Detecting epidemic coupling among geographically separated populations

Detecting epidemic coupling among geographically separated populations

Karsten Hempel, B.Sc., M.Sc.

A Thesis
Submitted to the School of Graduate Studies
in Partial Fulfillment of the Requirements
for the Degree
Doctor of Philosophy

McMaster University

DOCTOR OF PHILOSOPHY (2018)	McMaster University
(Mathematics & Statistics)	Hamilton, Ontario

TITLE:	Detecting epidemic coupling among geographically separated populations
AUTHOR:	Karsten Hempel M.Sc. (McMaster University) B.Sc. (Mount Allison University) B.A. (Mount Allison University)
SUPERVISOR:	Prof. David J. D. Earn
NUMBER OF PAGES:	[xiii], 110

ABSTRACT

The spread of infectious agents has been observed as long as their hosts have existed. The spread of infectious diseases in human populations, however, is more than an academic concern, causing millions of deaths every year, and prompting collective surveillance and intervention efforts worldwide. These surveillance data, used in conjunction with statistical methods and mathematical models, present both challenges and opportunities for advancements in scientific understanding and public health.

Early mathematical modeling of infectious diseases in humans began by assuming homogeneous contact among individuals, but has since been extended to account for many sources of non-homogeneity in human contact. Detecting the degree of epidemic mixing between geographically separated populations, in particular, remains a difficult problem. The difficulty occurs because although disease case reports have been collected by many governments for decades, case reporting is imperfect, and transmission events themselves are nearly impossible to observe.

The degree to which epidemic coupling can be detected from case reports is the central theme of this thesis. We present a careful, biologically motivated and consistent derivation of the transmission coupling (fully derived in Chapter 4). In Chapter 2 we consider the simple scenario of an epidemic spreading from one population to another, and present both numerical and analytic methodology for estimating epidemic coupling. Chapter 3 considers the problem of estimating epidemic coupling among populations undergoing recurrent epidemics, such as those of childhood diseases which have been widely observed. In Chapter 4 we present a method for estimating coupling among an arbitrary number of populations undergoing an epidemic, and apply it to

estimate coupling among the parishes of London, England, during the Great Plague of 1665.

ACKNOWLEDGEMENTS

It would be impossible to sufficiently thank Prof. David Earn for everything he has done in the years I have spent at McMaster University. When I first came in contact with David, I immediately recognized the combination of heartfelt enthusiasm and striking clarity of thought that has characterized his mentorship in all the years since then. His tireless generosity with his time and energy was compounded by his attention to detail, careful advice, and wise guidance. He has gone to tremendous lengths to encourage and facilitate contact with other researchers in my field, with other students at McMaster, including those in his research group (EarnLab), and with friends and family. His mentorship has been a profoundly positive force in my academic and personal growth, and for this I am deeply grateful.

I am also grateful for the helpful advice and careful comments provided by members of my advisory committee, Jonathan Dushoff and Ben Bolker. In addition, Prof. Dushoff's perspectives on my proposed research were of great help, and discussions with Prof. Bolker were invaluable in developing the ideas for this thesis.

I would also like to thank past and present members of Earnlab, and others with whom I have shared experiences at McMaster, namely Sarah Drohan, David Champredon, Irena Papst, Dora Rosati, Michelle Dejonge, Chai Molina, Alexandra Teslya, and Lindsay Keegan for interesting discussions, collaborative work, helpful presentation comments, and for all the good times we have shared in the many meetings and conferences we have attended together. I would also be remiss not to mention my office-mates and good friends Tyler Meadows and Alexander Chernyavsky, along with Adrien Thierry, for our many excellent adventures, and Stephen Murray for convincing me to follow him to study at McMaster in the first place. My time has been

made incalculably richer for the company of these and many others at McMaster.

I would finally like to express my gratitude to my family for their love and support throughout my time at McMaster. My father's warm and wise encouragement in every aspect of my life as I've followed in his footsteps toward a Ph.D., and my mother's continuing support and advice in building my career, have made all the difference. My siblings Andreas, Michael, Rosanna, Stephanie, Franzeska, and Thomas' infectious energy and openness have been a continuous source of motivation and joy.

I am also grateful for the funding I have received from the Department of Mathematics and Statistics, and for an Ontario Graduate Scholarship (OGS), for making my studies at McMaster possible.

My deepest gratitude to all of you.

DECLARATION OF ACADEMIC ACHIEVEMENT

The chapters of this thesis are formatted as separate manuscripts for the purpose of publication, and Chapter 2 is in preparation for submission for publication. The computer programming, mathematical analysis, and writing required for the preparation of these manuscripts was primarily undertaken by the author, with contributions in analysis and editing from David Earn.

Contents

List of Figures

List of Tables

Chapter 1

General Introduction

Human history is replete with epidemic events brought on by contact between geographically separated populations. The spread of the Black Death throughout Europe in the 14th century [1,2], the spread of smallpox, measles, and other diseases into the Americas during the colonial era [3], and the Spanish Flu beginning in the final year of World War I [4] are a few well-known and devastating examples. The increase in contact between people from different geographic regions has continued to the present day, raising the risk of explosive epidemic and pandemic events in the future. Set against this, recent decades have seen a dramatic increase both in cheap computing power and digitized epidemiological data available for research. There is both a profound need and opportunity to advance our ability to understand and predict the spatial spread of epidemics, and it is the purpose of this thesis to contribute methods in mathematical modeling for doing so.

The mathematical modeling of epidemiological systems is thought to have had its first expression in the 18th century with Daniel Bernoulli offering recommendations on the public health benefits of preventative measures against smallpox [5, 6]. A systematic approach to epidemic modeling arrived later with the concept, borrowed from physics, of approximating the contact among humans spreading measles in a population [7], or humans exposed to malaria-infected mosquitos, with the "law of mass action". This approach yielded a result that has become central to the field of mathematical epidemiology, first derived by Kermack and McKendrick [8] as the *epidemic threshold*[1]. Kermack and MacKendrick divided the population into susceptible, infected, and recovered individuals, an approach now widely referred to as the *susceptible-infected-removed* (SIR) model [9–11]. The SIR model, and variants derived from it, have been used in investigations of many characteristics of infectious disease spread in humans [9, 12–19]. It has been extended to account for heterogeneous population mixing due to separation into geographic regions [20–35], age structure [26, 36–40], and social network structure [41–45], to name a few.

This thesis is concerned with modeling the geographic spread of epidemics, focusing on the problem of estimating the degree of coupling between geographically separated populations. There is a large body of work studying spatially structured SIR models [20, 24, 46–48]. Spatial structure is sometimes represented with a *meta-population*, where a spatial region is separated into discrete areas with local populations [20–27, 27–35, 46, 47, 49]. Other times space is represented as continuous [50–54]. Grenfell et al. [24] implement a spatial version of a previously developed TSIR model [55, 56], a discrete time SIR model[2]. Among other things, they found that large population centres drove epidemics in smaller population centres among cities in England and Wales. Viboud et al. similarly studied the phase of recurrent

[1] The epidemic threshold threshold is now encapsulated in the basic reproduction number, $\mathcal{R}_0$. $\mathcal{R}_0$ is defined as the average number of new infections that will be caused by a single infection in a population which is otherwise completely susceptible to disease. Thus when $\mathcal{R}_0 > 1$, a small number of infections is expected to grow, resulting in an epidemic.

[2] The TSIR model used by Grenfell et al. [24] is a discrete time dynamical system model, where the time step is two weeks. This time-step was well suited for the spatially structure measles data the authors used, since measles has a combined latent and infectious period of approximately two weeks, and the data were weekly case reports.

influenza epidemics spreading through US cities, and used data regarding volumes of inter-city travel to replicate observed patterns [46, 47].

The approach presented in this thesis uses a continuous-time SIR meta-population model intended to be generalizable to any disease for which the SIR model is appropriate. The input data are assumed to be either case or mortality reports (simulated mock data throughout the thesis, and real-world data in Chapter 4). Our implementation of meta-population cross-coupling is formalized with a contact matrix [9], in which we define entries to be the proportion of time residents of any infected status in one geographic location spend visiting another.

Simulation models can be fitted to digitized real-world case or mortality reports, after which one can investigate interventions and future predictions theoretically without running real-world experiments. Such models are fitted by finding parameters which best predict the given data, where this best prediction is found using one of a few statistical frameworks [57]. The fitting method presented in this thesis is generally classified as maximum likelihood estimation with *probe-matching*, whereby optimal model parameters are found by fitting to a summary statistic that reduces the number of dimensions of the raw data [58]. We consider three types of data sets in Chapters 2, 3, and 4, with a different summary statistic in each case.

In Chapter 2 we investigate a simple scenario in which two coupled populations are invaded by infection. The first population begins with one or more infected individuals, and as the epidemic in the first population grows, infection spreads to the second population. We pose the question of how well the degree of coupling between these populations can be estimated merely from the time to invasion of the second population. We obtain analytic formulae for estimating coupling, which we compare with results from numerical methods. The analytic formulae have the advantage of being computationally cheap, and can quickly find initial estimates of coupling which can be refined afterward if necessary.

In Chapter 3 we investigate a more complicated scenario than in Chapter 2, wherein two populations undergoing *recurrent* epidemics are coupled. This chapter is motivated by the well-studied phenomenon of hierarchical recurrent epidemics, wherein an endemically infected large population re-infects and drives epidemics in smaller populations [24, 59–62]. Keeling and Rohani in particular examine coupling between two equally sized populations undergoing endemic recurrent epidemics [62], but note the difficulty of inferring coupling in the presence of the complex dynamics that such systems are known to exhibit [16, 63]. Chapter 3 explores the feasibility of estimating the degree of coupling between two differently-sized populations undergoing recurrent epidemics [17], and with regular fadeouts in the smaller of the populations.

Chapter 4 is a case study in the spread of plague throughout the city of London, England, in 1665. The so-called "Great Plague" was recorded in the London Bills of Mortality (LBoM), which have been completely digitized by David Earn's research group at McMaster University (see [64] for previous work based on these data). The Great Plague was the last and largest of many that had hit the city since the arrival of plague in Europe in the 14th century [65–67]. Thanks to the digitization of the LBoM, we have weekly plague death totals for 130 of London's parishes for the full duration of the epidemic. We investigate the importance of geographic location in the spread of the epidemic by fitting our coupled meta-population model to the distribution of times when parishes reported their first plague deaths.

Chapter 5 summarizes and discusses the major results of the thesis, and discusses potential avenues of future research.

Chapter 2

Estimating epidemic coupling between populations from the time to invasion

Abstract

Identifying the mechanisms by which diseases spread among populations is important for understanding and forecasting patterns of epidemics and pandemics. Estimating transmission coupling among populations is challenging because transmission events are difficult to observe in practice, and connectivity among populations is often obscured by local disease dynamics. We consider the common situation in which an epidemic is seeded in one population and later spreads to a second population. We present a method for estimating transmission coupling between the two populations, assuming they can be modeled as *susceptible-infected-recovered* (SIR) systems. We show that the strength of coupling between the two populations can be estimated from the time taken for the disease to invade the second population. Confidence in the estimate is low if only a single invasion event has been observed, but is substantially improved if numerous independent invasion events are observed. Our analysis of this simplest, idealized scenario represents a first step toward developing and verifying methods for estimating epidemic coupling among populations in an ever-more-connected global human population.

2.1 Introduction

Mechanistic mathematical models are powerful tools for understanding and predicting how infectious diseases spread in human populations [9, 15–18]. The spread of infections in well-mixed populations has been extensively studied, and continuing research is tackling the effects of seasonal forcing [13, 68, 69], intensity and duration of infectiousness [70–75], and contact network structure [41–44].

One area of research that is important for public health policy is forecasting the spatial spread of diseases, which can be greatly advanced by improving estimates of model parameters from real-world data. Estimating parameters of spatial epidemic models is especially difficult [24, 47, 48], even for the well-studied, highly idealized class of meta-population models [20–22, 28, 31, 34, 44, 63, 76–78]. Here, we consider the simplest meta-population consisting of individuals who reside in one of two "habitat patches" (*e.g.,* cities). We suppose an epidemic begins in one patch, and we attempt to estimate the degree of spatial coupling to the population in the second patch. In this situation, we investigate whether we can successfully estimate the magnitude of coupling using the observed time taken for the second patch to be infected (the *time to invasion*, t_{inv}).

The specific meta-population model that we use is a two-patch *susceptible-infectious-recovered* (SIR) model (§2.2). We consider both deterministic and stochastic versions of this model (§2.2) and show that the distribution of times to invasion can be approximated analytically from model parameters (§2.3.1). We then show how, in the presence of stochasticity, the degree of coupling can be estimated using a maximum likelihood approach based on one or more observations of t_{inv} (§2.3.4).

2.2 Two-population SIR model

In the absence of coupling, we assume that disease dynamics in each patch evolve according to the standard SIR model,

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta S \frac{I}{N} \tag{2.1a}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \beta S \frac{I}{N} - \gamma I \tag{2.1b}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I \,. \tag{2.1c}$$

The three state variables represent the numbers of individuals who are susceptible to infection (S), currently infected and infectious (I), and recovered and immune (R). The total population size, $N = S + I + R$, is necessarily constant (since $\mathrm{d}N/\mathrm{d}t = 0$). The two disease parameters are the rate of transmission (β) and the rate at which infected individuals recover (γ). The ***force of infection*** is

$$\Lambda = \beta \frac{I}{N} \,. \tag{2.2}$$

The ***basic reproduction number***, the average number of secondary cases that result from a single primary case in a completely susceptible population [9], is

$$\mathcal{R}_0 = \frac{\beta}{\gamma} \,. \tag{2.3}$$

If we take the time unit to be the mean infectious period ($1/\gamma$) then $\mathcal{R}_0$ is the only disease parameter. Implicit in Equation (2.1) are assumptions that recovered individuals remain immune permanently and that vital dynamics (births and deaths) can be ignored (both these assumptions are reasonable for most infectious diseases on the timescale of invasion that concerns us here). In addition, the population in any given patch is assumed to be homogeneously mixed.

2.2.1 Form of transmission coupling

We assume that coupling of disease dynamics between the two patches arises because residents of one patch sometimes visit the other patch temporarily. We model this with a ***coupling matrix*** $c = (c_{ij})$, where c_{ij} is the proportion of the residents of patch j visiting patch i at any time.[1] Since we are considering only two patches, and the entries are proportions, the most general coupling matrix is

$$c = \begin{pmatrix} 1-m_1 & m_2 \\ m_1 & 1-m_2 \end{pmatrix}, \tag{2.4}$$

where $0 \leq m_i \leq 1$. Note that with only two patches, if the focal patch is i then the other patch is $j = 3-i$. Thus, using subscripts on state variables to identify *populations* (*i.e.,* the patches in which individuals are *resident*), the number of individuals in patch i at any time is

$$(1-m_i)N_i + m_jN_j\,, \qquad i = 1,2\,, \qquad j = 3-i, \tag{2.5}$$

and the number of those that are currently infected is

$$(1-m_i)I_i + m_jI_j\,, \qquad i = 1,2\,, \qquad j = 3-i. \tag{2.6}$$

[1] Similar formulations of cross-coupling can be found in literature, such as Murray and Cliff, 1977 [27], Lloyd and May, 1996 [35], Lloyd and Jansen [79]. We derive our formulation of coupling on a meta-population fully in §4.3.2, which we omit here since we are dealing only with two populations.

The force of infection on *residents* of patch i arises from interactions that occur in both patches. For the $(1-m_i)S_i$ susceptibles who are resident in patch i and currently located in patch i, the force of infection is

$$\beta\frac{(1-m_i)I_i+m_jI_j}{(1-m_i)N_i+m_jN_j}\,, \qquad i=1,2\,, \qquad j=3-i. \tag{2.7}$$

whereas the force of infection on the m_iS_i susceptible residents of patch i who are currently in patch j is

$$\beta\frac{m_iI_i+(1-m_j)I_j}{m_iN_i+(1-m_j)N_j}\,, \qquad i=1,2\,, \qquad j=3-i. \tag{2.8}$$

The total force of infection on residents of patch i is the sum of these two contributions, namely

$$\begin{aligned}\Lambda_i \;=\; \beta\left[(1-m_i)\frac{(1-m_i)I_i+m_jI_j}{(1-m_i)N_i+m_jN_j}+m_i\frac{m_iI_i+(1-m_j)I_j}{m_iN_i+(1-m_j)N_j}\right]\\ i=1,2\,, \qquad j=3-i.\end{aligned} \tag{2.9}$$

This formulation avoids the need to explicitly model the movements of individuals among populations (as is sometimes done [34]).

2.2.2 Deterministic model

Our two-population model is, for $i=1,2$,

$$\frac{\mathrm{d}S_i}{\mathrm{d}t}=-S_i\Lambda_i\,, \tag{2.10a}$$

$$\frac{\mathrm{d}I_i}{\mathrm{d}t}=S_i\Lambda_i-\gamma I_i\,, \tag{2.10b}$$

$$\frac{\mathrm{d}R_i}{\mathrm{d}t}=\gamma I_i\,, \tag{2.10c}$$

where Λ_i is defined in Equation (2.9) and the (constant) size of each population is $N_i = S_i + I_i + R_i$ for $i = 1, 2$.

If all individuals are initially susceptible and a resident of patch i is infected then an epidemic will occur (in population i) if the number of cases in population i is initially increasing, *i.e.*, if $\mathrm{d}I_i/\mathrm{d}t > 0$ in the limit that $S_i \to N_i$ and $I_i \to 0$ (given $S_j = N_j$ and $I_j = 0$). Retaining the notation $\mathcal{R}_0$, as in Equation (2.3), for the basic reproduction number of the uncoupled model ($m_1 = m_2 = 0$), and defining $\mathcal{R}_{i,j}$ via

$$\mathcal{R}_{i,i} = \mathcal{R}_0 \left[\frac{(1-m_i)^2 N_i}{(1-m_i)N_i + m_j N_j} + \frac{m_i^2 N_i}{m_i N_i + (1-m_j)N_j} \right], \tag{2.11a}$$

$$\mathcal{R}_{i,j} = \mathcal{R}_0 \left[\frac{(1-m_i) m_j N_i}{(1-m_i)N_i + m_j N_j} + \frac{m_i (1-m_j) N_i}{m_i N_i + (1-m_j)N_j} \right], \tag{2.11b}$$

we can rewrite Equation (2.10b)

$$\frac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} I_1 \\ I_2 \end{pmatrix} = \left(\begin{bmatrix} \mathcal{R}_{1,1} & \mathcal{R}_{1,2} \\ \mathcal{R}_{2,1} & \mathcal{R}_{2,2} \end{bmatrix} \gamma - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \gamma \right) \begin{pmatrix} I_1 \\ I_2 \end{pmatrix}, \tag{2.12}$$

from which it follows that the next generation matrix [80, 81] is

$$\begin{bmatrix} \mathcal{R}_{1,1} & \mathcal{R}_{1,2} \\ \mathcal{R}_{2,1} & \mathcal{R}_{2,2} \end{bmatrix}. \tag{2.13}$$

The spectral radius of this matrix, *i.e.*, the basic reproduction number of the two-patch system, is

$$\rho = \frac{\mathcal{R}_{1,1} + \mathcal{R}_{2,2}}{2} + \sqrt{\mathcal{R}_{1,2}\mathcal{R}_{2,1} + (\mathcal{R}_{1,1} - \mathcal{R}_{2,2})^2}\,. \tag{2.14}$$

In the special case that $N_1 = N_2$ and $m_1 = m_2\,(\equiv m)$, Equation (2.11) reduces to

$$\mathcal{R}_{i,i} = \mathcal{R}_0[1 - 2m(1-m)]\,, \tag{2.15a}$$

$$\mathcal{R}_{i,j} = \mathcal{R}_0 2m(1-m)\,, \tag{2.15b}$$

and the spectral radius (2.14) simplifies to

$$\rho = \mathcal{R}_0\,, \tag{2.16}$$

i.e., the basic reproduction number of the two-patch system is the same as that of the single patch system. In this case, there is a simple partitioning of $\mathcal{R}_0$:

$$\mathcal{R}_0 = \mathcal{R}_{i,i} + \mathcal{R}_{i,j}\,. \tag{2.17}$$

In addition, note that

$$\mathcal{R}_{i,j} \;=\; \mathcal{R}_{i,i} - (1-2m)^2\mathcal{R}_0 \;\le\; \mathcal{R}_{i,i}\,, \tag{2.18}$$

i.e. the reproduction number is higher when considering transmission within a patch as opposed to between patches.

2.2.3 Stochastic model

If the ODEs are not solved directly, but are instead used to define event rates for the corresponding stochastic process, then there is a distribution of possible times to invasion (t_{inv}). We simulate the stochastic model using the standard “tau-leaping” adaptive time-step algorithm [82].

We define the time between the first appearance of one infection in the first population ($I_1 = 1$, $t = 0$), and the first appearance of one infection in the second population ($I_2 = 1$, $t > 0$), to be the ***time to invasion***, t_{inv}. Since the ordinary differential equations (ODEs) in Equation (2.10) have a unique solution associated with any given initial state, there is exactly one value of t_{inv} associated with each parameter set ($\{\beta, \gamma, N_1, N_2, m_1, m_2\}$). In Figure 2.1, we show a single realization of the model, and the corresponding time to invasion t_{inv}.

2.2.4 Notation summary

Our notation for variables and parameters, and the initial conditions used in all simulations and analyses, are summarized in Tables 2.1, 2.2, and 2.3. All our simulations were performed with equal populations in the two patches ($N_1 = N_2$). We also restrict attention to symmetric coupling ($m_1 = m_2$), so there is only one ***coupling parameter*** m.

Variable	Description
t	Time in units of the mean infectious period, $1/\gamma$
S_1, S_2	Number of susceptible individuals in each population
I_1, I_2	Number of infected individuals in each population
R_1, R_2	Number of removed individuals in each population

Table 2.1

Parameter	Range	Description
β	> 0	Transmission rate
$\mathcal{R}_0$	> 0	Basic reproduction number of the disease
γ	> 0	Rate of recovery from infection
m_1, m_2	$\in [0, 1]$	Transmission coupling between populations
N_1, N_2	10^5	Total number of individuals in each population

Table 2.2

Initial Condition	Value
$S_1(0)$	$N_1 - I_1(0)$
$S_2(0)$	N_2
$I_1(0)$	≥ 1
$I_2(0)$, $R_1(0)$, $R_2(0)$	0

Table 2.3

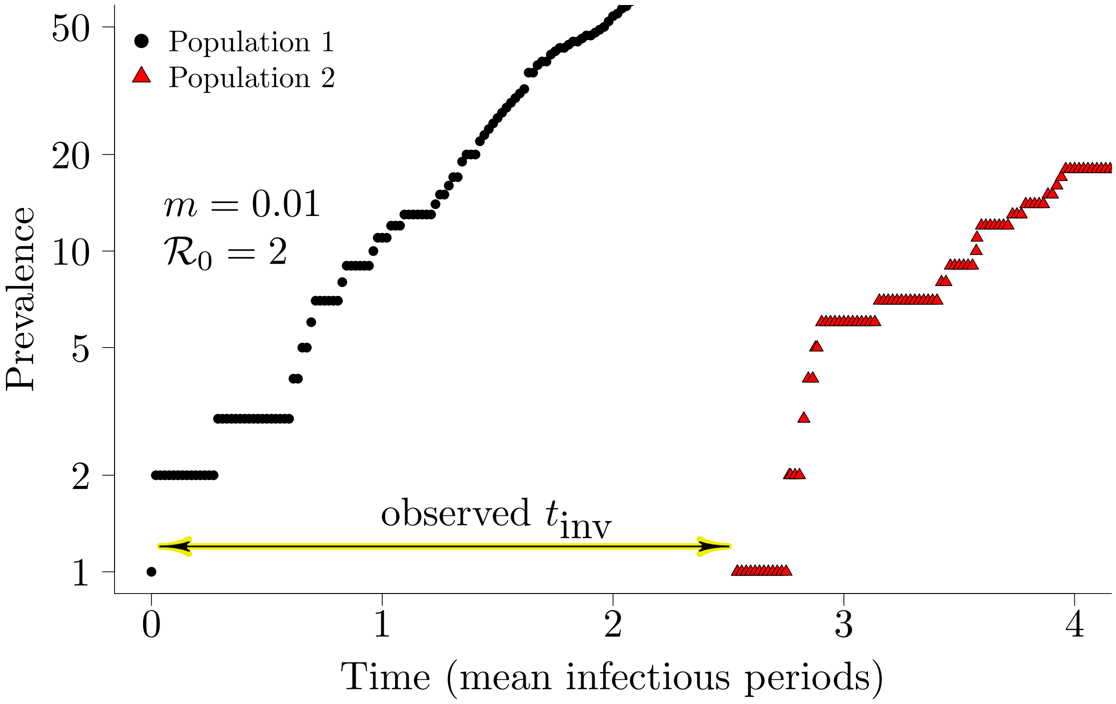

Figure 2.1: The ***time to invasion***, t_{inv}, is the time between an initial infection in one population and the first case that appears in the other population. The figure shows a single realization of the stochastic SIR model, generated using the Gillespie algorithm [83, 84] (see §2.2). Parameter values were $m = 0.01$, $\mathcal{R}_0 = 2$, $N_1 = N_2 = 10^5$.

2.3 Stochastic time to invasion

The distribution of the time to invasion (t_{inv}) is shown in Figure 2.2 for four parameter sets ($\mathcal{R}_0 = 2, 4$, $m = 0.01, 0.1$). The histograms are each based on 10,000 stochastic simulations [82]. The red curves show an analytical approximation that we derive below in §2.3.1. We present numerically computed and analytically approximated maximum likelihood estimates (MLEs) for the coupling parameter m, given observation(s) of t_{inv}, in §2.3.4 and §2.3.5.

