Spatial Demography and the Epidemiology of Measles

Hannah Michelle Korevaar

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Program in
Population Studies
Advisers: C. Jessica E. Metcalf and Bryan T. Grenfell

June 2020
Abstract

Infectious diseases remain a significant cause of death worldwide, in spite of advances in medical technology and treatment in the last century. The life cycles of infectious diseases depend on host populations producing hospitable conditions for survival and proliferation. Secular demographic events such as births, local contact rates and regional migration are necessary for the persistence of acute immunizing infections. Each chapter uses measles incidence data from pre-vaccination England and Wales to investigate the relationship between disease dynamics and demographic events.

In the first chapter, I use the time-series susceptible-infected-recovered (TSIR) model to investigate transmission rates and measles persistence in urban and rural areas. I use a matched-pair analysis to separate the influence of space and population size to isolate the difference in disease dynamics between urban and rural areas. I find that population size, more than population density, influences the size and persistence of outbreaks; however, there is some evidence that population density may impact the transmission of measles.

In chapter two, I challenge previous findings that transmission of measles is uncorrelated with population size. I also address the potential of population-correlated bias in TSIR estimates. In this chapter I leverage high volume stochastic simulations in order to determine if biases in estimates of transmission scale with population size. The results indicate that, in general, large populations have higher transmission rates and bias is greatest for small communities where outbreaks are highly stochastic.

In the final chapter, I expand on previous frequency domain results by introducing tensor decomposition as a method for analyzing multiple oscillatory time-series simultaneously. Using this method of dimensionality reduction I am able to extract the dominant periodic signals in measles incidence data. I use this descriptive technique to verify a previous finding: the baby boom surge in births in England and Wales resulted in larger annual peaks in incidence.
Together these results highlight the close relationship between demography and infectious disease dynamics. Furthermore, they demonstrate the importance of increasing and maintaining vaccination coverage for measles, particularly in an increasingly urban world where local populations and population density continue to rise.
Acknowledgements

A great many people have helped, supported, and advised me during my five years at Princeton, and I will attempt to mention as many as I can here.

First and foremost, I would like to thank my advisors, Jess Metcalf and Bryan Grenfell, for listening to all my ideas and gently but honestly telling me which were worthwhile and which were flops. I am particularly grateful that Bryan accepted my poorly formed pitch for investigating heterogeneous mixing rates and disease dynamics and sculpted it into an extensive exploration of urban and rural measles epidemiology. I learned more from reading the papers Bryan recommended and attempting the analyses Bryan cavalierly suggested than I did in many of the classes I took. I would also like to thank Heather Howard, Andrea Graham and Cecilia Rouse for serving on my committee, all tremendous academics and mentors with whom I am very grateful to have formed relationships. In particular, teaching with Dean Rouse remains one the most formative experiences of my time at Princeton.

I have been fortunate to have had a number of great student collaborators, all smarter and more generous than I am, from whom I have learned a tremendous amount. In particular Alex Becker, who taught me so much about disease modeling, simply by talking to me as if I already understood what he was saying. A subset of our collaborative efforts appear in the second chapter of this dissertation. Though none of our collaborative work is featured here, I would also like to thank Celeste Marin and Rebecca Johnson with whom I worked on a project very close to my heart. Celeste Marin served as my unofficial advisor until I found Bryan, and without whom I would not have survived my first two years, or scheduled a single general exam; Rebecca Johnson who is brilliant and humble and learned to program in Java just to make our survey design dreams come true.

I benefited from the immense support of the capable staff of the Office of Population Research (OPR) and Ecology and Evolutionary Biology (EEB). Lynne Johnson,
without whom OPR could not function and without whom I would have been lost. Lynne is the heart and soul of the department and the map that made navigating administrative obstacles possible. Ksenia Rodionova who always found time for Bryan to meet with me, always had the most appropriate Russian proverb ready for every situation, and made me feel at home in Eno. Wayne Appleton, Steven Mars, and Jesse Saunders made all my computing and virtual conferencing needs possible and answered my many many questions.

I must mention Betsy Armstrong who served on too many committees and still agreed to serve as Director of Graduate Studies (DGS); Betsy always listened to my complaints, always found time to advise me, or write letters for me, in spite of always having too much to do and literally running out of space in her inbox. Additionally, Alicia Adsera, who served as DGS during my first and last years at Princeton and who took an interest in my well-being as well as my progress toward my PhD; Alicia was a sounding board for many of my early ideas and her feedback was invaluable.

To the students of OPR and Ecology and Evolutionary Biology (EEB), you are a delightful and eclectic band of misfits, don’t ever change. A special thanks to Yo-yo Chen, Gracie Himmelstein and Becky Kruetter - my roommates who made living in Princeton as enjoyable as it could be and with whom I am grateful to have lifelong friendships. A big thank you to Amy Winter who taught me how to talk to Bryan, helped me build an illusion of a research agenda, and who accompanied me to many EEB events so I would feel at home. I am also grateful to officemate(s) Caroline Wagner (and Bertie), who made coming to work such a joy. To Jonathan Tannen who encouraged my love for models, recommended classes and faculty, gave me advice on the importance of open and practical science, and once did me the honor of calling me his intellectual doppelganger (I am so glad to have this opportunity to immortalize that compliment in print). A huge thank you to everyone in the Grenfell
and Metcalf labs and the Princeton Disease Group for being so intelligent, welcoming and supportive.

I presented and received feedback on an early version of work in chapter one at the annual conference for the Population Association of America (PAA). I received valuable suggestions which influenced the final form of the project.

My family has been a critical support system through my five years at Princeton. My father, Nick Korevaar, volunteered to drive me and my dog between New Jersey and Utah for several summers — and one particularly ill-advised winter holiday — because I could not handle staying in the Princeton bubble for twelve months of the year. My mother and step father, Karen and John Martinez, were always interested to hear what I was working on and watched my presentations; they offered unyielding praise for my accomplishments and asked many questions. To my partner, Sasha Frankel who read my drafts and supported all my ideas, tolerated my ranting and reminded me that I could quit at any time (though she knew I would not).

Finally, to the health workers across the globe that work to vaccinate and treat infectious diseases (particularly in these challenging times), thank you.
Abstract . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . iii
Acknowledgements . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . v
List of Tables . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . xi
List of Figures . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . xii

Introduction 1

1 Structure, Space, and Size: Competing Drivers of Variation in Urban and Rural Measles Transmission. 29
  1.1 Abstract . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 29
  1.2 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 30
    1.2.1 Time Series SIR Model . . . . . . . . . . . . . . . . . . . . . . 32
    1.2.2 Potential Urban and Rural Differences . . . . . . . . . . . . . 35
  1.3 Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 37
  1.4 Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 38
    1.4.1 The TSIR Model . . . . . . . . . . . . . . . . . . . . . . . . . . . 38
    1.4.2 Principal Components Analysis on Paired Data . . . . . . . . . 42
    1.4.3 Epidemic Exchange in Paired Locations . . . . . . . . . . . . . 44
  1.5 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 44
    1.5.1 Aggregate Urban and Rural Comparison . . . . . . . . . . . . . 44
    1.5.2 Urban and Rural District Pairs: Principal Components Analysis 45

viii
2 Estimation Bias and Transmission Across Population Scales: a machine learning approach to correcting time-series SIR parameters

2.1 Introduction

2.1.1 Infectious Disease Modeling

2.1.2 Transmission and Population Size

2.1.3 Estimation Bias

2.2 Materials and methods

2.2.1 Time-Series Susceptible-Infected-Recovered Model

2.2.2 Measuring the Time-Series

2.2.3 Machine Learning and Bias Prediction

2.2.4 England and Wales Data

2.3 Results

2.3.1 In Silico Data

2.3.2 Biases: In Silico Data

2.3.3 Random Forest: Feature Importance

2.3.4 Random Forest: Bias Correction

2.3.5 Application to E&W Data

2.4 Discussion

2.5 Conclusion

3 Tensor Decomposition for Disease Incidence Data

3.1 Abstract

3.2 Introduction

3.3 Materials and Methods

3.3.1 Simulations
### Index

3.3.2  The Data: England & Wales Urban Districts 1944-1966 . . . . 84
3.3.3  Continuous Wavelet Transform . . . . . . . . . . . . . . . . . 84
3.3.4  Tensor Decomposition . . . . . . . . . . . . . . . . . . . . . . 86
3.4  Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 89
   3.4.1  Simulations . . . . . . . . . . . . . . . . . . . . . . . . . . . 89
   3.4.2  England & Wales Urban Districts 1944-1966 . . . . . . . . . . 91
3.5  Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 96

A  Chapter One Supplemental Information 98
   A.1  TSIR . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 98
   A.2  Pair selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 99
   A.3  PCA . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 100
   A.4  Pair Regression . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 100

B  Chapter Two Supplemental Information 104
   B.1  Simulation Details . . . . . . . . . . . . . . . . . . . . . . . . . . . . 104
      B.1.1  Alpha . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 105
      B.1.2  Beta . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 105
   B.2  Estimating With Fixed Alpha . . . . . . . . . . . . . . . . . . . . . . 105
   B.3  Estimating Transmission Shape . . . . . . . . . . . . . . . . . . . . . 106
   B.4  Time-Series Measures . . . . . . . . . . . . . . . . . . . . . . . . . . 108

C  Chapter Three Supplemental Information 114
   C.1  Crude Birth Rates and the Third Component . . . . . . . . . . . . . 114
   C.2  Simulation Raw and Reconstructed Wavelet Spectra . . . . . . . . . . 114
   C.3  Sample Location Raw and Reconstructed Wavelet Spectra . . . . . . . . . . 115

Bibliography 118
List of Tables

2.1 Descriptive Statistics of Pre- and Post-Correction Estimates . . . . . 74

A.1 Paired districts: descriptive statistics . . . . . . . . . . . . . . . 99

A.2 Loadings for All Principal Components . . . . . . . . . . . . . . . 100
List of Figures

1. Measles burden and resurgence in the post-vaccine era . . . . . . . . 4
2. Long and short-term drivers of measles transmission . . . . . . . . . 6
3. Illustration of compartmental disease models . . . . . . . . . . . . . 17
5. Wavelet spectra for London 1945-1994 . . . . . . . . . . . . . . . . . 23

1.1 Epidemic coupling and outbreak dynamics in urban and rural districts in pre-vaccination England and Wales . . . . . . . . . . . . . . 46
1.2 Principal components analysis on urban and rural district pairs . . . 49
1.3 Space versus size in urban and rural pairs . . . . . . . . . . . . . . 50
1.4 Epidemic interactions between neighboring urban and rural districts . 53
1.5 Patch mixing model simulations, fadeouts, and final size of outbreaks 57

2.1 Illustration of a random forest model . . . . . . . . . . . . . . . . . 69
2.2 Feature importance and the distribution of bias in simulations . . . 73
2.3 Bias and corrections in simulations . . . . . . . . . . . . . . . . . . 74
2.4 TSIR estimates and adjusted estimates from England and Wales data 75

3.1 Wavelet power spectra for London 1944-1994 . . . . . . . . . . . . . 86
3.2 The tensor as an array of wavelet spectra . . . . . . . . . . . . . . . 87
3.3 Theoretical Canonical Polyadic Decomposition . . . . . . . . . . . . . 88
3.4 Sample time-series from simulations and data . . . . . . . . . . . . . 90
3.5 Three-rank tensor decomposition of the simulated locations . . . . . . 91
3.6 Four-rank tensor decomposition of the England and Wales data . . . 92
3.7 Component by component reconstruction of London . . . . . . . . . . 93
3.8 Component by component reconstruction of Norwich . . . . . . . . . 94
A.1 Estimates of the basic reproductive number by population . . . . . . 99
A.2 Comparison of epidemic features of urban and rural districts . . . . . 101
A.3 Magnitude and significance for urban/rural dummy at different popu-
lation thresholds . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 103
B.1 Impact of fixing \( \alpha \) in the estimation procedure . . . . . . . . . . 107
B.2 Seasonal shape recovery from simulations . . . . . . . . . . . . . . . 108
B.3 Coverage of time series measures across simulations and data, plotted
against population (part one) . . . . . . . . . . . . . . . . . . . . . . 111
B.4 Coverage of time series measures across simulations and data, plotted
against population (part two) . . . . . . . . . . . . . . . . . . . . . . 112
B.5 Coverage of time series measures across simulations and data, plotted
against population (part three) . . . . . . . . . . . . . . . . . . . . . . 113
C.1 Crude birth rates for locations above the Critical Community Size and
scores on the third component . . . . . . . . . . . . . . . . . . . . . . . 115
C.2 Averaged and reconstructed power spectra for simulations . . . . . . 116
C.3 Reconstructed and original wavelet spectra for sample locations . . . 117
Introduction

Infectious diseases remain a significant cause of death worldwide in spite of advances in medical technology and treatment in the last century. More than half of the global deaths of children under 5 are due to vaccine preventable or treatable diseases [3]. In 2016 an estimated 4.3 million people died from communicable infections. The spread of these diseases depends on demographic factors such as birth and contact rates. The impact of these demographic rates on infectious disease transmission has been well-studied [10, 78, 9, 101]. A less studied approach to this relationship is how disease dynamics can illuminate hard-to-measure aspects of human population behavior. This thesis addresses the two-way interplay between disease and demography in the context of acute immunizing infections where a simple disease natural history and rich data aid our studies.

Camera Obscura: Disease Dynamics and Demography

The interaction between pathogens and host populations has been modeled extensively in ecology and epidemiology [10, 78, 98]. Such interactions can produce long-term epidemic cycles which reveal as much about the pathogen as its host population. Pathogens depend on secular host actions - such as births and contact rates - to pro-
liferate and persist. Pandemics, such as the 2009 influenza and 2020 coronavirus (COVID-19) outbreaks depend on local and global contact rates. 

The life cycles of infectious diseases depend on the host population producing conditions hospitable to pathogens’ survival - particularly for acute immunizing infections (See Sect 2). Changes in the host population that influence the rate at which a pathogen will encounter new susceptible individuals can alter the dynamics of the pathogen. These ripples in infection behavior can thus reveal changes in the underlying host population. A spike in births can cause a temporary shift in periodic outbreaks as a result of a sudden change in the susceptible population; variation in contact rates or population density may impact how quickly a disease can spread. Changes in population density can thus reveal changes in the underlying host population. A spike in births can cause a temporary shift in periodic outbreaks as a result of a sudden change in the susceptible population; variation in contact rates or population density may impact how quickly a disease can spread. These shifts are especially valuable for research when the underlying change in population dynamics is not directly observed. More pointedly, the intuitive question might be: what is the impact of demographic shifts on disease dynamics? But we could rather ask: what can disease dynamics reveal about demography?

Research has shown that essentially all infections - airborne or sexually transmitted - spread slower in populations where the contact of interest (physical, sexual, proximity) has dense and sparse patches. For example, the difference in prevalence of STIs across racial groups was found to be a result of strong assortative mixing - more frequent sexual contact between individuals of the same racial group and rare sexual encounters across racial groups. In contrast, well-mixed populations generally produce the fastest and largest epidemics. Many studies on the relationship between contact rates on the spread of disease rely on simulations to determine the impact of contact patterns on the rate of transmission and epidemic size. This is because highly detailed information on contact rates is necessary to improve forecasts relative to assuming a well-mixed population. This level of detail in contact rates is very difficult to obtain. In light of this difficulty, it seems prudent to let diseases tell us about human populations rather than the other way around.
In the following sections I describe the importance of demography for understanding infectious disease dynamics across space and time. I give both intuitive explanations as well as examples of relevant findings with particular attention to what has already been uncovered with regard to measles in pre-vaccination England and Wales (E&W), the specific context I examine. While much of this will be repeated in the following chapters, this introduction provides the foundation upon which the dissertation will build. Subsequently I give a brief introduction to the types of disease modeling used in each chapter. In the final section I outline the main questions of each chapter and the key findings therein.

**A Social Scientist’s Guide to Measles**

Acute immunizing infections are particularly revealing of host population dynamics as they tend to settle into an equilibrium state characterized by periodic oscillations \[22, 115, 53, 52\]. These boom and bust cycles are driven by the balance of infected individuals and susceptible individuals. When the host population is largely susceptible, a small number of infections can grow rapidly. Eventually, the population will be saturated with infections and the susceptible population will be insufficient to maintain the spread of disease. Infected individuals will recover and acquire lifelong immunity. The pathogen will then go extinct or persist at a low level until the susceptible population has become large enough for reinfection \[10, 78\].

