distributions of two larger median R_0 values and one smaller median R_0 value (Figure 18). These densities represent plots of the posterior densities of R_{0i} under the CAR random effects structure, using the coda function in WinBUGS. The last 1,000 iterations of the Markov chain for each tract are used for kernel estimation purposes. The standard error and spread of posterior density are strongly tied to value of R_0 - lower R_0 values equate to more spread in density. Again, we observe tighter densities and stronger estimation in the transition chain binomial model compared to the Reed-Frost model across all values of R_0. In addition, the Reed-Frost model has difficulty locating and estimating lower R_0 values, as is shown with the lower R_0 estimate of 0.69 and its corresponding flat distribution. It is clear that long strings of zero counts in a given tract over time make it difficult to create the posterior density of R_0 in the Reed-Frost model. As a result, we recommend the use of the transition chain binomial model when estimating model parameters spatially with our data.

However, if we relax some of our assumptions about the Reed-Frost model, in particular, how many individuals we consider susceptible to infection at the start of the study period, our modified Reed-Frost model can be more stable in its identifying core areas of disease transmission. We can successfully implement random effects structures in an effort to induce correlation if we reduce the number of susceptible individuals in the population. For example, we assume that the 336,551 individuals aged 15-49 are at risk for infection, which the literature suggests as a reasonable approximation for a susceptible population size. However, a more realistic estimate may be a lower, less conservative number, such as 60,000 individuals in the population, or 300 per tract, which is equivalent to about 10% of the overall population. When implementing these new numbers, we obtain much more reliable results with the Reed-Frost model.

The scatterplots generated in Figure 19 reveal more variability in the R_0 estimates comparing CAR, exchangeable, and convolution models, as we can now place vague gamma prior distributions on τ_{CAR} and τ_v. There are no extreme outlying points. As with the transition chain binomial model, the median tract estimates are very similar between each of the three random effects structures, and we generally find the strongest overlap for higher median R_0 tract estimates, and weaker overlap for lower estimates (Figure 20). We find the strongest evidence for model reliability in Figure 21, as we now observe the exchangeable and CAR random effects pulling the crude R_0 estimates back towards the overall global estimate and neighborhood estimates, respectively.

In the previous Reed-Frost example with the much larger susceptible class population size, the exchangeable, CAR, and convolution models produce unusually high or low R_0 estimates in tracts with small case counts compared to the crude model. This result is the reverse of what we would expect to occur when inducing correlation, and we are able to correct the model by reducing the number of individuals initially susceptible at the outset. The adjusted susceptible population size will increase the values of R_0 across the study space, since the transmission probabilities are now estimated to be higher; however, the underlying core area pattern that emerges is similar to the results generated by the transition model under the assumption of a larger susceptible population size (Figure 22). Thus, when adjusting the population at risk to smaller sizes in relation to the number of infectious individuals over the course of the study period, we can use either the transition or Reed-Frost chain binomial models without significant issues with estimation.

Model		**Prob. Model**	**Trans. Model**	**RF Model**
Exchangeable	$\bar{D}$	25183.60	14996.70	15099.90
	$\hat{D}$	24990.40	14900.00	14903.40
	pD	193.23	96.71	196.56
	DIC	25376.80	15093.40	15296.50
CAR	$\bar{D}$	25184.50	14955.60	15107.20
	$\hat{D}$	24997.90	14899.50	14904.00
	pD	186.59	56.07	203.139
	DIC	25371.10	15011.70	15310.30
Convolution	$\bar{D}$	25183.20	15017.50	15104.90
	$\hat{D}$	24996.50	14900.30	14903.40
	pD	186.64	117.28	201.542
	DIC	25369.80	15134.8	15306.40

Table 9: $\bar{D}$, $\hat{D}$, pD, and DIC values for probability, transition, and Reed-Frost models.

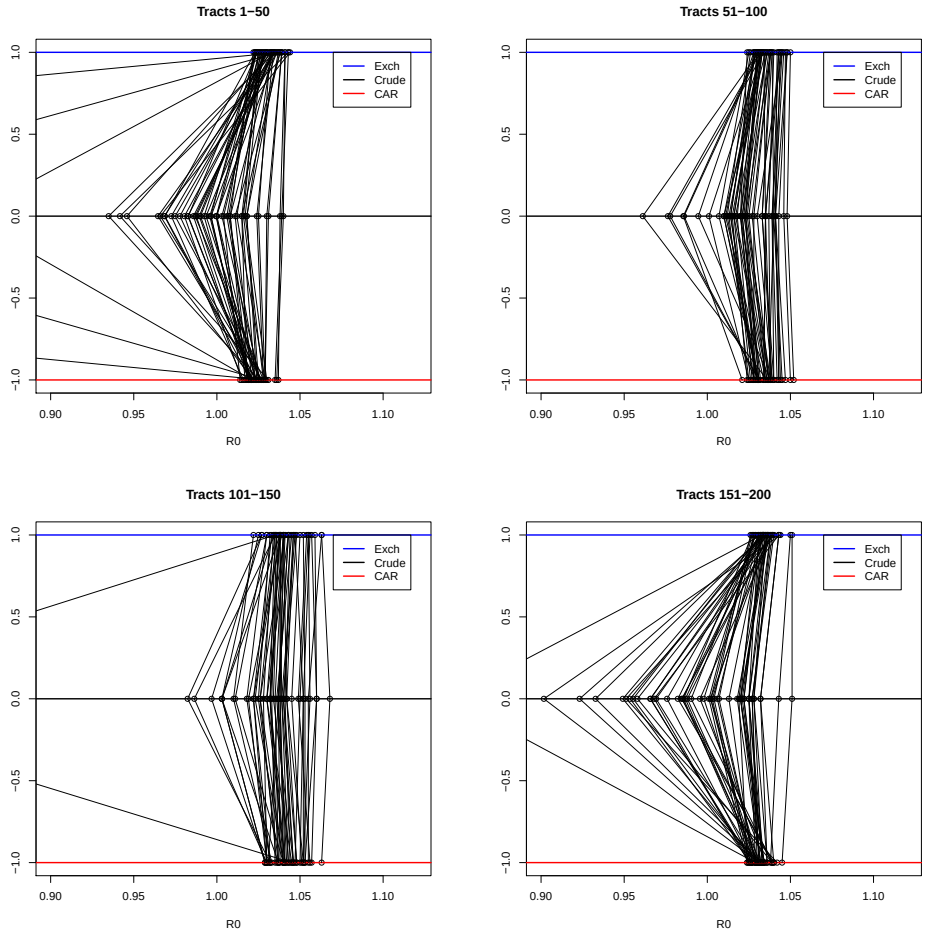

Figure 16: Linking the values of R_0 for exchangeable, CAR, and crude models (Transition model).

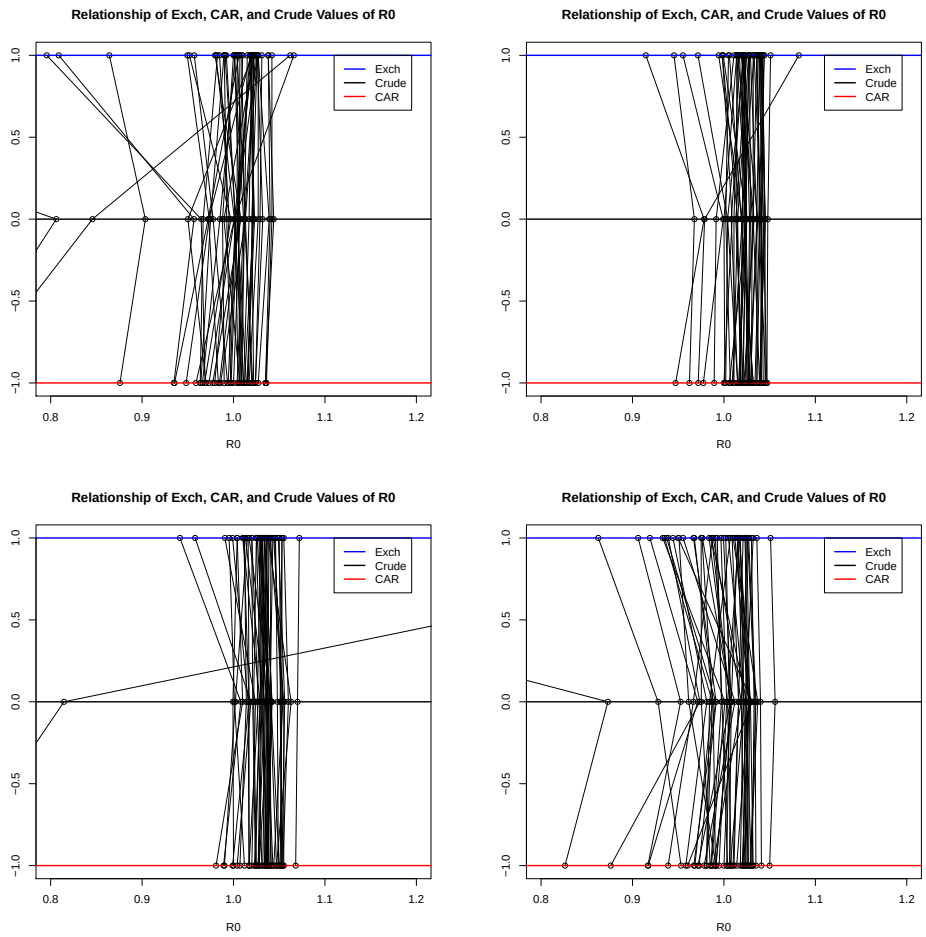

Figure 17: Linking the values of R_0 for exchangeable, CAR, and crude models (Reed-Frost model).