2.3.1 Analytical approximation of time to invasion distribution

Suppose that at time $t = 0$ the system is in the initial state specified in Table 2.3, *i.e.,* there is a small number of individuals infected in the ***source population*** (population 1). We are interested in the time t_{inv} at which a first infection occurs in the ***target population*** (population 2). Until that time, there are no infections in population 2 and we will assume that t_{inv} is sufficiently short that susceptible depletion in population 1 is negligible. Thus, for $0 \le t \le t_{\text{inv}}$ we have $I_2(t) = 0$ and $S_1(t) \simeq N_1$, so—if we ignore demographic stochasticity[2] in population 1—Equation (2.10b) with $i = 1$ implies that for $0 \le t \le t_{\text{inv}}$ we can approximate the population 1 dynamics with the single equation,

$$\frac{\mathrm{d}I_1}{\mathrm{d}t} = r_1 I_1 \,, \tag{2.19}$$

where

$$r_1 \equiv \gamma(\mathcal{R}_{1,1} - 1)\,, \tag{2.20}$$

and $\mathcal{R}_{i,i}$ is defined in Equation (2.11a). Our approximation is therefore

$$I_1(t) = I_1(0)\, e^{r_1 t}\,, \qquad 0 \le t \le t_{\text{inv}}. \tag{2.21}$$

Given Equation (2.21), and that no infections have occurred yet in population 2 (i.e., $S_2 = N_2,\ I_2 = 0$), Equation (2.10b) with $i = 2$ specifies the (mean field[3]) rate at

[2] In the stochastic setting, with probability $(1/\mathcal{R}_{1,1})^{I_1(0)}$, an outbreak in population 1 fizzles out without causing a full blown epidemic [85, §7.6.2, p. 321]. Nevertheless, the second population is sometimes infected before the outbreak fizzles out in the first population. This effect is larger for lower $\mathcal{R}_0$, and for sufficiently small $\mathcal{R}_0$ must be taken into account to understand the expected distribution of t_{inv}. We ignore fizzles in our analysis, but in Figures 2.2 and 2.3 we indicate the number of simulations that fizzled and were therefore ignored.

[3] The *mean field* refers to the ensemble mean of all stochastic realizations.

which infection events occur in population 2,

$$\mu(t) = \frac{\mathrm{d}I_2}{\mathrm{d}t} = N_2\Lambda_2 = \mu_0\, e^{r_1 t}\,, \tag{2.22a}$$

$$\text{where} \qquad \mu_0 = I_1(0)\,\gamma\,\mathcal{R}_{2,1}\,, \tag{2.22b}$$

and $\mathcal{R}_{2,1}$ is defined in Equation (2.11b).[4]

In a small time interval $[t, t+\Delta t)$, we can assume that rate $\mu(t)$ is constant so the probability that an infection occurs in population 2 in this time interval is

$$\int_0^{\Delta t} \mu\, e^{-\mu s}\, \mathrm{d}s = 1 - e^{-\mu \Delta t} \simeq \mu\Delta t\,, \tag{2.23}$$

and this is therefore also the probability that t_{inv} lies in the interval $[t, t+\Delta t)$ *given* that an infection in population 2 has not already occurred, *i.e.,*

$$\text{Prob}(t \leq t_{\text{inv}} < t + \Delta t \,|\, t_{\text{inv}} \geq t) \simeq \mu\Delta t\,. \tag{2.24}$$

If we now denote the probability that invasion of population 2 occurs *before* time t by

$$F(t) = \text{Prob}(0 \leq t_{\text{inv}} < t)\,, \tag{2.25}$$

i.e., F is the cumulative distribution function for t_{inv}, then the probability that invasion occurs *after* time t is

$$\text{Prob}(t_{\text{inv}} \geq t) = 1 - F(t)\,. \tag{2.26}$$

[4] In the derivation that follows, we assume that the incidence in population 1 must be approximated in order to estimate the distribution of the time to invasion, t_{inv}. However, if the actual trajectory of incidence in population 1 is known, then this distribution can be computed exactly, since the force of infection on population 2 can be calculated at each point in time.

[5] In general, we have

$$\mathrm{Prob}(t \le t_{\mathrm{inv}} < t + \Delta t) = \mathrm{Prob}(t_{\mathrm{inv}} \ge t) \times \mathrm{Prob}(t \le t_{\mathrm{inv}} < t + \Delta t \,|\, t_{\mathrm{inv}} \ge t)\,, \quad (2.27)$$

and hence

$$F(t + \Delta t) - F(t) \simeq \big[1 - F(t)\big]\mu(t)\Delta t\,. \quad (2.28)$$

Dividing by Δt and taking the limit $\Delta t \to 0$ we have

$$F'(t) = \big[1 - F(t)\big]\mu(t)\,, \qquad F(0) = 0. \quad (2.29)$$

This is a separable first order ODE for $F(t)$, the solution of which is

$$F(t) = 1 - \exp\Big[-\int_0^t \mu(s)\,\mathrm{d}s\Big]\,. \quad (2.30)$$

Consequently, we can approximate the probability density function for t_{inv} by $f(t) = F'(t)$, *i.e.*,

$$f(t) = \mu(t)\exp\Big[-\int_0^t \mu(s)\,\mathrm{d}s\Big]\,. \quad (2.31)$$

Inserting Equation (2.22a) in Equations (2.30) and (2.31) we obtain

$$F(t) = 1 - \exp\Big[\frac{\mu_0}{r_1}\big(1 - e^{r_1 t}\big)\Big]\,, \quad (2.32)$$

and

$$f(t) = \mu_0\,\exp\Big[r_1 t + \frac{\mu_0}{r_1}\big(1 - e^{r_1 t}\big)\Big]\,. \quad (2.33)$$

Recall from Equations (2.11), (2.20) and (2.22b) that r_1 and μ_0 depend implicitly on

[5] The derivation presented here follows along the lines of standard survival analysis, where our hazard function is characterized by the force of infection on population 2 by population 1. See, for example, Cox and Oakes, 1984 [86, pp. 13].

m_1 and m_2; this is important because we will need to think of f as a function of the coupling parameter(s) later.

2.3.2 Approximation error in time to invasion distribution

Our analysis leading to Equation (2.33) was based on the approximation of pure exponential growth of cases in the first population. We can better appreciate the approximation that is being made if we recognize that the underlying process is a continuous-time branching process in the early phase during which it behaves like a simple birth-death process. During this phase, the ensemble mean number of cases in population 1 can be approximated with Equation (2.21) and the associated variance is [85, p. 250]

$$\mathrm{var}[I_1](t) = I_1(0)\, e^{r_1 t}(e^{r_1 t} - 1)\,. \tag{2.34}$$

To approximate the standard deviation in the force of infection from population 1 to population 2 (which we denote by σ), we scale as in Equation (2.22), *i.e.*,

$$\sigma(t) = \sigma_0 \sqrt{e^{r_1 t}(e^{r_1 t} - 1)}\,, \tag{2.35a}$$

$$\text{where} \qquad \sigma_0 = \sqrt{I_1(0)}\, \big(\gamma\, \mathcal{R}_{2,1}\big). \tag{2.35b}$$

We can indicate uncertainty in our analytical approximation (2.33) by replacing

$$\mu(t) \quad \longrightarrow \quad \mu(t) + \alpha\, \sigma(t) \tag{2.36}$$

in Equation (2.31), and then, for each t, finding the maximum and minimum values of $f(t)$ for α in some specific range. Details of this calculation are given in Appendix A. The thin dashed blue lines in Figures 2.2 and 2.3 indicate uncertainty in $f(t)$

obtained for $\alpha \in [-0.5, 0.5]$. Note that while the dashed blue curves emphasize that the time to invasion distribution is only approximately given by the solid blue curve, they do not represent formal confidence limits; the "α level" specified in (2.36) does not translate into a confidence limit on $f(t)$.

2.3.3 Comparison of simulations and analytical approximation

For four different parameter sets, Figure 2.2 compares the approximate density function (2.33) with the t_{inv} distribution obtained from $10,000$ realizations of the fully stochastic model[6]. As expected from the approximate formula (2.33), the probability density for t_{inv} is sensitive to both the underlying transmissibility of the pathogen ($\mathcal{R}_0$) and the degree of transmission coupling between the two patches (m).

The discrepancy between the simulations and analytical approximation in Figure 2.2 results from variance in the epidemic curve in population 1, which is less important when the initial number of cases in population 1 is larger. To see this, note from Equations (2.22) and (2.35) that the coefficient of variation in the force of infection in population 2 is

$$\frac{\sigma(t)}{\mu(t)} = \frac{\sqrt{1 - e^{-r_1 t}}}{\sqrt{I_1(0)}}\,, \tag{2.37}$$

which decreases rapidly with $I_1(0)$. Figure 2.3 shows that as $I_1(0)$ is increased, the analytical approximation of the t_{inv} distribution converges to the histogram obtained from simulations. A standard measure of the difference between two continuous

[6] We keep a stochastic simulation only if two conditions are satisfied: (i) the second population is eventually infected ($I_2(t) > 0$ for some $t > 0$), and (ii) the first population does not fizzle. We consider the outbreak to have fizzled in population 1 if the prevalence in that population drops to zero before the cumulative proportion of the population infected reaches the level corresponding to the peak of the deterministic epidemic curve. The number of susceptibles in the first population, $S_1(t)$, does not increase, and decreases as individuals become infected. After the time t when the condition $S_1(t) < \frac{N_1}{\mathcal{R}_1}$ is satisified, $\frac{\mathrm{d}I_1}{\mathrm{d}t}$ remains strictly negative. Thus the condition to avoid fizzles is $I_1(t) = 0$ for $t > 0$ and $\frac{S_1(t)}{N_1} < \frac{1}{\mathcal{R}_1}$. (*cf.* Equations (2.10b) and (2.11)).

probability distributions p and q is the Kullback-Leibler (K-L) divergence [87, p. 6],

$$D_{\mathrm{KL}}(p\|q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx\,. \tag{2.38}$$

We define $q(x)$ to be the heights of the histogram bins, produced from stochastic simulations, in Figure 2.3. $p(x)$ is Equation (2.33) evaluated at the histogram bin midpoints. We use the K-L divergence to show the convergence of the analytic approximation of the t_{inv} probability distribution to the distribution obtained from simulations in Figure 2.4.

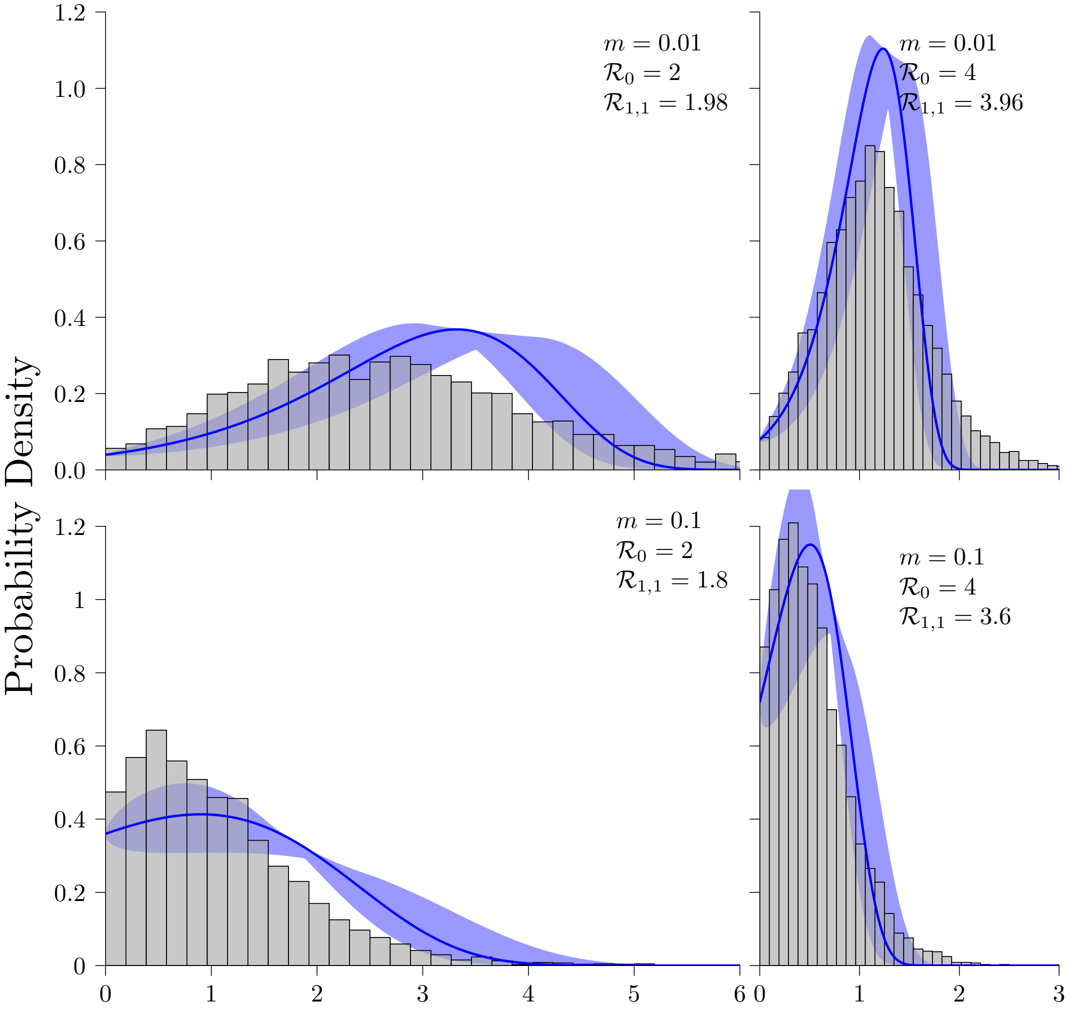

Figure 2.2: The probability density function for the time to invasion (t_{inv}, in units of the mean infectious period) estimated for four parameter sets ($\mathcal{R}_0 = 2, 4$; $m = 0.01, 0.1$; $N_1 = N_2 = 10^5$; $\mathcal{R}_{1,1}$ from Equation (2.11)). A single infectious individual is assumed in population 1 at time 0 ($I_1(0) = 1$). Grey bars show the estimated density based on a frequency histogram constructed from 10^4 stochastic simulations [82] that did not fizzle (see footnotes in §2.3.1 and §2.3.3). Solid blue curves show the analytical approximation (2.33). Pale blue bands indicate uncertainty in the approximation, based on Equation (2.46) with $\alpha \in [-0.5, 0.5]$.

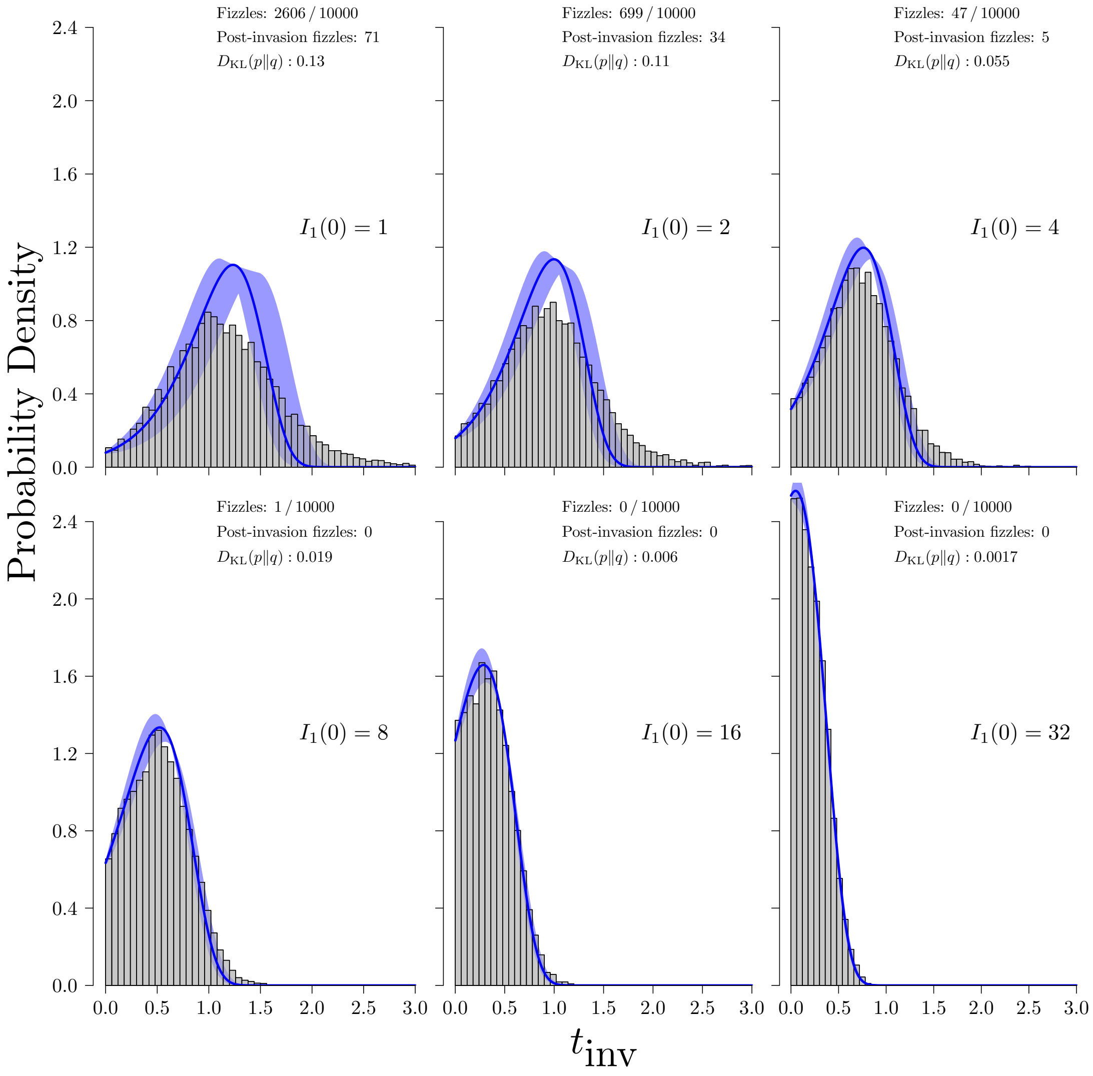

Figure 2.3: Probability density functions of the time to invasion t_{inv}, as in Figure 2.2, but for a single parameter set ($\mathcal{R}_0 = 4$, $m = 0.01$, $N_1 = N_2 = 10^5$). The six panels differ in the initial numbers of infectives in population 1 ($I_1(0) \in \{1, 2, 4, 8, 16, 32\}$). Only simulations in which infection successfully spread to the second population and did not fizzle out in the first population are shown (in grey); *cf.* footnote in §2.3.3. $D_{\mathrm{KL}}(p\|q)$ refers to the Kullback-Liebler divergence (*cf.* Equation (2.38) and [87]), and shows the analytical approximation error when compared to the probability density estimated from 10^4 stochastic simulations (2.33).

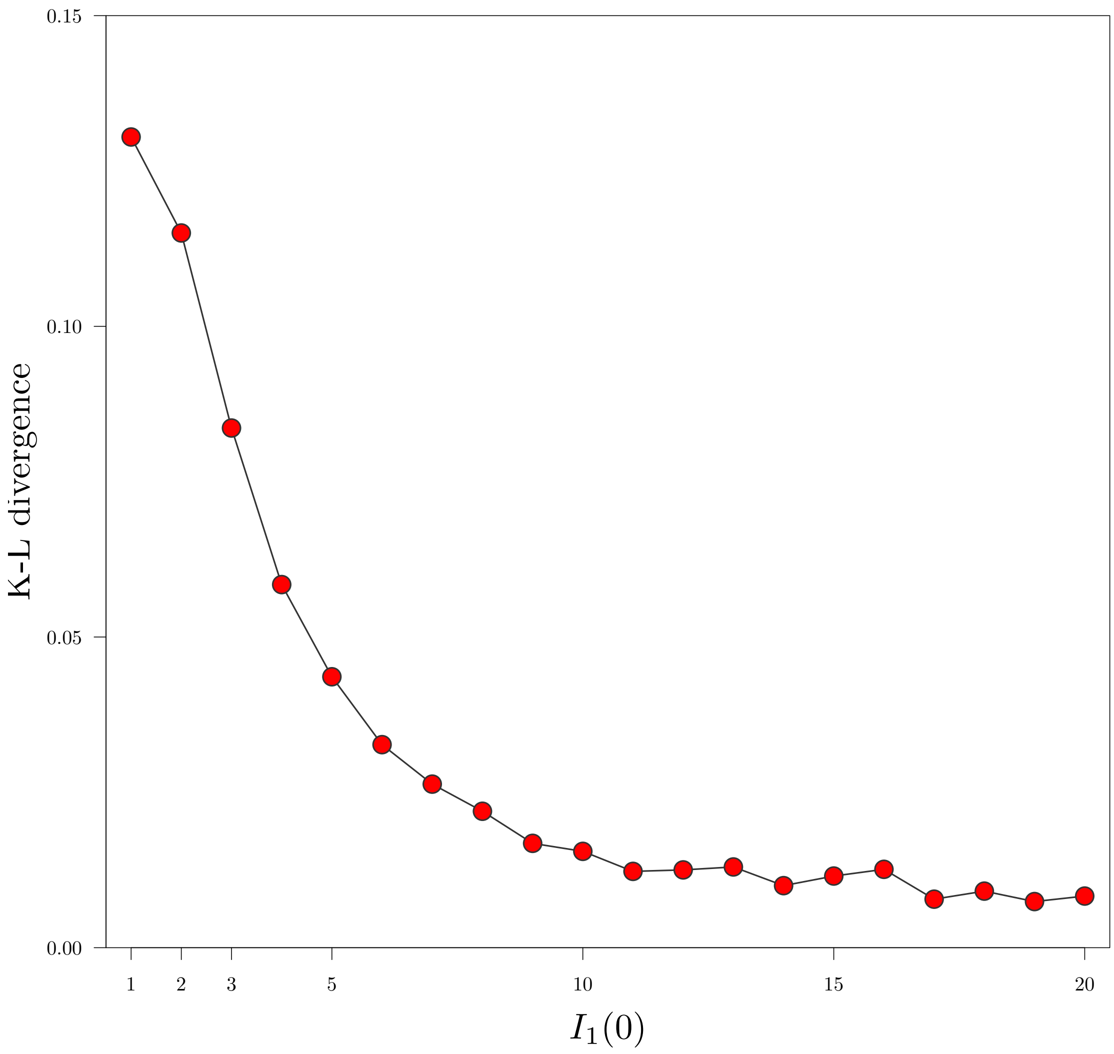

Figure 2.4: K-L divergence between t_{inv} distributions produced from simulations and from the analytic approximation (*cf.* Equation (2.31) and Equation (2.38)). The K-L divergence shows the degree of difference between observed and predicted probability density distributions. Parameters used were: $\mathcal{R}_0 = 4$, $m = 0.01$, $N_1 = N_2 = 10^5$.

2.3.4 Maximum likelihood estimation of coupling parameter m

If we know the values of the underlying parameters ($\mathcal{R}_0$, m, N_1, N_2), then Equation (2.33), or easily-computable histograms like those shown in Figure 2.2, allow us to estimate the probability of observing any particular time to invasion (t_{inv}) [58]. Our goal is to start with knowledge of

- the patch population sizes (N_1, N_2),
- the disease reproduction number of the uncoupled system ($\mathcal{R}_0$),
- the mean infectious period ($1/\gamma$),

and

- one or more observations of the time to invasion (t_{inv}),

and then *estimate* the underlying transmission coupling m between the two patches. To that end, in standard fashion, we interpret the probability density of observing t_{inv} given knowledge of the underlying parameter set as the likelihood of observing m given an observation of t_{inv}. If we use our approximation (2.33), we have[7]

$$\mathcal{L}(m \,|\, t_{\text{inv}}) \simeq f(t_{\text{inv}})\,. \tag{2.39}$$

Based on this approximation, Figure 2.5 shows the maximum likelihood estimate (MLE) of the coupling parameter m as a function of the observed time to invasion t_{inv}, for several reproduction numbers.

We can also approximate $\mathcal{L}(m \,|\, t_{\text{inv}})$ by constructing many simulation-based histograms like those in Figure 2.2, for a range of values of m [58]. In Figure 2.6 we show (as a heat map) a likelihood surface constructed in this way. To obtain an MLE

[7] Note that the likelihood is not a probability density, since it is not normalized by $\int_0^1 f(t_{\text{inv}})\,\mathrm{d}m$.