Measles is a paradigmatic example of one such acute immunizing infection. Measles has a long and intimate relationship with human populations. Though it was pushed to the brink of elimination in Western countries by mass vaccination campaigns beginning in the 1960s, vaccine hesitancy has allowed the disease to regain a foothold (as shown in Figure 1) \[88, 71, 94, 19, 29\]. In 2019 The United States experienced a surge of measles cases as a result of large outbreaks in American
Figure 1: **Measles burden and resurgence.** (A) The global burden of measles cases in the last decade by WHO region [2]. (B) Measles cases from the last decade in England and Wales. The caseload of measles in England and Wales has remained quite high in spite of vaccine availability. Still, this is approximately 1% of the incidence during the pre-vaccination era [1]. (C) The last ten years of measles Cases in the United States. The 2010s has seen a precipitous rise in measles cases in the United States, particularly in 2019 with large outbreaks in New York and American Samoa [29].

Samoa and New York state [40 87 89]. In fact, in 2019 measles cases in the United States reached their highest levels in over 25 years. [102]. Furthermore, measles has remained a significant cause of infant mortality in the developing world [55 82 2]. Nearly a quarter of a million people were infected with measles in the Democratic Republic of Congo in 2019, an epidemic that lead to 4,723 deaths in just ten months [40].
Measles is one of the most contagious endemic pathogens in human populations [50]. It is an airborne virus and can survive for up to two hours in the air or on surfaces. As a result of this transmission mechanism, the spread of measles both depends on (and can reveal) patterns in population mixing. Its long-term dynamics are also highly sensitive to changes in birth rates [48, 42]. Due to its high contagiousness, in the pre-vaccine era measles would infect essentially every person by age 10 [51]. If an individual survives infection, they are guaranteed lifelong immunity, therefore, the primary susceptible population is young children [18, 53, 22, 48]. Changes in this population as a result of increasing birth rates - or in the case of post-vaccination era, decreasing vaccination - will alter the dynamics of the disease.

Before widespread vaccination, measles incidence in Western countries generally settled into annual or biennial cycles (epidemics every year or every other year respectively) [53, 48, 52]. These outbreaks would be driven by a combination short-term seasonal forcing - such as school calendars or seasonal migration - and long term susceptible recruitment and depletion. The timing of outbreaks is driven by these short term seasonal changes in contact rates, while the spacing of the outbreaks - annual, biennial and occasionally more exotic multi-annual gaps - is driven by susceptible replenishment. Figure 2 shows these long and short-term forces by plotting cases, estimated susceptible populations and the seasonally varying transmission rate.

While understanding how measles responds to population dynamics is critical for public health interventions, it is also true that these infections reveal significant information about the populations they inhabit - particularly in the pre-vaccine era when measles cases would wane but never disappear. In this way, the study of measles specifically is really a study of human society. This dissertation investigates the relationship between demographic shifts in space and time and the impact of these shifts on measles ecology. Each chapter will use subsets of the same pre-vaccination England and Wales (hereafter E&W) measles data. The data contains weekly case
reports from 954 urban and 468 rural districts for the years 1944-1965, recorded on a weekly basis by the UK Registrar General (OPCS). National notification was made mandatory in 1944. In addition, the data contains annual births and populations as well as geographic coordinates for each district. The unparalleled granularity and diversity of this data set make it ideal for investigating the relationship between demography and measles transmission.

Figure 2: **Long and short-term oscillations in measles transmission.** Panel A shows measles incidence for pre-vaccination London (top) and the estimated susceptible population (bottom). The oscillations between outbreaks and susceptible recruitment are clear, with a build up of susceptibles occurring before each outbreak, each outbreak then drastically reduces the number of susceptibles. Panel B shows the estimated seasonal transmission rates with the grey rectangles representing the school term and dotted lines indicating standard errors of the estimate. We see that transmission rates spike when children return to school and transmission is highest in general when school is in session.
Demographics and Disease Dynamics

Demographic and epidemiological models have been linked since Bernoulli modeled smallpox mortality in 1766 [98]. As an argument for inoculation, Bernoulli calculated the increase the life expectancy at birth if smallpox were eradicated as a cause of death. However, in the years since this seminal paper, the importance of demographics methods in the study of epidemic dynamics has been relatively neglected by researchers [10, 78].

Though epidemiologists understand the importance of key demographic features such as population size, migration, and birth rates, the fields do not frequently combine methods or literature. Historically, it has been the spatial statisticians, epidemiologists, and geographers that point out the importance of spatial and demographic processes for the emergence and spread of disease [80, 113, 45, 46, 35, 91, 101]. Some of this work draws attention to the importance of sociodemographic features such as population flux, vaccination coverage, and population density on whether and how disease emerge, explode, go extinct or persist [101, 91, 35]. Other research demonstrates the importance of local and regional population movement and vaccination coverage in the dynamics and persistence of infectious disease [80, 113, 46, 45, 68]. Additionally, epidemiologists have used mathematical and statistical models to provide insights into how key demographic features such as birth and seasonal contact rates impact measles persistence in England and Wales [53, 22, 23, 115]. While many of these findings demonstrate a critical link between demography and epidemiology, the majority of these researchers are not demographers by training. This is an area in which demography stands to gain a lot by being in an interdisciplinary conversation.

The most important demographic events, for our purposes are births and (local and regional) contact rates. The primary driver of pathogen survival is encountering susceptible hosts. For acute immunizing infections, the most significant source of susceptibles is births - allowing for the waning of short-lived maternal immunity.
In areas when diseases go locally extinct access to new susceptibles can come from migration of infected individuals. Population mixing rates can impact how quickly a disease will spread within a community [15, 39, 52, 20].

Both local and global population mixing rates can impact the spread of disease. For example, local sexual network structures have a large impact on transmission rates and the total number of infections [5, 83]. Global connectivity is critical for the spread of emerging infections or seasonal influenza outbreaks [36, 70]. In the sections below I provide explanations and examples for the importance of each demographic feature and its impact on disease dynamics. In particular, I will focus on measles in the context of pre-vaccination England and Wales as these will be most relevant to the dissertation.

**Birth Rates**

Birth rates are critical for influencing the life cycle of infectious diseases, particularly in the case of acute immunizing infections [42]. Infections that grant the host lifetime immunity after recovery require a supply of new susceptibles from births. Seasonality in births can drive differences in the seasonal transmission of acute immunizing infections, but is not as dynamically important as seasonal contact rates [42]. However, birth rates are typically responsible for long-term oscillations (multi-year boom and bust cycles) as the infection awaits an increase of the susceptible population to reach sufficient levels to sustain transmission [48].

**Birth Rates in England and Wales**

As the rate of susceptible replenishment drives long term dynamics, we often see changes in these long-term oscillations as a result of changes in birth rates. Finkenstad and Grenfell showed using simulated epidemics that low crude birth rates (below .07) produce small annual outbreaks. As the birth rate increases, between .08 and 1.2,
the outbreaks become predominantly biennial with small annual outbreaks. Finally, at very high birth rates, the epidemics return to an annual pattern (above 1.2). In pre-vaccination E&W, the aggregated epidemics were predominantly biennial after 1950. Before 1950, the national dynamics demonstrate some larger annual outbreaks. Finkenstad and Grenfell posit that this annual period is a result of the post-war surge in births known as the "Baby Boom" [48].

Local Contact Rates

This section focuses specifically on ways contact rates can vary locally rather than regionally. Factors such as population density and heterogeneous mixing rates within a single location are a particular concern. The migration section will consider contact rates between locations. Local contact rates can be explored using two different concepts: assortative mixing and and population density.

Assortative mixing describes a pattern by which similar individuals mix more frequently - or have greater contact rates - with each other than with dissimilar individuals [83]. One of the most commonly documented such patterns is age assortativity. In general, individuals mix most frequently with individuals of similar ages, and the very young also mix with adults (e.g. parents and children). This pattern of mixing can alter the trajectory of an outbreak by spreading most quickly among individuals in certain age classes [84, 95]. Due to the patterns of population mixing and the agglomeration of youth in schools, school-age children tend to be the most susceptible to spreading infections quickly. In other words, once an infection reaches a school population it will spread rapidly through that population and filter slowly into other populations via comparatively rare contacts in other age groups [84].

Regarding population density, epidemiologists classify pathogen transmission as either frequency dependent or density dependent [53, 47]. Density dependence, as the name suggests, means that greater population density affords faster transmission.
The density of individuals corresponds to environments that can be conducive or inhibitive for pathogen proliferation. In general, the number of new hosts a pathogen will encounter in a single time unit will be greater in a higher-density population relative to a sparse populations [28, 9]. For this reason, high density populations typically require higher rates of vaccination to eradicate infectious diseases [9, 11, 11]. This is particularly true for diseases like influenza or smallpox where proximity plays a key role in transmission [11].

Frequency dependent transmission means one infected individual is to infect the same number of people independent of the host population density. In an abstraction, this is like saying people have only eight friends whether they live in a large dense city or a rural town; a sick person will therefore only infect eight people regardless of local population density. Frequency dependent transmission may occur if the density of the susceptible population is independent of population size. In the cases of measles, for example, the primary susceptible population is school children and the primary transmission location is in schools. In this case, what matters is the density of classrooms and child-to-child contact rates. These may be relatively constant over a variety of cities and towns regardless of local density [22]. Another example is sexually transmitted infections. In this case, the relevant type of contact is likely to be relatively constant across population densities because individuals are not necessarily more likely to have more sexual partners as a result of local population density [53, 106, 93].

**Local Contact Rates in England and Wales**

Estimates of transmission rates for measles demonstrate the log of transmission is inversely proportional to the log of location size. This scaling produces flat estimates of the number of secondary infections, estimated in this case as the product of the transmission rate and the population size. For density-dependent infections, we would
expect greater numbers of secondary infections as a result of an increase in population size. Previous research on measles suggests that the number of secondary infections does not scale with population size. This is because the primary population of interest for pre-vaccination measles is school aged children. So the population density of school age children (i.e. classroom and school size) is the more important driver of epidemic dynamics, and this population density may not scale with population size [53]. Some of the research in this dissertation will complicate and expand on these findings.

In regard to the seasonal shape of transmission (driven by temporal variation in contact rates), districts in E&W are extremely consistent. Across districts in E&W the timing of epidemics is consistently tied to the school calendar. Estimates of transmission rates show they are highest during the school-term with peaks in transmission when children return from holidays [92, 53, 28].

Population Size

The size of the host population influences pathogen life-cycle as it determines the limits of pathogen spread. If a population (particularly the population of susceptible individuals) is not sufficiently large, the disease will die out after infecting all possible hosts. If the population is large enough, the disease can persist until births or migration provide access to new hosts. The threshold at which a disease can persist without going extinct is called the critical community size (CCS). This threshold varies and is typically smaller the less contagious a pathogen is. The more contagious the pathogen, the more rapid the spread, the larger the population necessary to keep it from going extinct.

Population Size in England and Wales

Measles is highly contagious and thus quickly depletes the susceptible pool. Therefore, to remain endemic, it requires a large population to produce a large number of births.
In E&W during 1944-1965, measles only remains endemic in populations greater than 250,000 to 300,000 [52, 23, 16].

Locations of sufficient size to support endemic measles typically act as epidemic pacemakers for the surrounding region. For areas below the CCS threshold, measles will invariably go extinct. These areas depend on imported cases from large, endemic locations to spark new epidemics. Though only about twelve districts in E&W during this time period are of a sufficient size, these locations are able to disseminate cases across the country. So epidemics happen dependably (with the exception of very small, isolated places) every two years. Even though the disease will go locally extinct, it will be reintroduced as a result of cross-location contact. In endemic districts the disease never completely disappears and these larger districts act as a pacemakers, spreading new cases into the surrounding areas [52].

Migration

The movement of individuals across locations provides the opportunity for pathogens to encounter new susceptible populations. Small or geographically sparse populations are not conducive to the survival of a highly contagious acute immunizing disease. In such locations the disease will go locally extinct. Migration presents an opportunity for the pathogen to be reintroduced [34, 101].

Migration is a measure of connectivity between regional or international locations. Where humans move, pathogens will necessarily follow. Therefore, both routine and exceptional changes in population movement provide an opportunity for disease dynamics to change. Migration comes with a variety of temporal and spatial variations. These movements can be temporally novel such as those brought on by crises, or systematic and routine such as commuting for work, traveling between the city and the country for holidays, and migrating for seasonal agricultural work [45, 20, 101]. Migration can be regional, say from a city to a neighboring rural district; or interna-
tional as a result of routine air travel (e.g. the spread of influenza or COVID 19); or as a result of mass movement in response to environmental or political instability (e.g. Venezuela and the reintroduction of childhood illnesses to Colombia) [34, 101, 100].

Though the movement of populations is dynamic, a large subset of that movement is regular and consistent. Both these regular movements as well as the aberrations are opportunities for a disease to move from one host population to another. Regular movement does not necessarily mean the same individuals travel back and forth between two locations, but rather that a consistent volume of individuals flow between locations across time. This produces a kind of connectivity that can be conceptualized in the form of probability. If two places have a steady stream of migration across them, the likelihood that a pathogen will spread from one such location to another is high. For example, a suburb outside a large city will almost certainly import an infection from a city via regular commuters. In contrast, a rural farm town many miles from a large city will have a reduced likelihood of receiving an infection. Or, more relevant for global outbreaks, two cities with many flights between them are more likely to communicate infectious pathogens with each other than two locations with few or no direct flights between them [115, 20, 45, 80].

Migration in England and Wales

The impact of population connectivity in E&W has been studied in two ways. First, using the timing of epidemics, researchers have found (controlling for population size) locations nearer to cities above the CCS experience more frequent and more synchronized outbreaks than those farther away [23]. Additionally, larger locations tend to have greater rates of cross-location transmission. In this way, disease cases demonstrated that large locations generally have more individuals moving between them and therefore a greater probability of reintroducing infections. Additionally, research has shown that infections start in endemic areas and travel, much like a
ripple effect, to surrounding districts. Thus the movement of measles in E&W from one location to another is a story of both proximity and population size.

**Seasonality**

Seasonal fluctuations are common in a variety of demographic processes, many of which I have already mentioned. For example, seasonality drives local contact rates (school calendar), migration (seasonal employment), births and deaths [26, 79, 50, 12, 44, 100, 42, 75]. In the United States for example, births tend to occur in warmer months while deaths tend to occur in colder months [76, 44]. For acute immunizing infections however, seasonality of births tend not to impact cases much as the seasonal contact rates driven by the school calendar which are more influential [42]. The seasonality of deaths does not have a significant impact on measles transmission. This is because measles infects school aged children (5-10) [51], a population for whom death rates are typically low - particularly during the relevant time period for this research.

**Measles Seasonality in England and Wales**

Measles incidence in E&W are subject to both short-term (fluctuations within a year) and long term (annual, biennial, triennial). Long term cycles are produced by the comparatively slow build up of susceptible hosts. Short term cycles and the timing of epidemics are driven by the variation in contact rates across the year. The primary driver of epidemic timing in E&W is the school calendar, because this determines when the main population of interest (susceptible school-aged children) will be in contact with one another [48, 22].
Infectious Disease Analysis

We know human populations and their interactions with infectious disease reveal much about the diseases as well as the host populations, but how do we measure and interpret this relationship? To quantify transmission rates and identify necessary vaccination rates to maintain herd immunity, we must have a way to translate disease surveillance into these key estimates. There are two domains of data analysis relevant to this dissertation: the time domain and the frequency domain. The time domain involves using infectious disease time-series to evaluate the dynamic transmission of a pathogen. Time domain analyses use mechanistic (or semi-mechanistic) models to compare hypothesized structures of transmission with observed time-series data. The frequency domain focuses on oscillatory behavior and relates these periodic fluctuations to population dynamics. Frequency analyses provide detailed descriptive information regarding periodic disease behavior.

The unparalleled granularity of the E&W data set has produced a wide range of results and insights into the epidemic dynamics of measles including metapopulation dynamics, spatial infection hazards, local transmission rates and school term seasonality. These results have been obtained using a combination of time-domain and frequency-domain techniques. I will build off many of these findings in this dissertation. What follows is a brief review of the methods that will appear in the dissertation.

Time Domain

Compartmental Models

In 1766, Daniel Bernoulli published the first mathematical model detailing the spread of an infectious agent in a human population. Since then, this work has been extended to include mechanistic models, host-population age structure, and models that can
be calibrated against observational data. In particular, the mass action Susceptible-Infected-Recovered (SIR) family of models has proven its worth in illuminating disease dynamics in a number of contexts from short-term outbreaks to centuries of epidemic patterns [98].