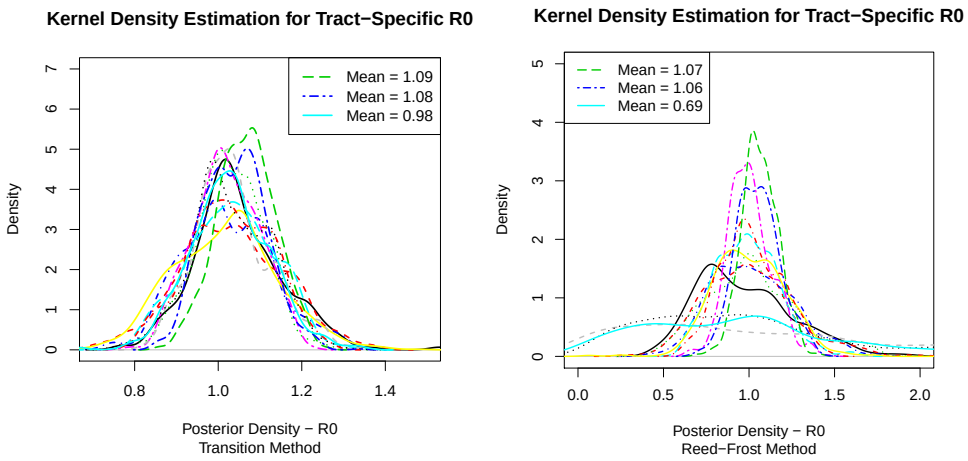

Figure 18: Posterior densities of R_0 for the first ten tracts as well as two higher median R_0 tracts, and one lower median R_0 tract. CAR random effects - Chain Binomial Models.

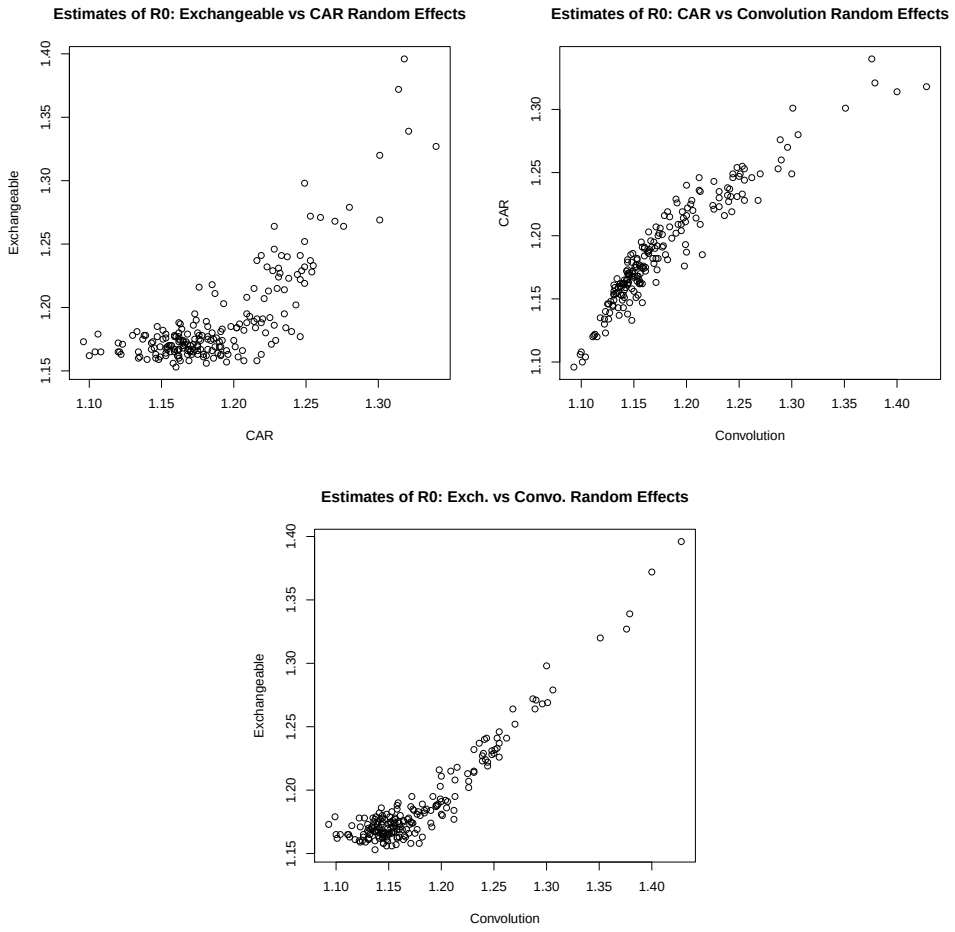

Figure 19: Local median estimates for R_0. Comparing median estimates of exchangeable, conditionally autoregressive, and convolution random effects with the Reed-Frost chain binomial model. Assumption of $S_{i0} = 300$ susceptible individuals per tract.

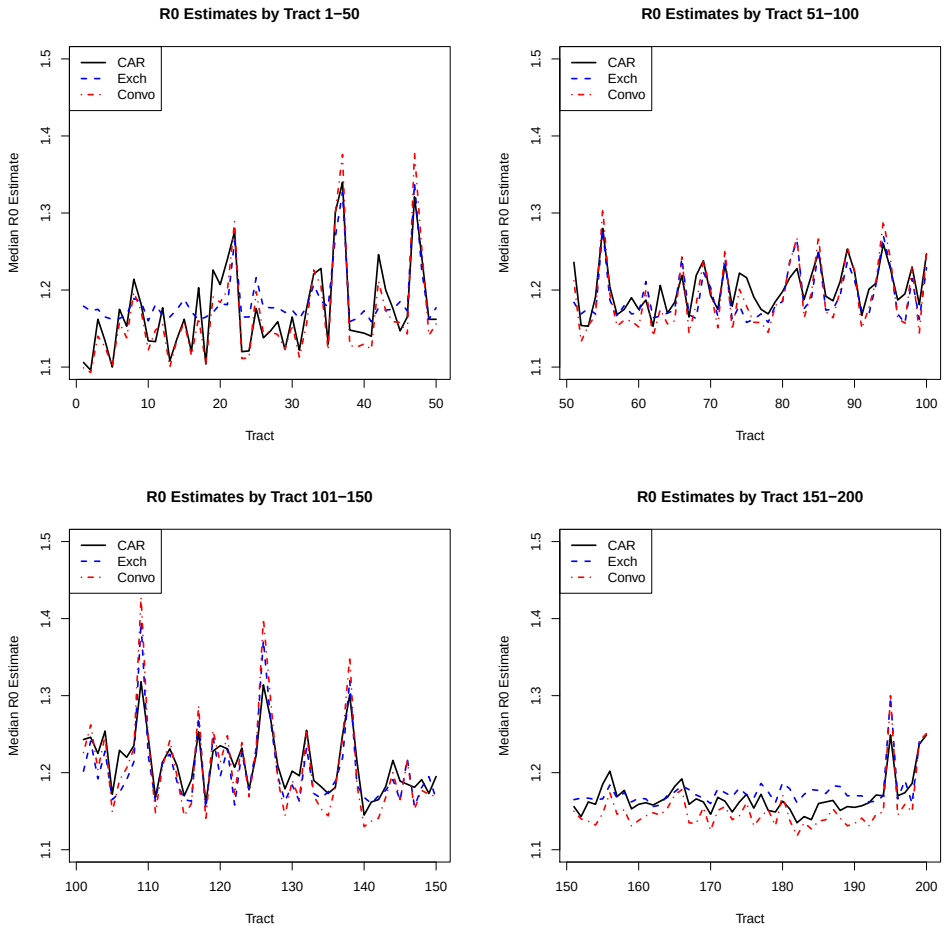

Figure 20: Local median estimates for R_0. Comparing median estimates of exchangeable, conditionally autoregressive, and convolution random effects across tract number using the Reed-Frost chain binomial model. Assumption of $S_{i0} = 300$ susceptible individuals per tract.

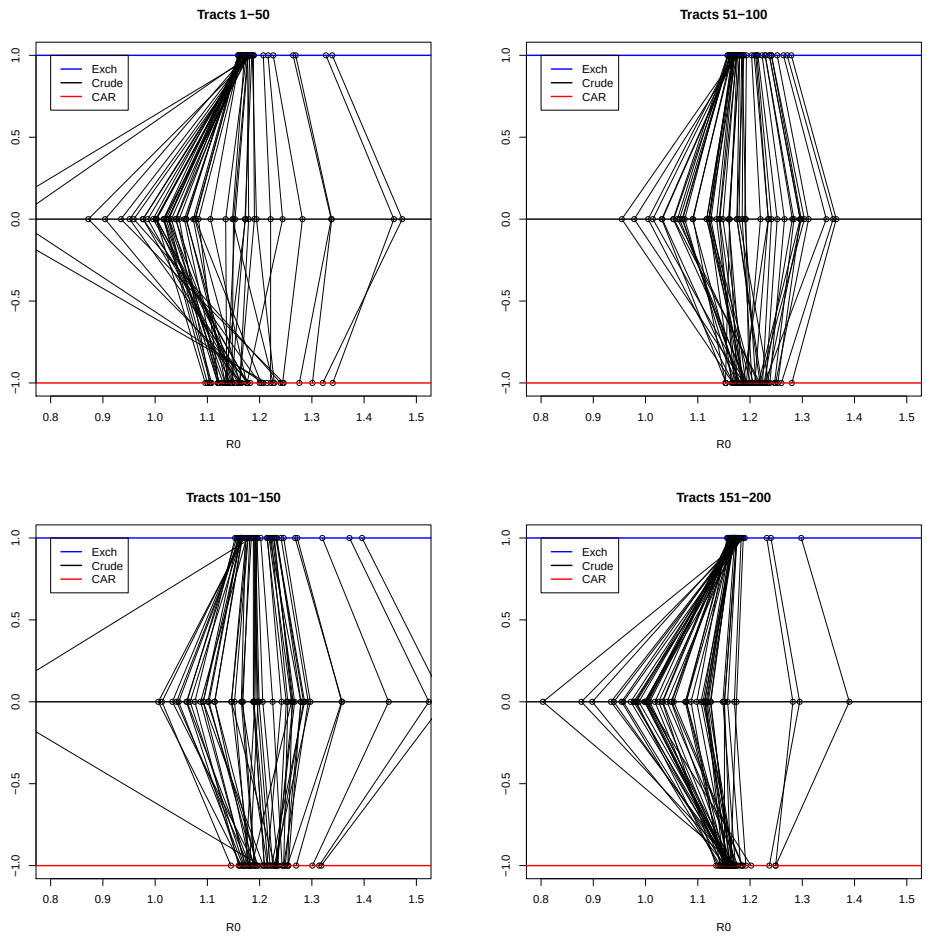

Figure 21: Linking the values of R_0 for exchangeable, CAR, and crude models (Reed-Frost model). Assumption of $S_{i0} = 300$ susceptible individuals per tract.