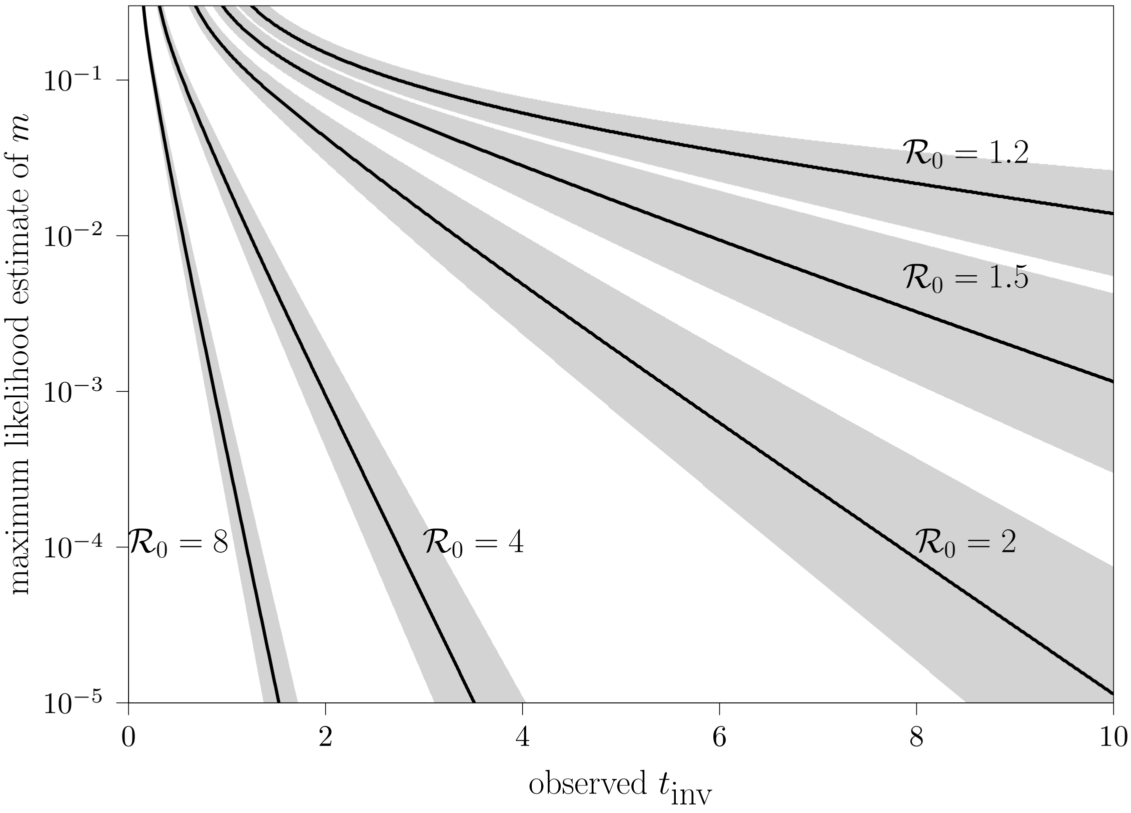

Figure 2.5: Maximum likelihood estimates (MLEs) of coupling m *vs.* observed time to invasion t_{inv} (in units of the mean infectious period), according to our analytical approximation (*cf.* Equations 2.33 and 2.39). The population sizes are $N_1 = N_2 = 10^5$, and the initial number of infections in population 1 is $I_1(0) = 1$. Grey bands under the black MLE curves indicate the effect of 10% uncertainty in the value of $\mathcal{R}_0$.

of m for a given t_{inv} from this simulation-based likelihood surface, we (i) obtain a likelihood profile as a function of m by slicing the surface at t_{inv}, (ii) smooth the profile with a cubic spline, and then (iii) find the maximum point of the smoothed profile (see Figure 2.7).

Whether we use the analytically approximated or simulation-based likelihood, we compute confidence limits based on the likelihood ratio test (LRT) [57, Ch. 6, pp. 254–258]. The LRT, applied to our estimate m_{est}, assumes that the *deviance*,

$$-2\log\left[\frac{\mathcal{L}(m_{\text{est}}\,|\,t_{\text{inv}})}{\mathcal{L}(m\,|\,t_{\text{inv}})}\right] = -2[\log\mathcal{L}(m_{\text{est}}\,|\,t_{\text{inv}}) - \log\mathcal{L}(m\,|\,t_{\text{inv}})]\,, \tag{2.40}$$

is approximately chi-squared distributed with one degree of freedom. In order to compute 95% confidence limits, we find the interval along the likelihood profile of m for which

$$\log\mathcal{L}(m_{\text{est}}\,|\,t_{\text{inv}}) - \log\mathcal{L}(m\,|\,t_{\text{inv}}) < \chi^2_1(0.95)/2 = 1.92\,. \tag{2.41}$$

The MLE and confidence interval for m for a particular observation of t_{inv} are shown with a black dot and error bars in Figure 2.6 (see Appendix B for computational details). The solid blue curve shows the MLE as a function of t_{inv} obtained from our analytical approximation (2.39), and the dashed blue curves show confidence bands.

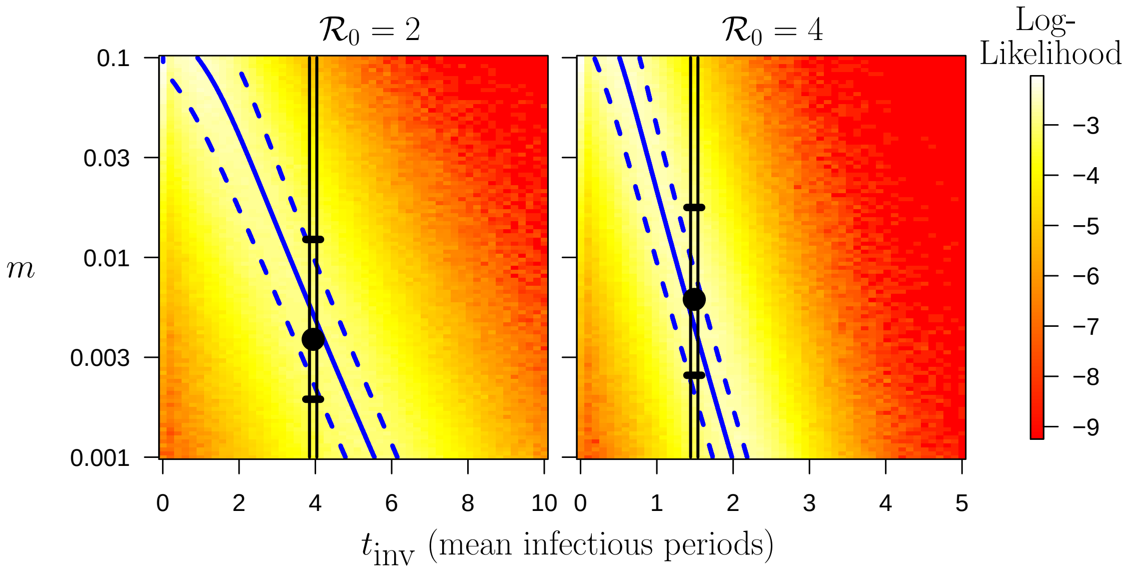

Figure 2.6: Likelihood of coupling parameter m given observed t_{inv}, $\mathcal{L}(m \,|\, t_{\text{inv}})$, computed from stochastic simulations. The fixed parameters are $N_1 = N_2 = 10^5$ and $\mathcal{R}_0 = 2$ (4) in the left (right) panel. The heavy black dot shows the maximum likelihood estimate (MLE) of m given an observed $t_{\text{inv}} = 4$ (1.5) infectious periods on the left (right). The vertical black lines enclose likelihood profiles of m for the observed t_{inv}, and are shown in further detail in Figure 2.7. 25% and 75% confidence limits are shown with horizontal black bars. The solid blue curves in each panel show the MLE of m according to the analytical approximation Equation (2.39) and correspond to particular curves in Figure 2.5. The dashed blue curves show 25% and 75% confidence limits for the analytical approximation (see Appendix B for details).

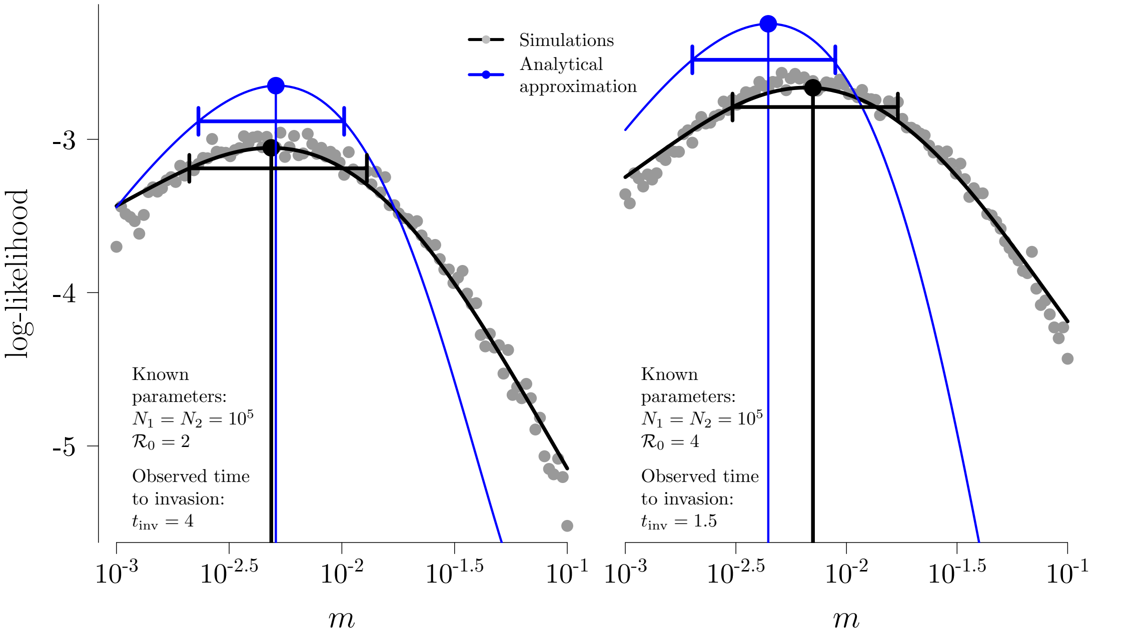

Figure 2.7: Likelihood profiles for the coupling parameter m. Black curves show the likelihood profile obtained from stochastic simulations (*cf.* Figure 2.6) and blue curves are obtained from our analytical approximation Equation (2.39). Heavy dots show the MLE and error bars show the 25% and 75% confidence limits. The grey dots correspond to the column enclosed with vertical black lines in the heat map in Figure 2.6; we smooth these log-likelihood values with a cubic spline and define the MLE and confidence limits using the spline.

2.3.5 MLE based on multiple observations of time to invasion

If multiple events of disease spread from one population to the other have been observed then much more accurate estimation of the transmission coupling parameter m is possible. It is important to emphasize in this context that since we are aiming to estimate a parameter of the social contact network—as opposed to a disease parameter—there is no need to restrict attention to repeated invasions by a single pathogen. Independent invasions by unrelated infectious diseases with the same mode of transmission could, in principle, be just as valuable for this purpose. Estimates of

m from independent invasions would require the assumption that m does not change between events, along with accurate estimates of disease parameters, $\mathcal{R}_0$ and γ, for each invading disease.

Suppose n independent invasions have been observed and let θ_i denote the set of observations $\{\mathcal{R}_0, \gamma^{-1}, t_{\text{inv}}\}$ associated with the ith invasion event. Then the likelihood of the coupling parameter being m, given this sequence of n observed invasions, is

$$\mathcal{L}(m \,|\, \{\theta_1, \ldots, \theta_n\}) = \prod_{i=1}^{n} \mathcal{L}(m \,|\, \theta_i)\,. \tag{2.42}$$

Each factor $\mathcal{L}(m \,|\, \theta_i)$ can be approximated using Equation (2.33) or via a simulation-based, smoothed likelihood profile, as in Figure 2.7.

Figure 2.8 shows four examples of how an estimate of m using the simulation-based approach improves as the number of observed invasions increases from 1 to 64. In each of four panels, the 64 invasions are assumed to be by the same disease (so the same $\mathcal{R}_0$ and mean infectious period). Exactly how the MLE and 95% confidence intervals change as additional invasions are observed depends on the sequence in which the observations occur. Each panel of Figure 2.8 shows three extreme cases, in which the 64 t_{inv} observations occur from (i) shortest to longest, (ii) longest to shortest, and (iii) from the median of the 64 observations to median of the remaining 63, and so on. The equivalent figure based on the analytical approximation (2.39) is shown in Figure 2.9.

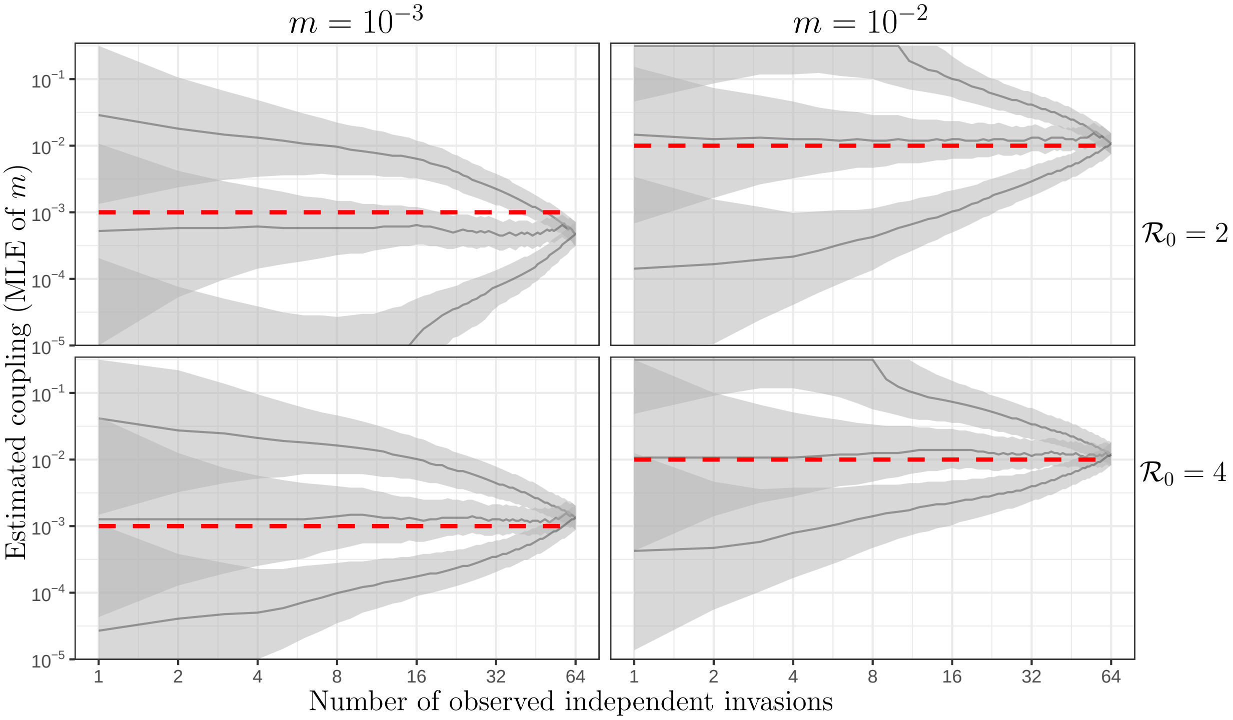

Figure 2.8: Estimates of the coupling parameter (m) improve as more independent invasion events are observed. The underlying $\mathcal{R}_0$ and coupling (m) are indicated above and to the right of the panels, and the underlying m is shown with a red dashed line. Populations sizes are $N_1 = N_2 = 10^5$ in all panels. In each case, 64 invasion events were simulated with the stochastic model (§2.2.3). The lower and upper curves show the MLE of m estimated from the subset of the 64 simulations corresponding to the largest and smallest observed times to invasion (note that high observed t_{inv} implies low coupling m, and vice versa). The MLEs shown with the middle curve correspond to the subset of simulations for which the observed t_{inv} was closest to the median. The shaded regions shows 95% confidence limits. In this figure we show estimation of coupling m using stochastic simulations (*cf.* Figures 2.6 and 2.7, and §2.3.5). See Figure 2.9 for the equivalent graphs based on the analytical approximation (2.39).

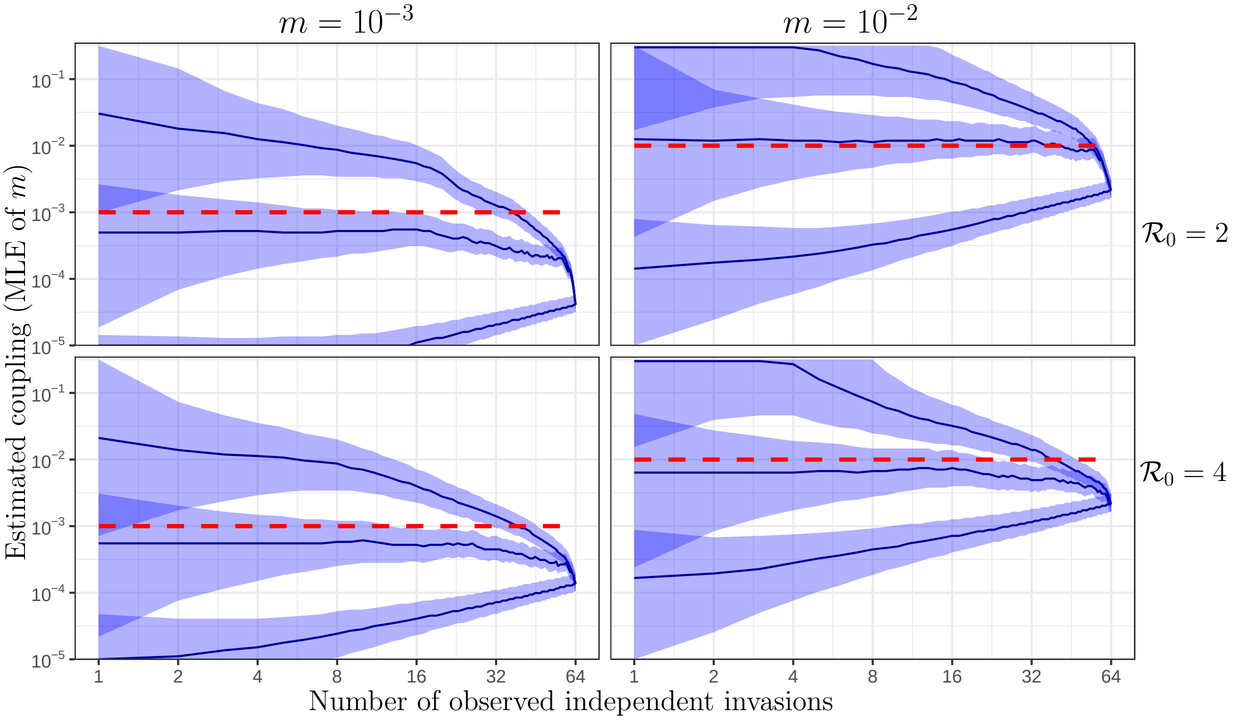

Figure 2.9: The equivalent of Figure 2.8 based on the analytical approximation (2.33) rather than simulations.

2.4 Discussion

We have explored the feasibility of using the time taken for an infectious disease to spread from one population to another (the time to invasion, t_{inv}) to estimate the degree of social contact between two populations. We quantified the degree of social contact with the proportion (m) of time that individuals typically spend outside their home region.

We have considered only the most idealized situation in which there are only two populations and the basic reproduction number, $\mathcal{R}_0$, and mean infectious period, $\frac{1}{\gamma}$, of the disease are known precisely. Even so—if based on a single observed disease invasion—the confidence intervals we obtain for the degree of coupling (m) stretch over an order of magnitude (Figure 2.7), which therefore provides only crude information about the social connectivity of the two populations. However, if multiple invasions are observed, much more accurate estimation of m is possible (Figure 2.8), and the independent invasions need not be of same disease (§2.3.5).

We estimated the likelihood profile for the coupling parameter m in two ways (Figure 2.7), one based on large numbers of stochastic simulations and the other based on an analytical approximation that we derived in §2.3.1. The simulation approach is more accurate (Figure 2.2 and Figure 2.8 *vs.* 2.9), but significantly so only if the number of cases in the seed population is very small when the estimate is made (Figure 2.3). The large computational expense of the simulation approach could be reduced by, for example, iterated filtering [88] beginning from the analytically derived maximum likelihood estimate (MLE), but simulations would be hard to justify if $\gtrsim 10$ cases had already occurred in the seed population (Figure 2.3).

Our analytical approximation facilitates exploration of how the relationship between observed t_{inv} and MLE of m depends on underlying disease characteristics—such as $\mathcal{R}_0$ and the mean infectious period—and on uncertainty in estimates of those properties (Figure 2.5).

Limitations

If attempts are made to apply our methodology to real epidemics, a number of limitations are important to bear in mind.

- The time to invasion t_{inv} can be difficult to estimate because of incomplete or inaccurate reporting, reporting delays, asymptomatic cases, and lack of temporal resolution in reporting (especially for historical data).

- If multiple invasions are observed, with long breaks between them, the possibility of changes in population characteristics in the times between epidemics should be considered. This can be a particularly significant concern when examining historical epidemics separated by decades or centuries.

- In general, changes in human behaviour and other factors may alter the social contact network *during* an epidemic and consequently the coupling of subpopulations of a meta-population.

Possible further developments

There are several natural directions for enhancement of the methods developed in this paper.

- Rather than relying on the exponential growth approximation, as in §2.3.1, the actual time series of observed cases in the seed population could be used instead of Equation (2.21) (for example, by assuming each case is infectious for exactly the mean infectious period). This would lead to a (presumably more accurate) estimate of $\mu(t)$, the expected rate at which new infections occur in the target population; this estimate would replace Equation (2.22a) and, after insertion in Equation (2.31), lead to an alternative version of Equation (2.33) for the probability density of the time to invasion.

- In a meta-population with more than two populations, the time at which a first case occurs in each subpopulation could be used to inform the overall coupling in the system. In principle, it could turn out to be easier to estimate the *average* inter-population transmission coupling when there are more subpopulations. On the other hand, potentially different degrees of coupling between each pair of subpopulations increases the range of possible contact networks.

- In Figure 2.5, we indicated the effect of uncertainty in $\mathcal{R}_0$. A more systematic and complete analysis of the effects of uncertainty in estimates of non-coupling parameters would be valuable.

- We have focussed on the time to invasion, but if there are more than two subpopulations then the locations of the source subpopulations that seed each invasion could also be used to constrain estimates of connectivity.

- If age-stratified incidence or mortality data are available, more detail about transmission coupling could be extracted, in principle. Different age-groups have been observed to make contact at different rates [89], and the age distribution of infections in the source population along with the age of the first case in the target population could better inform estimations of inter-population coupling than the time to invasion alone.

- In some situations, information about travel volumes and destinations may be available, in which case ways to use such data to constrain connectivity estimates (such as with the use of Bayesian priors [90]) could be useful.

- In a situation where multiple independent invasions can be observed, an estimate of m from earlier events, along with another from later events, may have non-overlapping confidence intervals. This would be evidence of changes in the underlying social contact network.

Our analysis in this paper has shown that while estimating coupling from the time to invasion is difficult, it is possible. Enhancing methods of doing so will advance understanding of the mechanisms and predictability of infectious disease outbreaks in meta-populations.

2.5 Acknowledgments

KH was supported by an Ontario Graduate Scholarship (OGS). DE was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Appendix A: Approximation error on t_{inv} distribution

The ensemble mean and variance of the force of infection from the source to the target (population 1 to population 2) are given in Equations (2.22) and (2.35), respectively. To quantify uncertainty on the distribution of the time to invasion of population 2, we must evaluate the integral in Equation (2.31) for $\mu(t) + \alpha\,\sigma(t)$ rather than $\mu(t)$, *i.e.*, we must calculate

$$f_\alpha(t) = \big[\mu(t) + \alpha\,\sigma(t)\big] \exp\Big\{ -\int_0^t \big[\mu(s) + \alpha\,\sigma(s)\big]\,\mathrm{d}s \Big\}\,. \tag{2.43}$$

(Note that $f(t)$ in Equation (2.31) corresponds to $f_0(t)$ in this notation.) To evaluate the integral in Equation (2.43) explicitly, we use

$$\int_0^t \sqrt{e^{rs}(e^{rs}-1)}\,\mathrm{d}s = \frac{1}{r}\Big[\sqrt{e^{rt}(e^{rt}-1)} - \log\big(\sqrt{e^{rt}-1} + \sqrt{e^{rt}}\big)\Big]\,. \tag{2.44}$$

Thus, with μ and σ given by Equations (2.22) and (2.35), respectively, and writing r for r_1 to reduce clutter, we obtain the explicit expression,

$$\begin{aligned} f_\alpha(t) = \Big[\mu_0 e^{rt} + \alpha\,\sigma_0\sqrt{e^{rt}(e^{rt}-1)}\Big] \times \exp\Big\{\frac{\mu_0}{r}\big(1-e^{rt}\big)\Big\} \\ \times \exp\Big\{ -\alpha\frac{\sigma_0}{r}\Big[\sqrt{e^{rt}(e^{rt}-1)} - \log\big(\sqrt{e^{rt}-1} + \sqrt{e^{rt}}\big)\Big]\Big\} \end{aligned} \tag{2.45}$$

For a given α range ($\alpha_{\min} \le \alpha \le \alpha_{\max}$, where normally $\alpha_{\min} = -\alpha_{\max}$), we then define upper and lower error estimates,

$$f_{\mathrm{U}}(t) = \max_{\alpha}\{f_\alpha(t) : \alpha_{\min} \le \alpha \le \alpha_{\max}\}\,, \tag{2.46a}$$

$$f_{\mathrm{L}}(t) = \min_{\alpha}\{f_\alpha(t) : \alpha_{\min} \le \alpha \le \alpha_{\max}\}\,, \tag{2.46b}$$

which correspond to the dashed blue curves in Figures 2.2 and 2.3. For any given t, at least one of the upper and lower estimates is obtained at an edge of the α range; solving $\partial f_\alpha/\partial\alpha = 0$ for α, we find a single critical point,

$$\alpha_{\text{crit}}(t) = \frac{\sqrt{e^{rt}-1}\,(r-\mu_0 e^{rt}) + \mu_0 e^{\frac{rt}{2}} \log\!\left(\sqrt{e^{rt}-1} + e^{\frac{rt}{2}}\right)}{\sigma_0\left[e^{\frac{rt}{2}}\,(e^{rt}-1) - \sqrt{e^{rt}-1}\,\log\!\left(\sqrt{e^{rt}-1} + e^{\frac{rt}{2}}\right)\right]}\,. \tag{2.47}$$

Appendix B: Numerical details of simulation-based likelihood

This appendix relates to the construction of Figure 2.6, as described in §2.3.4.