In practice, the SIR model assumes that individuals can be classified into one of the three (S-I-R) states. Individuals move from class S to class I at a transmission rate (usually $\beta$) and from class I to class R at a recovery rate (typically $\gamma$). For any given values of $\beta$ and $\gamma$, the infectious disease dynamics can be modeled using ordinary differential equations. Adaptations can be made to the SIR for more complex dynamics, such as: age-structured mixing, the addition of an exposed class (SEIR), or a return to the susceptible class rather than recovery (SIS), i.e. in the case on non-immunizing infections [98]. Figure 3 shows the standard SIR model with SEIR and SIS variants.

While both the SIR and SEIR model have been used to study measles, [48, 21, 66] predictions and analysis are consistent between models. Though the SIR model is a simplification of the SEIR model, modeling results of measles are robust to model choice. For this reason, the analyses herein uses the SIR for simplicity and consistency with previous work [66].

Compartmental models have been immensely valuable in quantifying transmission rates and providing concrete thresholds for understanding epidemics. The basic reproductive number $R_0$ is one such threshold value. $R_0$ is defined as number of secondary infections expected when infection is introduced to a susceptible population. It is estimated as $R_0 = \beta \tau N$, where $\tau$ is the duration of the infectious period and $N$ is population size. According to the SIR model, a disease will spark an epidemic if $R_0 > 1$ and the disease will die out if $R_0 < 1$. [98] The goal of max vaccination or quarantine is to reduce the value to $R_0$ to less than unity by reducing $N$ (or $S$, the susceptible population).
Figure 3: Compartimental disease models. The SIR model (top row), the model we will use for measles, has three states: susceptible, infected and recovered. Individuals move between these states by rates $\beta$ and $\gamma$. A large part of infectious disease analysis is estimating the rate of transmission, $\beta$. The SEIR model includes an exposed class and the SIS model returns individuals to the Susceptible class post-infection. Births supply new individuals to the Susceptible class with the rate $M$ representing the waning of maternal immunity. Each class can be subject to a state-specific death rate ($\mu$). These models can also be modified to include vaccination rates.

In spite of providing intuitive parameters and a simple mathematical framework, these models have been called “qualitative caricatures;” an illustrative abstraction of disease dynamics but without the ability to capture the stochastic and complex nature of observed disease phenomena [98]. In particular, critics point out that it does not capture heterogeneities in mixing rates, space or pathogen genetics. Still, this class of models is immensely popular and much of what we know and understand about disease transmission comes from studies and simulations based on compartmental models. Furthermore, the SIR model has had routine success in fitting and predicting outbreaks (particularly for measles) and is surprisingly robust for such a simple framework [22, 53, 62, 10, 78, 9]. Recently, an SEIR model was used to estimate the transmission rate of the new coronavirus (COVID-19) [114].
Time-series SIR Model

While the simplicity of the SIR model is beneficial for interpretability, calibrating the seasonality-forced SIR model against data can be statistically challenging [48, 59, 18]. The main challenges of fitting the SIR model to data results from two sources: only one state variable is observed (the number of cases) and this is affected by unobserved rates of under-reporting. These are further compounded by stochasticity in the underlying system. A computationally efficient option for addressing both these challenges is the time-series SIR (TSIR) model [53, 18]. The TSIR model relies on two main assumptions: first, that the infectious period is fixed at the sampling interval of the data (e.g., bi-weekly for measles) and that over a long enough time (e.g., 10-20 years), the sum of births and cases should be approximately equal if pathogens are sufficiently infectious. Both of these assumptions have been thoroughly tested for measles and found to be largely appropriate for pre-vaccine era data [53, 18, 17]. For large places, the model produces very robust results; for small places, researchers need to account for spatial coupling and imported infections to explain dynamics [23].

Quantifying Spatial Coupling of Epidemics

Epidemic coupling is a measure of how connected, in an epidemiological sense, a location is with its surrounding area. One approach to estimating this connectivity is an infection-specific spatial hazard model [23]. The spatial hazard approach uses disease incidence of a single location, relative to the surrounding region to deduce the connectivity of this population to the regional population [23]. This is, in essence, an estimate of migration. By design, epidemic coupling uses a spatial hazard model to determine how likely contact is between this location and surrounding locations. This is a key example of how disease incidence data can give an abundance of demographic information. Using disease incidence data, we are able to deduce an underlying probability of disease importation, which is really an estimate of migration rate. We
can estimate this connectivity using a spatial hazard model. We first need to estimate location-specific parameters (i.e. transmission rate) as well as the reconstructed susceptible dynamics; we can obtain these from the TSIR method [23].

Once a disease goes locally extinct, the district dynamics are converted into a waiting time distribution. The probability this fade-out will end is governed by (1) the probability of contact between a local susceptible individual and a nonlocal infected individual and (2) the probability that an epidemic will result from such contact. Spatial contact depends on the probability that a local individual is susceptible, the probability a regional individual is infected, and the spatial isolation of the local community. We want to know the probability that contact occurs and that an epidemic is sparked. In other words:

\[ P(A \cap B) = P(A|B)P(B) \]

Here A is the probability of an epidemic occurring and B is the probability of contact between a local susceptible individual and a regional infected individual. From here, using parameters obtained from TSIR or other time-series models, we can estimate the spatial or epidemic coupling between locations as a measure of the probability of infection reaching a single location when the disease is present in the surrounding area [23].

We can use statistical distributions and maximum likelihood estimation to find a value for the strength of coupling between locations [23, 115]. For measles, coupling is log-linear with log-population, large populations demonstrate more connectivity than small locations. Furthermore, locations near large cities demonstrate larger coupling than average for their population size [23].

A second approach to estimating epidemic coupling is the population gravity model [115, 20]. This model assumes migration between two locations \( j \) and \( k \), \( m_{j \rightarrow k} \),
is proportional to the product of the population sizes of both locations and inversely proportional to distance:

$$m_{j \rightarrow k} \propto \frac{\Theta N_j^{\tau_1} N_k^{\tau_2}}{D_{j,k}^\rho}$$

$N_j$ and $N_k$ are the populations of locations $j$ and $k$, $\tau_1$ and $\tau_2$ are immigration and emigration exponents, $D_{j,k}$ is the Euclidean distance between locations $j$ and $k$, and $\rho$ is a power describing how flux decays with distance, finally $\Theta$ is the coefficient of coupling. This model has effectively captured the major spatiotemporal properties of measles dynamics in pre-vaccination in E&W [115]. In the case of coastal cities however, gravity models underestimate migration and do not explain the observed influx of cases [20].

**Frequency Domain**

Signal processing techniques allow us to investigate seasonality and longer period oscillations. Relevant techniques encompass methods such as Fourier and wavelet analyses, both of which fit periodic functions (sinusoids and wavelets, respectively) to time series data to identify periodic behavior in complex signals. For example, a strum on a guitar creates a single note via the vibration of six strings. Signal processing techniques allow us to isolate the influence of each of those six strings (component frequencies) on the full signal (the note the guitar plays) [8]. Spatiotemporal variations in these signals can identify differences in demographic rates within locations and differences in connectivity and synchronicity between places [52]. Frequency domain analyses can be done directly from the case data itself, so we need not worry as much about estimation biases.
Fourier Analysis

We can fit sinusoidal functions to a complex signal to identify the periodicities with the most influence. To make disease incidence data more amenable to Fourier analysis, we typically scale the log transformation of the cases (adding one to avoid zeros). This creates a more sinusoidal shape, which fits better to a sine wave [24]. In essence, Fourier analysis estimates what periods create the most correlation when the time series is lagged. A Fourier series is an expansion of periodic functions, a function of many sines and cosines weighted to correspond to the original signal. This series contains all the necessary information to reconstruct the original signal from its frequency parts. It converts a time series from a function of time to a function of frequency; or equivalently, it is a projection from time-space into frequency-space.

To obtain an estimate of the dominance of different frequencies in a complex signal a spectral density is needed. A periodogram provides an estimate of the spectral density from the Fourier series. The periodogram is proportional to the squared correlation between the centered time series and the sine/cosine waves of a given frequency. The raw periodogram is not a consistent estimator of the true spectral density - in fact, variance in the estimate increases with the true power - but the variance can be improved using smoothing. Adjacent values are asymptotically independent so a consistent estimator can be derived by smoothing the periodogram. This method typically uses a moving average with half weight given to endpoints [24]. Figure 4 shows a smoothed periodogram for the London measles time series. We can see clear peaks in the spectral density and the one and two year periods (52 and 104 weeks respectively).

Continuous Wavelet Analysis

Unlike Fourier analysis, wavelets can describe nonstationary signals by providing a measure of power in both frequency and time. The major limitation of Fourier
Figure 4: **An estimated spectral density for measles periodicity in London, 1944-1965.** We can see clear peaks at the one and two year (52 and 104 weeks) periods. This indicates that one and two year cycles are dominant relative to other oscillations.

Analysis is that if a signal has periodicity A for half the observation window and periodicity B for the other half, the power spectrum will look identical to one where a signal alternates periodicity A and B rapidly for the entire observation period. In other words, Fourier analysis is a powerful tool for demonstrating which periods are present, but not for what duration of time. Continuous wavelet analysis fits wavelets locally in time and thus can estimate which periods are dominant at each time step [21 107 115 52].

Wavelets allow us to both examine changes in periodic oscillations over time as well as to reconstruct the phase of a dominant period [115 52]. Using phase we can measure temporal lag in major epidemics. For example, biennial epidemics were most
Figure 5: **London wavelet spectrum for measles cases 1945-1994.** Here we see dominant one and two year (after 1950) cycles before the implementation of mass vaccination in the late 1960s. After vaccination, cycles become annual before falling into low period oscillations. It is clear in this diagram the massive interruption in epidemic behavior caused by mass vaccination.

common in E&W during the pre-vaccination era. Wavelets enable a reconstruction of just that biennial signal, without the noise generated by smaller outbreaks. The phase allows us to isolate the timing of an epidemic for a given location. Bjornstad and Grenfell have shown, using the biennial phase of epidemics surrounding London, that the phase difference (i.e. temporal lag) between districts’ epidemics increases with distance from London [52].

**Dissertation Prospectus**

In this section I describe how the dissertation explores the interactions of population dynamics and disease ecology. Every chapter uses data from E&W during the pre-
vaccination era. My analyses investigate what measles dynamics can reveal about regional migration, local mixing rates, and birth rates. The first two chapters focus on methods of analyzing measles transmission from the time domain. The third and fourth chapters demonstrate applications of frequency-domain methods to measles incidence data.

**Time Domain**

**Chapter One**

In chapter one, I look at the impact of population density on disease transmission by focusing on urban and rural areas. Research on transmission among a subset of districts in E&W suggested that transmission did not vary systematically with population size. This is because school children are the primary transmitters of measles and schools are the primary location of transmission, so it is the number of schools and the density within schools that should matter \[22\]. However, we would still expect that transmission should still be more rapid if children have more opportunities to interact outside of school or live in more densely populated areas. For this reason, we expect urban areas to have both higher rates of spatial coupling and higher transmission rates. In fact, we find at the aggregate, urban and rural areas are remarkably similar controlling for population size.

Previous research has also demonstrated the importance of proximity to large cities in determining epidemic dynamics, so I control for spatial influence by examining urban and rural neighboring districts. Examining these neighbors, I find marginal evidence that rural areas are less well-mixed, indicating the possibility that population density may impact disease transmission. In particular, rural areas have longer epidemics while urban areas have bigger epidemics. This may be evidence that epidemics spread more slowly in rural areas and ultimately infect fewer individuals than in urban counterparts.
Chapter Two

Chapter two focuses on whether or not transmission rates scale with population size. Previous research using a subset of the E&W urban districts has suggested it does not. However, extending these calculations to the full data set I find evidence of increasing transmission rates with population size \[53\]. These raw estimates also demonstrate evidence of estimation bias in very small populations. Without a tail in the smallest locations, the trend with population size is less robust. This chapter features work and results from a collaborative project with Alexander Becker. We leverage high volume stochastic simulations to separate process error from a true trend. We then use machine learning to correct parameter estimates from the E&W data to evaluate whether a trend with population size still exists. We select a random forest model as it is less susceptible to over-fitting than other models but maintains strong predictive power; additionally the random forest provides intuitive measures for feature importance \[73\]. These measure of importance help us build concrete recommendations for deciding whether a time-series is suitable for the TSIR procedure.

We find that locations with more stochastic, irregular outbreaks are consistently difficult to fit using TSIR. This is due primarily to a lack of data over the observed time period. When we apply our findings to the E&W data set, we find that predicted bias is highest for the most isolated locations. A lack of demographic connectivity results in fewer outbreaks, and with less data we predict a consistent underestimation of transmission rates.

We also demonstrate that transmission does scale with population size, contradicting previous results. We find transmission estimates to be highest in mid-range populations and lower in the smallest and largest locations. One hypothesis for this relationship is that small locations (below 10,000) tend to be sparse as well and this reduces population mixing rates and slows disease spread. Large populations (above 100,000) may be an agglomeration of many communities with near-random mixing
within and weaker mixing across communities, this will create lower transmission rates on average across the entire reporting area. We anticipate that mid-sized cities (between 10,000 and 100,000) are closest to well-mixed and thus demonstrate the most rapid transmission rates.

**Frequency Domain**

**Chapter Four**

In chapter four I introduce a method for concisely summarizing frequency-based disease transmission across many locations. This novel method allows us to describe the wavelet spectra of all 954 urban areas without collapsing the time dimension. In this chapter I adapt tensor decomposition, a method commonly used in neuroscience for use with epidemic data [69, 38, 37].

Tensor decomposition is a form of dimensionality reduction for multidimensional data sets. It is a n-dimensional generalization of principal component analysis (PCA) and singular value decomposition (SVD) [37, 65]. PCA and SVD convert a large matrix into vectors of variable values and observation-specific scalars (called scores). Tensor decomposition (in this case) takes a three-dimensional data set and converts into an observation-specific score and two vectors for each of the other dimensions.

Using this method, I confirm past results as well as identify new relationships between birth rates and temporal shifts in outbreaks periodicity. In particular, I isolate a temporal shift in the dominance of annual outbreaks across locations; I find that locations with higher birth rates during the baby boom have larger annual outbreaks before 1950 and settle into biennial outbreaks as birth rates drop.
Conclusions and Future Work

Demographics are as much a feature of disease dynamics as diseases are an attribute of humans. Dynamics of an endemic pathogen can reveal volumes about the demography of a host population. In this thesis, I have explored what measles can tell us about population mixing rates, birth rates, regional migration and isolation. Though my research is geographically and temporally specific, the findings have broad implications. In particular, my findings regarding the impact of population growth and mixing rates on measles outbreaks is a critical concern in a more connected, rapidly urbanizing world [58, 58]. These concerns are ever more pressing in light of emerging infections [114] and waning vaccination coverage [19]. Though measles is only one pathogen, its method of transmission is shared by many other infections (influenza, smallpox, and tuberculosis) and is similar to many more (coronavirus, mumps, pertussis, rubella). Therefore, measles has the capacity to reveal dynamics that are pertinent to a variety of other diseases.

The E&W data set provides the granularity necessary to undertake many such studies. There is an abundance of qualitative data available that could accompany these data to improve our understanding of local conditions. For example, if we wish to know more about the physical structure of many of these districts through time, such data exists in archives and would undoubtedly be a valuable companion data set. Additionally, this method could be easily applied to modern infectious disease data sets as surveillance methods improve and we see a continued increase in large data sets. One feature of the annual health reports digitized by the Wellcome Trust is annual births, ages of infection, and age-specific death rates. On a very rare occasion, there is documentation of schools. All of these records could provide vast improvements for our understanding of age specific mixing rates, transmission rates, and case fatality; these data points could also improve the reconstruction of an age-specific susceptible class.
Additionally, comparing measles with diseases that are less contagious or transmitted via a different kind of contact (e.g. STIs) may help uncover heterogeneities in contact rates. One reason heterogeneities may be so difficult to spot with measles is both its school-room transmission as well as its contagiousness. Once measles enters a susceptible population, it moves quickly, limiting the impact of variation in contact rates unless the variation in extreme. Coupling these analyses with data on a less infective pathogen may reveal more about heterogeneous or patch mixing patterns.

Though I introduce tensors to the study of infectious disease, there are many opportunities for improvements and extensions. For example, I limit my analysis to dominant frequencies present in the data; an obvious extension would be to include the phase of wavelets so that lag and synchrony could be examined across an array of locations simultaneously.

Finally, extending these results into the vaccine era to increase their relevance to the modern age of disease dynamics is also an obvious area for future work. For example, using estimates of spatial coupling, and migration rates may help improve disease forecasting for emerging infections. Furthermore, incorporating the use of serological data on immunity to explicitly characterize the susceptible population could improve modeling assumptions and better guide public health interventions. With the growing improvement of contact data, the use of high power computing with agent-based models and contact network data may improve estimates and predictions when heterogeneous contact patterns exist.
Chapter 1

Structure, Space, and Size: Competing Drivers of Variation in Urban and Rural Measles Transmission.