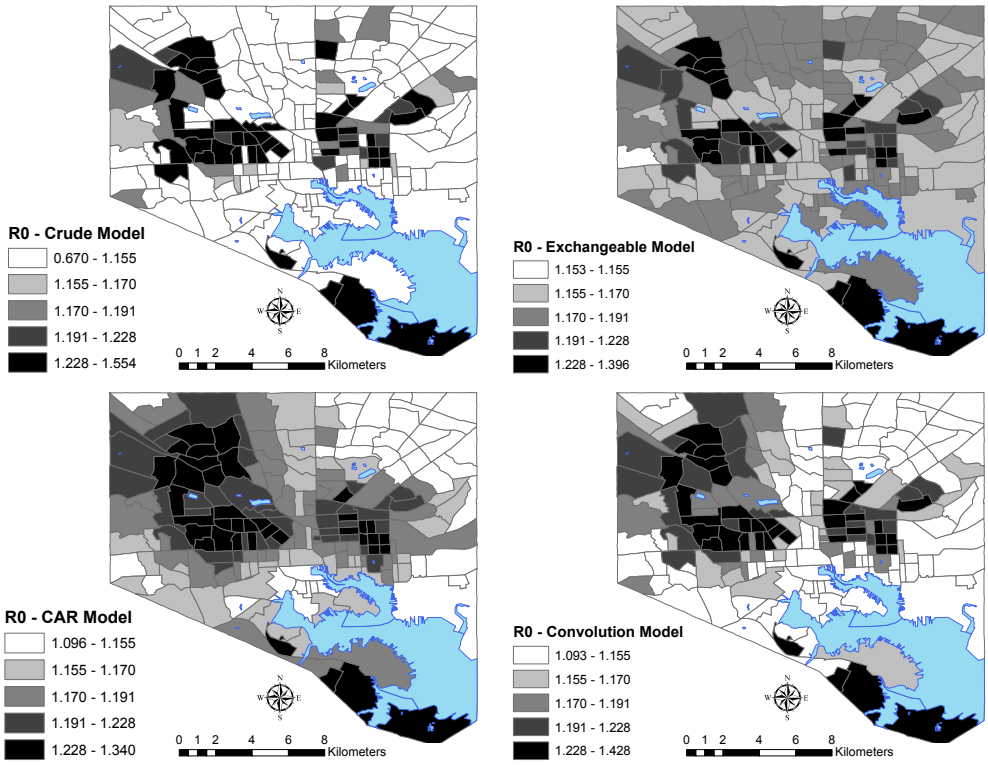

Figure 22: Local median estimates for R_0. Estimates obtained using assumption of a binomially distributed set of infected individuals, with exchangeable, CAR, and convolution random effects correlation induced in the transmission probability within the Reed-Frost model. Crude model map also included. Assumption of $S_{i0} = 300$ susceptible individuals per tract.

5.7 Results: General Epidemic Model - Spatial Estimation

The general epidemic model and corresponding frailty model extensions, introduced earlier, are a novel approach in the spatial estimation of R_0. Previous research estimated R_0 and its component parameters through this type of counting process [2, 8, 43, 44], and we develop an initial method for extending this process into spatial analysis. We run the model using the Metropolis-Hastings algorithm in **R**, and derive the chains for R_0 by dividing the values of β and γ at each iteration, thus providing the posterior density of R_0.

When running the MCMC algorithm over the entire study space, we estimate R_0

to be 1.022 (Figure 23), which is very close to the R_0 estimated by both the transition and Reed-Frost chain binomial models - 1.019. Based on the histogram of the last 500 iterations as well as the autocorrelation function, we conclude that the algorithm has converged appropriately.

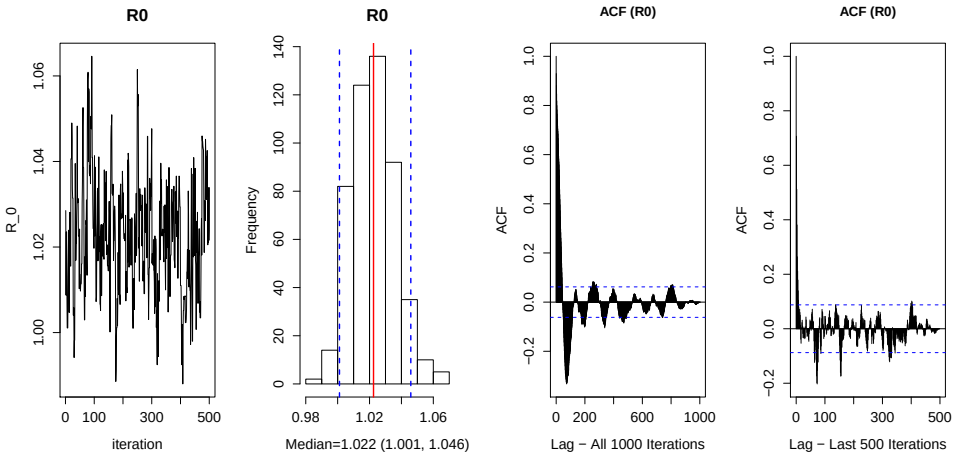

Figure 23: Histogram of R_0 for last 500 iterations of the general epidemic model (Left). Autocorrelation functions over all 1000 iterations and last 500 iterations also shown (Right)

Within our spatial estimation of R_0 using the general epidemic model, we map local estimates of R_0 with no random effects structure, an exchangeable random effects structure, a conditionally autoregressive random effects structure, and a convolution random effects structure. We assign vague gamma priors to τ_{CAR} and τ_v, and base estimates on 10,000 iterations of the MCMC algorithm. We fix the number of susceptible individuals in our population to be 300 per tract, since we had difficulty once again with larger, more conservative, estimates of the susceptible population size. In addition, we note that values of R_0 will be higher for these local estimates based on the reduction of initial susceptibles and increase in the transmission probabilities. However, as with the Reed-Frost model with relaxed assumptions, we should still be

able to detect the pattern of disease transmission under a different scale. We again create choropleth maps in ArcMap, using cut-offs established by the CAR quantiles.

From Figure 24, we find similar core areas of disease transmission across Baltimore compared to the transition and Reed-Frost models. Central and western parts of Baltimore tend to have higher values of R_0 than areas in the northern and eastern regions. We note that the CAR model will smooth estimates based on neighboring regions, as seen with the southernmost census tract. Its estimate has been smoothed to a lower quantile level based on the effects of its surrounding regions, which is not the case with the other random effects structures. When comparing estimates across random effects structures, we produced scatterplots of median R_0 estimates of exchangeable random effects vs. CAR random effects, CAR vs. convolution, and exchangeable vs. convolution (Figure 25). We find that estimates across each set of pairs are highly correlated when comparing the different types of random effects structures, although we notice a few outliers. In addition, we track median exchangeable, CAR, and convolution R_0 estimates across census tract number (Figure 26). Aside from a couple of outlying points in Tracts 20 and 130 where the numbers of cases are small, we generally find strong association between each of the three random effects structure.

As with the two chain binomial models we evaluated earlier, we examine the effect of correlation-inducing random effects on estimated median R_0 values. We would expect a crude model to have more varied values for R_0, while CAR and exchangeable random effects models should have a smoothing influence on the parameter estimates. From Figure 27, we find that the CAR and exchangeable random effects pull several extreme values from the crude model towards the global and neighborhood means, while the random effects have a smaller influence on less extreme values. This was also observed in the chain binomial model analysis on the influence of random effects.