For each of 100 m values, we measured time to invasion t_{inv} from 10^4 stochastic simulations using the `adaptivetau` package in R [82], and grouped these t_{inv} values into 100 bins on the t_{inv} axis. More precisely, our 100 m values, which we refer to as m_i, were spaced logarithmically between 0.001 and 0.1. For each m_i, and for $\mathcal{R}_0 = 2, 4$, we produced $n_{\text{sim}} = 10^4$ simulations and measured the corresponding t_{inv} for each simulation. We then divided the full range of resulting t_{inv} values into 100 bins, b_j. We produced a grid where Cell(i, j) contained the number of simulations with $m = m_i$ and t_{inv} in bin b_j. We used the grid of m vs. t_{inv} simulation frequencies to produce likelihoods of t_{inv} given m,

$$\mathcal{L}(t_{\text{inv}}|m_i) \approx \frac{\text{Cell}(i,j)}{n_{\text{sim}}} \,. \tag{2.48}$$

We produced a full grid of log-likelihoods, *i.e.*, $\log \mathcal{L}(t_{\text{inv}}|m_i)$ (see Figure 2.6). We select the bin b_j that contains the observed t_{inv}. The log-likelihoods of column j yield the likelihood profile of the observed t_{inv} with respect to m, and the cell with the maximum likelihood indicates the maximum likelihood estimate (MLE) of m given t_{inv} (see Figure 2.7).

Chapter 3

Estimating transmission coupling from fadeout times of infectious diseases

Abstract

Advancing our understanding of the mechanisms by which infectious diseases spread within and between human populations is critical in efforts to understand and predict widely spread epidemics and pandemics. Mathematical modeling provides many tools to understand disease spread, but parameterizing transmission between populations is a difficult problem, since the process itself is not practically observable. We present a method for estimating coupling between one large and one small population, each undergoing recurrent epidemics, and modeled as *susceptible-infected-recovered* (SIR) systems. We show that the strength of coupling between the two populations can be estimated from the time the small population spends uninfected. Confidence in the estimate is increased the longer recurrent epidemics are observed. The method presented, though simple, shows that information about epidemic coupling can be successfully inferred from spatiotemporal disease data, which is becoming ever more widely available in digital form.

3.1 Introduction

Mathematical models provide a powerful range of tools for understanding and predicting the spread of infectious diseases in human populations [9,13,15–18,68]. In particular, the mechanistic SIR model (*susceptible-infected-recovered*), which approximates a population as being well-mixed (contact occurs uniformly at random) and where infection confers permanent immunity upon recovery, has had remarkable success explaining observed dynamics. Various areas of study aim to address oversimplifications inherent in the basic model, including the effects of seasonal forcing [9, 13, 16, 68, 91], intensity and duration of infectiousness [70–75], vital dynamics [69], network structure within populations [41–45], and others. This area of research has been motivated in part by large quantities of digitized disease data which have become available in recent decades [13, 19, 68, 92, 93].

Many infectious disease data sets are spatiotemporal in nature, and show evidence of epidemic coupling between populations. However, one of the central difficulties of modeling infectious diseases is the unobservable nature of the transmission process, necessitating the development of methods for indirectly inferring transmission parameters [94]. This problem is compounded when considering epidemic coupling between geographically separated populations.

In this paper, we focus on the latter problem, and present a method for estimating the degree of coupling between a large and a small population from case report data alone. Our goal is to show how well the degree of coupling between two populations undergoing recurrent epidemics can be estimated in an ideal scenario. To this end, we construct a theoretical scenario in which two populations undergoing recurrent epidemics differ in size such that only the smaller of the two populations sees occasional disease fadeouts[1][2]. We then show that the degree of coupling between the two populations, formalized with a single parameter (specified with a parameter m, defined in §3.2), can be estimated from the proportion of total time the small population spent faded out, t_{f} (*time faded out*)[3]. We furthermore show that the quality of the estimate is improved the longer the system is observed, as more fadeout events in the small population are observed.

Recurrent epidemics in a host population typically occur when periods of low disease prevalence allows a build-up of susceptible individuals, either through births, immigration, or waning immunity. These periods are then followed by epidemics due to the re-introduction of disease or to an increase in disease transmission. Seasonal patterns in contact rates between individuals [13, 68], birth rates [69], changing weather [97], and other seasonally varying factors can be drivers of seasonally varying disease prevalence. We model seasonally recurring epidemics with seasonal variation in transmission, which is sufficient to generate recurrent epidemics, and represents realistic phenomena such as increased contact rates between children during the school term in the winter. We model the susceptible recruitment required to generate recurrent epidemics as births, which occur at a rate relative to the total population size. Finally, we model the scenario stochastically in order to capture the phenomenon of randomly occurring disease fadeouts in the troughs between recurring epidemics. The frequency and duration of disease fadeouts in a population undergoing recurrent

[1] The recurrent reintroduction of disease in small populations by large population centres has been noted in previous research [20, 24, 59, 95].

[2] We refer to the temporary absence of disease in populations undergoing recurrent epidemics as either a 'fadeout' or an 'endemic fadeout', avoiding the term 'epidemic fadeout', which has been used to refer to the extinction of an invading pathogen in the trough after the first epidemic wave [96].

[3] The time faded out, t_{f}, is connected conceptually to the concept of the time to invasion, t_{inv}, presented in Chapter 2. After a fadeout in the small population, there is a time to *re*-invasion, and the total time taken for re-invasion across one or more fadeouts is measured by t_{f}. The state of the system at the beginning of a fadeout is almost certainly different than the initial conditions considered in Chapter 2, but this does not preclude a potential theoretical bridge between the concepts.

epidemics is negatively correlated with the size of the population [75, 98]. We make use of this property of fadeouts to choose parameters in which fadeouts in the smaller population are common, and fadeouts in the larger population are virtually absent (see §3.2).

Parameter estimation methods vary greatly depending on the natural phenomenon a model is intended to capture. Our use of time faded out, t_{f}, to estimate degree of coupling m between two populations undergoing recurrent epidemics is motivated by several key features of coupling between populations. We note first that individual members of two populations separated geographically typically interact far more with their respective local populations than with members of the other population. Assuming this holds true for disease transmission, we expect the amount of transmission between populations to be low relative to the amount of local transmission. As a result, when disease prevalence in a population is high, the effect of coupling can be difficult to observe and distinguish from stochasticity. Without detectable features in the data driven by coupling, coupling parameters can be practically unidentifiable. However, when one population's prevalence is low, infection from another population is detectable. In the case of a disease fadeout in one population, re-infection is driven completely by coupling with another infected population, and the duration of the fadeout is negatively correlated with the degree of coupling with the infected population, all else being equal. Estimating coupling parameters without observing low prevalence is difficult, and requires the observation of other dynamical patterns or transitions caused by coupling, such as synchrony in recurrent epidemics [63]. Our aim is to present the best possible case for estimating a single coupling parameter, m, with the methodology presented. To this end we assume perfect knowledge of all parameters except m in the estimation process. In §3.3, we test the methodology presented on stochastic simulations, and can thereby compare the effectiveness of estimation with known true values of m. This approach furthermore has the advantage of showing the degree of error present in estimates of m that results only from the methodology, absent the additional uncertainty in other parameter estimates.

3.2 Two population recurrent epidemics

We model a two-patch meta-population stochastically, where each population has an SIR (*susceptible-infected-recovered*) compartmental structure, and coupling takes place in the transmission term. We first define deterministic rates of state transition as a system of ordinary differential equations (ODE), and then define the stochastic system by interpreting the deterministic transition rates as probabilistic event rates. The system of ODEs for a single population is given as follows

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \nu N - \Lambda S - \mu S \tag{3.1a}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \Lambda S - (\gamma + \mu) I \tag{3.1b}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I - \mu R \tag{3.1c}$$

The state variables S, I, and R are the numbers of susceptible, infected, and recovered individuals, with the total population $N = S + I + R$. All births enter the susceptible compartment at the rate νN, where ν is the *per capita* birth rate. All compartments lose individuals at the *per capita* death rate μ. Throughout this paper, we set the death rate equal to the birth rate, $\mu = \nu$.

New infections occur according to the assumption of uniform mixing of susceptible and infected individuals, where the rate per unit time of susceptibles becoming infected is the ***force of infection***

$$\Lambda = f(t)\beta\frac{I}{N}\,. \tag{3.2}$$

where β is the transmission rate. We modify this definition of Λ later in §3.2.1 to incorporate cross-coupling in the meta-population, using the coupling parameter m.

The only non-autonomous component of the system is the forcing function $f(t)$, which we define as a sinusoidal function with amplitude α and a one-year period

$$f(t) = 1 + \alpha\cos(2\pi t) \tag{3.3}$$

The oscillation of $f(t)$ is intended to represent the realistic phenomenon of higher transmission in the winter and lower transmission in the summer. While sinusoidal forcing is sufficient for our purpose of driving seasonally recurring epidemics, real-world seasonal forcing, especially in childhood infectious disease, is often caused by school terms, and term-time forcing is a realistic alternative to the sinusoidal form of $f(t)$ we use [99]. Infected individuals recover at constant rate γ, which results in an exponentially distributed period of infection with mean $1/\gamma$. The basic reproduction number of an infectious disease, $\mathcal{R}_0$, is defined as the mean number of new infections caused by a single infected individual in an otherwise completely susceptible population. Throughout this paper, we make use of $\mathcal{R}_0$ as defined for one population without seasonal forcing or coupling ($m = 0$, $\alpha = 0$), i.e.

$$\mathcal{R}_0 = \frac{\beta}{\gamma + \mu} \tag{3.4}$$

We use $\mathcal{R}_0$ for the definition of initial conditions in the model, noting that in the deterministic case for a population in isolation ($m = 0$) and without seasonal forcing

($\alpha = 0$), the system yields an endemic equilibrium of

$$(S^*, I^*) = \left(\frac{N}{\mathcal{R}_0}, \frac{N(\mathcal{R}_0\nu - \mu)}{\mathcal{R}_0\gamma} \right) \tag{3.5}$$

We initialize state variables in stochastic simulations in each population to be the closest whole numbers to these quantities, $(S_0, I_0) \approx (S^*, I^*)$. These initial conditions result reliably in endemic disease prevalence with recurrent epidemics in the large population.

3.2.1 Coupling in Transmission

Coupling between host-populations in an epidemiological system can be modeled in many ways, including—though not limited to—any combination of implicitly or explicitly defined movement of susceptible or infected individuals between the geographic regions ("patches"), and with rates of contact between members of the metapopulation occurring proportional to a static or dynamic social network, or proportional to geographic distance between individuals or population centers [48, Ch. 4]. We implement a coupling framework in which two patches each have a resident population, and residents of each patch visit one another some proportion of the time. We express this by means of a coupling matrix

$$c = \begin{pmatrix} 1-m & m \\ m & 1-m \end{pmatrix}, \tag{3.6}$$

This formulation of coupling is more fully developed in §2.2.1. At any given time, the proportion of population i present in patch j is given by c_{ij}, and we refer to m throughout the paper as the ***coupling parameter***. Each patch j has a local force of infection, Λ_j, to which all susceptibles present are exposed, and which is given by

$$\Lambda_j = \beta f(t) \frac{\sum_{i=1}^{2} c_{ij} I_i}{\sum_{i=1}^{2} c_{ij} N_i}\,, \qquad j = 1,2 \tag{3.7}$$

The susceptibles of population i are distributed between patches j according to the matrix c, and thus the rate of new infections in population i is given by

$$\sum_{j=1}^{2} c_{ij} S_i \Lambda_j = S_i \sum_{j=1}^{2} c_{ij} \Lambda_j\,, \qquad i = 1,2 \tag{3.8}$$

The complete system of rates with population cross-coupling is therefore given by

$$\frac{\mathrm{d}S_i}{\mathrm{d}t} = \nu N_i - S_i \sum_{j=1}^{2} c_{ij} \Lambda_j - \mu S_i \tag{3.9a}$$

$$\frac{\mathrm{d}I_i}{\mathrm{d}t} = S_i \sum_{j=1}^{2} c_{ij} \Lambda_j - (\gamma + \mu) I_i \tag{3.9b}$$

$$\frac{\mathrm{d}R_i}{\mathrm{d}t} = \gamma I_i - \mu R_i\,, \qquad i = 1,2 \tag{3.9c}$$

We produce stochastic simulations with the rates in Equation (3.9) to produce event probabilities, using an adaptive time-step approximation algorithm. The standard Gillespie algorithm [83, 100] for computing exact realization of the stochastic process requires event rates to remain fixed while no event occurs, which is only approximately true in our system on account of the seasonal forcing function $f(t)$. An exact stochastic simulation algorithm for the seasonally forced case does exist [101], but sampling one event at a time is far too computationally costly for the population sizes and time-scales we consider. We therefore use adaptive time-step methodology, or "tau-leaping" [100], which samples numerous events over some time step from either Poisson or Binomial distributions parameterized by the rate questions. These methods are approximations, and balance the trade-off between accuracy and computational cost by adjusting time step length while simulating[4]. We use the methods implemented in the `adaptivetau` package in R [82].

When the seasonal forcing amplitude α is positive, trajectories of the deterministic SIR system shown in Equation (3.9) converge to periodic orbits or more complicated attractors. Realizations of the stochastic model also approach these periodic attractors, but in the stochastic case trajectories are perturbed by demographic stochasticity, and disease fadeouts are possible since the number of infecteds may randomly reach zero. Trajectories in the deterministic case can approach periodic attractors after a transient period. Demographic stochasticity prevents close asymptotic approach to attractors in the stochastic case, resulting in more complicated dynamics in stochastic realizations [17, 103].

3.2.2 Duration of endemic fadeouts

When disease prevalence reaches low levels, fluctuations due to demographic stochastic may result in prevalence reaching zero. Once no infections remain in a population, no new local infections can occur, and prevalence remains zero until external re-infection of the population. Populations undergoing recurrent epidemics, such as those driven by seasonal forcing, reach low levels of prevalence in the troughs between epidemics. The closer the troughs in prevalence are to zero, the higher the probability of extinction, thus the probability of extinction is negatively correlated with population size, and positively with the magnitude of fluctuations. The relationship between the magnitude of seasonal prevalence fluctuations and the magnitude

[4] The accuracy of approximation for tau-leaping realizations can be affected by the inclusions of non-homogenous terms such as our seasonal forcing function, $f(t)$. However, since the relative change in event rates over the τ-step is held below a threshold [102], the loss of accuracy is small if $f(t)$ does not change significantly within the τ-step, which is the case in our simulations.

of the seasonal forcing that drives them is not straightforward. It depends on disease parameters $\mathcal{R}_0$ and γ, magnitude of seasonal forcing α, and birth rate ν, on demographic stochasticity, and on dynamical resonance [17, 93, 104]. An analytical examination of characteristics of fadeouts during prevalence troughs, such as when they begin and how long they last, could be a useful direction for future research (see §3.4). When extinction events occur in the small population, the fadeouts are ended by a re-infection by infected individuals in the large population. Therefore, the duration of endemic fadeouts in the smaller population is negatively correlated with the degree of coupling m. In our model, we set the larger population to be the first ($i = 1$), and smaller population to be the second ($i = 2$), i.e. $N_1 > N_2$, where i refers to the index used in Equation (3.9). Given a time-series of observed prevalence in two populations, we define t_f to be the *proportion* of total time during which prevalence in the small population is 0. We show an example of t_f observed for a simulated time-series in Figure 3.1.

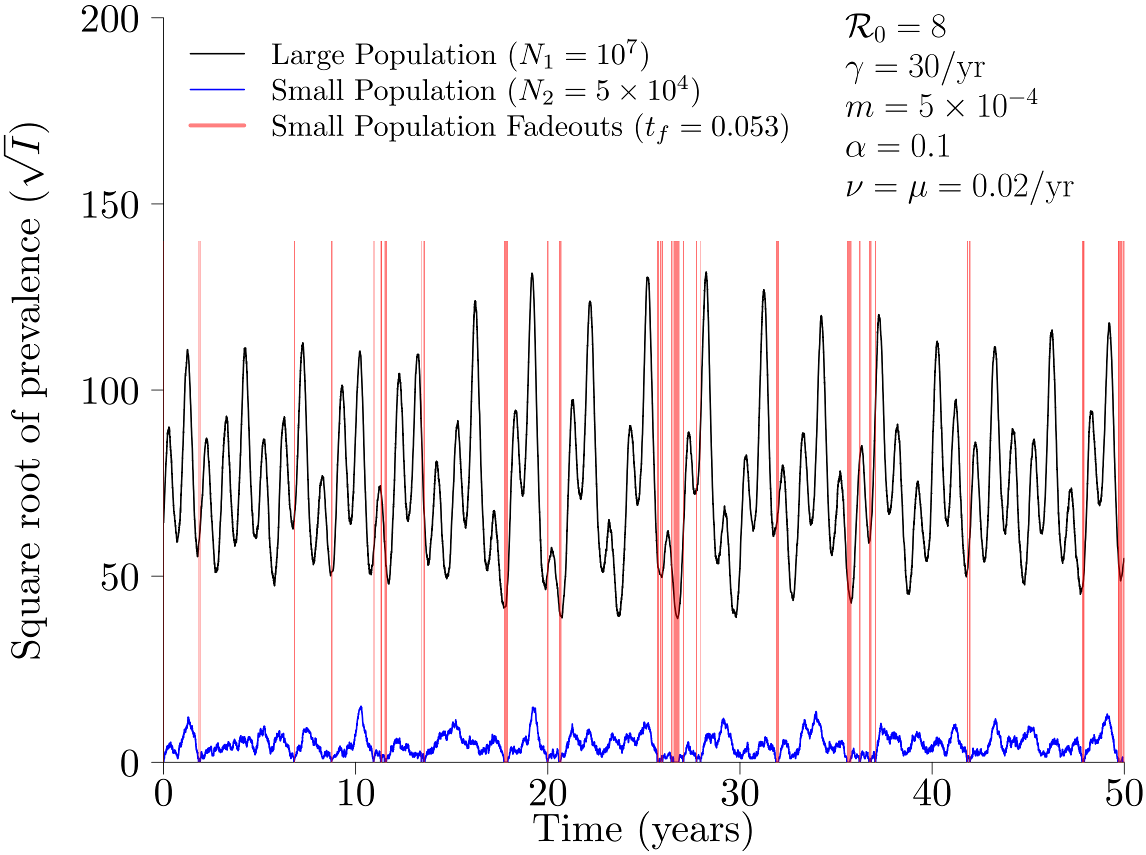

Figure 3.1: Example of two-population recurrent epidemics showing periods of fadeout in the smaller population. The simulation shown was run for a 150 year burn-in period prior to the 50 years shown. Red bands show periods of fadeout in the small population.

For a single parameterization of the model, repeated stochastic realizations will produce a distribution of observed t_f. We show examples of this distribution in Figure 3.2 for different numbers of years.

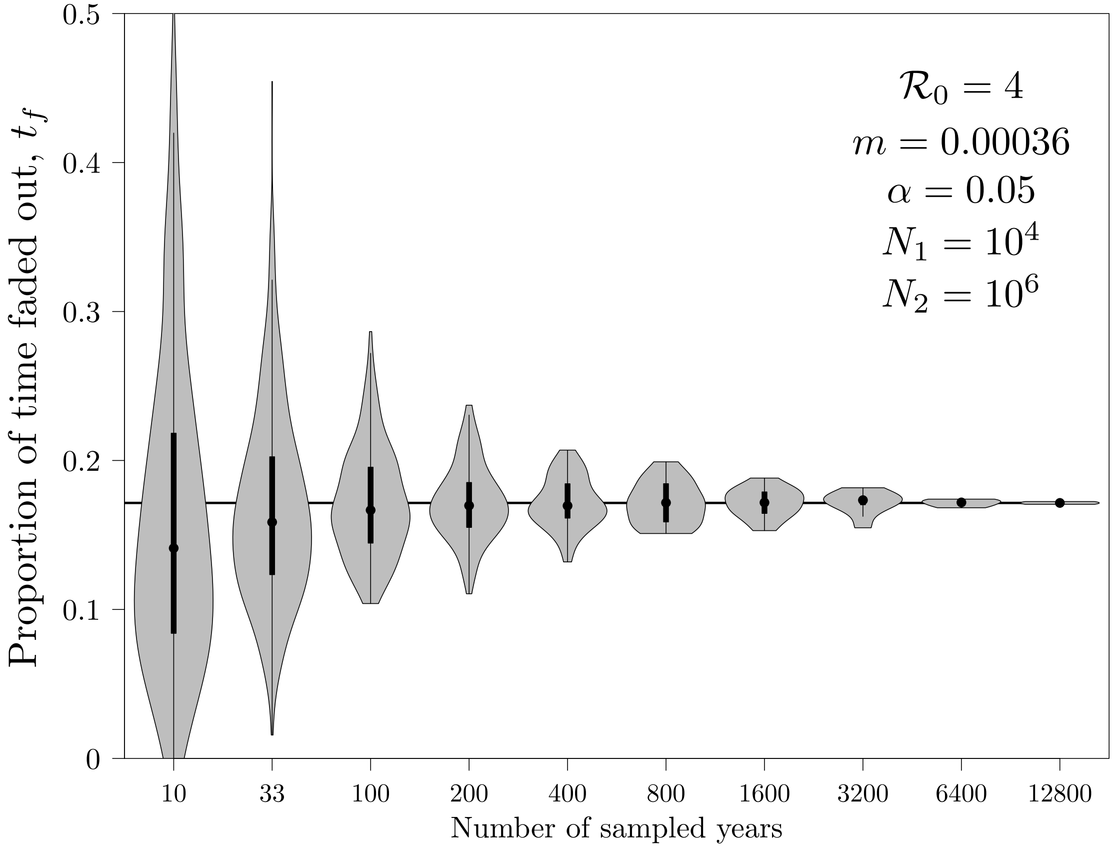

Figure 3.2: Distributions of time population 2 spends faded out, t_{f}, as a proportion of total time. For a window of a given number of years (x-axis), the distribution of t_{f} is shown as a violin plot (y-axis). Plotted data were produced from 256 simulations, each run for a 100 year burn-in period followed by another 100 years. t_{f} value for 200 year windows were produced by averaging t_{f} from two 100 year simulations, likewise from 200 to 400, and so on. The horizontal black line shows the average t_{f} across all 256 simulations.

3.3 Estimating coupling with MLE

We use maximum likelihood estimation to estimate the coupling parameter m from large numbers of simulations [58]. The distributions shown in Figure 3.2 are an approximate probability distribution of the proportion of time population 2 spent faded out, t_{f}, given chosen parameters. Fixing all parameters except for m, we write

$p(t_\mathrm{f}|m)$ as the probability of observing some t_f given m. The inverse relationship of this p is the likelihood of m given t_f, $\mathcal{L}(m|t_\mathrm{f})$. The m that maximizes the likelihood $\mathcal{L}(m|t_\mathrm{f})$ for a given observed t_f is the maximum likelihood estimate (MLE). We compute approximate probability distributions $p(t_\mathrm{f}|m)$, as in Figure 3.2, for a set of fixed parameters, by simulating n_sim realizations. To find the MLE of coupling m for a given observation of t_f, we select n_m values of m spaced logarithmically within a fixed range, $m \in [m_\mathrm{min}, m_\mathrm{max}]$, and compute $\mathcal{L}(m|t_\mathrm{f})$ in each case (see Figure 3.3 for an example). In addition to locating the MLE of coupling m by this method, we can also show the precision of the estimate from the relationship between $\mathcal{L}(m|t_\mathrm{f})$ and m, referred to as the *likelihood profile* (see Figure 3.4).

We compute confidence limits on MLEs based on the likelihood ratio test (LRT) [57, Ch. 6, pp. 254–258]. The LRT approximates the *deviance*, $-2[\log\mathcal{L}(m_\mathrm{est}\,|\,t_\mathrm{inv}) - \log\mathcal{L}(m\,|\,t_\mathrm{f})]$, to be chi-squared distributed with one degree of freedom. We then compute 95% confidence limits by cutting off m above and below the MLE such that

$$\log\mathcal{L}(m_\mathrm{est}\,|\,t_\mathrm{f}) - \log\mathcal{L}(m\,|\,t_\mathrm{f}) < \chi^2_1(0.95)/2 = 1.92\,. \qquad (3.10)$$

We show an example of maximum likelihood estimation of m along with corresponding confidence intervals for a given observed t_inv, assuming different durations of observation of the time series (10, 33, and 100 years), in Figure 3.4. We note that increasing the duration of observation of the time-series narrows the confidence intervals of the m estimation, thus improving the estimate with more data.