1.1 Abstract

A key concern in public health is whether disparities exist between urban and rural areas. One dimension of potential variation is the transmission of infectious diseases. In addition to potential differences between urban and rural local dynamics, the question of whether urban and rural areas participate equally in national dynamics remains unanswered. Specifically, urban and rural areas may diverge in local transmission as well as spatial connectivity, and thus risk for receiving imported cases. Finally the potential confounding relationship of spatial proximity with size and urban/rural district type has not been addressed by previous research. It is rare to have sufficient data to explore these questions thoroughly. We use exhaustive weekly case reports
of measles in 954 urban and 468 rural districts of the UK (1944-1965) to compare both local disease dynamics as well as regional transmission. We employ the time series SIR model to estimate disease transmission, epidemic severity, seasonality and import dependence. We also examine variation in periodicity via spectral densities. Congruent with past results, we observe a clear dependence on population size for the majority of these measures. We use a matched-pair strategy to compare proximate urban and rural districts and control for possible spatial confounders. This analytical strategy reveals a modest difference between urban and rural areas. Rural areas tend to be characterized by more frequent, smaller outbreaks compared to urban counterparts. The magnitude of the difference is slight and the results primarily reinforce the importance of population size, both in terms of local and regional transmission. In sum, urban and rural areas demonstrate remarkable epidemiological similarity in this recent U.K. context.

1.2 Introduction

Though widespread vaccination has greatly reduced global transmission of measles since the mid-1960s, it continues to be a major cause of death among children in sub Saharan Africa [110]. Additionally, re-emergence of measles in many parts of the world due in part to vaccine hesitancy emphasizes the importance of continued attention to measles [94, 31]. More broadly, in an increasingly urbanized world, understanding the impacts of urbanization and population density on transmission of highly contagious infections such as measles are increasingly urgent [7]. A simple epidemic clockwork and detailed and reliable notification systems across urban and rural settings, makes measles one of the best documented spatiotemporal consumer-resource model systems generally and a particularly apt candidate for examining disease interactions across diverse population densities [52, 48, 64, 112, 33].
Urban and rural disparities in health have been studied in a variety of contexts. In the United States, the urban-rural gap in life expectancy has widened over the last four decades. This is due in large part to mortality of individuals under the age of 25 and correlates with accidental injury and reduced access to high quality medical care [99, 104]. The incidence of Dengue is known to be higher in urban areas, [96, 30] whereas the burden of malaria is higher in rural areas [58]. In rural areas, increased likelihood of concurrent sexual contacts increase risk for HIV and other STIs [5]. Seasonal migration between urban and rural areas also impacts transmission of infection [111]. While these papers have investigated differences in urban and rural health outcomes, disease burden, and migration few have explored the urban/rural hierarchy in transmission across a metapopulation or attempted to quantify differences in transmission due to urban/rural environment. This study allows an investigation into differences within urban and rural areas in addition to differences between urban and rural areas as part of a larger, connected population.

Ferrari et al show that population size is the most consistent driver of the magnitude of epidemic seasonality in measles epidemics in Niger across urban and rural districts [45]. Rainfall and agriculturally induced variation in contact rates also impact the amplitude of seasonality. The authors find reduced seasonal forcing in sparsely populated areas with the highest seasonal amplitudes present in large and/or densely populated areas. A comparable investigation of urban and rural districts in a context where school calendar forcing is the dominant mechanism has not been done. The consistency of seasonality in E&W as a result of the school calendar provides a unique opportunity to compare transmission rates and epidemic dynamics while isolating urban/rural status from other potentially confounding factors such as climate, variation in seasonal contact rates, population size and proximity to epidemic pacemakers. Additionally, the granularity of the data as a result of the number of
districts (1422) and the duration of notification data before vaccination (20 years) is unparalleled.

1.2.1 Time Series SIR Model

Measles is a paradigmatic infection for investigating nonlinear dynamics of disease transmission [25, 63, 53]. Infection with strongly immunizing pathogens such as measles results in either death of the host, or more often, recovery and lifetime protection. Compartmental models, such as the SIR (Susceptible-Infected-Recovered) model are useful as simple models of disease dynamics but can be difficult to adapt to data. The SIR model assumes a well-mixed population, and in the most basic form balances demographic processes (e.g., births, deaths, and immigration) with properties such as contact rate, $\beta$, and infectious period particular to a given pathogen. In general the transmission coefficient, $\beta$, varies seasonally, in the case of measles this seasonality is largely driven by the school calendar.

While the simplicity of the SIR model is beneficial for interpretability and cross-setting comparisons of transmission, calibrating a the seasonality-forced SIR model against data can be statistically challenging [48, 59]. The main challenges of fitting the SIR model to data result from two sources: only one state variable is observed (the number of cases) and unobserved rates of under-reporting. A computationally efficient option for addressing both these challenges is the time-series SIR (TSIR) model. The TSIR model relies on two main assumptions: first, that the infectious period is fixed at the sampling interval of the data (e.g., bi-weekly for measles) and that over a long enough time (e.g., 10-20 years), the sum of births and cases should be approximately equal due to the high infectivity of pathogens such as measles and other childhood infections. Both of these assumptions have been thoroughly tested and found to be largely appropriate for this pre-vaccine era data [53]. Fixing the infectious period to
be equal to the sampling period means we can assume an individual that is infected at the $n$ time step will be recovered by $n+1$.

**Metapopulation Dynamics and Spatial Coupling**

Highly transmissible childhood infections, such as measles, can spread quickly through a community until the susceptible population is depleted, to the point that it can lead to local extinction. Thus measles requires a steady stream of new susceptible hosts (mainly from births) to remain endemic. For this reason, large populations, typically above 300,000, are required for sustained transmission – this threshold is called the critical community size (CCS). For communities below the CCS, future outbreaks are dependent on imported infections from other locations [16]. In these small populations, the susceptible proportion will increase until the pathogen is reintroduced through spatial transmission from a neighboring community. These metapopulation dynamics, the reintroduction of infections via migration, are a characteristic component of measles transmission in England and Wales (E&W) during pre-vaccination years (1944-1965) [52, 25, 20]. Endemic areas such as London would act as epidemic pacemakers, replenishing infections for communities below the CCS threshold. Echos of London’s strong biennial epidemic pattern radiate across the surrounding region, creating biennial epidemics in locations which would otherwise be too small to experience such regular outbreaks [52]. Quantifying rates of spatial transmission is a crucial challenge for epidemiologists as the spread and persistence of pathogens depends on this connectivity between large and small places [63, 80, 20, 23].

Population gravity and spatial hazard models have been used to estimate movement of individuals across locations [80, 20, 23, 115]. Population gravity models assume that movement between locations can be approximated by the product of the population size normalized by a transformation of the distance between them. So we expect movement between large places to be high, movement between small places to
be low, and for migration between locations to decline with distance. These models have been able to successfully capture population movement in the context of disease transmission in some cases [115]. However, these models do not assume different migration rates by urban or rural location type. Further these models underestimate migration rates for coastal locations [20]. For this reason we calculate spatial coupling using a spatial hazard model which is informed by the data and assumes no predetermined functional forms. The spatial hazard model relies on susceptible and infectious dynamics to determine infection probabilities and is not constrained by location size or distance [23].

Spatial coupling provides an estimate of how much a district’s epidemic dynamics are influenced by the influx of new cases from other locations. Previous hazard-based coupling estimates suggests that transmission across locations is strongly correlated with population size [23]. Bjornstad et al use a spatial hazard model to estimate spatial coupling for all 954 urban areas in the dataset. The authors show larger places exhibit larger coefficients of spatial coupling than smaller, more isolated places. Large places exhibit more coordination with national epidemics. Additionally, the authors use residuals from linear regression of spatial coupling on population size to show that locations near large endemic locations (such as London, Manchester and Birmingham) have higher than expected estimated rates of coupling than other locations of comparable size. Similarly, locations very far from these population centers produce lower than expected coupling rates. This highlights the importance of both size and space in cross-location measles transmission.

In contrast, research on data from E&W suggests transmission within locations does not scale with location size. Bjornstad et al [22] use a subset of sixty cities in E&W to show that while transmission rates demonstrate some variability across locations, they do not vary uniformly with population size. The authors select sixty locations of various sizes and calculate the basic reproduction number based on the
epidemic data. The basic reproduction number ($R_0$) is a parameter commonly used to quantify the contagious power of disease, it is defined as the number of secondary infections resulting from a single infected individual if everyone else in the population is susceptible. Population level estimates of $R_0$ for measles are commonly between 18 and 30 [56]. Though the estimate can vary, Bjornstad et al posit that it does not vary systematically by population size. This is likely because schools act as transmission hot spots and the importance of these focus points outweighs any impact of population size. Furthermore, measles has a particularly high transmission rate: infected individuals are contagious for up to four days before they show symptoms and the measles virus is airborne and can survive up to two hours in airspace. These factors make the disease highly contagious. For this reason, once an infection is introduced to a susceptible population it will spread rapidly. As these infections will largely spread in schools, any differences due to population density are believed to be marginal, particularly in the case of E&W [22].

### 1.2.2 Potential Urban and Rural Differences

Estimates of transmission allow us to measure within-community epidemic dynamics. Spatial coupling allows us to estimate how these locations differ in their connection to metapopulation dynamics. We can therefore measure potential differences between urban and rural locations locally as well as contextually. However, we know spatial proximity plays a key role, both in terms of the local population dynamics as well as the number of imports a location can expect to receive. We also know urban and rural areas are not distributed randomly in space (both rural and small population locations are more likely to be farther from large urban areas) this indicates a need to control for spatial effects.

Though the relationship between population size and measles transmission has been the subject of many studies, previous E&W work has only focused on analyzing
data from urban districts, leaving a rich data set of 468 rural districts untouched. We may expect to see differences in the disease ecology of urban and rural locations for several reasons. It is possible that the decreased population density in rural areas leads to a fewer contacts within these locations, resulting in slower transmission of measles in rural areas relative to their urban counterparts. Variation in birth rates between urban and rural locations may impact transmission by replenishing the pool of susceptible individuals at different rates.

Differences in the number or size of schools - the primary location of outbreaks - may also impact the transmission of the disease. As susceptible contacts are generally driven by the school calendar, measles transmission in E&W typically has a consistent seasonality. We see peaks in transmission when students return from holidays, when susceptible populations are at their highest and when susceptible individuals are coming into frequent contact. If there are differences in the spatial proximity of schools we may see different transmission rates or different outbreak patterns. In particular when schools are farther apart and mixing between them is relatively weak, we might observe either multiple small epidemics or slower progression of the disease through the district [41, 13, 45, 25].

Finally, if the migration of individuals occurs at uneven rates between location types, this may impact the probability of introducing new infections, and therefore spatial coupling estimates. If population movement between locations depends on more than just population size we may see differential case import frequency between urban and rural areas. For example, people may move between urban locations with more frequency than from urban to rural or between rural locations, in which case we will see lower estimates of spatial coupling in rural locations than in urban locations.

We examine both spatial connectivity and within city measles transmission in urban and rural areas in 1422 locations in E&W. Using estimates of transmission rates and spatial coupling, we compare all urban and rural districts. At the aggregate,
we find a surprising amount of coherence: both internal dynamics and spatial coupling show consistent dependence on population size. We further restrict our sample to neighboring urban and rural districts to isolate the potentially confounding associations between size, location and urban/rural designation. We find that size is consistently a more significant driver of epidemic dynamics than location type. The exchange of outbreaks between neighbors is dependent on population size, with larger locations frequently introducing outbreaks to their smaller neighbors. In this way, these mini-communities mirror national metapopulation epidemic cascades. However, we do find slight distinctions in the epidemic behavior of urban and rural areas, namely that rural areas are characterized by more frequent outbreaks which infect fewer individuals. This suggests rural areas may sustain epidemics through internal rescue effects [80] and highlights the importance of accounting for heterogeneous mixing patterns to uncover subtle differences in epidemic spillovers.

1.3 Data

To explore differences in urban and rural areas, we analyzed pre-vaccination weekly measles incidence data from 1944 to 1965 E&W [52, 67]. This data set is unusually rich, with 954 urban cities and towns and 457 rural districts. In addition, we used annual births, population, and geographic location of each district.

The classification of districts as urban or rural was not strictly scientific at this time. The system of the era involved a combination of considerations such as population density (measured in people per acre), level of urban development, and the type of local government (e.g. urban council or parish) [105]. Historical documentation indicates that this system of classification was at times arbitrary and resulted in a misclassification rate of approximately 20% according to contemporary standards [105]. Still, this classification is a feature of the dataset and likely represents some
amount of structural difference between locations. Additionally, if the misclassification occurred at random or resulted in more frequent classification of small sparsely populated districts as “urban,” this would attenuate any differences we detect. Even so, as we subset our data to the paired locations, we do expect some variation to exist as they represent one urban and one rural component of a single regional community. Even if differences in population density to match extremes, if there existed a different epidemic signature for urban and rural locations, we would expect to see a trace of it here. The slight variation between urban and rural areas we find here is likely an attenuated version of differences we may see in more extreme contexts.

Incidence data were aggregated to the biweekly scale for modeling analysis (described below). The diversity of locations in terms of geographic space and population size, as well as the temporal detail of the incidence data provide an unparalleled and uniquely apt data set for investigating urban and rural differences in transmission. The twenty year epoch covered by the data allows for a robust study of outbreaks as well as sufficient opportunities to compare urban and rural epidemics. Furthermore, pre-vaccination data allows us to understand transmission patterns without uncertainty related to vaccination coverage, this provides the most direct estimates of transmission rates and mixing dynamics.

1.4 Methods

1.4.1 The TSIR Model

We compared population dynamics and transmission within urban and rural using a number of metrics. We first examined epidemic fadeouts (time between epidemics) to see if urban and rural areas differ in the pathogen extinction rates. As these estimates may be subject to bias due to systematic differences in reporting rates, we also calculated the number and length of three-week fadeout which previous research
has shown to be robust to under-reporting [22]. We also computed average birth rates as well as the coefficient of variation in births. Births may impact disease dynamics by altering the yearly influx of susceptible individuals. Finally, directly from the incidence data, we calculated epidemic growth rates which we expect may correspond to differences in population mixing. To further assess local disease dynamics, we use a time series Susceptible Infected Recovered model (TSIR) [22, 53, 17], to obtain estimates of seasonal transmission. The TSIR model is a discrete time mechanistic model where the susceptible dynamics can be modeled as:

\[ S_{t+1} = B_t + S_t - I_{t+1} \]  

(1.1)

The susceptibles at \( t + 1 \) (\( S_{t+1} \)) are simply the previous susceptibles (\( S_t \)) plus births (\( B_t \)) minus the new infections (\( I_{t+1} \)). The associated deterministic infected dynamics are:

\[ I_{t+1} = \beta_t S_t I_t^\alpha \]  

(1.2)

The seasonally varying transmission rate is estimated as \( \beta \) and \( \alpha \) is a tuning parameter fixed to be .97, in line with previous analysis [22]. The primary assumption of the TSIR model is over a sufficient period of time, due to the high transmission rate of measles, everyone should acquire the infection. This allows us to assume that cumulative cases and cumulative births will be approximately equal, yielding an estimation of reporting rate. We can then reconstruct the susceptible population at each time step. With estimates of both the infected (reported cases divided by reporting rate) and susceptible dynamics, equation 2.2 can be log transformed into a linear model:

\[ \log[I_t + 1] = \log\beta_t + \log S_t + \alpha \log I_t \]  

(1.3)
From equation 2.3 we can estimate both the seasonal transmission rate ($\beta_t$) and an approximate measure of $R_0$ ($= \beta_t N$ where $N$ is mean population size). We can also evaluate the seasonality by calculating the coefficient of variation in $\beta_t$, this allows us to measure whether transmission is variable over the year or relatively constant [28, 41, 17]. A full discussion of the implementation of TSIR can be found in references [17, 48, 28]. A discussion of result sensitivity to estimation procedure (such as the regression type selected for susceptible reconstruction) can be found in Appendix A.