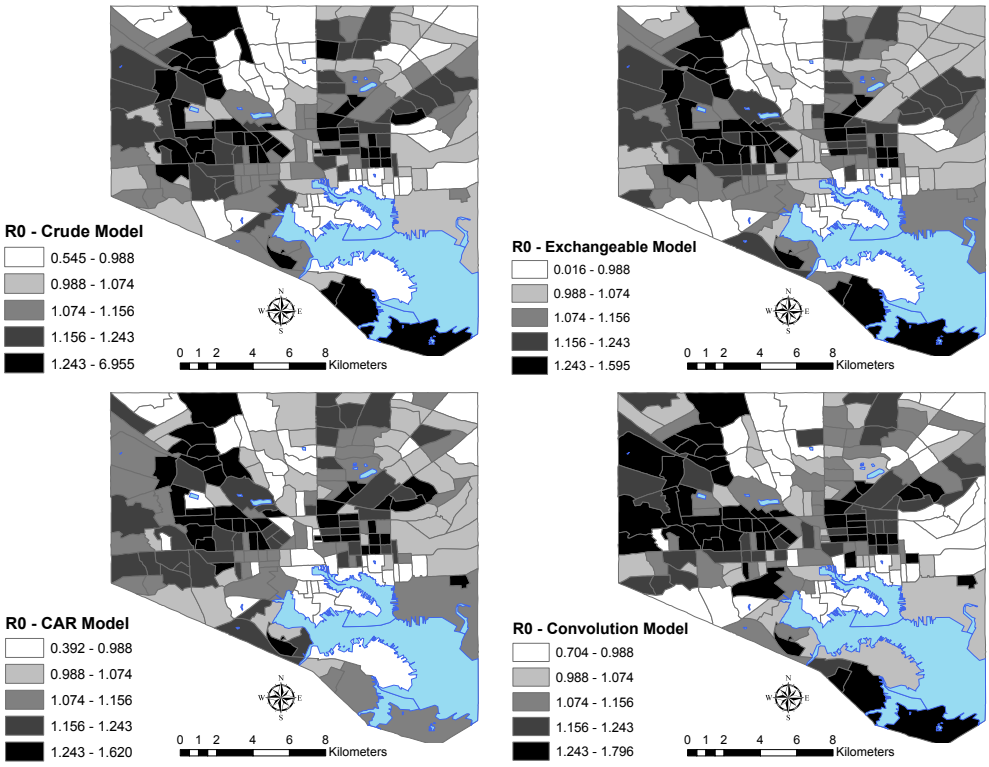

Figure 24: Local median estimates for R_0. Estimates obtained using the general epidemic model, with exchangeable, CAR, and convolution random effects correlation. Crude model map also included. Assumption of 300 susceptible individuals per tract.

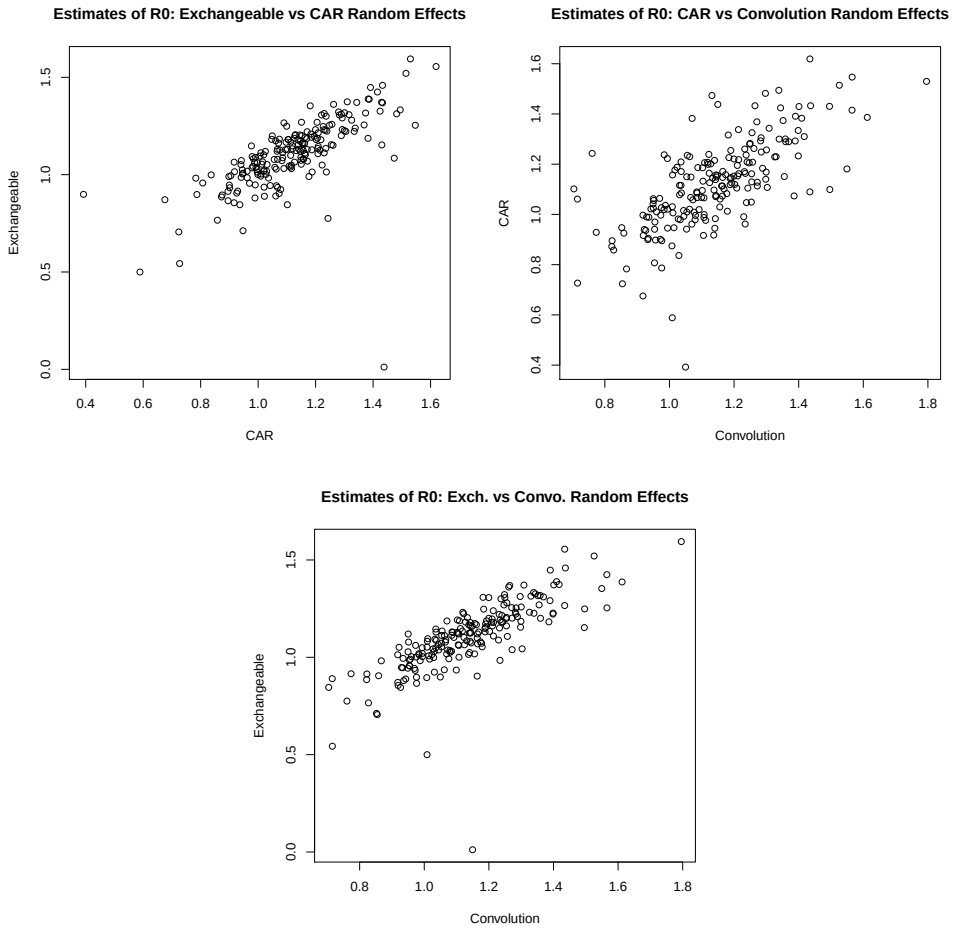

Figure 25: Local median estimates for R_0. Comparing median estimates of exchangeable, conditionally autoregressive, and convolution random effects with the general epidemic model. Assumption of 300 susceptible individuals per tract.

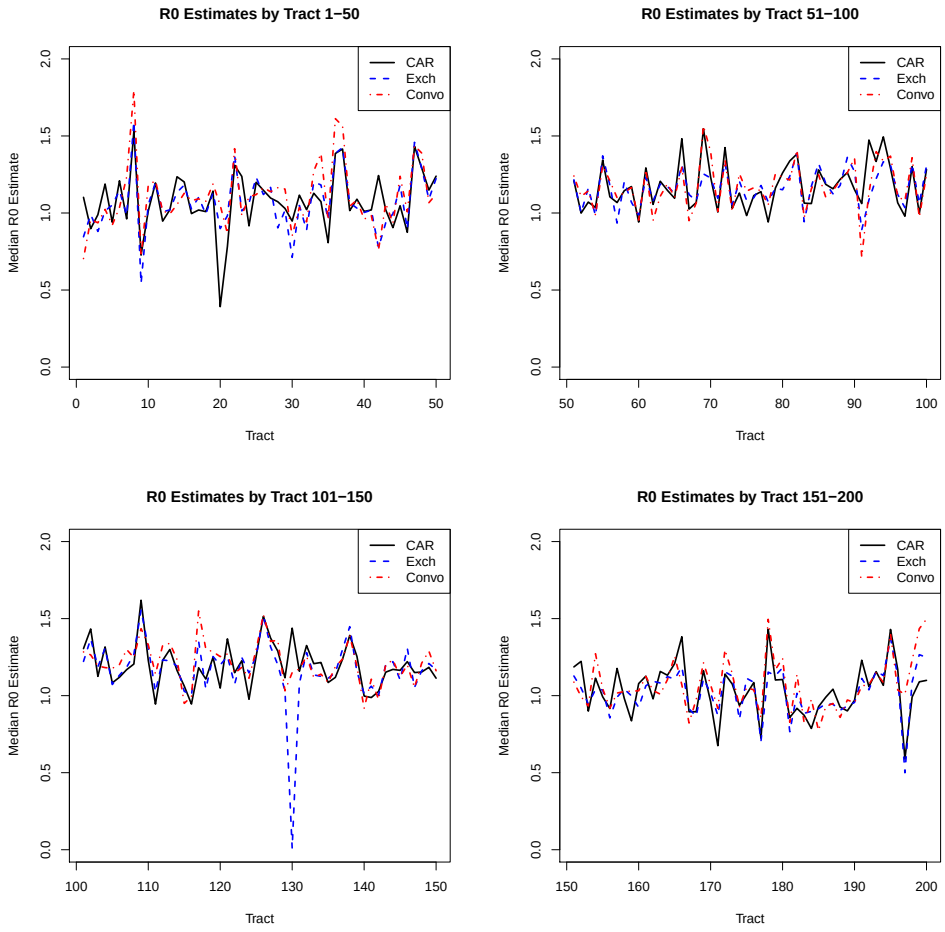

Figure 26: Local median estimates for R_0. Comparing median estimates of exchangeable, conditionally autoregressive, and convolution random effects across tract number using the general epidemic model. Assumption of 300 susceptible individuals per tract.

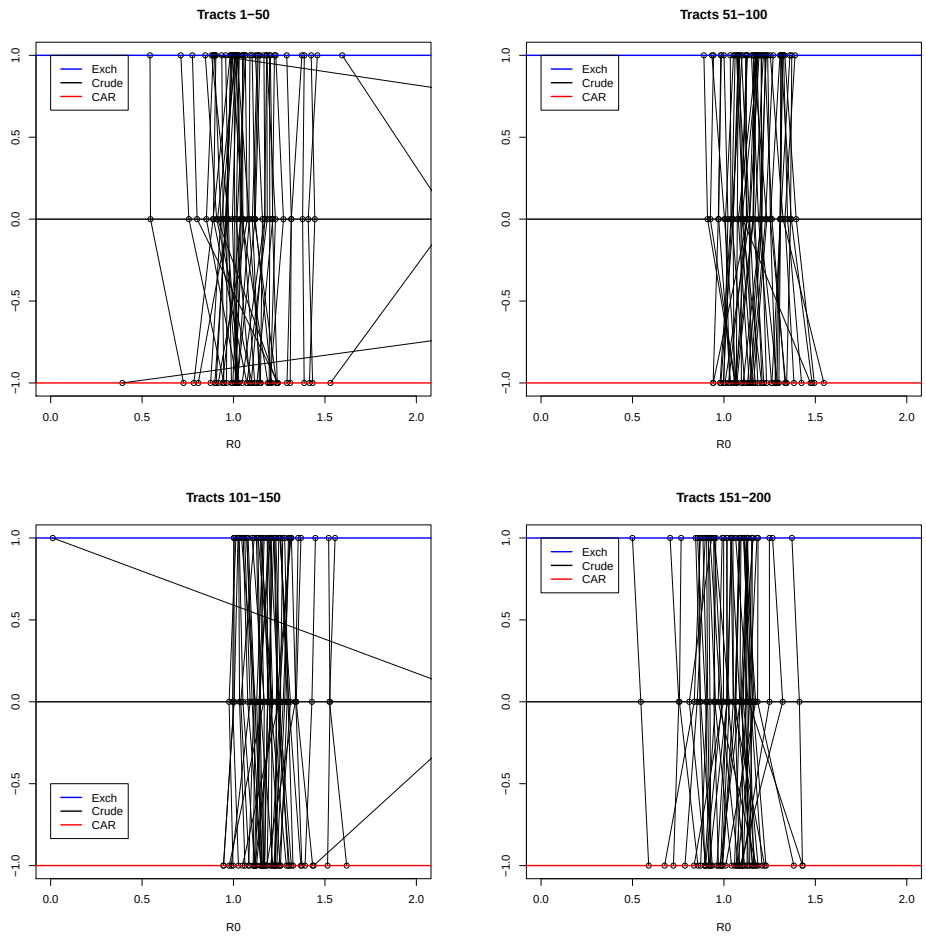

Figure 27: Linking the values of R_0 for exchangeable, CAR, and crude models (General epidemic model). Assumption of 300 susceptible individuals per tract.