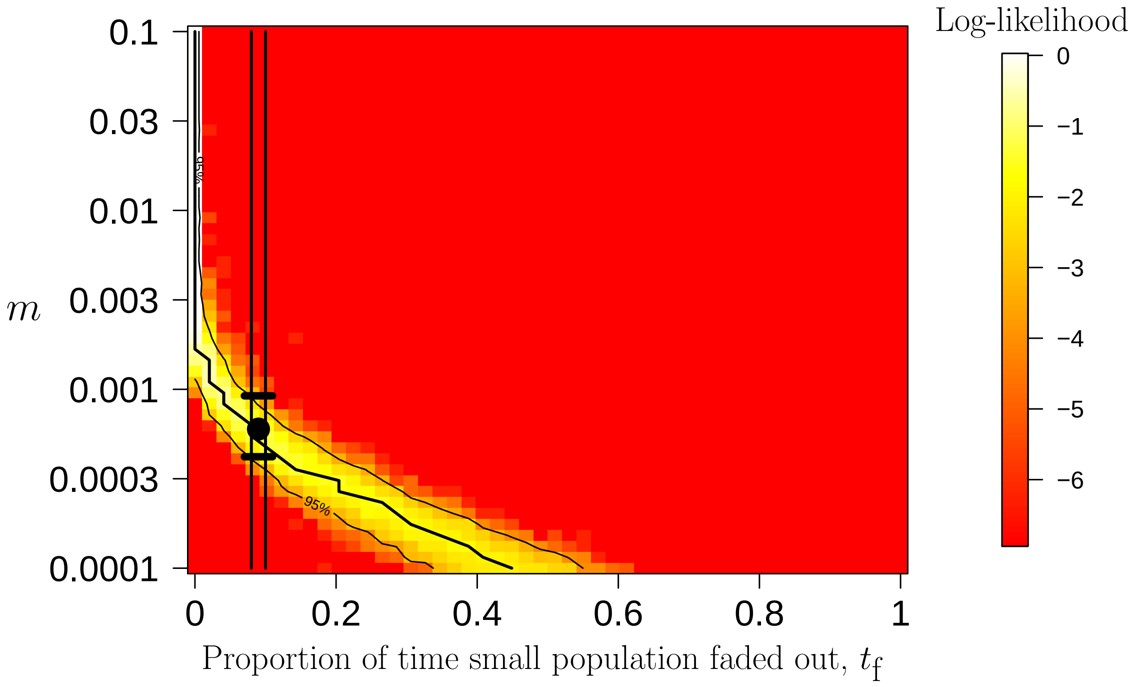

Figure 3.3: Likelihood of coupling parameter, m, given fadeout time, t_{f}: $\mathcal{L}(m|t_{\mathrm{f}})$. Parameters: $\mathcal{R}_0 = 4$, $\alpha = 0.05$, $N_2 = 10^3$, $N_1 = 10^6$, $\frac{1}{\gamma} = 10\,\mathrm{yr}$. Duration of time-series: 100 years. Each vertical slice is a likelihood profile for observed fadeout time, $t_{\mathrm{f,obs}}$, vs m. Produced from $n_m = 50$ different m values and $n_{\mathrm{sim}} = 500$ simulations each. Likelihood profiles are shown for 50 t_{f} values spaced uniformly from $[0, 1]$. Contours are shown for the maximum likelihood and 95% confidence intervals.

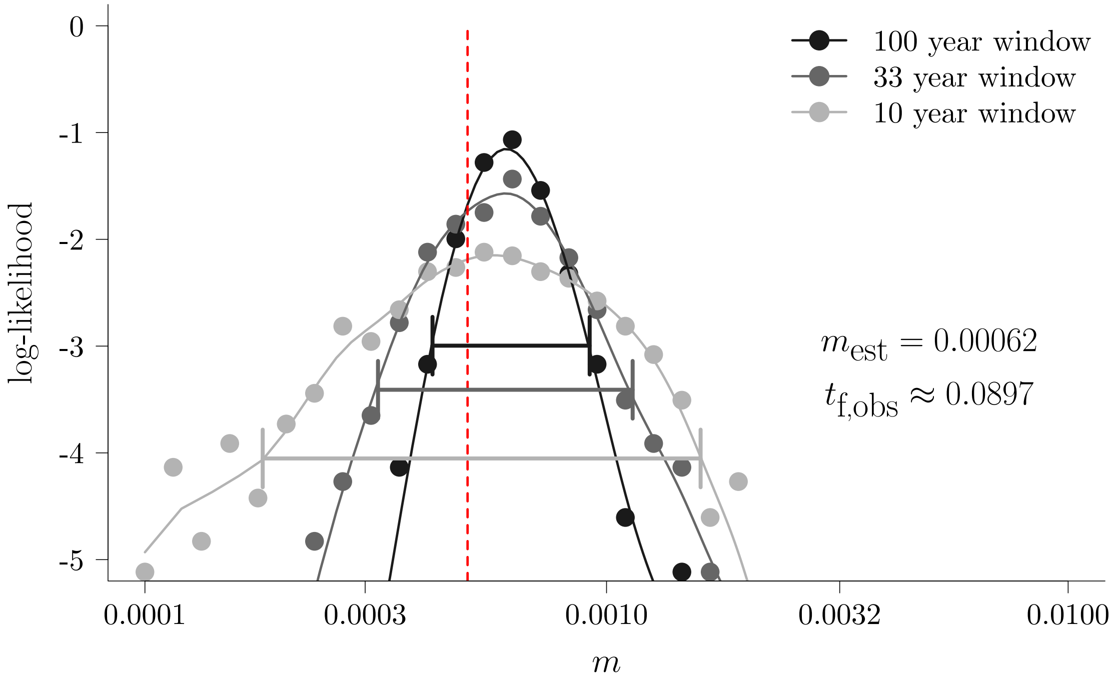

Figure 3.4: Likelihood of coupling parameter m given observed proportion population 2 spent in fadeout, $t_{\text{f,obs}} \approx 0.0897$. Solid dots show approximate log-likelihoods of m spaced logarithmically from $[10^{-4}, 10^{-2}]$, and solid lines show spline fits to approximate likelihood points used for estimation. Likelihood profiles shown for observed time-series lasting 10, 33, and 100 years, along with associated 95% confidence intervals. $t_{\text{f,obs}}$ was generated from a simulation using a true value $m = 0.0005$, shown as the red dotted line. Other parameters: $\mathcal{R}_0 = 4$, $\frac{1}{\gamma} = 10\,\text{yr}$, $\nu = \mu = 0.02\,\text{yr}^{-1}$, $\alpha = 0.05$, $N_2 = 10^4$, $N_1 = 10^6$.

3.3.1 Effect of Parameters on Estimation

Estimating parameters using MLE depends on the feasibility of locating global maxima in the likelihood profiles of those parameters. Under certain conditions, the coupling parameter m cannot be estimated from an observed t_{f}. In order to understand the preconditions for producing an estimate of m, we show the likelihood surface over a range of m and t_{f} (see Figure 3.3, and note that the likelihood profile shown in Figure 3.4 for a 100 year window is enclosed in black lines). Each vertical column of the grid shown is a likelihood profile computed in the same manner as in Figure 3.4. In order to obtain an estimate of m for a given t_{f}, the likelihood profile must contain a distinct maximum, and can fail to do so for reasons described in §3.4.

Other grids similar to Figure 3.3 for $\mathcal{R}_0 \in \{2, 4, 8\}$, $N_2 \in \{10^3, 10^4, 10^5\}$, and $\alpha \in \{0.01, 0.05, 0.1\}$ are shown in Figures 3.5, 3.6, and 3.7.

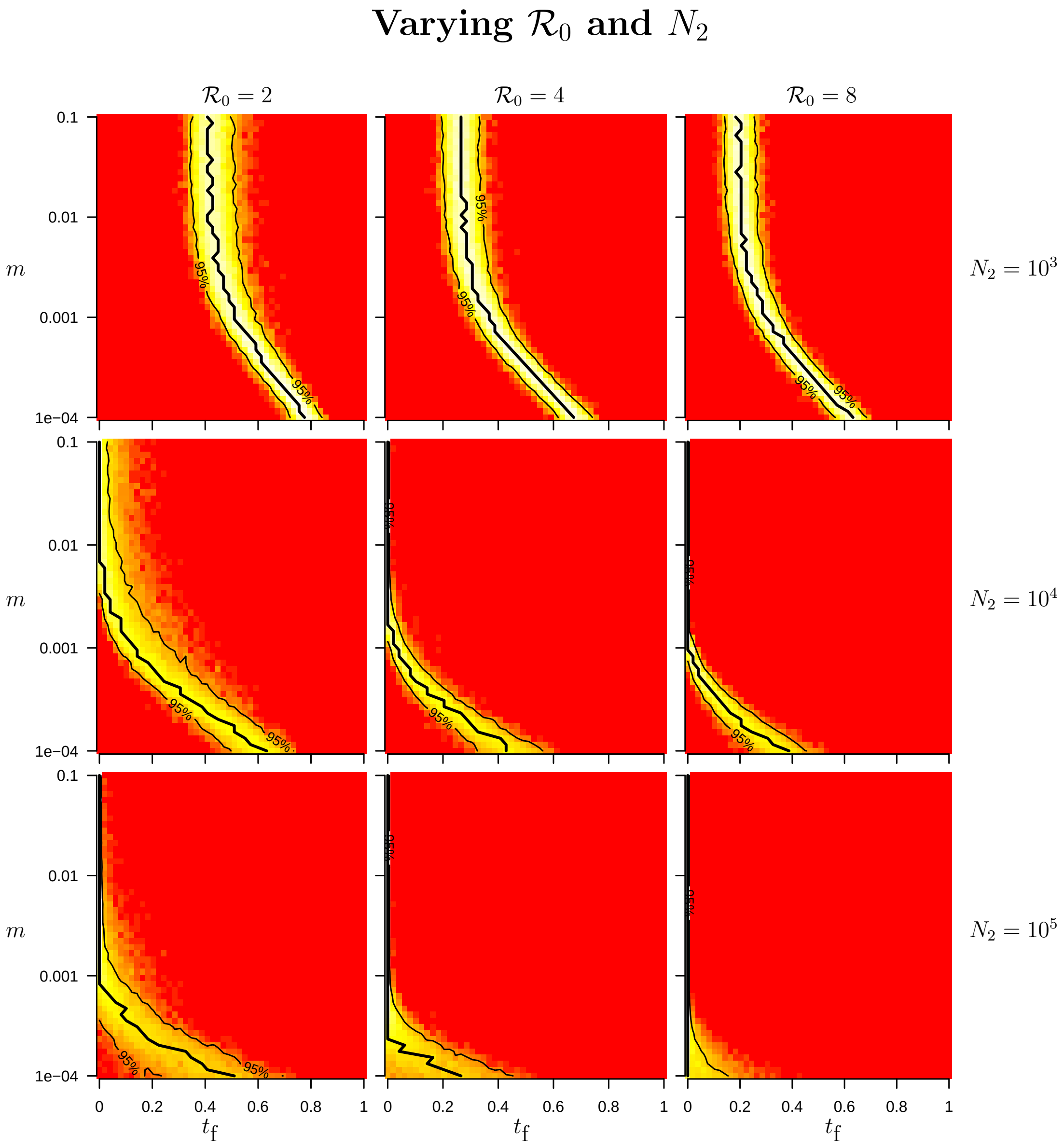

Figure 3.5: Likelihood of coupling parameter, m, given fadeout time, t_f: $\mathcal{L}(m|t_\mathrm{f})$ (Similar to Figure 3.3, with the same scale). Parameters: $\mathcal{R}_0 \in \{2, 4, 8\}$ (columns), $N_2 \in \{10^3, 10^4, 10^5\}$ (rows), with fixed $\alpha = 0.1$, $N_1 = 10^6$, and $\frac{1}{\gamma} = 10\,\mathrm{yr}$. Duration of time-series: 100 years. Each vertical slice is a likelihood profile for observed fadeout time, $t_{\mathrm{f,obs}}$, vs m. Produced from $n_m = 50$ different m values and $n_{\mathrm{sim}} = 500$ simulations each. Likelihood profiles are shown for 50 t_f values spaced uniformly from $[0, 1]$. Contours are shown for the maximum likelihood and 95% confidence intervals.

Varying α and N_2

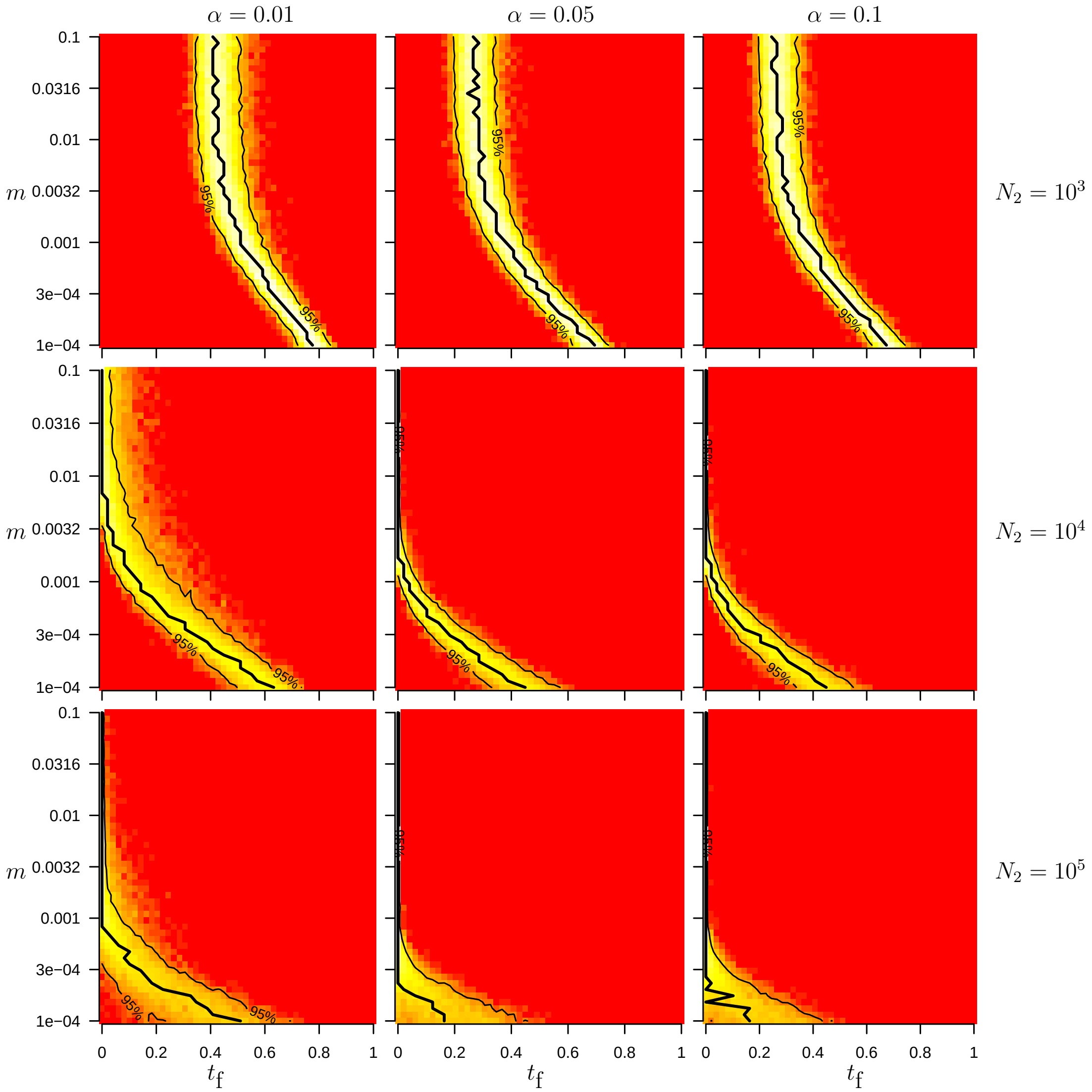

Figure 3.6: Similar to Figure 3.5, with $\alpha \in \{0.01, 0.05, 0.1\}$ (columns), $N_2 \in \{10^3, 10^4, 10^5\}$ (rows), and fixed $\mathcal{R}_0 = 4$.

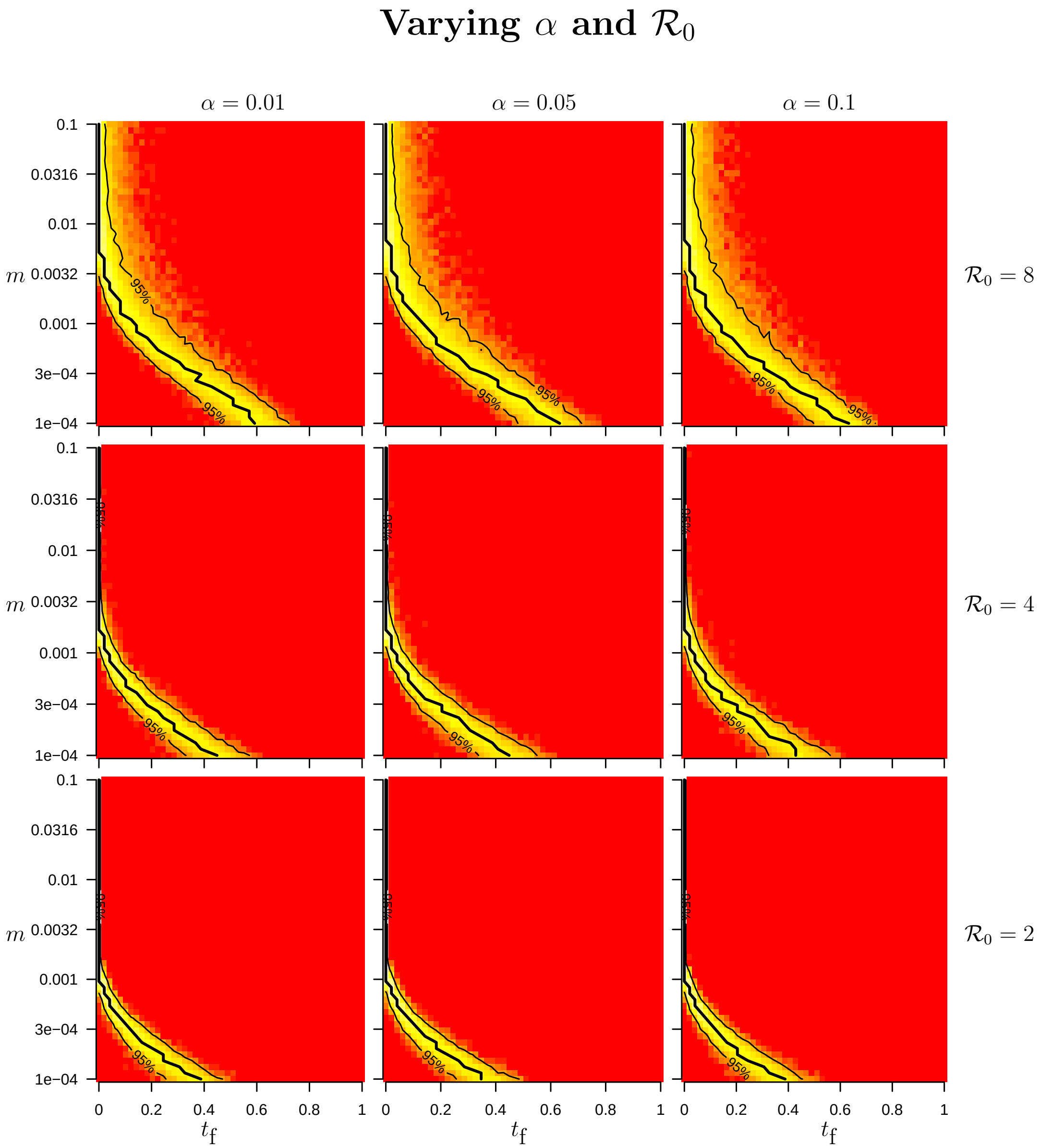

Figure 3.7: Similar to Figure 3.5, with $\alpha \in \{0.01, 0.05, 0.1\}$ (columns), $\mathcal{R}_0 \in \{2, 4, 8\}$ (rows), and fixed $N_2 = 10^4$.

3.4 Discussion

The use of the time the smaller population (population 2) spent faded out, t_f, as a probe to inform estimates of the coupling coefficient m can be successful under certain conditions. We show in various regions of parameter space (see Figures 3.5, 3.6, and 3.7) that likelihood profiles yield clear maxima. However, for all parameterizations displayed in the figures in §3.3.1, high values of the coupling parameter m are indistinguishable above a threshold that depends on the other underlying parameters. We separate these instances of unidentifiability into two cases.

Small population too large. Referring to the bottom left panel of Figure 3.6 ($\alpha = 0.01$ and $N_2 = 10^5$), we note that for all $m \geq 0.001$, the likelihood remains at its highest value for $t_\mathrm{f} \approx 0$. This phenomenon arises when the small population does not fade out in the observed time period for the majority of simulations throughout the upper range of m. If the small population does fade out it appears to be reinfected very quickly regardless of variation in m. Consequently, the small population rarely fades out.

Small population too small. Referring to the top left panel of Figure 3.6 ($\alpha = 0.01$ and $N_2 = 10^3$), we note that for $m \geq 0.01$, the likelihood remains at its highest value for $t_\mathrm{f} \approx 0.45$. In this case, for all values of the coupling parameter above some level, the small population remains faded out for some fixed amount of time (on average) despite the presence of some force of infection all of the time. This occurs in particular when N_2 is small (in our example, $N_2 = 1000$), and results from the depletion of susceptibles following outbreaks, preventing further reinfection despite the force of infection from the large population.

These two cases show a limitation of the method presented, namely that above some threshold, levels of coupling cannot be distinguished. The complete absence of fadeouts in a time-series naturally precludes use of this method, but for sufficiently small populations, t_f is uninformative even in presence of fadeouts.

The method we have presented shows the best possible case for using t_{f} as a probe for coupling, having assumed all other parameters are known and held fixed. It is evident from the results shown in §3.3.1 that the size of the population undergoing fadeouts strongly affects the relationship between likelihood of m and observed t_{f}. However, spatiotemporal disease case report data are usually accompanied by relatively accurate population and vital statistics, so population sizes can usually be estimated fairly accurately. The amplitude of the seasonal forcing driving the recurrent epidemics, α, does not strongly affect the relationship between likelihood of m and observed t_{f}, suggesting that accurate estimates of this amplitude are not needed to estimate coupling (this is fortunate, since α is difficult to estimate accurately). The disease parameters, $\mathcal{R}_0$ and γ, do affect the relationship between likelihood of m and observed t_{f}, and accuracy of coupling estimates will depend on accuracy of estimates of disease parameters. This cannot be avoided, since coupling between populations depends on the transmission rate of the disease.

The presented method explores the potential of the proportion of time faded out, t_{f}, as a tool for estimating coupling between large population centers and smaller satellite populations undergoing recurrent epidemics, and we identify key considerations in doing so. Other methods for estimating coupling could focus on the brief time period when infection re-invades the small population following a fadeout. However, aside from measuring the time of the re-invasion, the only other information informing the magnitude of the force of infection is the rate of growth of the outbreak in the small population. This depends on, among other things, the number of susceptibles present in the small population at the moment of invasion, which is not an observable quantity. Estimating the proportion of the population that is susceptible at any given time requires the reconstruction of the susceptible time series [105]. Susceptible reconstruction depends on consistently accurate statistics regarding susceptible recruitment and case reports throughout the time series, since sampling error accrues in the reconstruction process. If the relationship between serological markers of immunity and level of protection against infection is known, then susceptibility in a population can be assessed with serological surveys (for example, see [106]). Reporting inefficiency is much less likely to affect the time when a first case of infection is observed following a fadeout. A natural extension of this research would be using the distribution of the number of cases between observed fadeouts as a probe. Another potential alternative for the estimation of coupling in the presence of recurrent epidemics is observing the degree of synchrony between multiple populations [40, 59, 60]. Such a method would have the advantage of not requiring observed fadeouts, and thus being constrained by the sensitivity of fadeout patterns to population sizes. However, the driving causes of recurrent epidemics, such as seasonal changes in human contact rates, are typically common between coupled populations, and could produce synchrony independent of coupling. Moreover, once two populations are synchronized, coupling is likely very difficult to detect, and only observations of the populations becoming synchronized could inform estimates of coupling strength. An additional method for estimating coupling has been suggested by Schneeberger and Jansen, 2006 [107], who propose using covariance of fluctuations in prevalence to detect coupling.

3.5 Conclusion

Techniques for estimating epidemic coupling from spatiotemporal disease case report data are promising avenues of research for understanding and forecasting spatial epidemics. The effect of epidemic coupling between weakly coupled populations is largely obscured by local dynamics, but focusing on characteristics of the data that inform coupling through probe statistics can yield estimates. Total time spent with the disease absent in the smaller of two populations undergoing recurrent epidemics can inform estimates of the coupling strength between the populations, provided coupling is sufficiently weak. In all cases, levels of coupling between the populations above some threshold are indistinguishable.

Though the research presented here deals only with the estimation of coupling, assuming all other parameters are known, and assuming only two populations, the results are easily extended to encompass a larger scope of problems. The methods can be applied to real data for which disease and population parameter estimates are available, with sensitivity analyses measuring the dependence of estimates on error in parameters. Additionally, while we assume a large population and only one small population, t_{f} is a useful probe to estimate the force of infection that a small population is receiving in general. Future research could examine examples where this infection originates from numerous sources, or where numerous satellite populations are reinfected by one large population center. Finally, the estimates of coupling produced with this methodology is not, in principle, disease dependent. The predictive power of these methods could be tested in a context where recurrent epidemics of two or more diseases coincide, assuming the diseases share similar modes of transmission. In general, the exploration of more advanced methodology for estimating epidemic coupling from case reports alone, despite the notable difficulties in doing so, can nonetheless provide useful improvements in our understanding of and capacity to predict disease transmission.