**Epidemic Coupling**

We use the reconstructed susceptible dynamics as well as estimates of $\beta$ to calculate epidemiological coupling for each location. Following extinction the local dynamics are converted into a waiting time distribution, for which the probability that a fade-out will end is governed by the probability of contact between local susceptibles and regional infectives as well as the probability that an epidemic will result from such contact. Spatial contact depends on the probability that a local individual is susceptible, the probability a regional individual is infected, and the spatial isolation of the local community ($1/c_j$ where $c_j$ is the coefficient of coupling). We want to know the probability that contact occurs and that an epidemic is sparked. In other words: $P(A \cap B) = P(A|B)P(B)$. Here $A$ is the probability of an epidemic occurring and $B$ is the probability of contact between a local susceptible individual and a regional infected individual. In order to estimate the probability of an epidemic we estimate the number of susceptible individuals at each time step as in equation 1.1. We use a modified version of equation 1.2 in which the expected number of infections is given by:

$$\lambda_{t,j} = \beta_{u,j}(I_{t,j} + i_{t,j})^\alpha S_{t,j}$$

(1.4)
\( I \) are local infected individuals and \( i \) are infections arising from migration. The probability of spatial contact between a local susceptible and a nonlocal infectious individual is:

\[
1 - \exp(-c_j x_{t,j} \bar{y}_t)
\]  

(1.5)

The proportion of local susceptibles is \( x_{t,j} \) and \( \bar{y}_j \) is the proportion of infectious nonlocals and \( c \) is the coupling coefficient. So the distributions of infection is:

\[
I_{t+1,j} \sim \text{NegBin}(\lambda_{t,j}, I_{t,j} + i_{t,j})
\]  

(1.6)

Infections resulting from migration are determined by the spatial hazard:

\[
i_{t,j} \sim \text{Bin}(1, 1 - \exp(-c_j x_{t,j} \bar{y}_t))
\]  

(1.7)

Given contact has occurred, the probability that an epidemic does not occur is given by \(1/(1 + \beta_{t,j} S_{t,j})\), this is given by the null probability of the binomial distribution when \( I = 0 \) and \( i = 1 \). An epidemic will occur by the complementary probability:

\[
1 - \frac{1}{1 + \beta_{t,j} S_{t,j}} = \frac{1 + \beta_{t,j} S_{t,j} - 1}{1 + \beta_{t,j} S_{t,j}} = \frac{\beta_{t,j} S_{t,j}}{1 + \beta_{t,j} S_{t,j}}
\]  

(1.8)

Putting together the probability of spatial contact and the probability of an epidemic we obtain the discrete-time hazard:

\[
(A \cap B) = P(A|B)P(B) \rightarrow h(t, j) = \frac{\beta_{t,j} S_{t,j} (1 - \exp(-c_j x_{t,j} \bar{y}_t))}{1 + \beta_{t,j} S_{t,j}}
\]  

(1.9)
Conditional on the local susceptible population and regional prevalence of infection, the theoretical waiting time distribution can be written as the expectation of a binomial process for which the log-likelihood of the fade-out is given by:

\[ l(c_j|I_{t-1,j} = 0) = \sum_{I_{t-1,j}} \ln(h_{t,j}^{z_{t,j}}(1-h_{t,j})^{1-z_{t,j}}) \] (1.10)

The binary indicator \( z_{t,j} \) is equal to 1 if \( I_{t,j} > 0 \) and equal to zero otherwise, and \( h_{t,j} \) is given by equation 1.9. We use Newton-Raphson maximum likelihood estimation to obtain an estimate of \( c \) from equation 1.10. Note that it is not possible to estimate \( c \) for those communities in which measles is endemic (i.e. there are no fadeouts). There were 29 communities (out of 1422) that did not have a sufficient number of fadeouts to estimate the coupling coefficient, these locations are dropped from the coupling analysis but are included in the comparison of other measures of epidemic behavior.

In keeping with previous findings [23], we expect to see high coefficients of coupling for locations below the critical community threshold that are geographically proximate to larger districts which provide the reintroduction of pathogens. We expect small isolated areas to have the lowest estimates of coupling because migration to and from these areas will be relatively rare and thus the reintroduction of pathogens will have correspondingly lower probability.

### 1.4.2 Principal Components Analysis on Paired Data

Probing differences further, we then subset our data to neighboring urban and rural districts to isolate space from size and location type. These districts are adjacent and non-overlapping such that they allow us to control for spatial proximity and measure the relative influence of urban versus rural status and population size. These districts
are sampled from a variety of spatial locations across E&W so the results are not a feature of a single area.

We selected a representative sample of 179 pairs (a total of 358 districts) below the CCS and utilized principal components analysis (PCA) to uncover the correlations between demographic characteristics (birth rates) and estimated parameters (susceptible fluctuations, transmission rates) and subsequently to see how urban and rural areas vary across these numbers. PCA is the eigenvalue decomposition of the covariance matrix of scaled covariates. We scale the data so that variables with larger values are not weighted higher simply because they have higher variance. Eigenvalue decomposition factorizes a matrix into its canonical form. It produces the vectors which (ranked by their eigenvalues) explain the most variance within the data. PCA uses an orthogonal transformation to project a matrix possibly correlated covariates onto a new uncorrelated basis space. PCA has been used identify the genes which are responsible for the most population-level variation populations [85] as well as to isolate dominant frequencies in complex signals [24].

This method demonstrates (1) how variables are related to each other, (2) which variables are most influential in terms of looking for differences in the data and (3) whether population size or urban/rural designation influence how locations score on these maximal variance vectors. PCA enables an investigation of multiple variables simultaneously as well as isolating the importance of variables rather than testing each covariate separately. We withhold urban and rural indicators as well as population size so we can test their influence on the projections. After obtaining our principal components (eigenvectors), we project each location onto the first two principal components (the vectors responsible for the two dimensions of most variance). We compare each city’s score with its rural neighbor to assess the influence of space. If space is the primary driver we expect each location to be similar to its neighbor. We also compare the score’s by population size. Finally we calculate the euclidean
difference between each pair, this gives us a measure of how different each location is from its neighbor across the dimensions of highest variance. We then check the association of this distance with their difference in size. These comparisons together demonstrate the comparative influence of space, size and environment.

1.4.3 Epidemic Exchange in Paired Locations

We further attempt to disentangle the importance of district type by investigating the timing and duration of epidemics between pairs. Many of the paired districts are small (median population 15,000; range 700-250,000), and fitting the TSIR model to locations can be challenging due to frequent and lengthy extinctions. Making comparisons directly from the time series enables us to concretely measure timing and coordination of epidemics to assess how and if districts interact with each other. In particular we evaluate the proportion of rural epidemics which occur during a simultaneous epidemic at its urban neighbor. Similarly, we assess which of each pair leads or lags in local epidemics. For pair of districts, we evaluate the proportion of it’s epidemics which are preceded by an epidemic in it’s neighboring district. These proportions provide a measurement of how many epidemics can be attributed to the urban or rural component of each pair of districts. We also compare the total number of outbreaks and the number of large epidemics (> 14 weeks) between pairs.

1.5 Results

1.5.1 Aggregate Urban and Rural Comparison

Investigating aggregate differences, such as epidemiological spatial coupling and fade-outs, between urban and rural districts, we find a consistent relationship with population size but no obvious difference by urban or rural designation. These findings are consistent with previous estimates for urban districts [23]. Coupling increases
log-linearly with population size for both urban and rural areas indicating that imports increase with population size (Figure 1.1). Additional comparisons of TSIR parameters such as transmission rate ($\beta$) and $R_0$ reveal variation with population size, but urban and rural locations remain consistent (Figure A.1). The proportion of biweeks without cases correspondingly decreases with population size for both urban and rural areas. However, Figure 1.1 further demonstrates the importance of spatial proximity to large metropolitan areas such as London, Birmingham, Manchester, Liverpool, and Leeds in terms of both urban/rural designation as well as population size. Due to the importance of spatial proximity in regard to imported cases, it is necessary to isolate the effect of proximity from population size in order to further understand any potential urban and rural differences.

1.5.2 Urban and Rural District Pairs: Principal Components Analysis

When we subset the data to the selected paired districts (mapped in Figure 1.2 A) we observe a modest difference between urban and rural areas when applying principal components analysis to the estimated parameters. This decomposition shows that the two most variable axes of difference are (1) high coupling versus long fadeouts and (2) variation in seasonal transmission versus growth rates.

The first principal component (hereafter PC1) accounts for 36% of the total variance in the data (Appendix Table A.3). It is an axis which measures the data with coupling at one extreme (positive values) and fadeout length at the other (negative values). In other words, the spectrum which accounts for the most variance in the data is merely the separation of dynamics between locations which receive regular imports and demonstrate synchrony with the larger metapopulation, and isolated districts which experience long droughts of infection.
Figure 1.1: Epidemic coupling and outbreak dynamics in urban and rural England and Wales 1944-1965. (A) Plots the longitude and latitude of each of the 1422 locations color coded by urban-rural status and scaled by mean population size. The five largest locations (London, Birmingham, Manchester, Liverpool, and Leeds) are marked in red. (B) Patterns of measles incidence show strong spatiotemporal correlation across E&W with a strong dependence on distance. While correlation of outbreaks is high among near neighbors, this correlation decreases rapidly with distance until about 200km where it slowly falls below the average temporal correlation for both urban (.12) and rural (.14) districts. The spatial correlation of all locations combined is consistent with these separate estimates. (C) Analyzing coupling as a function of population yields a strong log-log linear relationship across both urban and rural cities. This is consistent with previous findings we do not see significant differences between urban and rural areas controlling for population size. At the large populations, we see a large increase in the size of confidence intervals. This is due to the small number of interepidemic periods in large locations. With few opportunities to calculate coupling, the standard errors increase drastically. (D) Similarly, we find strong agreement between the urban and rural relationships for fadeouts and population. These figures demonstrate the importance of population size as well as proximity in determining epidemic dynamics.
The second component (hereafter PC2) explains 26% of the total variation in the data (Appendix Table A.3). This component projects districts on an axis with variation in transmission ($\beta$) at one end (negative values) and growth rates on the other (positive values). In qualitative terms, this suggests that epidemics generally exhibit rapid epidemic spread or strong seasonal variation in transmission. In other words, locations with explosive epidemics tend to have less seasonal variation, implying outbreaks are more randomly spread throughout the year. Similarly, locations with more seasonal transmission experience epidemics which spread at a relatively slower pace. A district with a negative score on this axis will be characterized by high seasonality, indicating regular epidemics fed by relatively constant susceptible pools. A district with a positive score on the second component will likely have stochastic and explosive epidemics rather than annual or biennial school-based outbreaks. Figure 1.2 shows the projections on PC1 and PC2 (Panel B) along with the variable loadings for each component (Panel C).

We observe slight and statistically insignificant differences between urban and rural districts. On average, rural areas have slightly higher coefficients of coupling compared to urban areas. Conversely urban areas have on average fewer, more potent outbreaks.

With regard to where districts fall on PC 1 and PC 2, Figure 1.2 B shows districts either tend to have higher coupling estimates (positive on PC 1) accompanied by high variation in beta (negative on PC 2), or long fadeouts (negative on PC 1) and high growth rates (positive on PC 2). Therefore if we interpret the two-dimensional space created by PC 1 and PC 2 we see that the majority of variance in the data can be described as a spectrum from areas with strong epidemic coupling and consistent seasonality and one extreme and infrequent violent epidemics at the other extreme. This is consistent with previous studies of large and small urban locations [16, 20, 52].
The space created by the first and second principal components has a plausible association with population size. We expect large places to receive more import cases and thus to have larger coupling estimates and shorter time between epidemics. We also expect large places to have more regularity in seasonal transmission as dictated by the school calendar, while smaller places are more vulnerable to random outbreaks. We also see that the largest outliers in the data (third quadrant) have long fadeouts and high susceptible population remaining after each epidemic. This suggests that a few locations in the data have very long interepidemic periods with few outbreaks which are not sufficient in size to diminish the susceptible population. This is consistent with previous studies of measles dynamics in E&W [16, 20, 52].

Principal component results demonstrate no statistically significant difference between urban and rural areas. If space had been the primary driver of epidemic dynamics, we would expect the points in Figure 1.3 C and E to follow the identity line. If pronounced differences existed between urban and rural locations, we would not expect pairs to look like each other or be well predicted by population size alone. The projection of the pairs on the first two components appears to be well determined by population size (B, D) and the distance between each pair and its neighbor is well determined by the percent difference in their populations. This would imply that spatial location has a marginal impact on epidemic dynamics and population size is a stronger driver of dynamics than urban or rural status.

When comparing adjacent districts we see district neighbors do not resemble each other (Figure 1.3 C, E) and that population is the main driver of differences both in terms of raw projections (Figure 1.3 B, D) as well as in determining the difference between urban and rural pairs (Figure 1.3 A). We classify the difference between pairs as the euclidean distance between the district’s scores on the first and second principal components. We see that this difference is well-explained by the percent
Figure 1.2: **Principal components analysis on urban and rural district pairs.**
(A) shows the selected urban and rural districts, color coded for urban/rural status with each pair joined in a box. We obtain pairs in a variety of spatial locations. (B) Projection on the first two principal components demonstrates some difference in the urban and rural districts. Though the scores follow the same general trend, rural areas have slightly less variation in their scores as shown by the 95% confidence ellipses. (C) The loadings of the first and second principal components show the parameters plotted at their scores on PC 1 (x-axis) and PC 2 (y-axis). The first principal component separates the districts based on one axis which correlates positively with epidemic coupling and negatively with fadeout length and seasonality in beta; the second component provides an axis with larger numbers of fadeouts weeks and high average transmission rates at one end, and low transmission rates and low transmission at the other. A full table of scores is included in the appendix.
Figure 1.3: Untangling space and size in urban and rural districts. Panel (A) shows the euclidean distance between pairs when projected on the first and second principal components as a function of the percent difference in population size. Panels (B) and (D) show the relationship between population size and scores on each of the components. Panels (C) and (E) plot urban and rural district pairs against each other on the basis of their scores on the first (C) and second (E) principal components. Scores on or near the plotted line of identity indicate matching scores for the pairs. The score of an urban district has little to no relationship with the score of its rural counterpart in general (C,E). However, as shown by (A) we see when pairs are of comparable size, they tend to have similar scores.
difference in population size. This suggests size, rather than space is the predominant
driver of variation in disease dynamics.

Though urban and rural areas may not differ systematically, the analysis to this
point does not describe how neighboring towns and cities interact with each other.
Although pairs do not demonstrate coherence in their scores on PC1 and PC2, we
know infections move through space and expect to see some evidence of epidemic
interactions between neighbors. In order to investigate these district pairs in more
detail, we examine the case data directly.

1.5.3 Urban and Rural District Pairs: Epidemic Exchange

Projections on the first component demonstrate that urban areas may have longer
fadeouts and lower estimates of coupling than rural neighbors. Since uncertainty
around estimates of coupling can be large, particularly for larger districts, we exam-
ine the differences between urban and rural epidemics directly from the case data.
Consistent with the dynamics suggested by the results of PCA, we find that rural
areas fade out less and for shorter periods of time, resulting in more frequent, smaller
epidemics (Figure 1.4 C, D). In contrast, urban areas are characterized by more reg-
ular short epidemics (Figure 1.4 C, D). Urban and rural areas do not differ in the
number of large (final number of infections greater than the mean) outbreaks, which
increases consistently with population size for both district types (Appendix Figure
A.2 B).

We additionally find that larger districts tend to lead the epidemics of their smaller
neighbors. The relationship between the difference in population and the proportion
of epidemics lead in each location is strong (Figure 1.4 A). Larger places appear to
act as a importer of cases to their smaller neighbors, mirroring national patterns of
epidemic spillover at a small scale. An example of such a pair can be seen in Figure
1.4 B. This suggests transmission cascades from larger places to smaller places in
concurrency with previous findings, but replicated at a local scale. We do find a nominal, though significant difference between urban and rural areas particularly at smaller sizes. Urban areas have fewer, larger outbreaks when compared to rural areas of comparable size (Figure 1.4 C, D). This difference is slight though statistically significant if we look exclusively at small areas. Large urban and rural areas do not demonstrate a statistically significant difference. This indicates small urban populations may be nominally more well-mixed than small rural populations. Though this may seem intuitive and obvious it is important to keep in mind the crucial role of schools in measles dynamics. The population mixing rates relevant to this system are those of school-age children. These results therefore indicate that urban school-children may be better mixed, with more cross-school mixing, than rural counterparts. Alternatively urban areas may have fewer schools compared with rural counterparts, creating more concentrated contagion hot spots relative to rural districts.