5.8 Assessing Model Performance through Simulations of the Chain Binomial Models

In this section, we attempt to test the MCMC algorithm through the application to simulated epidemic data. The goal of simulations is to evaluate how well the model performs with data generated using the same chain binomial model described above. After generating simulated data with a set of initial values and parameters, which we assume to be true, we can then apply our models and their associated MCMC algorithm to these data. We seek a model which can accurately identify the true

values used to generate the data.

From the temporal results of the transition chain binomial model based on the observed data, we estimated a median R_0 value of 1.01. This value was used to generate 100 epidemic chains and the results are displayed below in Figure 28. We assume that 336,551 individuals are susceptible at the outset, and 340 individuals are initially infectious. Overall, the generated chains of infection follow the observed chain well, although the observed chain does not experience the same early peak of cases that is predicted by the model. Using a vague gamma prior of $Gamma(0.001, 0.001)$ with 10,000 iterations and a 2,000 iteration burn-in, we estimate an R_0 of 1.021 (95% credible set: (0.987, 1.054)). From Figure 29, the posterior medians of R_0 are distributed around the true value driving the simulations.

In order to evaluate how well the model performs over the study space of the 200 census tracts of Baltimore, we use the 2000 Census population numbers from each of the tracts, and take the estimates of those at risk - i.e. those ages 15-49. Those at risk are assumed to be the number of individuals susceptible in each tract, and we assume that 2 individuals are infectious at the outset in each tract. In order to evaluate which values of R_0 are reliable at the tract level, we assign three zones of true values of R_0 that are separated by 0.1, and we vary the window of true values from 1.0-1.2 to 1.9-2.1 in groups of 3 (see Figure 30). We look for the ranges of values which generate maps and produce statistics which accurately reflect the true values.

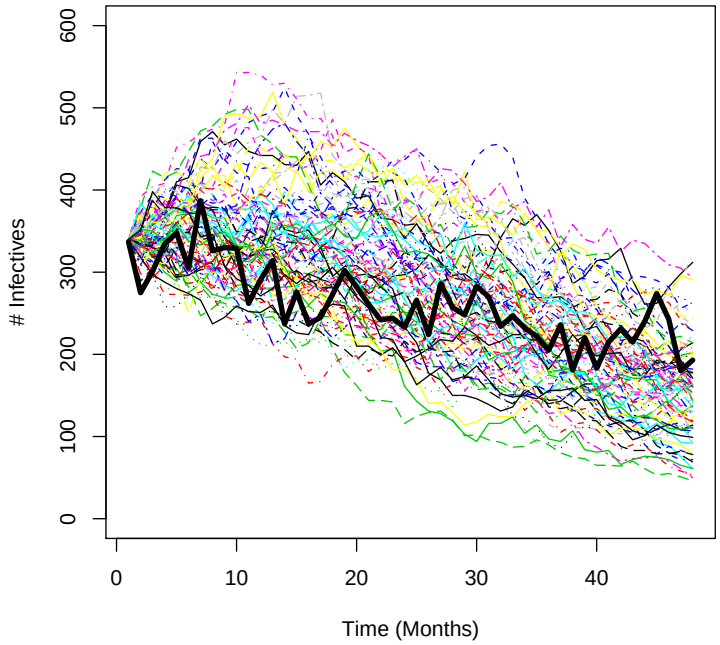

Figure 28: Simulated epidemic curves generated with the transition chain binomial model using a value of R_0 of 1.01, a susceptible population of 336,551 individuals, and 340 initially infectious individuals. The observed epidemic curve is in bold.

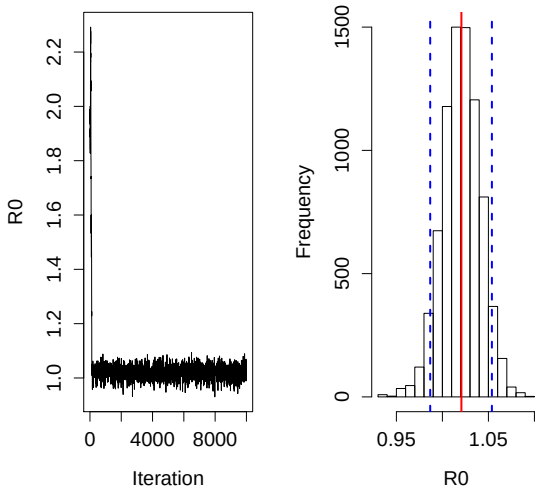

Figure 29: Iterations of MCMC algorithm, and the estimated posterior density of R_0 over the study space, with a median = 1.021, and 95% credible set: (0.987, 1.054)

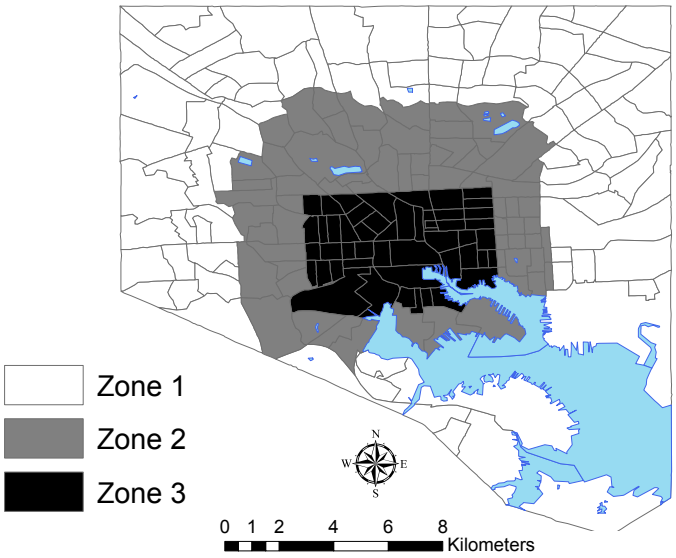

Figure 30: Baltimore City County census tracts divided into three zones. The zones are assigned the following R_0 values: $\{1.0, 1.1, 1.2\}$, $\{1.3, 1.4, 1.5\}$, $\{1.6, 1.7, 1.8\}$, $\{1.9, 2.0, 2.1\}$

From the results in the simulation analysis, the MCMC algorithm begins to have difficulty converging to the fixed true values as R_0 approaches 1 (Figure 31). For values ranging from 1.3 to 2.1, the MCMC chains converge well to posterior distributions containing the true values. This is true regardless of population size in a given tract, which has little influence on how well the algorithm estimates R_0 according to the simulations. From Table 10, we see that the algorithm generates the correct estimates from 1.3 to 2.1 with low variability and posterior standard deviations ranging from 0.024 to 0.204. The table shows the summary statistics for tract-specific R_0 medians. The higher standard deviation corresponding to R_0 of 1.9 is the result of an unusually low R_0 estimate for one tract. However, it is obvious that the algorithm has difficulty estimating values of R_0 less than 1.2.

We can also assess model performance across the geographic space by comparing maps of these sets of R_0 values to Figure 30. Figure 32 demonstrates the breakdown of structure as we lower the value of R_0. Maps displaying R_0 estimated values of

$\{1.3, 1.4, 1.5\}$, $\{1.6, 1.7, 1.8\}$, $\{1.9, 2.0, 2.1\}$ mirror the bullseye effect seen in the map of the true R_0 values; however, the map of the lowest estimates does not display the expected pattern.

In addition, we run simulations for the Reed-Frost chain binomial model under the assumption of 300 initially susceptible individuals per tract, in contrast to the population considered at risk for the transition chain binomial model. We assume that issues with estimation with the observed data and large susceptible population class will also occur when simulating from true values to assess model performance. Thus, along with the assumption of 300 susceptible individuals per tract, we also assume 5 initially infectious individuals, and use the same window of four sets of three R_0 values in order to determine how well the Reed-Frost model performs when varying the values of R_0. Since all tracts are assumed to be identical in terms of population, we will not analyze estimated R_0 across population or attempt to map the results.

Under these new assumptions, we find that the algorithm also has difficulty converging when R_0 approaches 1 as observed with the transition chain binomial model. However, for values ranging from 1.3 to 2.1 in Table 11, we discover that the MCMC chains converge well to posterior distributions containing the true value. From Table 11, median and mean R_0 estimates are close to their corresponding true values and the posterior standard deviations are relatively low, ranging from 0.029 to 0.058 for distributions containing true R_0 values of 1.3 to 2.1. We conclude that the Reed-Frost model provides adequate R_0 estimates across higher values of true values of R_0, comparable to the transition chain binomial model.