3.6 Acknowledgements

KH was supported by an Ontario Graduate Scholarship (OGS). DE was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Chapter 4

Inferring contact patterns from observed mortality during the Great Plague of London, 1665

Abstract

Developing methods to understand and predict the manner in which infectious diseases spread within and among human populations is critical not only for the advancement of scientific understanding, but for the development of public health measures to control harmful transmissible infections. Since the transmission process itself is largely unobservable, methods for inferring patterns of transmission are extremely useful for epidemic modeling efforts. We consider the problem of estimating transmission coupling between populations, and estimate coupling in the city of London, England, during the Great Plague of 1665. Estimates are produced from weekly mortality reports for 130 parishes contained in the London Bills of Mortality. We model each parish as a compartmental SIR (*susceptible-infected-recovered*) system, where the parishes are coupled through the transmission process with one of four spatial coupling schemes. We show that the degree of coupling among parishes and the basic reproduction number can be estimated, with better fits for the two least geographically constrained coupling schemes.

4.1 Introduction

Mathematical models are widely used to describe biological systems, and have greatly enhanced our ability to understand and forecast the spread of infectious diseases [9, 13, 15–18, 68]. In particular, mathematical modeling of transmission, whether within or among geographically separated populations, provides useful opportunities to increase our understanding of how diseases spread, since the transmission process is very difficult to observe in practice. We focus on the problem of estimating coupling between geographically separated populations, using methodology that exploits spatiotemporal incidence or mortality reports to produce estimates of the degree of coupling in a population over the course of an observed epidemic.

Restricting the type of data used to only spatiotemporal incidence or mortality reports has numerous advantages. Recent years have seen a dramatic increase in the quantity of available digitized spatiotemporal infectious disease data [13, 24, 64, 68, 92, 93], making the development of methodology to exploit such data valuable. Even if methodology for estimating spatial parameters includes other data regarding spatial transmission (such as data regarding host movement, see [46, 47] for example), a better understanding of the degree to which such data can inform estimates is important. Finally, in circumstances when no other quantitative information descriptive of coupling is available, as is the case presented in this paper, the only approach available is methodology applicable to incidence or mortality reports.

Our goal in this paper is to present the application of methodology capable of estimating coupling strength in a meta-population using reported mortality data during a single epidemic. The data in question are of an epidemic of plague that took place in the city of London, England, in the year 1665. The data are contained within the London Bills of Mortality (LBoM), an extensive and diverse set of records detailing the deaths of residents of the city of London from 1662–1829, recently digitized [64]. The LBoM contain weekly reports of deaths from plague in the 130 parishes of the city during the 1665 so-called Great Plague of London (GPL), which killed approximately 20% of the city's residents [108]. These data show a devastating epidemic spreading through its many geographically distributed parishes in sufficiently high spatial and temporal resolution to facilitate the estimation methodology we present.

We use a meta-population model where each population is defined as an SIR system (*susceptible-infected-recovered*). The SIR model approximates contact within a population as being well-mixed (contact occurs uniformly at random), and where infection confers permanent immunity upon recovery [9]. Various areas of study aim to further develop components of the basic model, such as the effects of seasonal forcing, intensity and duration of infectiousness [70–75], vital dynamics [69], network structure within populations [41–44], and others. We model coupling between parishes through the transmission process by parameterizing the proportion of time individuals spend interacting with individuals distributed throughout the meta-population. We implement four different contact structures in our model (See §4.3).

Methods for estimating model parameters vary greatly depending on the characteristics of real world data that the model is intended to capture. We are primarily interested in estimating the degree of coupling among parishes in the city of London, which we capture with a single parameter m (see §4.3). We also, simultaneously, estimate the basic reproduction number $\mathcal{R}_0$, which quantifies the potential a disease has to spread within a population (see §4.3 for description of $\mathcal{R}_0$). $\mathcal{R}_0$ has been estimated for pneumonic plague in modern settings [109], but we do not know if these estimates are appropriate for the study of an outbreak over 350 years ago in a pre-industrial population. We use a *probe-matching* [58] method to estimate both m and $\mathcal{R}_0$ (see §4.4), comparing the real-world LBoM data with large numbers of stochastic simulations. We complete the estimation procedure for each of the four spatial contact structures to investigate the significance of geographic distance in the spread of the GPL §4.5.

4.2 Data describing the GPL

The plague, or Black Death, arrived in and spread throughout Europe in the 14th century, resulting in the death of approximately one third of its population [67]. The city of London, England, sustained repeated epidemics of plague over centuries since the initial European pandemic of Black Death in 1348, and saw the last of these epidemics in 1665 [65,66] during what is commonly referred to as the GPL. Based on reports in the London Bills of Mortality (§4.2.2), this epidemic killed approximately 70,000 people of a total population of approximately 400,000 [110], accounting for nearly 17% of the population[1]. The weekly reports of Great Plague deaths available in the Bills of Mortality are distributed among 130 parishes. The fine spatiotemporal detail in these digitized data permit the analysis of spatial spread of the epidemic presented in this paper. We begin by describing the nature of the disease and sources of data used.

4.2.1 Causative agent and natural history of infection

Plague is caused by the bacterium *Yersinia pestis*, shown to have been responsible for the Plague of Justinian, the European Black Death, and modern plague [1, 2]. The infection of humans by this pathogen is categorized in one of three ways: bubonic, septicemic, and pneumonic plague [111]. Bubonic and septicemic plague refer to

[1] The true percentage was almost certainly higher, since only Christian burials are recorded in the LBoM.

the infection of the lymphatic system and blood stream, respectively, and can have numerous causes, including infections from flea bites, which in turn can carry the pathogen from small rodents such as rats. Pneumonic plague refers to the infection of the respiratory tract through airborne droplets containing pathogen particles, and can be spread directly from human to human. Bubonic plague is fatal in 40-70% of cases, and virtually always fatal in its septecemic and pneumonic forms. A single epidemic may contain one or more types of plague, and may spread by numerous modes of transmission [112]. It is not known which types of plague and mode of transmission were present or dominant during the GPL.

We use estimates for pneumonic plague [109] to obtain a mean infected period of 6.8 days (summing the estimated mean latent period of 4.3 days and mean infectious period of 2.5 days). We do not explicitly represent vector transmission in our model, since we are not aware of parameter estimates necessary to produce such a model. Our results are, therefore, heavily contingent on the assumption that the primary driver of spread during the Great Plague was human-to-human transmission[2]. The SIR model we use removes both recovered and deceased individuals from the transmission pool, and thus our results are unaffected by the accuracy of estimated disease-induced mortality. The difference between the types of plague are practically very significant, and differences in the nature of transmission intuitively impact patterns of spatial spread. However, we are not able, in this study, to distinguish between these types of transmission (see §4.3).

[2] An example of modeling plague with a subpopulation of rats can be found in Keeling and Gilligan, 2000 [113]. It would be interesting to investigate the effect of a rat population on our results. This would require either data or assumptions regarding the number and spatial distribution of rats, the rates of transmission between rats and humans (which can occur through fleas as well as directly), and the spatial transmission dynamics among the rats themselves.

4.2.2 The London Bills of Mortality

In the 16th century, frequent outbreaks of plague in and around London prompted efforts by London's city administrators to record deaths during these outbreaks [65, Ch. 6]. Few of these early bills of mortality survive. They generally follow the resurgence of plague in the city, and were discontinued soon after the temporary fadeout of plague. However, plague observed throughout England along with cases in the vicinity of the city resulted in the commencement of weekly record-keeping in the Bills, at first sporadically in 1563, and then continuously in 1662.

Though records were not kept for all parishes in the country-side around London, records for 130 parishes—including all parishes within the city walls—were kept for the full duration of the epidemic, including the first recorded death of the epidemic in late 1664. The early commencement of record-keeping during this epidemic is particularly relevant to our case-study, since most of the information regarding the spatial spread of the epidemic is found in the early stages of the outbreak. We show spatial coverage of the LBoM in Figure 4.1, including the location of the first reported death of the epidemic. We furthermore make use of published estimates of parish populations [108] to produce initial conditions needed to generate stochastic simulations (see §4.3 for details regarding our simulation model).

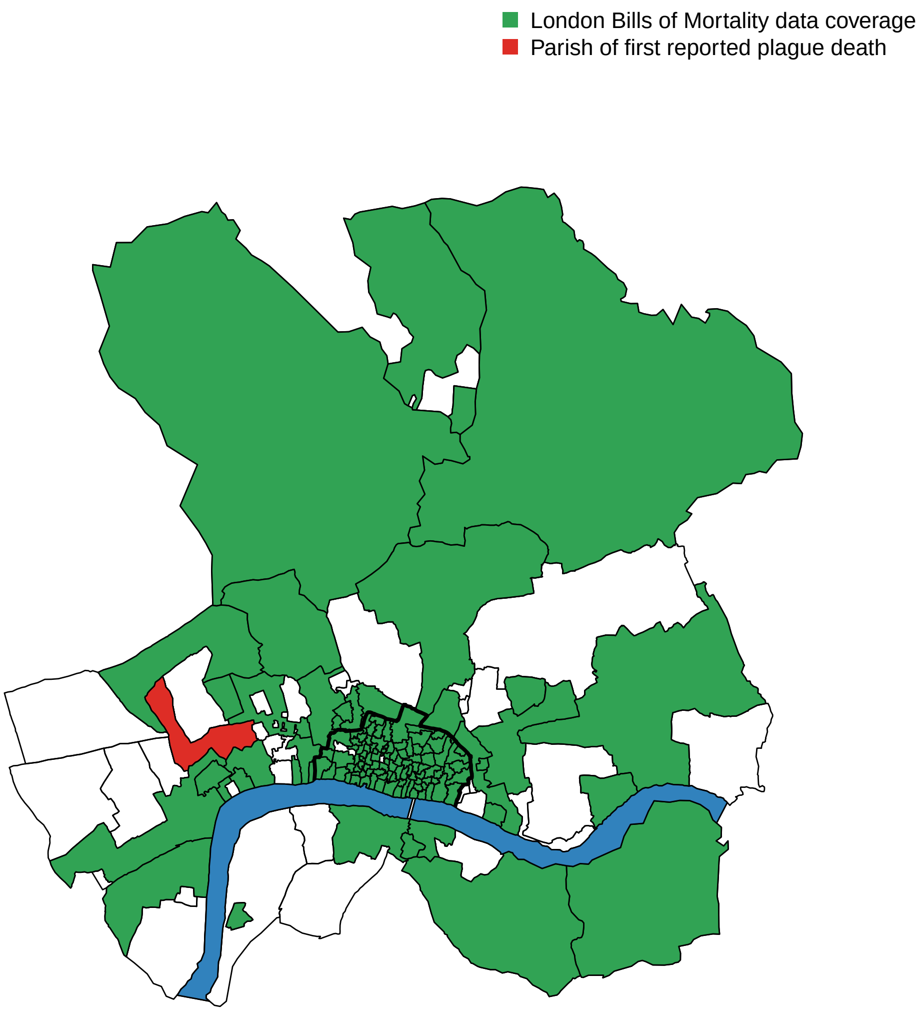
🟩 London Bills of Mortality data coverage
🟥 Parish of first reported plague death

Figure 4.1: Map of data coverage throughout the parishes of London in 1665. We use parish-level imputed population [108] and weekly plague mortality reports from our LBoM plague data (also used in Tien et al. [64]). The parish St. Giles in the Fields (shown in red) saw the first plague death of the great plague in late 1664 [65, Ch. 12, pp. 679–682]. Thick black lines show the city walls. The Thames river is shown in pale blue.

4.2.3 Epidemic onset

In §4.4, we show the process of estimating the degree of coupling from the LBoM data, but we make an assumption in advance of using this methodology. We assume that in the geographically distributed population of London, there is significantly more contact between individuals living in the same parish than between individuals living in different parishes. Thus from the outset, we expect the degree of coupling among parishes to be small relative to local contact. As a result, once the plague has begun to spread within a parish, it becomes very difficult to detect the effect of infections from other parishes. Thus the most useful information concerning the spatial spread of the epidemic is found from the times of observing first cases of plague throughout the parishes of London. We therefore use a *summary statistic* of the data for the estimation of coupling, comparing this statistic of the LBoM data to that of stochastic simulations. The summary statistic we use is the number of parishes reporting their first death due to plague in each week of 1665, which we refer to throughout this paper as the *epidemic onset distribution*[3]. We show the epidemic onset distribution for the data from the LBoM in Figure 4.2.

[3] spatiotemporal data describing epidemic onset has been used to characterize spatial transmission rates elsewhere. See, for example, Smith et al. 2002, which examines the spatial spread of rabies [114].

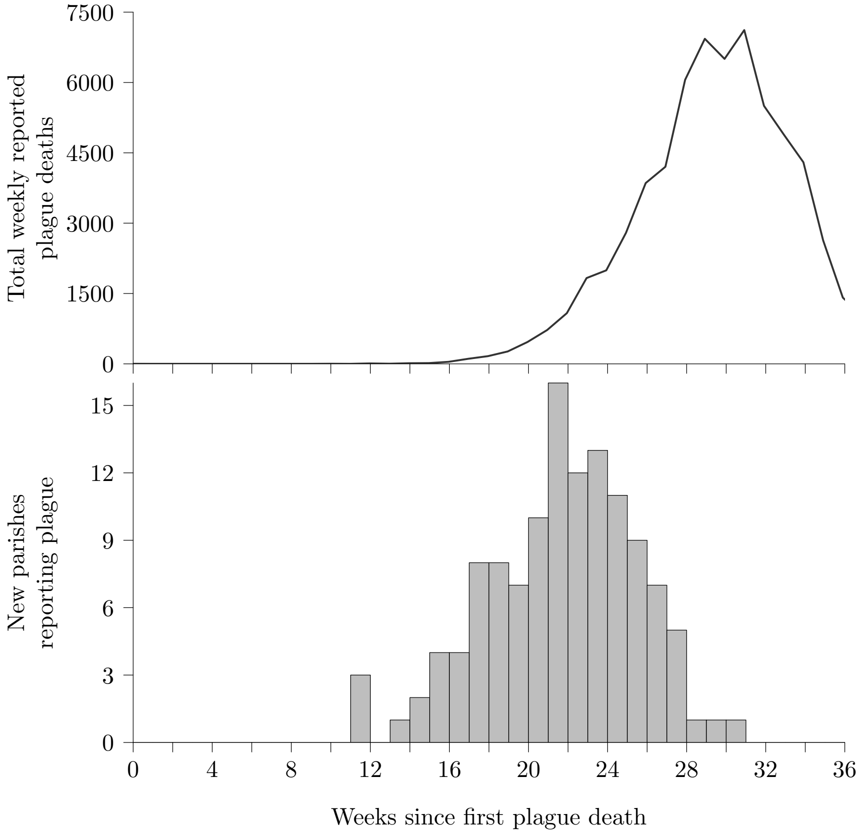

Figure 4.2: *Top:* Weekly deaths from plague during the Great Plague as reported in the London Bills of Mortality.
Bottom: Distribution of parishes by week of first reported plague death during the Great Plague of London, 1665. The first recorded plague death occurred in the parish St. Giles in the Fields the week of December , 1664 [65, Ch. 12, pp. 679–682].

4.3 Modeling the Spread of the Great Plague

In order to estimate coupling m and basic reproduction number $\mathcal{R}_0$, we construct a stochastic simulation model that takes these parameters as input, and produces data resembling the GPL for comparison using probe-matching (*cf.* §4.2.3 and §4.4). We begin by defining the meta-population compartmental model as a deterministic system of ordinary differential equations, and then use the transition rates of this system to define event rates in a stochastic simulation model.

4.3.1 Deterministic simulation model

We represent the 130 parishes of the city of London in 1665 as $n_{\mathrm{P}} = 130$ coupled populations in a meta-population model. The dynamics of disease spread within each population are modeled using an SIR (*susceptible-infected-recovered*) system, and coupling between populations occurs through the transmission process. We define the rates of change governing dynamics for the resident populations of each parish using the following system of ODEs

$$\frac{\mathrm{d}S_i}{\mathrm{d}t} = -S_i\Lambda_i \,, \tag{4.1a}$$

$$\frac{\mathrm{d}I_i}{\mathrm{d}t} = S_i\Lambda_i - (\gamma + \mu_{\mathrm{d}})I_i \,, \tag{4.1b}$$

$$\frac{\mathrm{d}R_i}{\mathrm{d}t} = \gamma I_i \tag{4.1c}$$

where S_i, I_i, and R_i represent the number of susceptible, infected, and removed individuals in population i, respectively, and $N_i = S_i + I_i + R_i$. The ***force of infection*** acting on susceptible members of population i, Λ_i, depends on meta-population cross-coupling, which we define precisely in §4.3.2. γ is the rate of recovery from infection, and μ_{d} is the rate of death from infection. These two rates of leaving the infected class result in a mean time infected of $\frac{1}{\gamma+\mu_{\mathrm{d}}}$. We fix the mean time an individual spends in the infected class to be 6.8, noting that this combines both latent and infectious periods (see §4.2.1). The latent and infectious stages of infection can be modeled explicitly, but they are short relative to the weekly temporal resolution of the LBoM data, and so we consider only a single infected class I.

We do not model births or natural deaths, due to the short time-period studied in this paper, and as a result the total population of each infected parish decreases over the course of the epidemic due to disease-induced mortality. The basic reproduction number $\mathcal{R}_0$ is defined as the mean number of new infections caused by a single infected individual in a completely susceptible population. We emphasize here that our definition of $\mathcal{R}_0$ is for an individual parish in the absence of coupling, rather than for the meta-population as a whole.

The SIR model represents situations in which individuals become infected with a disease at most once. It is appropriate in situations where individuals either die or acquire immunity, or when the time interval being considered is sufficiently short to preclude waning immunity and reinfection. It is appropriate for the GPL because the greater part of the epidemic took place in the span of five months in 1665. The SIR model assumes human-to-human transmission, which can occur in the spread of pneumonic plague (see §4.2.1). We note the omission of any mechanism representing the potential of vector transmission, through small rodents such as rats, of plague during the GPL. We cannot distinguish types of plague infections from the LBoM, and have no empirical information for the inclusion of vector transmission mechanisms in our model.

4.3.2 Form of transmission coupling

We implement coupling by assuming n_{P} distinct geographic *patches*, along with n_{P} distinct *populations*, where a member of population i is defined as a resident of patch i. We assume that infection within patch i is driven by mixing according to the law of mass action, so the rate at which new infections occur (incidence) is

$$\beta S_i I_i / N_i \tag{4.2}$$

where β is the transmission rate [115]. For clarity, we index the compartments S_j, I_j, and N_j always to mean members of *population* j. Members of population j are residents of patch j, but may be visiting other patches at a given time. We define the levels of individual movement among patches with the contact matrix (c_{ij}), where

$$c_{ij} = \text{proportion of members of population } j \text{ visiting patch } i \text{ at any time.} \tag{4.3}$$

We do not explicitly model movement, but use the contact matrix (c_{ij}) to define rates of infection.[4] We note that (c_{ij}) is column-stochastic, *i.e.,* all elements of a column sum to 1. Considering patch i, and taking into account members of the local population currently absent, and visiting members of other populations present, the total number of individuals in patch i at any given time is

$$\sum_{k=1}^{n_\mathrm{P}} c_{ik} N_k \,. \tag{4.4}$$

Likewise, the total number of infected individuals in patch i is

$$\sum_{k=1}^{n_\mathrm{P}} c_{ik} I_k \,. \tag{4.5}$$

Now considering only the proportion of susceptible individuals from population j that are currently visiting patch i, the rate of infection is

$$-\frac{\mathrm{d}}{\mathrm{d}t}(c_{ij} S_j) = \beta c_{ij} S_j \frac{\sum_{k=1}^{n_\mathrm{P}} c_{ik} I_k}{\sum_{k=1}^{n_\mathrm{P}} c_{ik} N_k}. \tag{4.6}$$

[4] Our formulation of implicit movement allows an infected individual to simultaneously affect a force of infection on all other individuals in the meta-population, and maybe therefore infect two individuals in difference patches closely in time. Coupling could be implemented such that individuals only interact with individuals in the same patch as they are resident or visiting. The difference between these implementations is analogous to that between deterministic and stochastic simulation in that we model individuals mix partially in all patches simultaneously, rather than completely in one patch at a time.

Members of population j are distributed throughout the patches, and we can obtain the total rate of new infections for population j by summing up the rates of infection for each of the patches i.

$$-\frac{\mathrm{d}S_j}{\mathrm{d}t} = -\frac{\mathrm{d}S_j}{\mathrm{d}t}\sum_{i=1}^{n_\mathrm{P}} c_{ij} = -\sum_{i=1}^{n_\mathrm{P}} c_{ij}\frac{\mathrm{d}S_j}{\mathrm{d}t} = \sum_{i=1}^{n_\mathrm{P}} \beta c_{ij} S_j \frac{\sum_{k=1}^{n_\mathrm{P}} c_{ik} I_k}{\sum_{k=1}^{n_\mathrm{P}} c_{ik} N_k}\,, \tag{4.7a}$$

$$= \beta S_j \sum_{i=1}^{n_\mathrm{P}} c_{ij} \frac{\sum_{k=1}^{n_\mathrm{P}} c_{ik} I_k}{\sum_{k=1}^{n_\mathrm{P}} c_{ik} N_k}. \tag{4.7b}$$

From this we complete our definition of the force of infection introduced in §4.3,

$$\Lambda_j = \beta \sum_{i=1}^{n_\mathrm{P}} c_{ij} \frac{\sum_{k=1}^{n_\mathrm{P}} c_{ik} I_k}{\sum_{k=1}^{n_\mathrm{P}} c_{ik} N_k}, \tag{4.8}$$

where Equation (4.8) refers to the rate, per unit time, at which susceptible members of population j become infected.

This formulation simplifies to n_P uncoupled SIR systems if we take (c_{ij}) to be the identity matrix ($c_{ij} = 1$ if $i = j$, and $c_{ij} = 0$ if $i \neq j$), since in that case

$$-\frac{\mathrm{d}S_j}{\mathrm{d}t} = \beta S_j \sum_{i=1}^{n_\mathrm{P}} c_{ij} \frac{\sum_{k=1}^{n_\mathrm{P}} c_{ik} I_k}{\sum_{k=1}^{n_\mathrm{P}} c_{ik} N_k}\,, \tag{4.9a}$$

$$= \beta S_j \frac{\sum_{k=1}^{n_\mathrm{P}} c_{jk} I_k}{\sum_{k=1}^{n_\mathrm{P}} c_{jk} N_k}\,, \tag{4.9b}$$

$$= \beta S_j \frac{I_j}{N_j} \tag{4.9c}$$

noting the change of index from i to j in the fraction in the second step. This formulation also simplifies to a single SIR system, such that members of all populations are indistinguishable, if $c_{ij} = \frac{1}{n_{\mathrm{P}}}$ for all i, j, as follows

$$
\begin{aligned}
-\frac{\mathrm{d}S_j}{\mathrm{d}t} &= \beta S_j \sum_{i=1}^{n_{\mathrm{P}}} c_{ij} \frac{\sum_{k=1}^{n_{\mathrm{P}}} c_{ik} I_k}{\sum_{k=1}^{n_{\mathrm{P}}} c_{ik} N_k}\,, &\text{(4.10a)}\\
&= \beta S_j \frac{\sum_{k=1}^{n_{\mathrm{P}}} c_{ik} I_k}{\sum_{k=1}^{n_{\mathrm{P}}} c_{ik} N_k}\,, &\text{(4.10b)}\\
&= \beta S_j \frac{\sum_{k=1}^{n_{\mathrm{P}}} I_k}{\sum_{k=1}^{n_{\mathrm{P}}} N_k}. &\text{(4.10c)}
\end{aligned}
$$

Note that the force of infection does not depend on j, so if we take $S = \sum_{j=1}^{n_{\mathrm{P}}} S_j$, $I = \sum_{j=1}^{n_{\mathrm{P}}} I_j$, $N = \sum_{j=1}^{n_{\mathrm{P}}} N_j$, we have

$$
\begin{aligned}
-\frac{\mathrm{d}S}{\mathrm{d}t} = -\sum_{j=1}^{n_{\mathrm{P}}} \frac{\mathrm{d}S_j}{\mathrm{d}t} &= \sum_{j=1}^{n_{\mathrm{P}}} \beta S_j \frac{\sum_{k=1}^{n_{\mathrm{P}}} I_k}{\sum_{k=1}^{n_{\mathrm{P}}} N_k}\,, &\text{(4.11a)}\\
&= \beta \sum_{j=1}^{n_{\mathrm{P}}} S_j \frac{I}{N}\,, &\text{(4.11b)}\\
&= \beta \frac{SI}{N}. &\text{(4.11c)}
\end{aligned}
$$

We also verify that the total number of effective contacts[5] per unit time between individuals of population j and population k is the same, whether viewed from the perspective of population j or k. It follows from the definition of c_{ij} in Equation (4.3) and the specification of the infection rate in Equation (4.6) that the number of effective contacts per unit time between population j and population k in patch i is given by

$$
\frac{c_{ij} N_j c_{ik} N_k}{\sum_{\ell=1}^{n_{\mathrm{P}}} c_{i\ell} N_\ell}\,. \tag{4.12}
$$

The total number of effective contacts between populations j and k is obtained by

[5] We say a contact event between two individuals is effective if transmission will occur if one of the individuals is infectious and the other is susceptible.

summing Equation (4.12) over all patches i,

$$\sum_{i=1}^{n_{\mathrm{P}}} \frac{c_{ij} N_j c_{ik} N_k}{\sum_{\ell=1}^{n_{\mathrm{P}}} c_{i\ell} N_\ell} = \frac{1}{\sum_{\ell=1}^{n_{\mathrm{P}}} c_{i\ell} N_\ell} \Big(\sum_{i=1}^{n_{\mathrm{P}}} c_{ij} c_{ik} \Big) N_j N_k \,. \tag{4.13}$$

we obtain equivalent expressions whether we consider the number of effective contacts with population k seen by population j or vice-versa.