1.6 Discussion

Understanding how transmission may vary between rural (or sparsely populated) and urban (or densely populated) areas is a critical area of research in a rapidly urbanizing world. The United Nations predicts that nearly 70% of the global population will live in urban areas by 2050. Though previous analysis on this unique detailed data set has suggested measles transmission is size and density-independent with a strong seasonality in transmission and signature of contagion movement between locations, the urgency of contemporary changes necessitates a more complete understanding of potential differences across settings. Previous analyses have been limited to urban areas, including rural areas provides a more complete understanding of metapopulation dynamics and variation across space and urban/rural district type. The complete
Figure 1.4: Epidemic interactions among neighboring urban and rural districts. (A) Large places consistently lead epidemics more often. When rural areas are larger, they lead their urban outbreaks. Likewise when urban places are larger they lead rural outbreaks. This confirms previous findings that large areas provide over to proximate smaller areas, this highlights the cascading effect of epidemics at a local level. (B) Guilford provides an example of this subtle dynamics wherein an urban area has fewer small outbreaks and larger major epidemics even when the neighbors are of similar sizes. The dotted lines indicate the average final outbreak size (134 total average infections for the urban district and 115 for the rural). In the case of Guilford, both the urban and rural district have 57 outbreaks, the rural district has 13 larger outbreaks while the urban district has 11. These 13 larger rural outbreaks are smaller in terms of final size than the 11 outbreaks in the urban district. (C) Rural areas have more outbreaks than urban areas of comparable size. (D) Rural areas tend to have smaller outbreaks on average though the standard errors are large.
and rich nature of this data set make it uniquely suited to be an initial case study for such investigations.

This analysis shows that infectious dynamics are not uniform across locations. However while we find an inverse relationship between fadeouts and coupling, and between growth rates and seasonality, urban and rural locations follow the same pattern in spite of potential structural differences. Although we see a slight difference on average between urban and rural areas when controlling for location and population, the overall patterns are consistent. Population size is the most significant driver of epidemic dynamics (though total number of births is a comparable predictor and highly correlated with population size). Additionally, while location does appear to impact dynamics, the similarity by pairs is not what we would expect if differences were entirely spatial (Figure 1.3). The difference in population sizes appears to explain many of the differences we observe. These findings are generally consistent with Ferrari et al’s results for Nigerian measles epidemics. Ferrari et al find a rural/urban gradient characterized by reduced seasonal amplitude in sparsely populated settings as well as climatically driven contact rates. Cross-location contact rates are even more sporadic in the Nigerian context, indicating that much of the consistency between urban/rural locations in E&W is likely driven by a consistent seasonal forcing mechanism (school calendar) as well as more frequent cross-location mixing.

In addition to confirming similarity between urban and rural areas, principal components analysis shows an important difference in large (above 10,000) and small (below 10,000) populations. Larger places can be characterized by more frequent epidemics with a typical seasonal signature, while small places are characterized by stochastic epidemics which are slower and do not deplete susceptible populations. This confirms that small places inherit epidemics as spillover from bigger neighbors (Figure 1.3). When we investigate epidemic interactions between urban and rural
areas we find size is the most important when determining which location will kick off a local epidemic (Figure 1.4).

Therefore we see large metapopulation dynamics mirrored in these urban and rural pairs. The larger member of each pair seems to serve as an epidemic feeder for its smaller neighbor. In this case it does appear to be size which drives the influx of cases rather than space or urban/rural status. The differences are most profound when population sizes are substantially divergent (on the order of 100-300%). When population sizes are similar, there is not a clear epidemic leader (Figure 1.4).

Even when controlling for population size the epidemic behavior between urban and rural areas are characterized by a subtle, statistically significant difference (Appendix Figure A.3). While we expect pairs to share epidemics as a result of spatial proximity, rural areas have more frequent, smaller outbreaks while urban neighbors have relatively fewer, larger epidemics. This explains the slight difference in scores on the first principal component (Figure 1.4). Urban areas have fewer, larger outbreaks followed by longer waiting time until a subsequent epidemics. Rural areas have slightly higher coupling estimates, either as a result of more frequent imports or as a result of internal rescue effects, the a reintroduction of infection from within the community. We illustrate this pattern by simulating a number of communities at a fixed total size and examine fadeout proportions and final sizes across a combination of patch numbers and mixing rates (Figure 1.5). We simulate a TSIR transmission model with multiple patches. We assume the population and birth rates within each patch are equal and the within-patch transmission rate is constant for an $R_0$ of 20. Across-patch transmission rate is fixed at 1%, 5%, 10%, 15%, 20%, 40%, and 80% of the within-patch rate. From the simulated incidence data we calculate the average number of fadeouts and the average final size of outbreaks. Increasing patches reduces the final size of each epidemic and reduces the the fadeout proportions. Additionally, increasing the across-patch mixing rate increases the final size and increases the
proportion of fadeouts. This illustrates epidemic behavior congruent with the slight
difference we observe between small urban and rural districts.

This could be due to heterogeneous contact patterns in rural areas. Rural districts,
particularly small ones, may have several transmission hot spots (schools) distributed
over a greater area. This could lead to variation in transmission rates within and
between hot spots, resulting in several small outbreaks rather than a single outbreak.
In contrast, urban areas, which we expect to be more dense for a given size, may have
multiple schools but stronger mixing across schools. In a comparatively dense setting
children may come into contact with one another at a greater rate outside of school.
So while a rural area may have several schools separated by a substantial distance,
urban schools may be closer together and children from different households may live
in closer quarters. In this case, when infection is introduced in an urban area, most
susceptibles will get it while in rural areas epidemics will move more slowly between
schools. This difference in built environment could shift the transmission pattern
of this highly contagious pathogen just enough to allow one location to have several
smaller outbreaks. This internal rescue effect will inflate estimates of coupling as each
outbreak is assumed to be the result of an import.

**Future Work**

School-level data is crucial for further disentangling differences between urban and
rural areas. Specifically school data could elucidate whether rural areas are receiving
more cases than urban areas or if multiple hot spots are producing multiple epidemics
from within the community. However, to our knowledge adequate data for this time
period does not exist in an appropriate scale to address this question. Furthermore,
even if we optimistically assume urban and rural designations are substantive in this
data, it is also true that urban and rural distinction in E&W is not comparable to
In panel A, each tile plot shows mean final sizes of outbreaks for a given mixing rate across patches ("across") and number of patches ("patches") for a single population size. Panel B shows the proportion fade out for each combination of mixing rates and patches. The across mixing rates were reduced such that the represented a fraction (given by the "across" axis of the internal mixing rate. Panel C shows example time series for a population of 100,000 with an across patching mixing rate of 20% of the within patch mixing rate. Cases are shown for sample simulations with two and eight patches. Within patch mixing rates were held constant to produce an $R_0$ of 20, a typical estimate for measles. Simulations show the number of outbreaks increases as the number of patches increases (Panel B) and the final size of the outbreaks decreases as the number of patches increases (Panel A). We also see as mixing rates between patches increase, the proportion fadeout biweeks increases and the final size increases. This supports the conclusion that rural areas may be characterized by patch mixing with relatively weak mixing between leading to more, smaller outbreaks when compared with urban counterparts.
global differences in urban and rural environments. That is, rural areas are denser than global rural extremes and urban areas are smaller and less dense than contemporary megacities. Even still, this exhaustive analysis of urban and rural disease dynamics provides a strong first examination of possible differences.

In the wake of contemporary measles outbreaks and declining vaccine coverage, comprehension of measles transmission has gained renewed urgency. Understanding transmission over metapopulation structures is vital for predicting outbreaks and planning interventions. Additionally, understanding the spread of disease over different population densities and mixing patterns is crucial in a rapidly urbanizing world. This analysis illustrates the cascading of disease transmission even at local levels, suggesting the larger of two populations is at greater risk of infection holding geographical location relatively constant. Further it suggests transmission may be slightly more rapid in larger areas but that persistence may be greater in smaller areas. The strength of transmission across locations highlights the potency of measles infection across scales. In addition, case data demonstrates that infections cascade from endemic areas to places of next-largest size, and that this pattern persists even at extremely local scales. In general this suggests the importance of targeting interventions in large population centers were disease outbreaks can grow to epidemic levels and instituting control strategies to prevent disease from traveling to subsequent locations. Finally, results on rural transmission highlight the importance of understanding local population mixing patterns and maintaining records on the number and spatial distribution of community hot spots.

Further research is necessary to build a comprehensive understanding of transmission in urban and rural areas. In particular, more detailed data on population densities within urban/rural areas as well as mixing patterns will be critical in untangling the pace and persistence of epidemics. In addition to highlighting slight differences between urban and rural districts within a metapopulation, this work
demonstrates the importance of the spatial scale of reporting for estimates of disease transmission. Aggregating several transmission zones into one reporting region may reduce estimates of contagion and over estimate import rates.
Chapter 2

Estimation Bias and Transmission Across Population Scales: a machine learning approach to correcting time-series SIR parameters

Understanding the relationship between local population size and transmission rate of infectious diseases is vital as global population continue to grow and concentrate in urban areas. Compartmental models have been invaluable in estimating transmission rates of infectious diseases. Using a semi-mechanistic compartmental model (the time-series susceptible-infected-recovered model) to estimate the transmission rate of measles in pre-vaccination England and Wales, there is evidence that transmission rates scale with population size. However, estimation bias also likely scales with population size due to the stochasticity of outbreaks in small, isolated locations. To test the bounds of the TSIR model, and isolate trends in bias from true trends in
transmission, we leverage simulations to compare estimated parameters with true parameters. We use machine learning to determine which types of time-series are appropriate for fitting with TSIR. This allows us to build a bias-correction algorithm for data. We find the TSIR model tends to underestimate transmission in small locations due to stochasticity and sparsity of data in these locations. We also present a set of guidelines for determining whether TSIR is likely to produce unbiased estimates as a function of time-series qualities.

2.1 Introduction

2.1.1 Infectious Disease Modeling

In 1766, Bernoulli published the first mathematical model detailing the spread of an infectious agent in a human population [98]. Since then, this work has been extended to include mechanistic models, host-population age structure, and models which can be calibrated against observational data. In particular, the mass action Susceptible-Infected-Recovered (SIR) family of models has proven its worth in illuminating disease dynamics in a number of contexts from short-term outbreaks to centuries of epidemic patterns [53, 74, 17, 41, 20, 80, 62]. In practice, the SIR model assumes that individuals can be classified into one of the three (S-I-R) states. SIR models have been immensely valuable in quantifying transmission rates and providing concrete thresholds for understanding epidemics. This class of models has been used recently to estimate the transmission rate of the emerging coronavirus (COVID-19) epidemic [114].

While the simplicity of the SIR model is readily for interpretable, calibrating the seasonality-forced SIR model against data can be statistically challenging [48, 59]. The main challenges of fitting the SIR model to data results from two sources: only one state variable is observed (the number of cases) and this is affected by unob-
served rates of under-reporting. These challenges then are further compounded by stochasticity in the underlying system or low numbers of incidence. A computationally efficient option for addressing these aforementioned challenges is the time-series SIR (TSIR) model. The TSIR model relies on two main assumptions: first, that the infectious period can be well-approximated by the sampling interval of the data (e.g., bi-weekly for measles) and that over a sufficient time period (e.g., 10-20 years), the sum of births and cases should be approximately equal if pathogens are sufficiently infectious. [53, 115].

2.1.2 Transmission and Population Size

Previous work with data from England and Wales (E&W) consisting of pre-vaccination weekly measles cases as well as annual births and population sizes has established the reliability of these estimates for large cities [53, 22, 48]. The unparalleled granularity contained in the full data set (22 years, 954 districts) has produced a wide range of results and insights into the epidemic dynamics of measles including metapopulation dynamics, spatial infection hazards, local transmission rates and school term seasonality [23, 48, 53].

A 2002 paper closely examines a subset of sixty diverse urban districts from the England and Wales data [22]. The authors demonstrate the predictive accuracy of the TSIR model for each district and show that the corresponding estimates of $R_0$ - the number of secondary infections expected as a result of a single infection - do not scale systematically with population size [22]. The authors assert this is a result of school-based transmission of measles. Since schools dominate transmission opportunities, it is the size of classrooms that matters more than the size of the local population. However, we might expect size to correlate with other features - such as population density - that may impact opportunities for contact among susceptible and infected individuals outside of school. Furthermore, areas with more schools in close proximity
may have more opportunities for children to interact outside the classroom, increasing transmission potential. For this reason, we extend the analysis conducted in the 2002 paper to all urban districts in the England and Wales data set.

2.1.3 Estimation Bias

Much of the early work with this data set focused on large urban districts with epidemics in a seasonal equilibrium [22, 48, 53]. These locations have been consistently well-described by TSIR methods but fitting becomes more challenging for small locations which are prone to long periods of local extinction punctuated by violent and stochastic outbreaks. While the challenges of fitting TSIR to smaller locations have been known to epidemiologists, the limitations of TSIR in this context have not been quantified.

As we investigate estimates of $R_0$ across population scales, the difficulty of obtaining reliable estimates is a paramount concern. Since we know bias is likely to correlate with population size, we want to disentangle this from any true variation in $R_0$. Our goal in this regard is to (1) identify the most influential features of a time series for producing bias in the TSIR estimates and (2) predict model bias accurately for a given time series. Machine learning has recently proliferated as a strategy for identifying influential variables among a plethora of options. Machine learning also has the capacity to drastically improve predictive accuracy relative to traditional regression methods [86, 73].

To assess the limitations of the TSIR model, we leveraged the power of high-volume stochastic simulation. We compare parameters estimated using the TSIR fitting process with the parameters used for the simulations. We then take eighteen measures on the incidence time-series to describe the stochasticity and persistence of the disease and compare these measures with the bias in our estimates. We use machine learning techniques to determine which features of the incidence data are
most influential in determining bias. We use these to generate guiding principles for identifying appropriate time-series for this fitting procedure. Furthermore, we extend our machine learning algorithm to generate a bias correction and apply this correction to the estimates we obtain from the well-studied E&W measles data set.

In addition to the E&W data, recently made publicly available, an open source program for fitting the TSIR model [18] presents an opportunity both for rich new results as well as generation of estimates without attention to the limitations of TSIR. In this paper we investigate the potential pitfalls of fitting this model without thinking critically about potential biases. We also present an approach to navigating these challenges that can be applied to a variety of research settings when data is sparse or the suitability of a model is unknown.

2.2 Materials and methods

In the following subsections, we describe our simulation and estimation framework, our in silico data set, and finally, an application of our methods to the uniquely rich spatio-temporal 1944-1966 England and Wales measles incidence data set.

2.2.1 Time-Series Susceptible-Infected-Recovered Model

The time-series Susceptible-Infected-Recovered (TSIR) model forms the backbone of both our simulation and fitting procedure. This well-studied and oft-implemented model balances susceptible recruitment (e.g., births) with susceptible depletion (e.g., infection) on a discrete time scale equivalent to the generation time of measles; here, assumed to be biweekly [53 22 18]. The susceptible dynamics can therefore be modeled as:

\[ S_{t+1} = B_t + S_t - I_{t+1} \]
where susceptibles at $t+1$ (i.e., $S_{t+1}$) are simply the previous susceptibles (i.e., $S_t$) plus births (i.e., $B_t$) minus any new infected individuals (i.e., $I_{t+1}$). The associated deterministic infected dynamics are:

$$E[I_{t+1}] = \beta_{t+1} \ast S_t \ast I_t^\alpha$$  \hspace{1cm} (2.2)

Where $\beta_t$ is the time-varying contact rate, while $\alpha$ is a tuning parameter often fixed (or inferred) to be slightly below unity. While each new infected cohort can be drawn from a range of distributions, we choose a Poisson distribution for our simulation framework.

**TSIR: simulation framework**

To explore how epidemic variability impacts parameter identifiability, we simulated multiple epidemics across 30 population scales ranging from 2,000 to four million, i.e., effectively the range observed in the E&W data set. In line with empirical estimates of measles transmission, we utilized an $R_0$ ranging between 15 and 25 (citations). Consistent with previous analyses, we explored the range of $\alpha$ between between 0.95 and 1. Additionally, we enforced a sinusoidal seasonality function which has been shown to qualitatively produce the same results as more complicated school-term forcing patterns ranging from cubic splines to on-off functions (earn citations). We utilized combinations of population sizes, $R_0$, and $\alpha$ and ran twenty simulations with each permutation. For simplicity, we assumed a constant birth rate $(=14$ per $1,000)$. Reporting rate was fixed to be approximately 50%, in line with most estimates from E&W during the pre-vaccination time period [22, 33]. For each time step we assumed $I_{t+1}$ follows a Poisson distribution with the expected value given by equation [2.2]. We allowed a burn-in period of 15 years before using 22 years of in silico data. In total, this provided us with 7,200 time-series, containing biweekly cases, population sizes, and births. These temporal scales were chosen so that our simulations are analogous.
in structure to the E&W data we will eventually apply our machine learning model to.