Overall, we can expect the model to estimate effectively and efficiently higher values of R_0, but we will encounter difficulties with smaller values of R_0 at the tract level for both models. This is the result of the unpredictability of binomial chains generated with lower transmission probabilities and small initial infectious counts. In some situations, the initial cases generate a sustainable chain of infectious individuals

where R_0 is well defined. In other situations, the initial cases generate almost no future cases, and the resulting chain is a series of zeros, which makes estimation of the model parameters difficult. We had no issues estimating R_0 values close to 1 for the entire study space since we had a large number of initial infectious individuals. With smaller initial susceptible population sizes and slightly more initial cases, we find that the MCMC algorithm in coordination with Reed-Frost model correctly identifies most true R_0 values.

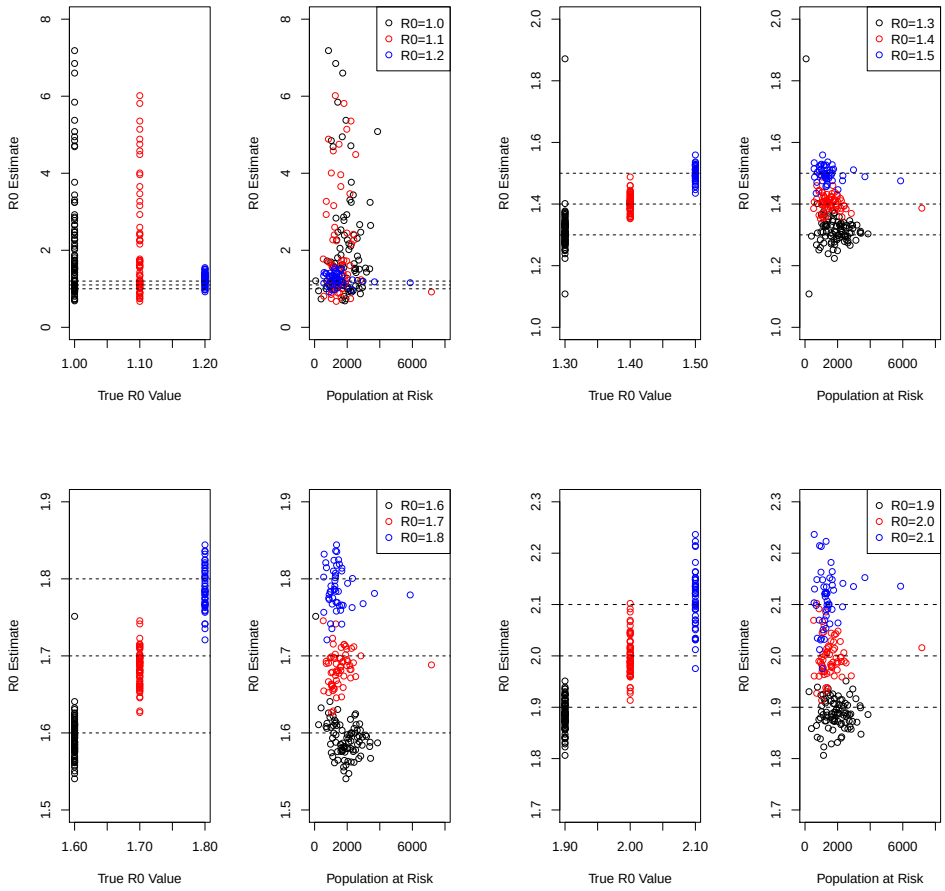

Figure 31: Comparing estimates of R_0 to the true R_0 value used to simulate binomial chain. Effects of tract population size on R_0 estimation also noted.

R_0 Value	N	Mean	St. Dev.	Median	(Min, Max)
1.0	87	2.124	1.480	1.548	(0.687, 7.184)
1.1	64	2.103	1.416	1.586	(0.674, 6.015)
1.2	49	1.249	0.154	1.262	(0.922, 1.549)
1.3	87	1.316	0.072	1.309	(1.108, 1.871)
1.4	64	1.405	0.027	1.406	(1.352, 1.488)
1.5	49	1.498	0.024	1.496	(1.435, 1.559)
1.6	87	1.593	0.027	1.589	(1.541, 1.751)
1.7	64	1.684	0.024	1.687	(1.626, 1.746)
1.8	49	1.788	0.024	1.784	(1.721, 1.844)
1.9	87	1.865	0.204	1.886	(0.000, 1.951)
2.0	64	1.999	0.040	1.991	(1.913, 2.102)
2.1	49	2.112	0.055	2.121	(1.975, 2.236)

Table 10: Measures of model performance - Comparing the fixed true R_0 to estimated R_0 value using summary statistics with the transition chain binomial model.

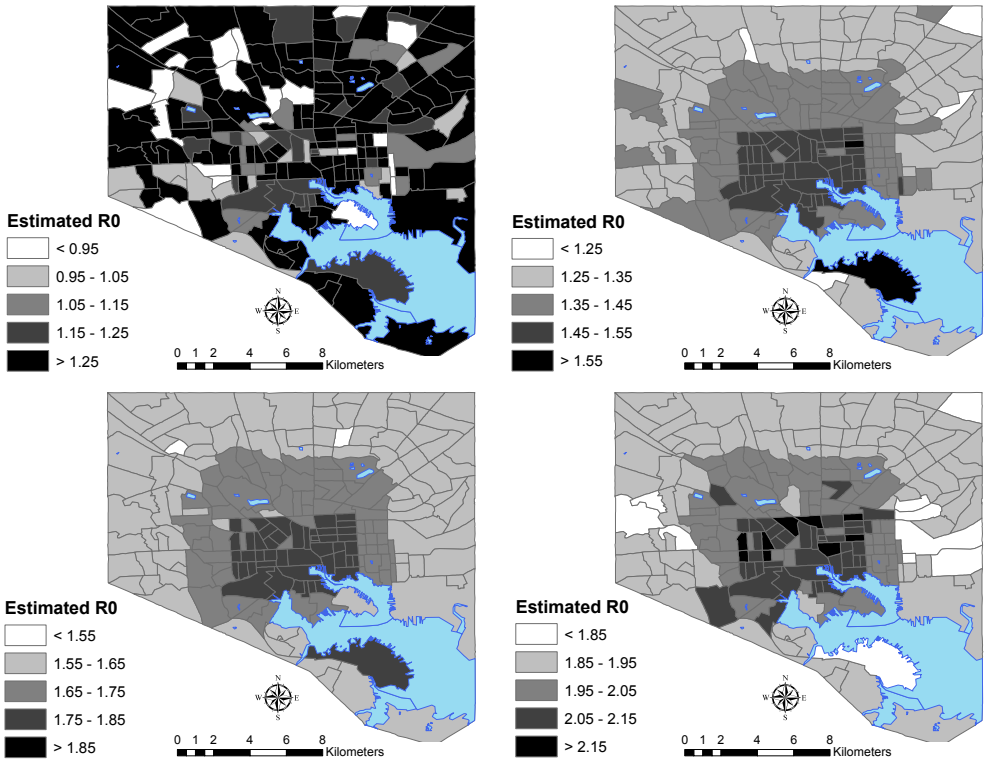

Figure 32: Estimates for R_0 across Baltimore using the following set of fixed R_0 values: $\{1.0, 1.1, 1.2\}$, $\{1.3, 1.4, 1.5\}$, $\{1.6, 1.7, 1.8\}$, $\{1.9, 2.0, 2.1\}$

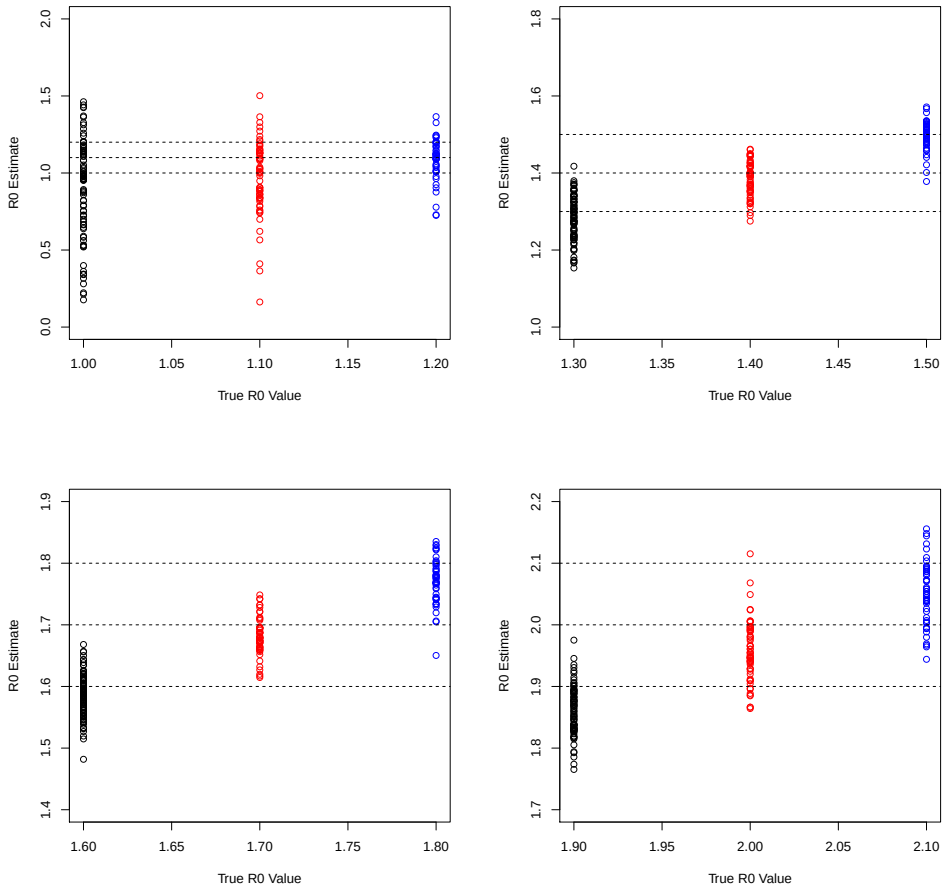

Figure 33: Comparing estimates of R_0 to the true R_0 value used to simulate Reed-Frost binomial chain. $S_{i0} = 300$ susceptible individuals per tract.