We define the elements of the coupling matrix (c_{ij}) by means of the parameter m, which represents the average proportion of time residents of one parish spend in any other parish. Thus we define the diagonal entries of the contact matrix $c_{ii} = 1 - m$ for $1 \leq i \leq n_{\mathrm{P}}$. The sum of all entries in a column not found on the diagonal is therefore the degree of parish cross-coupling,

$$m = \sum_{j=1}^{n_{\mathrm{P}}} c_{ij} \,, \qquad j \neq i. \tag{4.14}$$

The precise values of off-diagonal entries of the contact matrix (c_{ij}) are defined depending on the type of contact structure used. As noted in §4.1, a central aim of this paper is to determine the importance of geographic location in the spread of the GPL throughout the city as detectable from the mortality reports alone. We incorporate geographic information in the modeled contact structure by filling the off-diagonal entries using the three schemes. We begin by defining the off-diagonal entries of a matrix (a_{ij}) based on each scheme, and we then scale the rows of (a_{ij}) such that Equation (4.14) is satisfied, thus obtaining (c_{ij}),

$$c_{ij} \equiv \begin{cases} m \, \frac{a_{ij}}{\sum_{k=1,k\neq i}^{n_{\mathrm{P}}} a_{ik}} & i \neq j \\ 1 - m & i = j \end{cases} . \tag{4.15}$$

The three contact schemes we use are as follows:

1. *Uniform:* All off-diagonal entries of (c_{ij}) are equal:

$$a_{ij} = \frac{1}{n_{\mathrm{P}} - 1}\,, \qquad i \neq j. \tag{4.16}$$

   This uniform coupling scheme ignores distance between parishes, and thus assumes distance has no effect on disease spread.

2. *Gravity:* Off-diagonal entries scaled inversely with the square of the distance between parish i and parish j [21, 116]:

$$a_{ij} = \frac{1}{d_{ij}^2}\,, \qquad i \neq j. \tag{4.17}$$

   Where d_{ij} refers to euclidean distance between parish i and parish j (computed using the centroids of the parishes as shown in the map in Figure 4.1). Gravity coupling is typically defined as proportional to $\frac{N_i N_j}{d_{ij}^2}$, but standard transmission already contains factors in units of the coupled populations, namely S and I (see Equation (4.2)). Gravity coupling takes geographic proximity into account while ignoring the city layout.

3. *Near-Neighbour:* Off-diagonal entries are scaled with a power law through nearby parishes

$$a_{ij} = \begin{cases} m^p\,, & p \leq 4 \\ 0\,, & p > 4. \end{cases} \tag{4.18}$$

   Where p refers to the degree of separation between parish i and j, and $p = 1$ between parishes that share an edge (see Figure 4.1). Note that we limit the degrees of separation for which coupling is non-zero in this scheme. This results in coupling being heavily constrained by local neighbourhood, and geographic barriers such as the River Thames and the city walls become relevant. We test two implementations of this scheme, both including and precluding infections across the city wall.

4.3.3 Stochastic Simulations

We produce stochastic simulations using the rates in Equation (4.1) as event probabilities, using an adaptive time-step approximation algorithm. Methods for computing exact stochastic simulations from rate equations exist [83, 100], which require event rates to remain fixed while no event occurs. For our purposes, however, sampling one event at a time is far too computationally costly. Adaptive time-step methodology, or "tau-leaping" [100], samples many events over some time step from either Poisson or Binomial distributions parameterized by the rate equations. These methods are approximations, and balance the trade-off between accuracy and computational cost by adjusting time step length while simulating. We use the "tau-leaping" methods implemented in the `adaptivetau` package in R.

The information available to us about the spread of the plague in London is mortality data reported weekly by parish, and thus the observable quantity in our simulation model is disease-induced mortality. The stochastic simulation model produces unobserved states, and samples the total number of disease induced deaths at the desired weekly interval, for each parish. Disease incidence can often be significantly under-sampled since not all instances of infection are reported or documented. In the case of the London parishes, officials were tasked with recording a cause of death for burials, and though the plague was widespread and recognizable, it is likely that there is underreporting of plague-induced mortality in the LBoM. We rely on the week of the first plague reports being correct, which is affected by underreporting when a previously uninfected parish fails to report any of its first cases. It is possible that incidences of plague were, in some cases and for variable amounts of time, deliberately concealed, but we do not have information to control for this. We furthermore do not take into account a delay between the time of plague death and the time of its reporting. We have no information about the distribution of this delay, so we assume it to be roughly equal for all parish plague reports, that it is on the order of the weekly time resolution of reporting, and since we are concerned with the relative times of plague onset in the different parishes (see §4.4), that it does not significantly affect our results.

4.4 Estimating spatial transmission parameters

We estimate the coupling parameter m and the basic reproduction number $\mathcal{R}_0$ using maximum likelihood inference [58]. In §4.2.3, we describe the summary statistic of the epidemic onset distribution which we use for statistical inference. We now describe how we use this summary statistic in conjunction with simulated data to produce maximum likelihood estimates of m and $\mathcal{R}_0$.

We label the weeks since the first recorded plague death as $1 \leq k \leq n_{\text{weeks}}$, where we take the number of weeks, $n_{\text{weeks}} = 32$, to be the number of weeks prior to the end of the epidemic. If y is either the observed time series or a simulation of the GPL, we define the function g such that $g(y, k)$ is the number of parishes reporting their first plague death in week k. We define x to be a stochastic simulation sampled from X_θ, where the parameter set $\theta = \{m, \mathcal{R}_0\}$ is the subset of model parameters we wish to estimate, assuming all other parameters are held fixed, and X_θ is the set of all realizations possible from θ.

To estimate parameters θ from the GPL data y, we estimate a probability of observing y given θ. To this end, we generate $n_{\text{sim}} = 100$ stochastic simulations, $\{x_i\}_{i=1}^{n_{\text{sim}}}$. From these we obtain the mean number of new parishes reporting plague each week,

$$\overline{x_k} = \overline{\{g(x_i,k)\}_{i=1}^{n_{\text{sim}}}} \tag{4.19}$$

If we assume that deviations from $\overline{x_k}$ are approximately normally distributed [117][6], we can obtain an expression for the probability of observing $g(y,k)$,

$$p(g(y,k)\,|\,\theta) \approx \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(g(y,k)-\overline{x_k})^2}{2\sigma_k^2}} \tag{4.20a}$$

$$\sigma_k^2 = \operatorname{var}(\{g(x_i,k)\}_{i=1}^{n_{\text{sim}}}) \tag{4.20b}$$

To obtain an expression for the probability of observing y, if we assume independence of deviations from the mean, we take the product of Equation (4.20) over all the weeks of the GPL[7],

$$p(y\,|\,\theta) \approx \prod_{k=1}^{n_{\text{weeks}}} \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(g(y,k)-\overline{x_k})^2}{2\sigma_k^2}} \tag{4.21}$$

We use Equation (4.21) to estimate θ using maximum likelihood estimation (MLE). The likelihood of θ given y, $\mathcal{L}(\theta\,|\,y)$, is defined to be $p(y\,|\,\theta)$. We adhere to the

[6] Alternatively, one could use the observed probability distributions of $\{g(x_i,k)\}_{i=1}^{n_{\text{sim}}}$, provided they can be sufficiently sampled. We found that $n_{\text{sim}} = 100$ simulations per parameter set θ were insufficient to do so, and assumed normally distributed deviations from the mean due to computational limitations.

[7] The assumption that deviations from the mean number of onsets each week are independent is an approximation, since each realization has only a fixed total number of onsets in all weeks.

convention of minimizing the negative log-likelihood,

$$\begin{aligned}
-\log\mathcal{L}(\theta\,|\,y) &= -\log[p(g(y,k)\,|\,\theta)] && \text{(4.22a)}\\
-\approx \log&\left\{\prod_{k=1}^{n_{\text{weeks}}}\frac{1}{\sqrt{2\pi\sigma_k^2}}e^{-\frac{(g(y,k)-\overline{x_k})^2}{2\sigma_k^2}}\right\} && \text{(4.22b)}\\
&= -\sum_{k=1}^{n_{\text{weeks}}}\left\{\log(\sqrt{2\pi\sigma_k^2})+\frac{[g(y,k)-\overline{x_k}]^2}{2\sigma_k^2}\right\} && \text{(4.22c)}
\end{aligned}$$

Note that the likelihood function given in Equation (4.22) is a *synthetic likelihood* [58], comparing the epidemic onset distributions of simulations and LBoM data, rather than the spatiotemporal mortality reports themselves.

We find the maximum likelihood estimate of θ by computing $-\log\mathcal{L}(\theta\,|\,y)$ for a grid of values of θ, and identifying the θ with the least negative log-likelihood. For 21 values of $\mathcal{R}_0 \in [1.0625, 2]$ and 32 values of $m \in [10^{-3.5}, 10^{-0.5}]$, we compute $n_{\text{sim}} = 100$ simulations for each combination of $\mathcal{R}_0$ and m, and plot the corresponding $-\log\mathcal{L}(\theta\,|\,y)$ in Figure 4.3. To generate Figure 3, a total of $n_{\mathcal{R}_0}\times n_m\times n_{\text{sim}} = 67,200$ simulations were required[8]. We compute this grid of log-likelihoods for the four spatial coupling schemes: uniform, gravity, and near-neighbour with and without coupling between parishes on opposite sides of the city wall (*cf.* Figure 4.1 and §4.3.2).

[8] This took 253,232 CPU hours on the SHARCNET server “Orca”. Jobs were run on 2688 cores (168 nodes × 16 cores), where each core operates maximally at $2.6-2.7$ GHz, with $32-128$ GB memory. SHARCNET (`www.sharcnet.ca`) is a consortium of 18 colleges, universities and research institutes operating a network of high-performance computer clusters across south western, central and northern Ontario.

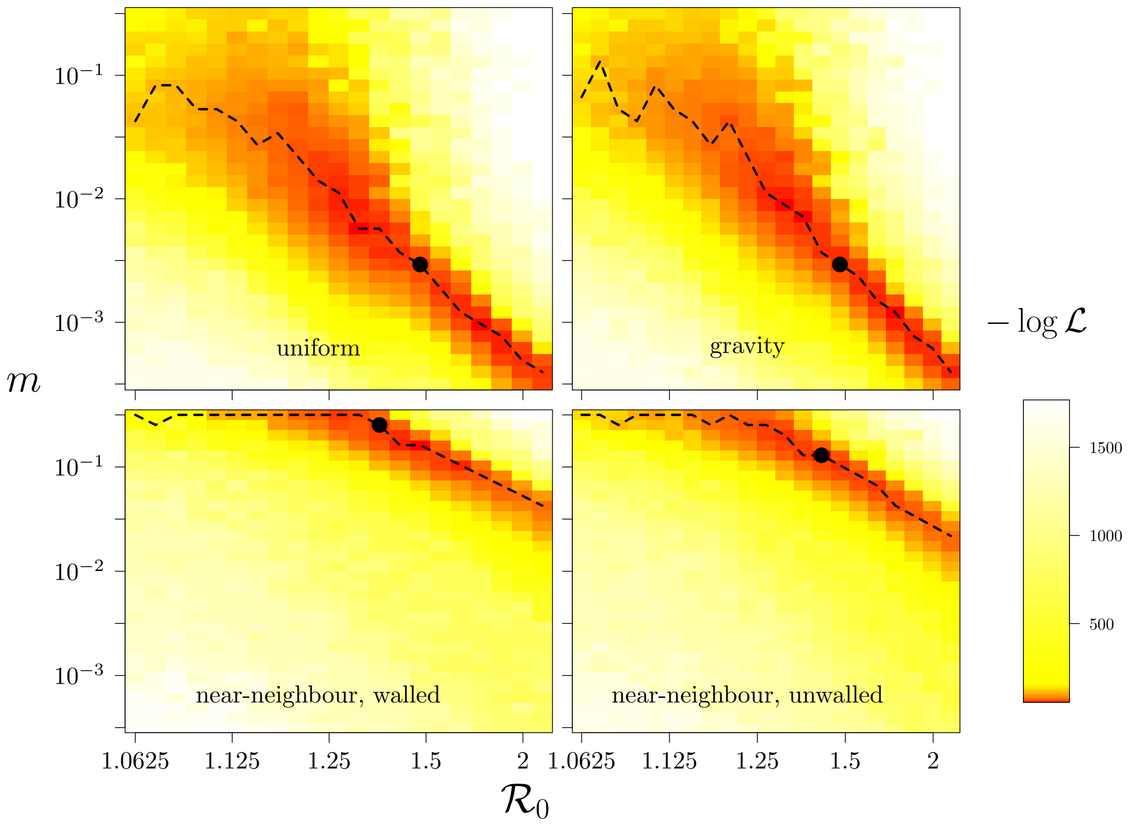

Figure 4.3: Negative log-likelihood of parameter pairs $\theta = \{m, \mathcal{R}_0\}$ given the LBoM data y, $-\log\mathcal{L}(\theta\,|\,y)$ (see Equation (4.22)). Each grid cell was produced from 100 simulations. Dotted black line shows the likelihood profile, with the solid dot showing the maximum likelihood estimate. The four panels shown correspond to the four coupling schemes used (see Equations (4.16), (4.17), and (4.18)).

To obtain confidence limits on our estimates of θ, we first compute likelihood profiles with respect to $\mathcal{R}_0$ and m. A *likelihood profile* is computed by holding one of the parameters in θ fixed while fitting the other parameter. This process is repeated for a range of values of the fixed parameter near the MLE. The likelihood profile for a given parameter shows how quickly the goodness of fit diminishes as one moves away from the MLE, thus producing confidence limits. We obtain these confidence limits on our estimate of θ using the likelihood ratio test (LRT) [57, Ch. 6, pp. 254–258].

The LRT assumes that the *deviance* along the likelihood profile of m (*i.e.* fixing $\mathcal{R}_0$ at the best estimate),

$$-2[-\log\mathcal{L}(m_{\text{est}}\,|\,y) - (-\log\mathcal{L}(m\,|\,y))]\,, \tag{4.23}$$

is chi-squared distributed with one degree of freedom. Thus, for the 95% confidence interval, we find the m along the likelihood profile above and below the MLE such that

$$-\log\mathcal{L}(m_{\text{est}}\,|\,y) + \log\mathcal{L}(m\,|\,y) < \chi^2_1(0.95)/2 = 1.92\,, \tag{4.24}$$

and similarly for for $\mathcal{R}_0$. We show likelihood profiles for MLEs of both $\mathcal{R}_0$ and m, for each of the four coupling schemes, in Figure 4.4.

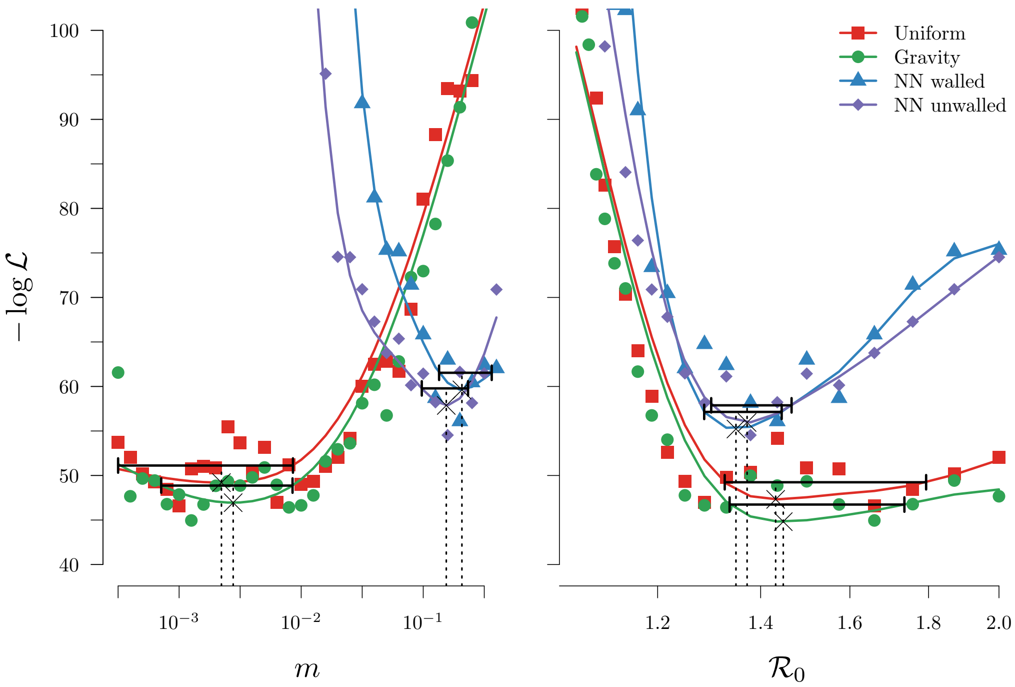

Figure 4.4: Likelihood profiles showing negative log-likelihood versus each parameter, m (left) and $\mathcal{R}_0$ (right). Profiles are obtained from grids such as that shown in Figure 4.3, with minima showing best estimates of m and $\mathcal{R}_0$ for the GPL from observed data in the LBoM. The profile corresponds to the greatest likelihood in the grid for each value of the focal parameter (dots), smoothed with a cubic spline (line). Profiles shown correspond to each of four coupling schemes: uniform, gravity, and near-neighbour with and without contact across the London city wall (see §4.3.1). 95% confidence intervals are shown for each profile based on Equation (4.24).

Maximum likelihood estimates and 95% confidence intervals for m and $\mathcal{R}_0$, as shown in Figure 4.4, are listed in Table 4.1. We also assess the fits with the Akaike information criterion (AIC) [118],

$$\mathrm{AIC} = 2k - 2\ln \hat{L}\,, \tag{4.25}$$

where $\hat{L}$ is the likelihood of the best fit parameters, and k is the number of parameters fit, which in all four cases is 2. We find gravity coupling to produce the best fit with

$\mathrm{AIC}_{\mathrm{fit}} = 97.9$, and refer to this value as $\widehat{\mathrm{AIC}}$. In Table 4.1, we compare other fits to gravity with the difference

$$\Delta\mathrm{AIC} = \mathrm{AIC} - \widehat{\mathrm{AIC}}\,. \tag{4.26}$$

.

Coupling Scheme	m	$\mathcal{R}_0$	ΔAIC
Uniform	$0.00222\,(0.000316, 0.00856)$	$1.43\,(1.33, 1.79)$	4.5
Gravity	$0.00277\,(0.000713, 0.00850)$	$1.45\,(1.35, 1.74)$	0
Near-Neighbour (walled)	$0.207\,(0.135, 0.364)$	$1.35\,(1.29, 1.44)$	25.3
Near-Neighbour (unwalled)	$0.154\,(0.0973, 0.233)$	$1.37\,(1.30, 1.47)$	21.8

Table 4.1: Maximum likelihood estimates of coupling m and basic reproduction $\mathcal{R}_0$, with 95% confidence limits. The best performing coupling scheme (in bold) is determined by applying the Akaike information criterion (AIC [118]), and we show ΔAIC for other models (see Equation (4.26)).

We additionally test the effectiveness of our estimation method by observing how well we are able to estimate θ from simulated data, for which we know the true values. We simulate $n_{\mathrm{test}} = 100$ stochastic realizations, $\{x_i\}_i^{n_{\mathrm{test}}}$, using $m = 0.00277$ and $\mathcal{R}_0 = 1.45$, our estimates from our best model fit (gravity), shown in Table 4.1. We then apply the same methodology to estimate m and $\mathcal{R}_0$ for these simulated data sets. Distributions of estimates m_{MLE}, $\mathcal{R}0_{\mathrm{MLE}}$, and ΔAIC, are shown in Figure 4.5.

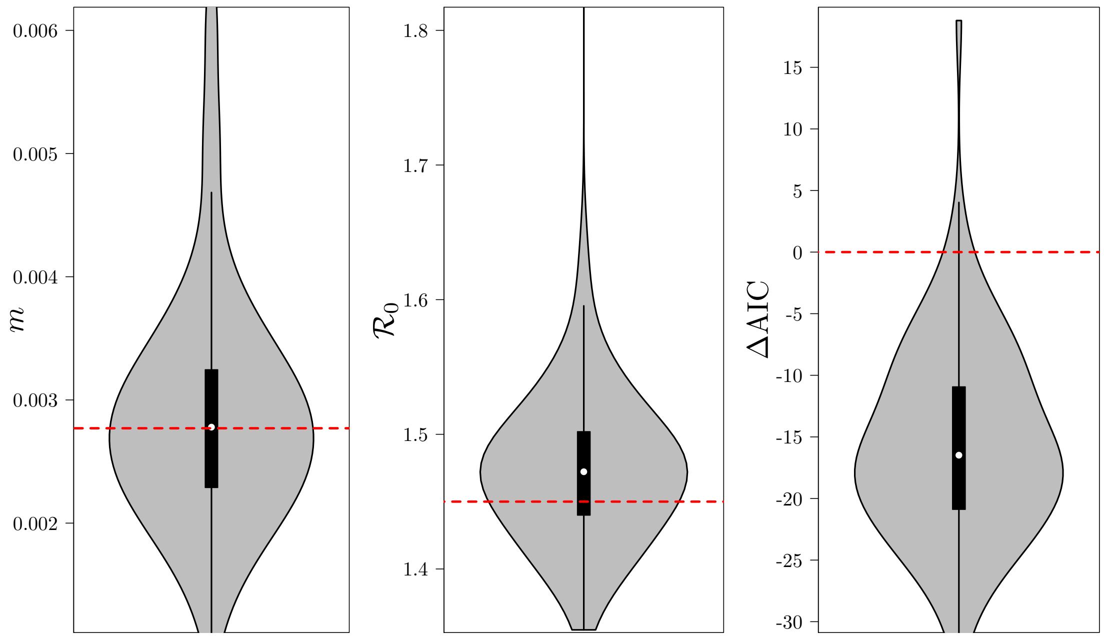

Figure 4.5: Distributions of estimates from $n_{\text{test}} = 100$ test simulations. Left and center panels show distributions of MLE of m and $\mathcal{R}_0$, respectively. The red dotted lines show values from our best fit to the LBoM data (gravity, see Table 4.1), which were used to generate test simulations. The right panel shows the distribution of ΔAIC for the test fits, with the red dotted line showing the negative log-likelihood of the best fit to LBoM data (see Figure 4.4). The violin plots shown are a combination of a vertical density plot with a boxplot, where the box shows 25%–75% quartiles.

4.5 Discussion

We asses the goodness of fit of each of the four coupling schemes with AIC, where the model with the least AIC is the best (see Table 4.1). Our results show clearly that both uniform and gravity coupling fit the LBoM data much better than near-neighbour coupling using our methodology. Furthermore, $\Delta\text{AIC} = 4.5$ for uniform coupling, which can be taken as weak evidence that gravity is more plausible [119]. In Figure 4.5 we test our methodology on simulations, and obtain high variation in

ΔAIC when the underlying parameters are known, suggesting that this evidence is at best weak.

We designed our test case in part to obtain information about the importance of geographic location in the spread of the plague throughout London. While our implementation of gravity coupling reduces contact between distant parishes compared with uniform coupling, infection is still able to spread directly from one end of the city to the other. On the other hand, our implementation of near-neighbour coupling precludes spread beyond a fixed number of parish connections, and was comparatively much worse in replicating the epidemic onset distribution in the GPL. Our results suggest that infections of parishes only by other nearby parishes is insufficient to explain the pattern of infection during the GPL. However, we cannot infer from our results to a precise degree what factor distance played, and further research is required to answer this question. Such research can include the fitting of p (see Equations (4.17) and (4.18)), since identifiability of these parameters would be evidence of some effect of distance on the epidemic onset distribution. We also note that our grid-search for the MLE as presented in §4.4 can be fine-tuned by means of stochastic optimization algorithms [58], and would be necessary for the estimation of more than two parameters simultaneously due to an increased computational cost.

A challenging aspect of parameter estimation is determining the characteristics of the data relevant to the parameters being estimated. The use of the epidemic onset distribution to fit parameters has the advantage of obscuring the precise order in which the epidemic spread throughout London's parishes, allowing for simulations to be "close" to the GPL while spreading to the city by substantially different routes. While facilitating estimation, a disadvantage of this method is that information regarding coupling in the particular sequence of onsets throughout the city could be lost in the summary statistic. An alternative probe could be matching the onset times of each parish as closely as possible. A different approach could calculate the probability of observing each onset, given the subset of parishes known to be infected up to the week of onset. Such an approach would make better use of information in the LBoM, but would be more sensitive to reporting efficiency, since accurate estimates of parish prevalence in each week would be required.

Estimates of the basic reproduction number for pneumonic plague exist [109], but we chose to fit $\mathcal{R}_0$ along with coupling m due to the inherent difficulty in comparing the population of 17th century London to other populations studied in the 20th century. However, for our best estimates of $\mathcal{R}_0$ using either gravity or uniform coupling, we find comparable estimates of the basic reproduction number.

If we take our best estimate of the coupling parameter m at face value, then we infer that typical residents of London in 1665 spent 0.28% of their time visiting other parishes. Future research could compare this estimate of population movement with other historical information, if other relevant data can be found.

Numerous avenues of further research beyond those mentioned can be pursued. We have altogether avoided the question of vector transmission, and it is not known which mode of plague transmission dominated the GPL. Our approach is consistent with a purely pneumonic epidemic, but the modeling of a rat population and estimating the parameterization of this additional mode of transmission may prove informative. We furthermore note that assuming uniformity in behaviour among parish populations significantly over-simplifies the historical reality, and while paucity of available information may preclude parish-specific parameter estimates, differences between rural and central city parishes could be made explicit in the model and fitted. Finally, we have assumed that a single initial case of plague sparked the epidemic, but the presence of plague elsewhere in England [66, Ch. 12] at the time of the GPL suggests the possibility of multiple exogenous infections throughout the epidemic. This could be investigated, and the inclusion of multiple exogenous infections in the model could significantly impact estimates of the coupling rate and the best fit spatial scheme.