**TSIR: estimation framework**

The primary assumption of the TSIR model is that over a sufficient period of time (e.g., 10+ years) everyone born in the population acquires infection; therefore cumulative cases and cumulative births will be approximately equal. A regression model fit between these two cumulative counts will have a slope equal to one in a perfectly reporting population, while any deviation below unity provides a measure of underreporting. The residuals of such a regression, $Z_t$, capture the fluctuations in the susceptible population, $S_t$, albeit centered around zero. Estimating a mean number of susceptible individuals via profile likelihood allows us to reconstruct the susceptible population at each time step. With estimates of both the infected (reported cases multiplied by reporting rate) and susceptible dynamics, equation 2.2 can be log transformed into a linear model:

$$
\log[I_t + 1] = \log \beta_{t+1} + \log S_t + \alpha \log I_t
$$

(2.3)

Although we know the true value of $\alpha$ for each simulation, this is an unobserved quantity in the E&W data. To keep the fitting procedure, and thus potential biases, consistent, we fit all data (both in silico and empirical) with the well-documented London estimate of 0.97. The basic reproductive ratio, $R_0$, can be approximated by $\bar{\beta} \times N$ where is the mean population size and $\bar{\beta}$ is the average contact rate across the year. More details on the implementation of TSIR can be found in and in Appendix 2.
2.2.2 Measuring the Time-Series

We compiled eighteen measures from the incidence time-series. We selected prudent metrics to determine the key characteristics of epidemics such as seasonality, persistence, and random noise. These primarily consisted of four main categories: (1) long-term seasonality – i.e., the tendency to exhibit annual and biennial cycles, (2) short-term seasonality – i.e., local fluctuations in transmission throughout the year, (3) epidemic growth rates, and (4) persistence – e.g., the average number of cases and frequency of local disease extinction. Each category required slightly different methods for producing relevant measures. Signal processing techniques enable us to calculate the power of different frequencies in the time-series. In particular, Fourier transform allows us to tease apart long-term seasonality (annual and biennial outbreaks) and random noise (i.e., high frequency signals under one year). To amplify the seasonal signature in the data, incidence counts were scaled the log transformed (with the addition of one). To estimate the influence of each yearly period, we calculated the spectral density of the data. With this information we examined the annual power, the biennial power, and the power of noise (high frequency signals). With these estimates in hand, we calculated signal to noise ratios by dividing the annual or biennial power by the power of noise [48, 52, 109].

In contrast, case burden and persistence measure a different type of disease regularity. For example, in small populations stochastic extinction dominates, while in large urban areas the disease remains endemic [28, 23]. These patterns may have critical consequences for model inference – for example, locations experiencing frequently extinction events may simply not provide enough data for accurate parameter estimation. Indeed, high variation in the yearly incidence totals may bias estimates of seasonal transmission. To quantify this hypothesis, we computed the number of individual extinctions as well as the total proportion of biweeks without cases.
Finally, we hypothesize the epidemic growth rate to be an important factor in estimating transmissibility from case data. We first isolated the beginnings of epidemics in the case data. To estimate growth rates, we log transformed the incidence data (again, adding one to avoid zeros) and calculated the difference in log cases from one time step to the next approximated growth rates. When epidemics are long we examined growth rates across the entire epidemic as well as in the first biweeks. We tabulated mean growth rates along with minimums, maximums and the standard deviations. A full list of these measures and the methods used to calculate them can be found in Appendix B. We cross referenced our measures of the simulations and measures of the E&W data to verify that we were effectively capturing the range of values. We found that the simulations covered the values we obtained from the measles surveillance data and additionally that trends of the measures with population size were consistent between simulations and the data. A full document detailing the coverage of these measures can be found in Appendix B (Figures B.3, B.4, B.5).

### 2.2.3 Machine Learning and Bias Prediction

Once we have used TSIR to estimate transmission, we fit a random forest regression model to examine biases in parameter estimates [73]. Specifically, we aim to (1) determine which features of the time-series are most influential for bias in estimates; (2) investigate thresholds for these features which can be used to guide future research; and (3) to predict bias and correct for it in both our simulations and our E&W data.

We trained the algorithm so it takes our time-series measures as inputs and uses them to predict bias in $R_0$. To be consistent with machine learning research we will refer to our time series measures as ”features.” When we discuss ”feature importance” we are simply describing which of the time-series measures are the most influential for the model. We trained the random forest with 60% of our data and tested it on the remaining 40%. We refer to these subsets as our training set and our test
Figure 2.1: **Illustration of random forest.** A random forest is an ensemble model, meaning it uses a combination of many machine learning models to produce one output. In this case it is an ensemble of decision trees. A tree uses values of the features to create nodes, or decision points, these push the observation one of two ways. After a series of decision points the observation is assigned a predicted value. Each tree produces its own prediction and the predictions are averaged (in the regression case) or a majority vote is taken (in the classification case).

set respectively. The training set contains all the features (time-series measures) as well as the outcome ($R_0$ bias) and the random forest uses this information to build a predictive model. We input the features from the test set to the algorithm and compare the predicted bias with the observed bias.

The random forest regression model consists of many (in our case 60) individual regression trees which all predict their own value for bias (Figure 2.1). These predictions are averaged to obtain a final predicted value. Random forests are advantageous to individual decision trees because they are less likely to be prone to over-fitting the training data set. In other words, the predictions should be more generalizable to diverse data sets [73].
An advantage of the random forest model is it provides measures of feature importance. Conceptually, both are measures of how much a feature improves the predictive power of the model. Random forests utilize bagging for the fitting process - each tree is grown on a subset of the data and the withheld portion can be used for cross-validation and measuring variable importance. The random forest uses two measures of variable importance: percent increase in out-of-bag (OOB) mean squared error (MSE) and increase in node purity. The prediction error (MSE) for the OOB sample is recorded for each tree. Each predictor variable is randomly permuted and the MSE is calculated again. The difference between the two is averaged over all trees and normalized by the standard deviation of the differences. This illustrates model performance, measured as out-of-sample prediction, as a result of including each variable in the tree [73].

Increase in node purity measures the reduction in the impurity (residual sum of squares) when that variable is chosen to split a branch in a tree. The goal of splitting is to be able to group similar outcomes together at the end of the tree. In the binary case, a pure node would be one for which every result down the tree is all 1s or all 0s. For a continuous case it is a measure of the variance in the branches below. Impurity is calculated before and after the node which uses that variable to split and averaged over all trees. This measures the improvement in prediction as a result of including that node as a decision point [73].

2.2.4 England and Wales Data

Ultimately, the aim of our simulation study is to determine drivers of biases in data sets that span populations and epidemic patterns. To estimate bias in real surveillance data, we apply our algorithm to E&W data to estimate potential bias. This unique data set contains weekly case reports from 954 urban districts in England and Wales between 1944 and 1966, along with annual births and populations. Due to
an unparalleled scope and scale (954 towns for 22 years), the data is exceptional for studying the population dynamics of childhood diseases.

2.3 Results

In the following subsections we present the results for our in silico data, the observed bias in estimates of $R_0$, and the random forest results including feature importance and bias correction. Ultimately we present raw and corrected estimates for the E&W data set.

2.3.1 In Silico Data

Utilizing the TSIR model framework, we produced 7,200 time-series with eighteen time-series measures. We used these 18 measures as features in our machine learning algorithm. We see that the measures correlate nicely with the measures we calculate for the E&W data (Appendix B).

2.3.2 Biases: In Silico Data

The initial bias in estimates from the in silico data is low on average. The mean bias is (.09) with 50% of the data having an absolute bias in $R_0$ of 2 or less, and 85% with a bias of 5 or less. Although there are a few notable outliers with estimation biases of 20 or more, only .02% of our estimates fall into this bracket.

2.3.3 Random Forest: Feature Importance

The random forest results indicate that the proportion fadeout, annual signal to noise ratio, annual variation in case counts, seasonal variation are among the most important features for determining bias in estimates of $R_0$. Figure 2.2 shows the top eleven features selected for importance. We selected eleven because for both MSE
and Node Purity the drop off between feature ten and eleven was minimal but the drop off between feature eleven and twelve was substantial. Figure 2.2 also shows the relationship between bias in $R_0$ and three of the top features by both measures of importance.

The average absolute bias increases as the disease fades out more (Figure 2.2 Panel C). This is consistent with what we know about the challenges of fitting the TSIR model in locations that are small and vulnerable to occasional but violent outbreaks. Though not linear, we see the average bias decreases as the annual to noise ratio increases, indicating that locations with higher annual power relative to the presence of noise in the signal produce less biased estimates (Figure 2.2 Panel D). Seasonality R-squared measures how well the biweekly cases approximate a sine curve when tallied over the 22 years of the time series. We see that the bias is lower on average when cases occur in a sinusoidal shape (Figure 2.2 Panel E). Since transmission is simulated with a sine-determined seasonality we expect locations with low r-squared have flat or erratic cases driven by frequent extinction and stochasticity. Together these trends indicate that time-series with more regularity and more case observations produce better estimates. In contrast, noisy time-series and time-series with few observations are more likely to produce biased estimates when approached with the TSIR procedure. In each case, we see the proportion of observations that fall into the high bias tails are relatively small (Figure 2.2 Panels C-E).

### 2.3.4 Random Forest: Bias Correction

We use predicted values of bias in $R_0$ to correct parameter estimates in our test set. Figure 2.3 compares original estimates with bias-corrected estimates based on the output of our random forest model. The variance in the predicted values indicates the possibility of over-fitting. Indeed, a number of our features are highly correlated. We also attempted the model after first eliminating some highly correlated features. This
Figure 2.2: **Feature importance and bias.** Panels A and B show the variables with the most influence in the random forest model. Panels C-E show the relationship between three of the most influential features and bias in the estimates. Panel C shows that bias increases as fadeouts increase, the mean bias increases as the disease goes extinct with greater frequency. Panel D shows that the bias generally decreases as the annual to noise ratio increases, though the relationship is nonlinear. Panel E shows that bias decreases when locations have higher seasonal $R^2$. This means locations which annual accumulations of cases that are well-approximated by a sine curve are less likely to be biased compared to those with flat or erratic cases by biweek.

had no substantive effect on the predictions or the variance explained by the algorithm so we opted to show results using the full set of features. Across population sizes and transmission rates, we are able to reduce bias. We see a consistent improvement in the average estimate across the time-series, as well as reduction in the variance of estimates. That is, we see an improvement in both bias and precision (Table 2.1).
Figure 2.3: Bias and corrections in simulations. This figure demonstrates the improvement in estimates after correcting for predicted bias. Panel A shows the bias in $R_0$ across population sizes. We see a tendency to underestimate transmission in small population sizes, and a tendency to overestimate in large populations. Panels B - D show the original estimates and their corrected values from the test set. We see an improvement in estimates across $R_0$ values and population sizes.

<table>
<thead>
<tr>
<th>$R_0$ value</th>
<th>original mean</th>
<th>original variance</th>
<th>adjusted mean</th>
<th>adjusted variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_0 = 15$</td>
<td>15.68</td>
<td>11.21</td>
<td>15.59</td>
<td>3.51</td>
</tr>
<tr>
<td>$R_0 = 20$</td>
<td>20.22</td>
<td>10.53</td>
<td>20.10</td>
<td>2.50</td>
</tr>
<tr>
<td>$R_0 = 25$</td>
<td>24.01</td>
<td>18.65</td>
<td>24.33</td>
<td>5.02</td>
</tr>
</tbody>
</table>

Table 2.1: Descriptive Statistics of Pre- and Post-Correction Estimates. The distribution of original and corrected estimates for $R_0$ from the simulated data.

2.3.5 Application to E&W Data

Figure 2.4 shows the predicted bias by population as well as the adjusted results. The range of adjustment varies from $-4$ to $8$ with an average predicted bias of $-0.5$ and a median of $-1.7$. There is a trend in the correction with population size, however we still see variance in the correction at each population size indicating the influence of other features of the data. Figure 2.4 also demonstrates the importance of
Figure 2.4: **TSIR estimates and adjusted estimates from E&W data.** Uncorrected and corrected estimates for $R_0$ for the 954 urban districts in the E&W data. Panel A shows original estimates for each of the 954 locations. Panel B shows the estimated bias in each estimate by population size. Panel C shows the estimates after correction from our random forest model. Panel D maps the estimated bias by latitude and longitude of each district with large, endemic locations plotted as triangles. While population size is a strong predictor of the estimated bias in $R_0$ we see that locations near large endemic populations require less correction. This is because proximity large cities provides more opportunity for epidemic spillover and thus more epidemics.

proximity to large epidemic centers. Small, isolated areas have the largest predictions of estimation bias whereas small locations near large population centers have lower predicted bias. This is likely due to the increased regularity of epidemics in these locations as a result of spillovers from neighboring cities. This increased regularity affords better data quality as a result of more frequent epidemics.

TSIR estimates and adjusted estimates from E&W data. Uncorrected and corrected estimates for $R_0$ for the 954 urban districts in the E&W data. Panel A shows original estimates for each of the 954 locations. Panel B shows the estimated bias in
each estimate by population size. Panel C shows the estimates after correction from our random forest model. Panel D maps the estimated bias by latitude and longitude of each district with large, endemic locations plotted as triangles. While population size is a strong predictor of the estimated bias in $R_0$ we see that locations near large endemic populations require less correction. This is because proximity large cities provides more opportunity for epidemic spillover and thus more consistent epidemics.

### 2.4 Discussion

The analysis presented in this paper suggests there is variation in $R_0$ that trends with population size. In particular, we see a weak relationship where $R_0$ increases with population size. We see the largest estimates of $R_0$ in mid-sized populations (between 10,000 - 100,000). One reason for this relationship may be that populations of interest approach well-mixed in these locations. Whereas small locations may also be sparsely populated or be limited in transmission by their size. Large locations may be an agglomeration of several smaller well-mixed areas. Diseases move more slowly through these kinds of heterogeneously mixed locations than they would through well-mixed locations of comparable sizes. [83, 39, 15, 84]

These results demonstrate that TSIR is most reliable when epidemic dynamics are at endemic equilibrium rather than violent and stochastic. However, we also show that it is possible to produce unbiased estimates even in locations with small populations and irregular outbreaks. The random forest was most useful in reducing variance in the estimates of $R_0$ across the time series. Though the geographic and temporal context of the data is specific, the results provide generalizable insights to measles dynamics across a variety of settings. In particular, our approach to estimating and correcting for process bias is applicable across contexts.
The output of the random forest indicates that the annual signal to noise ratio, the proportion fadeout, annual variation in cases, and seasonal case variation are the most important features. Qualitatively this indicates that regular, endemic time series are best suited for the TSIR procedure. Quantitatively, we see that time-series are best fit by TSIR when the annual to noise ratio is greater than 100 (the annual power is 100 times that of noise), the proportion fadeout is less than half (in this case this corresponds to 10 years or more of data), annual variation in cases is less than 1 (the standard deviation of yearly cases is no larger than the average annual number of cases) and the seasonal variation is less than 1 (the standard deviation of biweekly cases is no larger than the mean number of biweekly cases). We suggest these guidelines if researchers desire to avoid bias altogether.

In cases when time-series data is not ideally suited for the TSIR method, the trends in bias demonstrated here present some guidelines on the direction and magnitude of bias that can be expected and what adjustments or caveats may be prudent. It is the case, for example, that when using TSIR to estimate transmission for small locations, the value produced should be used as a lower bound for the true transmission rate rather than a point estimate. With regard to public health, understanding these potential biases, particularly the downward biases in small populations could have a significant impact on our understanding of herd immunity. Selecting vaccination thresholds which take into account estimation bias will better protect populations from outbreaks.

2.5 Conclusion

This paper presents an approach to testing and correcting for bias in estimates generated from disease incidence time-series. We find that initial estimates are fairly accurate on average though individual time series may produce large biases when
using the TSIR method to estimate transmission. We demonstrate that bias in parameter estimates may be low on average in the E&W dataset (around 3 for $R_0$) and that TSIR tends to underestimate transmission rates in small (below 10,000) populations and overestimate transmission in larger populations. We also present compelling findings for the robustness of a population-driven scaling of transmission rates as well as confirming that estimation bias correlates with population size. These trends along with the feature-specific thresholds we examine, provide a first order set of recommendations for whether a time-series is well suited for the TSIR estimation procedure.