R_0 **Value**	**N**	**Mean**	**St. Dev.**	**Median**	**(Min, Max)**
1.0	87	0.902	0.318	0.965	(0.177, 1.463)
1.1	64	0.970	0.239	1.004	(0.163, 1.502)
1.2	49	1.091	0.133	1.107	(0.725, 1.366)
1.3	87	1.285	0.058	1.288	(1.153, 1.418)
1.4	64	1.386	0.045	1.393	(1.275, 1.462)
1.5	49	1.497	0.038	1.501	(1.378, 1.571)
1.6	87	1.586	0.035	1.588	(1.482, 1.668)
1.7	64	1.681	0.029	1.679	(1.615, 1.749)
1.8	49	1.774	0.037	1.774	(1.650, 1.835)
1.9	87	1.860	0.039	1.860	(1.765, 1.975)
2.0	64	1.950	0.048	1.950	(1.864, 2.115)
2.1	49	2.051	0.051	2.050	(1.944, 2.156)

Table 11: Measures of model performance - Comparing the fixed true R_0 to estimated R_0 value using summary statistics. The Reed-Frost approach with $S_{i0} = 300$ susceptible individuals per tract.

5.9 Discussion - R_0 Estimation Models

Overall, we can expect the model to estimate effectively and efficiently spatial patterns in values of R_0, but we will encounter difficulties with smaller values of R_0 at the tract level. This is the result of the unpredictability of binomial chains generated with lower transmission probabilities, small initial infectious counts, and larger numbers of individuals who are susceptible. In some situations, the initial cases can generate a sustainable chain of infectious individuals where R_0 is well defined. In other situations, the initial cases generate almost no future cases, and the resulting chain is a series of zeros, considerably complicating estimation of the model parameters. We had no issues estimating R_0 values close to 1 for the entire study space since we had a large number of initially infectious individuals. In the Reed-Frost chain binomial model, larger estimates for the number of susceptible individuals in the population result in inconsistent estimation across the varying random effects structures. We also found this to be an issue with the general epidemic model. By lowering our susceptible population size to a reasonable 10% of the total population, we find that the Reed-Frost model and general epidemic model adequately identify and quantify patterns of disease transmission and correctly implement the smoothing techniques. Despite the differences in the models and the sets of accompanying assumptions, we find roughly the same estimated pattern of disease transmission across Baltimore in this time period (Figure 34).

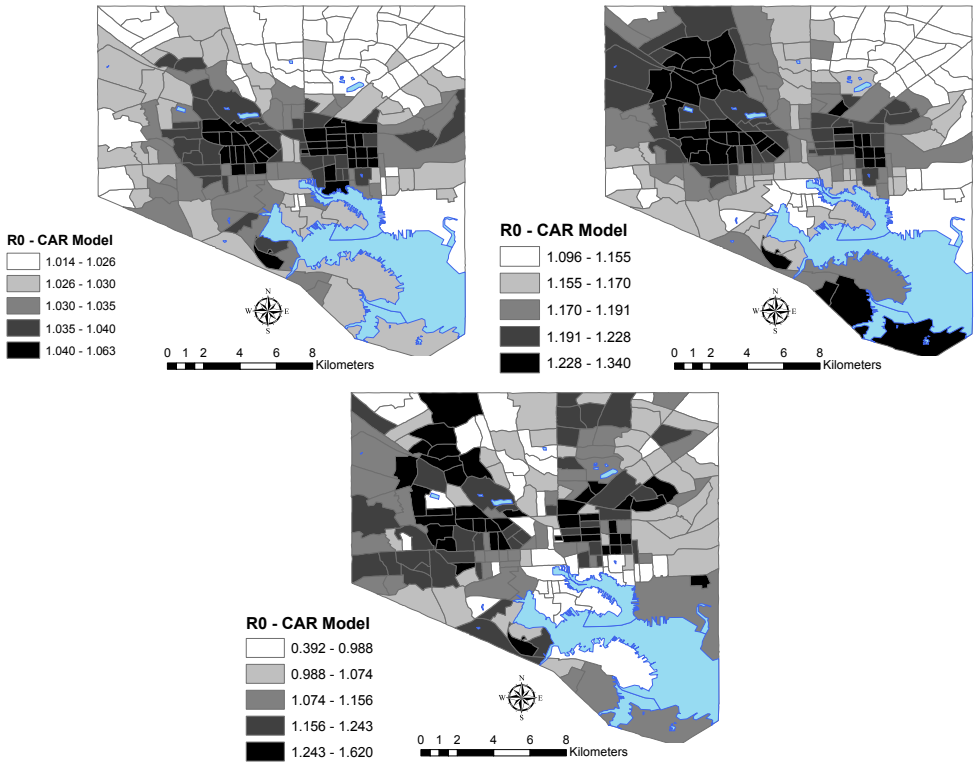

Figure 34: Local median estimates for R_0. Assessing the spatial pattern of disease transmission across model type using CAR random effects. The transition chain binomial model (upper left), Reed-Frost chain binomial model (upper right), and general epidemic model (bottom center) are shown above.

Our work demonstrates several advantages over existing models and models which do not implement correlation-inducing random effects. The Bayesian approach allows us to obtain posterior densities and credible sets for R_0 at each location. In addition, our random effects models incorporate local and global information which provide more statistically precise local estimates than crude estimation alone. One of the goals of our research is to obtain precise local estimates with adequate spatial resolution, and we believe correlation-inducing random effects provide more accurate, precise, and realistic local R_0 estimates. Our models also discovered a cluster of core area disease transmission in north-western Baltimore, which was not highlighted in previous spatial methods such as the Local Moran test and SaTScan.

Our results show that spatial estimation of epidemic model parameters such as the

transmission probability, transmission parameter β, or R_0 is feasible under the right conditions and data support. Chain binomial models require enough cases to sustain an endemic or epidemic disease for a long period of time, and more cases lead to stable estimation and less variability in our estimates, as is shown with both the observed data and simulated data examples. In our observed data model, we have shown that conditionally autoregressive random effects structures can effectively smooth the surface of parameter estimates over a study space without losing much spatial resolution from individual regions. In particular, for our motivating example of gonorrhea in census tracts in Baltimore City County, we have been able to capture clear spatial delineation of core areas of disease transmission through the use of chain binomial model estimation as well as general epidemic model estimation within the framework of Bayesian hierarchical modeling.

6 Future Work

6.1 Extensions to Existing Models

Although we chose vague priors for the precision parameters in every analysis, which yielded maps with only slight differences across random effect type, we could implement stronger priors to induce stronger correlation in our estimates. For our maps, the strength of correlation is a compromise between finding the underlying pattern of disease transmission, and allowing for accurate local estimation. We hope that the models used will assist in the local estimation of R_0 for regions with smaller numbers of cases.

Additionally, we have fixed our recovery parameter γ to be equal to 1 for each model used. This estimate was based on literature noting that the typical infectious period could range from two weeks to a few months depending on the severity of the

symptoms [32]. Future work could attempt to estimate infection times using MCMC techniques, although we note that estimation will be substantially more difficult with missing infection times to be estimated.

We also used an assortment of values for the number of initially susceptible individuals in each tract. In two of our models, the Reed-Frost chain binomial model and the general epidemic model, estimation of R_0 was more difficult assuming everyone aged 15-49 was susceptible to infection. Future analyses could look at results when varying the number of susceptible individuals from the least conservative estimate (let S_0 equal the total number of infections over time) to the most conservative estimate (let S_0 equal the number of individuals aged 15-49). This type of sensitivity analysis could be useful in determining accurate susceptible population sizes in future research.

Although we assumed an SIR model since only a very small number of individuals appeared in our dataset more than once, future work could transition towards an SIS (susceptible-infectious-susceptible) model where those who clear the infection class move back into the susceptible class. The accuracy of this model depends on having enough individuals appear more than once in the dataset, so a period of more than four years may be necessary to carry out this future analysis.

Hethcote and Yorke discussed models for control methods which involve classifying individuals based on gender, sexual activity, and whether the individual is symptomatic, resulting in 8 distinct groups [32]. A more effective and accurate model would account for these discrepancies within groups, and future analyses should focus on addressing this method of classification. They suggest closely following asymptomatics and sexually active individuals, assuming this level of information is available.

6.2 Spatially-varying Coefficient Models

Spatially-varying coefficient models represent the next step toward combining aspects of geographically weighted regression techniques with Bayesian inferential techniques. Instead of allowing the associations between disease rates and covariates to vary spatially, we can use spatially-varying coefficient models to predict transmission rates while incorporating different types of correlation structures. As with our random effects models from the chain binomial analysis, we estimate an overall mean intercept along with local deviations from the mean [53].

With more covariate baseline data on the subjects, we could attempt to adjust our epidemic model parameters for other variables. Although spatial heterogeneity in susceptible populations has been addressed, future steps could account for the effects of gender, age, and race on epidemic parameters in addition to space. Clearly, a number of factors influence the existence of varying levels of disease susceptibility in populations, and we hope to extend our epidemic models into the area spatially-varying coefficient models in order to provide additional details in the model-based inference of core areas of infectious diseases.