4.6 Conclusion

Our aim in this paper was to present a case study in the application of probe-matching to estimate coupling strength in a meta-population using reported mortality during an epidemic. We explored the degree to which these methods could determine the relevance of geographic location in the spread of the epidemic. We were able to successfully obtain fits of coupling m and the basic reproduction number $\mathcal{R}_0$, with the best fits corresponding to spatial coupling schemes that did not restrict the range of infection to nearby parishes. The use of a summary statistic of the epidemic onset distribution as a probe was able to facilitate estimation, while obscuring information about the precise path of invasion of the epidemic. Our estimates of $\mathcal{R}_0$ agree with estimates for modern data, while our estimate of m provides an insight into the level of intra-city movement in the 17th century London population.

Research in advancing our modeling tools for epidemics are invaluable in efforts to forecast and to understand the spread of diseases in human populations. The use of historical data sets such as the LBoM provide unique opportunities to develop and test such tools, while providing insights into the dynamics of disease spread during moments of historical interest. Our results show that spatiotemporal mortality reports during an epidemic are sufficient to obtain quantifiable information about population movement and the importance of geographic location to the spread of disease. Spatiotemporal disease reports, whether describing death or infection, are therefore a valuable and useful source of information for the understanding both of the dynamics of the disease, and of the behaviour of the population being infected.

4.7 Acknowledgments

KH was supported by an Ontario Graduate Scholarship (OGS). DE was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Chapter 5

General Conclusions

The combination of cheap and widely available computing power with researchers' increased access to digitized epidemiological data presents tremendous opportunities for advancing the science of epidemics. Our ability to explain phenomena observed in documented real-world epidemic events and to develop predictive models promises substantial public health utility, especially in forecasting and assessments of potential interventions. The contributions to this area of research presented in this thesis focus on our ability to estimate spatial coupling parameters from real-world data. We used maximum likelihood estimation with probe-matching, tested methods on simulated mock data in Chapters 2, 3, and 4 and a real-world data set in Chapter 4, as well as presented an analytic approximation for estimation in Chapter 2.

Chapter 2 focused on coupling between two populations undergoing an epidemic invasion, and presented both analytic and numerical methodology for estimating the degree of coupling from the *time to invasion* of the second population. Single invasion events produce estimates of coupling degree with broad confidence limits, but the observation of multiple independent invasions yields much more accurate estimates. Multiple invasion events can be observed, in principle, not only between the same two populations at different times, but at the same time for two different diseases with the same mode of transmission. Comparisons between analytic and numerical estimation methods show that numerical methods are more accurate, but analytic methods can produce initial estimates of coupling that are close to the correct values.

Future research can explore improvements in the quality of the analytic approximation, as well as extending it to encompass more general scenarios, such as an arbitrary number of spatial patches. These methods could also additionally be use to estimate coupling in real-world systems. In particular, it would be interesting to compare estimates of coupling produced from data describing two different diseases with similar modes of transmission.

Chapter 3 explored the possibility of estimating coupling from complex recurrent epidemics, which have been observed and studied extensively in real-world situations (see Chapter 1). We modeled two coupled populations, each undergoing recurrent epidemics, with only the second population small enough to experience disease fade-outs. We showed that estimates of the degree of coupling between the populations can be obtained from the proportion of time the smaller population spends faded out. In the idealized case where all non-coupling parameters are known exactly, the effectiveness of this method depends on potential of the smaller population to respond to re-infection by the larger population. When the small population is too small or too large, degree of coupling above a certain threshold ceases to affect the proportion of time the disease is faded out.

This research can be extended with examinations of the idealizing assumptions we made, such as sensitivity analyses of disease and population parameters, or additionally fitting unknown parameters parameters along with coupling. Applying these methods to real-world data is a natural extension of this research, since such data is becoming ever more widely available [13, 19, 68, 92, 93], but fitting efforts must be tailored for individual data sets. For example, reporting efficiency and immunization levels are important in modern data sets, and estimates of these and other factors are required for effective estimation of coupling. As with Chapter 2, expanding these methods to be applicable to an arbitrary number of populations is another avenue of future research. However, given the well-studied phenomenon in which a large population centre drives epidemics in smaller populations [24, 59–62], analyses using only the large population and one small population could be reasonable, even in a system with many coupled populations.

Chapter 4 presented a third probe-matching approach to estimating spatial coupling, this time applicable to an arbitrary number of geographically separated patches, and applied to the Great Plague of London, England, of 1665. We fitted four implementations of spatial coupling to weekly parish-level mortality data collected in the London Bills of Mortality. We were able to fit the data much more successfully with coupling formulations that did not constrain spread only to nearby parishes, but more research is required to determine the nature of geographic spread more precisely. Since we characterized coupling in our model as the proportion of time individuals spend visiting other parishes, our results, taken at face value, give this proportion to be approximately 0.28%. We furthermore obtained an estimate of the basic reproduction number for plague ($\mathcal{R}_0 \approx 1.45\,(1.35, 1.74)$, see Chapter 4, Table 4.1) that is comparable with modern estimates (see Gani and Leach, who found that $\mathcal{R}_0 \approx 1.3\,(0.96, 2.3)$ [109]).

Future research on the same data set could include vector transmission in the model, which can be significant in the spread of plague in humans [111]. The methods we present can also be extended by fitting additional spatial parameters[1], along with

[1] For example, our implementation of gravity coupling scales with the inverse square of the distance between parishes (see Chapter 4, Equation (4.17)), but this exponent could be made variable and estimated along with m.

performing sensitivity analyses on fixed parameters. Another interesting avenue of future research could compare results from our estimate of the volume of travel with independent information about such travel, where such data can be found. We are not aware of such data being available for London, England in 1665, but travel data have been used for spatial analyses of disease spread in modern contexts [46, 47].

The use of stochastic and analytic model fitting tools promises to substantially advance our understanding and capacity to forecast epidemics in human populations. This thesis presented numerical and analytic approaches to probe-matching, which we applied to both mock data and one real-world data set, and is part of a larger effort to expand the set of modeling tools available in mathematical epidemiology. It is our hope that this research contributes to further advances in a field promising both increased scientific understanding and utility to the public at large.

Bibliography

[1] Bos KI, Schuenemann VJ, Golding GB, Burbano HA, Waglechner N, Coombes BK, et al. A draft genome of Yersinia pestis from victims of the Black Death. Nature. 2011;478(7370):506–510.

[2] Wagner DM, Klunk J, Harbeck M, Devault A, Waglechner N, Sahl JW, et al. *Yersinia pestis* and the Plague of Justinian 541-543 AD: a genomic analysis. *Lancet Infectious Diseases*. 2014;14:319–326.

[3] Diamond J. Guns, Germs, and Steel. W. W. Norton & Company; 1999.

[4] Johnson NP, Mueller J. Updating the accounts: global mortality of the 1918-1920" Spanish" influenza pandemic. Bulletin of the History of Medicine. 2002;76(1):105–115.

[5] Bernoulli D. Essai d'une nouvelle analyse de la mortalité causée par la petite vérole et des advantages de l'inoculation pour la prévenir. Mém Mathematical Physics Academy Royal Science Paris. 1760;:1–45.

[6] Blower S. An attempt at a new analysis of the mortality caused by smallpox and of the advantages of inoculation to prevent it. Reviews In Medical Virology. 2004;14(5):275–288.

[7] Hamer WH. The Milroy lectures on epidemic disease in England: the evidence of variability and of persistency of type. Bedford Press; 1906.

[8] Kermack WO, McKendrick AG. A contribution to the mathematical theory of epidemics. Proceedings of the Royal Society of London Series A. 1927;115:700–721.

[9] Anderson RM, May RM. Infectious Diseases of Humans: Dynamics and Control. Oxford: Oxford University Press; 1991.

[10] Diekmann O, Heesterbeek JAP. Mathematical epidemiology of infectious diseases: model building, analysis and interpretation. Wiley Series in Mathematical and Computational Biology. New York: John Wiley & Sons, LTD; 2000.

[11] Allen LJS. An introduction to stochastic epidemic models. In: Lecture notes in mathematics. vol. 1945. Springer Berlin / Heidelberg; 2008. p. 81–130.

[12] Bartlett MS. Stochastic population models in ecology and epidemiology. vol. 4 of Methuen's Monographs on Applied Probability and Statistics. London: Spottiswoode, Ballantyne & Co. Ltd.; 1960.

[13] London W, Yorke JA. Recurrent outbreaks of measles, chickenpox and mumps. I. Seasonal variation in contact rates. American Journal of Epidemiology. 1973;98(6):453–468.

[14] Olsen LF, Schaffer WM. Chaos versus noisy periodicity: alternative hypotheses for childhood epidemics. Science. 1990;249:499–504.

[15] Hethcote HW. The mathematics of infectious diseases. SIAM Review. 2000;42(4):599–653.

[16] Earn DJD, Rohani P, Bolker BM, Grenfell BT. A simple model for complex dynamical transitions in epidemics. Science. 2000;287(5453):667–670.

[17] Bauch CT, Earn DJD. Transients and attractors in epidemics. *Proceedings of the Royal Society of London, Series* B. 2003;270(1524):1573–1578.

[18] Earn DJD. Mathematical epidemiology of infectious diseases. In: Lewis MA, Chaplain MAJ, Keener JP, Maini PK, editors. Mathematical Biology. vol. 14 of IAS/ Park City Mathematics Series. American Mathematical Society; 2009. p. 151–186.

[19] Krylova O. Predicting epidemiological transitions in infectious disease dynamics: Smallpox in historic London (1664-1930) [PhD]. McMaster University, Canada; 2011. Can be found online.

[20] Cliff AD, Haggett P. Atlas of Disease Distributions: Analytic Approaches to Epidemiologic Data. Oxford: Basil Blackwell; 1988.

[21] Cliff AD, Haggett P, Smallman-Raynor M. Measles: An Historical Geography of a Major Human Viral Disease, From Global Expansion to Local Retreat, 1840-1990. Oxford: Blackwell Publishers; 1993.

[22] Ferguson NM, May RM, Anderson RM. Measles: Persistence and synchronicity in disease dynamics. In: Tilman D, Kareiva P, editors. Spatial Ecology. vol. 30 of Monographs in Population Biology. Princeton: Princeton University Press; 1997. p. 137–157.

[23] Bolker B, Grenfell B. Space, persistence and dynamics of measles epidemics. Philosophical Transaction of the Royal Society of London Series B Biological Sciences. 1995;348:309–320.

[24] Grenfell BT, Bjornstad ON, Kappey J. Travelling waves and spatial hierarchies in measles epidemics. Nature. 2001;414(6865):716–723.

[25] Haggett P. Building geographic components into epidemiological models. Influenza Models. 1982;p. 203–212.

[26] Anderson RM, MAY RM. Spatial, temporal, and genetic heterogeneity in host populations and the design of immunization programmes. Mathematical Medicine and Biology: A Journal of the IMA. 1984;1(3):233–266.

[27] Murray G, Cliff AD. A stochastic model for measles epidemics in a multi-region setting. Transactions of the Institute of British Geographers. 1977;p. 158–174.

[28] Sattenspiel L. Population structure and the spread of disease. Human Biology. 1987;p. 411–438.

[29] Sattenspiel L. Epidemics in nonrandomly mixing populations: a simulation. American Journal of Physical Anthropology. 1987;73(2):251–265.

[30] Sattenspiel L, Simon CP. The spread and persistence of infectious diseases in structured populations. Mathematical Biosciences. 1988;90(1-2):341–366.

[31] Watts DJ, Muhamad R, Medina DC, Dodds PS. Multiscale, resurgent epidemics in a hierarchical metapopulation model. Proceedings of the National Academy of Sciences of the United States. 2005;102(32):11157–11162.

[32] Baroyan O, Genchikov L, Rvachev L, Shashkov V. An attempt at large-scale influenza epidemic modelling by means of a computer. Bull Int Epidemiol Assoc. 1969;18(22-31):107.

[33] Sattenspiel L, Herring DA. Structured epidemic models and the spread of influenza in the central Canadian Subarctic. Human Biology. 1998;70(1):91–115.

[34] Sattenspiel L, Dietz K. A Structured Epidemic Model Incorporating Geographic-Mobility among Regions. Mathematical Biosciences. 1995;128(1-2):71–91.

[35] Lloyd AL, May RM. Spatial heterogeneity in epidemic models. Journal of Theoretical Biology. 1996;179:1–11.

[36] Dietz K, Schenzle D. Proportionate mixing models for age-dependent infection transmission. Journal of mathematical biology. 1985;22(1):117–120.

[37] Hethcote HW. An Age-Structured Model for Pertussis Transmission. Mathematical Biosciences. 1997;145:89–136.

[38] Hethcote HW, Van Ark JW. Epidemiological models for heterogeneous populations: proportionate mixing, parameter estimation, and immunization programs. Mathematical Biosciences. 1987;84(1):85–118.

[39] Hoppensteadt F. An age dependent epidemic model. Journal of the Franklin Institute. 1974;297(5):325–333.

[40] Earn DJD, Levin SA. Global asymptotic coherence in discrete dynamical systems. *PNAS – Proceedings of the National Academy of Sciences of the U.S.A.*. 2006;103(11):3968–3971.

[41] Newman MEJ. Spread of epidemic disease on networks. Physical Review E. 2002;66(1):1–11.

[42] Newman MEJ. Networks: An Introduction. New York: Oxford University Press; 2010.

[43] Watts DJ, Strogatz S. Collective dynamics of 'small-world' networks. Nature. 1998;393(6684):440–442.

[44] Arino J, van den Driessche P. A multi-city epidemic model. Mathmatical Population Studies. 2003;10:175–193.

[45] Dangerfield C, Ross J, Keeling M. Integrating stochasticity and network structure into an epidemic model. Journal of the Royal Society Interface. 2009;6(38):761–774.

[46] Viboud C, Tam T, Fleming D, Miller MA, Simonsen L. 1951 Influenza Epidemic, England and Wales, Canada and the United States. Emerging Infectious Diseases. 2006;12(4):661–668.

[47] Viboud C, Bjornstad ON, Smith DM, Simonsen L, Miller MA, Grenfell BT. Synchrony, waves, and spatial hierarchies in the spread of influenza. Science. 2006;312(5772):447–451.

[48] Sattenspiel L. The geographic spread of infectious diseases: Models and applications. Princeton Series in Theoretical and Computational Biology. Princeton, New Jersey and Oxford, UK: Princeton University Press; 2009.

[49] Rushton S, Mautner A. The deterministic model of a simple epidemic for more than one community. Biometrika. 1955;42(1/2):126–132.

[50] Stanley E, Brown D. On the spatial spread of rabies among foxes. In: Proc. R. Soc. Lond. B. vol. 229. The Royal Society; 1986. p. 111–150.

[51] Murray JD. Modeling the spread of rabies. American Scientist. 1987;75(3):280–284.

[52] Murray J. Mathematical Biology. 1989. C271. 1989;.

[53] Mollison D. Dependence of epidemic and population velocities on basic parameters. Mathematical biosciences. 1991;107(2):255–287.

[54] Metz JA, Mollison D, Van Den Bosch F. The dynamics of invasion waves. IR-99-039; 1999.

[55] Finkenstädt B, Grenfell B. Time series modelling of childhood diseases: A dynamical systems approach. Journal of the Royal Statistical Society Series C (Applied Statistics). 2000;49(2):187–205.

[56] Bartlett M. Deterministic and stochastic models for recurrent epidemics. In: Proceedings of the third Berkeley symposium on mathematical statistics and probability. vol. 4; 1956. p. 109.

[57] Bolker BM. Ecological models and data in R. Princeton University Press; 2008.

[58] Hartig F, Calabrese JM, Reineking B, Wiegand T, Huth A. Statistical inference for stochastic simulation models–theory and application. Ecology Letters. 2011 Aug;14(8):816–827.

[59] Bolker BM, Grenfell BT. Impact of vaccination on the spatial correlation and persistence of measles dynamics. Proceedings of the National Academy of Sciences, USA. 1996;93:12648–12653.

[60] Cliff AD, Haggett P, Stroup DF. The geographic structure of measles epidemics in the northeastern United States. American Journal of Epidemiology. 1992;136(5):592–602.

[61] Cliff AD, Haggett P, Stroup DF, Cheney E. The changing geographical coherence of measles morbidity in the United States, 1962–88. Statistics in Medicine. 1992;11:1409–1424.

[62] Keeling MJ, Rohani P. Estimating spatial coupling in epidemiological systems: a mechanistic approach. Ecology Letters. 2002;5(1):20–29.

[63] Rohani P, Earn DJD, Grenfell BT. Opposite patterns of synchrony in sympatric disease metapopulations. Science. 1999;286(5441):968–971.

[64] Tien JH, Poinar HN, Fisman DN, Earn DJD. Herald waves of cholera in nineteenth century London. Journal of the Royal Society Interface. 2011;8(58):756–760.

[65] Creighton C. A history of epidemics in Britain. vol. 1. 2nd ed. London and Edinburgh: Frank Cass & Co. Ltd.; 1965.

[66] Creighton C. A history of epidemics in Britain. vol. 2. 2nd ed. London and Edinburgh: Frank Cass & Co. Ltd.; 1965.

[67] Ziegler P. The black death. Faber & Faber; 2013.

[68] Yorke JA, London W. Recurrent outbreaks of measles, chickenpox and mumps. II. Systematic differences in contact rates and stochastic effects. American Journal of Epidemiology. 1973;98(6):468–482.

[69] He D, Earn DJD. Epidemiological effects of seasonal oscillations in birth rates. *Theoretical Population Biology*. 2007;72:274–291.

[70] Anderson D, Watson R. On the spread of a disease with gamma distributed latent and infectious periods. Biometrika. 1980;67(1):191–198.

[71] Feng ZL, Thieme HR. Endemic models with arbitrarily distributed periods of infection I: Fundamental properties of the model. SIAM Journal on Applied Mathematics. 2000;61(3):803–833.

[72] Lloyd AL. Destabilization of epidemic models with the inclusion of realistic distributions of infectious periods. Proceedings of the Royal Society of London Series B-Biological Sciences. 2001;268(1470):985–993.

[73] Wearing HJ, Rohani P, Keeling MJ. Appropriate models for the management of infectious diseases. PLOS medicine. 2005;2(7):621–627.

[74] Nishiura H, Eichner M. Infectiousness of smallpox relative to disease age: estimates based on transmission network and incubation period. Epidemiology and Infection. 2007 10;135(7):1145–1150.

[75] Conlan AJK, Rohani P, Lloyd AL, Keeling M, Grenfell BT. Resolving the impact of waiting time distributions on the persistence of measles. Journal of the Royal Society Interface. 2010;7:623–640.

[76] May RM, Anderson RM. Spatial heterogeneity and the design of immunization programs. Mathematical Biosciences. 1984;72:83–111.

[77] Earn DJD, Rohani P, Grenfell BT. Persistence, chaos and synchrony in ecology and epidemiology. *Proceedings of the Royal Society of London, Series* B. 1998;265(1390):7–10.

[78] Earn DJD, Levin SA, Rohani P. Coherence and conservation. Science. 2000;290(5495):1360–1364.

[79] Lloyd AL, Jansen VAA. Spatiotemporal dynamics of epidemics: synchrony in metapopulation models. Mathematical Biosciences. 2004;188:1–16.

[80] Diekmann O, Heesterbeek JAP, Metz JAJ. On the definition and the computation of the basic reproduction ratio RO in models for infectious-diseases in heterogeneous populations. Journal of Mathematical Biology. 1990;28(4):18.

[81] van den Driessche P, Watmough J. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. Mathematical Biosciences. 2002;180(Sp. Iss.):29–48.

[82] Johnson P. adaptivetau: Tau-leaping stochastic simulation. URL http://CRAN R-project org/package= adaptivetau R package version. 2013;1.

[83] Gillespie DT. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. Journal of Computational Physics. 1976;22:403–434.

[84] Gillespie DT. Exact Stochastic Simulation Of Coupled Chemical-Reactions. Journal Of Physical Chemistry. 1977;81(25):2340–2361.

[85] Allen LJS. An introdution to stochastic processes with applications to biology. 2nd ed. New Jersey: Pearson education Inc.; 2010.

[86] Cox DR, Oakes D. Analysis of Survival Data. Chapman & Hall/CRC; 1984.

[87] Kullback S. Information theory and statistics. Courier Corporation; 1997.

[88] Ionides EL, Breto C, King AA. Inference for nonlinear dynamical systems. *PNAS – Proceedings of the National Academy of Sciences of the U.S.A.*. 2006;103(49):18438–18443.

[89] Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLOS medicine. 2008;5(3):381–391.

[90] Gelman A, Stern HS, Carlin JB, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis. Chapman and Hall/CRC; 2013.

[91] Stone L, Olinky R, Huppert A. Seasonal dynamics of recurrent epidemics. Nature. 2007;446(7135):533–536.

[92] Van Panhuis WG, Grefenstette J, Jung SY, Chok NS, Cross A, Eng H, et al. Contagious diseases in the United States from 1888 to the present. The New England Journal of Medicine. 2013;369(22):2152.

[93] Hempel K, Earn DJ. A century of transitions in New York City's measles dynamics. Journal of The Royal Society Interface. 2015;12(106):20150024.

[94] Farrington CP, Kanaan MN, Gay NJ. 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. 2001;50:251–283.

[95] Schwartz IB. Small amplitude, long period outbreaks in seasonally driven epidemics. Journal of Mathematical Biology. 1992;30:473–491.

[96] Ballard P, Bean N, Ross J. The probability of epidemic fade-out is non-monotonic in transmission rate for the Markovian SIR model with demography. Journal of theoretical biology. 2016;393:170–178.

[97] Patz JA, Graczyk TK, Geller N, Vittor AY. Effects of environmental change on emerging parasitic diseases. International journal for parasitology. 2000;30(12):1395–1405.

[98] Bartlett MS. Measles periodicity and community size. Journal of the Royal Statistical Society Series A. 1957;120:48–70.

[99] Schenzle D. An age-structured model of pre- and post-vaccination measles transmission. IMA Journal of Mathematics Applied in Medicine and Biology. 1984;1:169–191.

[100] Gillespie DT. Stochastic Simulation of Chemical Kinetics. Annual Review of Physical Chemistry. 2007;58:35–55.

[101] Lewis PW, Shedler GS. Simulation of nonhomogeneous Poisson processes by thinning. Naval research logistics quarterly. 1979;26(3):403–413.

[102] Cao Y, Gillespie DT, Petzold LR. Adaptive explicit-implicit tau-leaping method with automatic tau selection. The Journal of chemical physics. 2007;126(22):224101.

[103] Rand DA, Wilson HB. Chaotic stochasticity: a ubiquitous source of unpredictability in epidemics. Proc R Soc Lond B. 1991;246:179–184.

[104] Dushoff J, Plotkin JB, Levin SA, Earn DJD. Dynamical resonance can account for seasonality of influenza epidemics. *PNAS – Proceedings of the National Academy of Sciences of the U.S.A.*. 2004;101(48):16915–16916.

[105] Fine PEM, Clarkson JA. Measles in England and Wales — I: An Analysis of Factors Underlying Seasonal Patterns. International Journal of Epidemiology. 1982;11(1):5–14.

[106] Johnson H, Hillary IB, McQuoid G, Gilmer BA. MMR vaccination, measles epidemiology and sero-surveillance in the Republic of Ireland. Vaccine. 1995;13(6):533–537.

[107] Schneeberger A, Jansen VA. The estimation of dispersal rates using the covariance of local populations. ecological modelling. 2006;196(3-4):434–446.

[108] Cummins N, Kelly M, Ó Gráda C. Living standards and plague in London, 1560–1665. The Economic History Review. 2016;69(1):3–34.

[109] Gani R, Leach S. Epidemiologic determinants for modeling pneumonic plague outbreaks. Emerging Infectious Diseases. 2004;10(4):608–614.

[110] Finlay R. Population and metropolis: the demography of London, 1580-1650. vol. 12 of Cambridge Geographical Studies. Cambridge: Cambridge University Press; 1981.

[111] Perry RD, Fetherston JD. Yersinia pestis–etiologic agent of plague. Clinical microbiology reviews. 1997;10(1):35–66.

[112] Gage KL, Kosoy MY. Natural history of plague: perspectives from more than a century of research. Annu Rev Entomol. 2005;50:505–528.

[113] Keeling MJ, Gilligan CA. Metapopulation dynamics of bubonic plague. Nature. 2000;407:903–906.

[114] Smith DL, Lucey B, Waller LA, Childs JE, Real LA. Predicting the spatial dynamics of rabies epidemics on heterogeneous landscapes. Proceedings of the National Academy of Sciences of the United States of America. 2002;99(6):3668–72.

[115] Begon M, Bennett M, Bowers RG, French NP, Hazel S, Turner J. A clarification of transmission terms in host-microparasite models: numbers, densities and areas. Epidemiology & Infection. 2002;129(1):147–153.

[116] Erlander S, Stewart NF. The gravity model in transportation analysis: theory and extensions. vol. 3. Vsp; 1990.

[117] Von Mises R. Mathematical theory of probability and statistics. Academic Press; 2014.

[118] Akaike H. A new look at the statistical model identification. IEEE Transactions On Automatic Control. 1974;19(6):716–723.

[119] Burnham KP, Anderson DR. Model selection and multimodel inference: A practical information-theoretic approach. 2nd ed. New York: Springer; 2002.