6.3 Identifiability Issues

With MCMC, issues with reparameterization estimation can be applicable, since we are concerned with potential correlations between the epidemic parameters β and γ. If we sample and estimate β and γ, then the ratio of the two provides the posterior density of the function of interest: R_0. Since we have fixed γ in our previous analysis, we did not consider whether β or γ are varying spatially or whether both are varying spatially. Thus, the issue of identifiability arises when we are unable to distinguish the two parameters in a multilevel process. If the available data do not shed light on a parameter, then that parameter is said to be unidentifiable. Bayesians have the abil-

ity to place proper prior distributions on the unidentified parameters [55]. However, weakly identifiable parameters are even more common than perfectly unidentifiable parameters, as the model may be structured in a way that an enormous dataset is needed in order to have the power to differentiate ecological processes [13]. In MCMC, weak identifiability leads to weak convergence - an important issue to consider in estimating epidemic model parameters. Future work could address solutions to accounting for identifiability problems, including running models with known inputs through simulations.

In conclusion, the models represent a step toward the development of hierarchical spatial models of the spread of infectious diseases. The estimation problems are challenging, but manageable under the right data support. We believe the methodology of correlation-inducing random effects within stochastic infectious disease models can continue to provide accurate and precise local estimates for epidemic model parameters in the future.

Bibliography

[1] E.J. Amundsen, H. Stigum, J.A. Roettingen, and O.O. Aalen. Definition and estimation of an actual reproduction number describing past infectious disease transmission: application to HIV epidemics among homosexual men in Denmark, Norway and Sweden. *Epidemiologic Infections*, 132(06):1139–1149, 2004.

[2] P.K. Andersen. *Statistical Models Based on Counting Processes*. Springer Verlag, 1993.

[3] R.M. Anderson and R.M. May. Population biology of infectious diseases: Part I. *Nature*, 280(5721):361–367, 1979.

[4] L. Anselin, I. Syabri, and Y. Kho. GeoDa: An introduction to spatial data analysis. *Geographical Analysis*, 38(1):5–22, 2006.

[5] F. Ball and D. Clancy. The final size and severity of a generalised stochastic multitype epidemic model. *Advances in Applied Probability*, 25(4):721–736, 1993.

[6] S. Banerjee, B.P. Carlin, and A.E. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall, 2004.

[7] K.M. Becker, G.E. Glass, W. Brathwaite, and J.M. Zenilman. Geographic epidemiology of gonorrhea in Baltimore, Maryland, using a geographic information system. *American Journal of Epidemiology*, 147(7):709–716, 1998.

[8] N.G. Becker. *Analysis of Infectious Disease Data*. Chapman & Hall/CRC, 1989.

[9] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236, 1974.

[10] J. Besag and C. Kooperberg. On conditional and intrinsic autoregressions. *Biometrika*, 82(1):733–746, 1995.

[11] J. Besag, J. York, and A. Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1–20, 1991.

[12] H.L. Beyer. Hawth's analysis tools for ArcGIS, 2004.

[13] B.M. Bolker. *Ecological Models and Data in R.* Princeton University Press, 2008.

[14] P. Brëmund. *Markov chains: Gibbs fields, Monte Carlo simulation, and queues.* Springer-Verlag, 1998.

[15] T. Britton. Epidemics in heterogeneous communities: Estimation of R_0 and secure vaccination coverage. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4):705–715, 2001.

[16] R.C. Brunham. The concept of core and its relevance to the epidemiology and control of sexually transmitted diseases. *Sexually Transmitted Diseases*, 18(2):67–68, 1991.

[17] D. Clayton and J. Kaldor. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43(3):671–681, 1987.

[18] D.G. Clayton and J. Cuzick. Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society, Series A*, 148:82–117, 1985.

[19] A.C.A. Clements, S. Brooker, U. Nyandindi, A. Fenwick, and L. Blair. Bayesian spatial analysis of a national urinary schistosomiasis questionnaire to assist geographic targeting of schistosomiasis control in Tanzania, East Africa. *International journal for parasitology*, 38(3-4):401–415, 2008.

[20] O. Diekmann, J.A.P. Heesterbeek, and J.A.J. Metz. On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases

in heterogeneous populations. *Journal of Mathematical Biology*, 28(4):365–382, 1990.

[21] K. Dietz. The estimation of the basic reproduction number for infectious diseases. *Statistical Methods in Medical Research*, 2:23–41, 1993.

[22] S.P. Ellner and J. Guckenheimer. *Dynamic Models in Biology*. Princeton University Press, 2006.

[23] ESRI v9.3: AcrGIS. ArcInfo version. ESRI Inc. Redlands, CA, 2008.

[24] C.P. Farrington, M.N. Kanaan, and N.J. Gay. Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(3):251–292, 2001.

[25] Gonorrhea - CDC Fact Sheet. http://www.cdc.gov/std/gonorrhea/STDFact-Chlamydia.htm, June 2010.

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

[27] R.A. Gunn, S. Fitzgerald, and S.O. Aral. Sexually transmitted disease clinic clients at risk for subsequent gonorrhea and chlamydia infections: Possible 'core' transmitters. *Sexually Transmitted Diseases*, 27(6):343–349, 2000.

[28] M.E. Halloran. Concepts of infectious disease epidemiology. In K.J. Rothman, S. Greenland, and T.L. Lash, editors, *Modern Epidemiology*, pages 529–554. Lippincott Williams & Wilkins, 1998.

[29] M.E. Halloran, I.M. Longini Jr, and C.J. Struchiner. Estimability and interpretation of vaccine efficacy using frailty mixing models. *American Journal of Epidemiology*, 144(1):83, 1996.

[30] M.E. Halloran, I.M. Longini Jr, and C.J. Struchiner. *Design and Analysis of Vaccine Studies*. Springer Verlag, 2009.

[31] J.A.P. Heesterbeek. A brief history of R_0 and a recipe for its calculation. *Acta Biotheoretica*, 50:189–204, 2002.

[32] H.W. Hethcote, J.A. Yorke, and A. Nold. Gonorrhea modeling: A comparison of control methods. *Biosciences*, 58:93–109, 1982.

[33] M. Höhle, E. Jørgensen, and P.D. O'Neill. Inference in disease transmission experiments by using stochastic epidemic models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(2):349–366, 2005.

[34] J.M. Jennings, F.C. Curriero, D. Celentano, and J.M. Ellen. Geographic identification of high gonorrhea transmission areas in Baltimore, Maryland. *American Journal of Epidemiology*, 161(1):73–80, 2005.

[35] M.J. Keeling and P. Rohani. *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, 2007.

[36] W.O. Kermack and A.G. McKendrick. Contributions to the mathematical theory of epidemics - I. *Proceedings of the Royal Society of Medicine*, 115A:700–721, 1927.

[37] M. Kulldorff, K. Rand, G. Gherman, G. Williams, and D. DeFrancesco. SaTScan v 8.0: Software for the spatial and space-time scan statistics. *Bethesda, MD: National Cancer Institute*, 2009.

[38] D.C.G. Law, M.L. Serre, G. Christakos, P.A. Leone, and W.C. Miller. Spatial analysis and mapping of sexually transmitted diseases to optimise intervention and prevention strategies. *Sexually Transmitted Infections*, 80(4):294–299, 2004.

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

[40] R. Levins. The strategy of model building in population biology. *American Scientist*, 54(4):421–431, 1966.

[41] I.M. Longini Jr and M.E. Halloran. A frailty mixture model for estimating vaccine efficacy. *Applied Statistics*, 45(2):165–173, 1996.

[42] P.D. O'Neill. A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. *Mathematical Biosciences*, 180(1-2):103–114, 2002.

[43] P.D. O'Neill and G.O. Roberts. Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162(1):121–129, 1999.

[44] R Development Core Team. R: A language and environment for statistical computing, reference index version 2.9.2. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org, 2009.

[45] C.P. Robert and Casella G. *Monte Carlo statistical methods*. Springer-Verlag, 2004.

[46] SAS Institute Inc. SAS v9.2. Cary, NC, 2008.

[47] R.A. Scribner, S.A. Johnson, D.A. Cohen, W. Robinson, T.A. Farley, and P. Gruenewald. Geospatial methods for identification of core groups for HIV/AIDS. *Substance Use & Misuse*, 43(2):203–221, 2008.

[48] R.J. Smith? *Modelling Disease Ecology with Mathematics*. American Institute of Mathematical Sciences, 2008.

[49] TerraSeer Inc. ClusterSeer version 2. Ann Arbor, MI, 2006.

[50] J.C. Thomas and M.J. Tucker. The development and use of the concept of a sexually transmitted disease core. *The Journal of Infectious Diseases*, 174:134–143, 1996.

[51] U. S. Census Bureau. http://www.census.gov, April 2009.

[52] L.A. Waller and B.P. Carlin. Disease mapping. In A.E. Gelfand, P.J. Diggle, M. Fuentes, and P. Guttorp, editors, *Handbook of Spatial Statistics*, pages 217–244. Chapman & Hall/CRC, 2010.

[53] L.A. Waller, L. Zhu, C.A. Gotway, D.M. Gorman, and P.J. Gruenewald. Quantifying geographic variations in associations between alcohol distribution and violence: A comparison of geographically weighted regression and spatially varying coefficient models. *Stochastic Environmental Research and Risk Assessment*, 21(5):573–588, 2007.

[54] C.R. Warden. Comparison of Poisson and Bernoulli spatial cluster analyses of pediatric injuries in a fire district. *International Journal of Health Geographics*, 7(51), 2008.

[55] Y. Xie and B.P. Carlin. Measures of Bayesian learning and identifiability in hierarchical models. *Journal of Statistical Planning and Inference*, 136(10):3458–3477, 2006.

[56] J.A. Yorke, H.W. Hethcote, and A. Nold. Dynamics and control of the transmission of gonorrhea. *Sexually Transmitted Diseases*, 5:51–56, 1978.

[57] J.M. Zenilman, N. Ellish, A. Fresia, and G. Glass. The geography of sexual partnerships in Baltimore: Applications of core theory dynamics using a geographic information system. *Sexually Transmitted Diseases*, 26(2):75–81, 1999.