Our guidelines indicate that TSIR is best suited for time-series which demonstrate more regularity than noise and for which a decade of data is observed. We show that having fewer observed cases tends to result in a downward bias in the estimate of transmission. While these results suggest other methods may be better suited for highly stochastic time series, we emphasize that for its computational efficiency, the TSIR method provides largely accurate estimates. While we are encouraged by the ability of TSIR to estimate transmission rates on average, we encourage researchers to carefully consider the suitability of this method to each project and to keep in mind some of the trends in bias we have presented here. The machine learning findings presented in this paper are most applicable to the E&W data, as we used this data to guide our simulation parameters. Even still, we have presented a method for assessing potential bias which can be easily adjusted for other time-series.
Chapter 3

Tensor Decomposition for Disease Incidence Data

3.1 Abstract

Many demographic and ecological processes generate seasonal and other periodicities. Seasonality in infectious disease transmission can result from climatic forces such as temperature and humidity; variation in contact rates as a result of migration or school calendar; or temporary surges in birth rates. Seasonal drivers of acute immunizing infections can also drive longer-term fluctuations. Tensor decomposition has been used in many disciplines to uncover dominant trends in multi-dimensional data. We introduce tensors as a novel method for decomposing oscillatory infectious disease time series. We illustrate the reliability of the method by applying it to simulated data. We then present decompositions of measles data from England and Wales. This paper leverages simulations as well as much-studied data to illustrate the power of tensor decomposition to uncover dominant epidemic signals as well as variation in space and time. We then use tensor decomposition to uncover new findings and demonstrate the potential power of the method for disease incidence
data. In particular, we are able to distinguish between annual and biennial signals en masse and shifts in these signals over time. Tensor decomposition is able to isolate variation in disease seasonality as a result of variation in demographic rates. the method allows us to discern variation in the strength of such signals by space and population size. Tensors provide an opportunity for a concise approach to uncovering heterogeneity in disease transmission across space and time in large data sets.

### 3.2 Introduction

Seasonal and other oscillatory periodicities are widespread and important phenomena in ecosystem dynamics, climate science, health data, economic trends and many other important contexts. Seasonality is particularly critical in the study of infectious disease. Acute immunizing infections can manifest strong seasonal incidence as a result of the dynamic balance between susceptible recruitment (typically births) and herd immunity. This seasonal "forcing" can drive biennial or more exotic (multi-annual) epidemics. Seasonality could arise from climatic variation or periodic changes in crowding (e.g. school calendars) [42, 45, 53]. Though births often demonstrate strong seasonal cycles, fluctuations in births only impact seasonality of acute immunizing infections in the absence of other forcing mechanisms [42].

In the case of measles in England and Wales (E&W) the predominant force is transmission via contact in schools [42, 53, 48]. Largely deterministic measles dynamics occur in large populations where susceptible replenishment is substantial enough that infections will wane but never disappear altogether; the threshold for this epidemic equilibrium is approximately 300,000 in Europe and North America [42, 53, 16]. In small populations the pathogen will go extinct until it is reintroduced via imports; such an import will trigger a large epidemic. While large locations have outbreaks at
the same time every year or every other year, small towns experience more stochastic, violent outbreaks [16, 52, 25, 20].

Wavelet analysis is used frequently to explore nonstationary cyclicality and heterogeneity in ecological time series [21, 52, 54]. For measles in particular, wavelets have been used to determine seasonal and longer period outbreaks as well as spatial variation in the lag between epidemics [52]. Local dynamics of measles transmission in E&W are generally composed of seasonal (driven by the school term) and long-term cycles usually annual or biennial driven by susceptible replenishment. The strength of these longer term cycles can vary between places and across time.

Wavelets provide very detailed data on seasonal signals of individual localities; in this paper we present tensor decomposition as a method for characterizing the wavelet spectra of many places at once. Tensor decomposition is a multi-dimensional generalization of matrix decomposition methods such as Principal Components Analysis (PCA) or Singular Value Decomposition (SVD). Analogously to these matrix methods, tensor decomposition reduces the dimensionality of the data by providing lower-dimensional components which describe much of the variance in the data [65, 97, 90, 32].

Tensor decomposition has been used successfully in a number of fields to uncover trends in large, multi-dimensional data sets. The method has been used in neuroscience, [37, 69, 108, 38] text analysis, [4, 61, 116] and photogrammetry [57]. In neuroscience, tensors have been a useful method for feature extraction and pattern detection in electroencephalography (EEG) signals. Tensor decomposition has successfully isolated task-related brain activity from the mixture of unrelated brain activity, interference and noise [37, 108, 69, 38]. The method can also extract differences in EEG signals among individual subjects, experimental conditions, or tasks [108, 38]. These studies use data in some variation of a channel-time-subject format. Channel may be color or sound frequency and a subject may be an individual or a
task. Fields such as neurology and computer science have utilized tensors in processing signal data in a myriad of settings meanwhile applications to ecological data lag behind.

We validate the value of the method for disease incidence data by decomposing simulated epidemic data; we simulate epidemics under three different birth regimes and demonstrate the method’s ability to identify these differences. We then apply the method to the well-studied E&W measles data to uncover previously undocumented variation in measles seasonality.

3.3 Materials and Methods

3.3.1 Simulations

We simulate epidemics using a discrete-time stochastic susceptible-infected-recovered (SIR) model\cite{98, 23, 28, 20, 18}. At each time step, each individual is assumed to be either susceptible to infection, infected or recovered (dead or immune). Once an individual recovers we assume they can never be infected again. We simplify the simulation process by assuming that any individual infected at $t$ will be recovered at $t + 1$. New susceptibles are supplied by births determined by a pre-defined annual CBR which we distribute uniformly across each time step. Imported cases are determined by drawing from a binomial distribution with a 10% probability of importation. We use a starting population of 300,000 for all simulations, and initial infected population of 10 and initial susceptible population of 1000. The susceptible dynamics are determined by:

$$S_t = S_{t-1} - I_t + B_{t-1}$$ \hspace{1cm} (3.1)
We add births ($B_{t-1}$) and subtract new infections ($I_t$). The expected number of infected individuals at time $t$ is defined as a function of transmission rate ($\beta$), local susceptible ($S$) and infected individuals ($I$) as well as imported infections ($\iota$):

$$\lambda_t = \beta \ast S_{t-1} \ast (I_{t-1} + \iota_t)$$ \hspace{1cm} (3.2)

In equation 3.2, $\alpha$ is a tuning parameter, fixed to .97, consistent with previous analyses and simulations of measles [53, 17, 18]. We give transmission a seasonal shape consistent with what has been estimated for London, [53, 23, 18, 22] with an average number of secondary infections at 15 for all simulations, within the range of typical estimates for measles, [22, 53, 56]. We draw the number of infections using a poisson distribution to introduce stochasticity: [18, 23]:

$$I_t \sim \text{Poisson}(\lambda_t)$$ \hspace{1cm} (3.3)

We allow approximately 80 years for the epidemics to settle into equilibrium and evaluate the following 20 years so the scale is comparable to the 22 years of E&W data.

To alter the dynamics of each epidemic we vary the crude birth rates (CBR) used in each simulation. We use three different birth regimes: a constant CBR of .015, a constant CBR of .03, and an variable CBR which begins at .012 and increases to .036. The simulated birth rates cover the range of birth rates in the data we use; the tenth percentile of CBRs in E&W during this period is .012, .015 is approximately the median, and the max is .035. We collect 50 time series for each birth regime. A sample of the time series under different birth regimes is shown in Figure 3.4.
3.3.2 The Data: England & Wales Urban Districts 1944-1966

For our primary analysis, we fit wavelets to all 954 urban districts in England and Wales for the pre-vaccination period (1944-1966) [52, 22, 20, 28]. We used tensor decomposition to decompose the data into the time and period components which account for the most variation in the national epidemic dynamics. Sample case time-series is shown in Figure 3.4.

Post decomposition, we used annual births and population sizes to calculate the crude birth rate (CBR) to evaluate the relationship between time-frequency trends and demographic conditions. In particular, we considered the CBR for locations above the critical community size for measles [48]. The spatial influence of large cities on regional dynamics is substantial [52], and therefore we limited our analysis to larger locations as we have more confidence these locations are determining their own dynamics rather than echoing the dynamics of endemic neighbors [52, 16]. The post-war baby boom resulted both in a surge of birth rates as well as a large range in birth rates across locations, we examine these rates in particular.

To validate the decomposition for such a large data set, we reconstructed the time series for four diverse large places (Supplement Figure C.3). An illustration of the reconstruction for London (Figure 3.7) and Norwich (Figure 3.8) shows how each component contributes different frequency-time power. These reconstructions demonstrated the ability of a few components to explain a large amount of time-frequency variation.

3.3.3 Continuous Wavelet Transform

To ground our analysis in previous studies, we performed a local wavelet analysis to the log transformed data to assess the time-frequency variation in the signal [52]. Generalizing Fourier analysis, wavelets allow insight into a potentially non-stationary epidemic signal by decomposing it into multiple frequencies over time [107, 49]. Like a
Fourier Transform, wavelets decompose a complex signal into its component frequencies. However, in addition to learning which signals dominate the signal, wavelets allow us to determine whether and how those dominant frequencies change over time. Rather than using a sinusoid as in Fourier analysis, wavelet transformation uses wavelet basis functions which can explore local (in time) variations in frequency. In this analysis we used the Morlet wavelet function, essentially a damped complex exponential:

$$\Psi_0(\eta) = \pi^{-1/4} e^{-i\omega_0 \eta} e^{(-\eta^2/2)}$$ (3.4)

In equation 3.4, $\omega_0$ is the nondimensional frequency. For a discrete sequence $x_n$, the continuous wavelet transform (CWT) is defined as the convolution of $x_n$ with a scaled and translated version of $\Psi_0(\eta)$:

$$W_n(s) = \sum_{n'=0}^{N-1} x_{n'}\Psi^* \left[ \frac{(n' - n)\delta t}{s} \right]$$ (3.5)

In Equation 3.5 (*) indicates the complex conjugate. By varying the wavelet scale $s$ and translating along the localized time index $n$, one can show both the amplitude of any features versus the scale and how this amplitude changes over time. To approximate the CWT the convolution should be done $N$ times for each scale, where $N$ is the number of points in the time series.

An illustration of CWT power spectra can be seen in Figure 3.1. For each location (and simulation), we produce a matrix where the column indices represent time steps, and the row indices represent frequencies. The $(i,j)$-th element in the matrix represents the power of the $i^{th}$ frequency at the $j^{th}$ time step. We then assemble these matrices into a cube by stacking them as in Figure 3.2.
Figure 3.1: **Wavelet power spectra for London (1944-1994).** Darker blues indicate more power, red areas indicate time/frequencies that were determined to be statistically significant by Monte Carlo methods. The parabolic line indicates the cone of influence, the boundary of points that may be affected by edge effect artifacts. Similar to spectral analysis, errors will occur at the beginning or the end of time series. Padding the data with zeros introduces discontinuities into the data, as we increase in scale, the amplitude is decreased as more zeros enter the analysis. For the regions outside the cone of influence, it is not clear if decreases in variance are due to the additional zeros.

### 3.3.4 Tensor Decomposition

Once we have compiled our 3-dimensional data we can decompose it into vector components. Tensor decomposition can be understood as a multi-dimensional generalization of Principal Components Analysis (PCA) [65, 43, 32]. As with PCA, we seek to reduce the dimensionality of the data by expressing it in terms of components which capture the most variance in the data. In the CWT case, each component consists of location vector, a frequency vector and a time vector. The outer product of the frequency and time vector produce a general wavelet power spectrum as in Figure 3.1. For the $i^{th}$ location, the $i^{th}$ scalar in the location vector describes the amount the
After calculating the wavelet power spectra, we store each of the time × frequency matrices in a three-dimensional tensor with one matrix for each location. This creates a cube of dimensions \( n \times f \times t \) for \( n \) places, \( t \) time steps, and \( f \) frequencies.

Power spectrum contributes to that location’s original signal. In other words, each component describes a particular frequency and its power as a function of time. The location-specific scalars represent how much that signal is magnified or dampened within that location’s data. To reconstruct the original data for a specific location, we compute and add such a matrix for each component \([65, 57, 32]\). If our tensor decomposition had three components and we wanted to reconstruct an estimate of our original data we would calculate and sum three wavelet power spectra using the three time and frequency vectors along with the location-specific score (Figure 3.3).

To calculate the tensor decomposition, we use the Canonical Polyadic Decomposition (CPD). We can formalize CPD for a three-way tensor as follows:

\[
\min_{\hat{X}} \| X - \hat{X} \| \quad \text{where} \quad \hat{X} = \sum_{r=1}^{R} a_r \otimes b_r \otimes c_r = [A, B, C]
\]  

(3.6)
Figure 3.3: Theoretical Canonical Polyadic Decomposition for tensor $X$ from Figure 3.2. The outer product of each of the $f_i \times t_i$ vectors produces a wavelet power spectra and each of the $n$ places has a score in the $n_i$ vector specific to that wavelet power spectra. In this way we can use the three rank one tensors to approximate the original power spectra for each of the $n$ original places. We can think of the matrix formed by the outer product of $t_i \times f_i$ as a component in the principal component analysis sense. Where each entry in the matrix determines the loading of the that time, frequency value. Each entry in the $n_i$ vector represents a place-specific score which determines the contribution of that matrix in the final wavelet power spectrum.

In Equation 3.6, $R$ denotes the rank of the tensor [65]. This definition is illustrated in Figure 3.3 for a rank-three tensor. We use an alternating least squares algorithm to calculate $\hat{X}$. In the case when $\min_{\hat{X}} \| \hat{X} - X \| = 0$, $\hat{X}$ is referred to as an exact low-rank approximation of $X$. In that case, we can write out the matrix form of $\hat{X}$ as:

$$
\hat{X}(1) = (C \otimes B)A^T \\
\hat{X}(2) = (C \otimes A)B^T \\
\hat{X}(3) = (B \otimes A)C^T
$$

(3.7)

Each $\hat{X}(i)$ is a component, and each $a_r, b_r, c_r$ are factor vectors, and $A, B, C$ are factor matrices. These are analogous to components and loadings in PCA [32, 65, 43]. To estimate these components, we fix all except one of the factor matrices and optimize the remaining matrix. For example, we may fix matrices $B$ and $C$ and
optimize $A$ given these matrices. We repeat this for each matrix until we reach our stopping criteria \[72, 65\]. In our case we optimize until the Frobenius norm of the error matrix is below .0001 \[72\]. For the three-way tensor instance, this can be formalized as follows:

$$
A \leftarrow \arg\min_A ||X_{(1)} - (C \otimes B)A^T||
$$

$$
B \leftarrow \arg\min_B ||X_{(2)} - (C \otimes A)B^T||
$$

$$
C \leftarrow \arg\min_C ||X_{(3)} - (B \otimes A)C^T||
$$

(3.8)

Knowing the rank of a tensor is a complex mathematical problem without a simple solution \[6, 14, 103, 65\]. Therefore we attempt tensor decomposition beginning by selecting a single component, and increasing the number of components until the algorithm consistently converges.

### 3.4 Results

#### 3.4.1 Simulations

The CPD algorithm consistently converged at three components for the simulations. In Figure 3.5 each column corresponds to a component of the decomposition, each row represents the dimensions of the tensor: simulation group (A), period (B), and time (C).

The first row of Figure 3.5 shows the variation in score on each component by group. The third component is able to distinguish group two (higher birth rate) from group three (increasing birth rate). This component identifies a temporal shift where 1, 2, and 3 year periodicities shift in strength from the beginning to end of the time series (Figure 3.5 Panel C). The power of the signal peaks in the first quarter,
Figure 3.4: **Sample time-series from simulations and data.** Panel A shows a sample of simulations generated under different birth regimes. Group one is constant CBR of .015, group two is constant CBR of .03 and group three is CBR beginning at .012 and growing to .36. We see that group three begins with dynamics similar to those of group one; by the end of the twenty year period, group three more closely resembles group two. Panel B shows case data from three cities of different cities and periodicities from the E&W data.

changes sign at halfway, and reaches a nadir in the final quarter. A positive score on the third component indicates stronger signals in the first half of the time series and a reduction in the second half. A negative score on this component would indicate the opposite - weaker signals in the first half of the time series and increasing 1, 2, and 3 year periodicities in the second half.

Note that the period and time vectors for both the first and second component are negative, which means their products is positive. The first component therefore demonstrates an increase in biennial and annual signal over time, the second component describes an increase in 1.5 and 3 year signals over time.

Adding components with group specific scores results in a consistent annual and triennial periodicity for the first group (low birth rate); consistent annual and biennial
cycles in the second (high birth rate); and a transition from sporadic epidemics to annual and biennial signals in the third group (increasing birth rate). These results concur with samples of the time series (Figure 3.4) as well as reconstructions of the CWT for each group (Supplement Figure C.2).

Figure 3.5: 3-rank tensor decomposition of the 150 simulated places. Panel A shows the distribution of scores on each of the three components for each of the three groups. Panel B shows the dominant periods explained by each component. Panel C shows the temporal trends for the periods depicted across panel B. Each column gives a full description of each of the three components.

3.4.2 England & Wales Urban Districts 1944-1966

We found four components provided the most consistent convergence with the CPD algorithm for the E&W tensor (Figure 3.6). The most succinct way of classifying variation in measles dynamic across the country is variation in the strength of the annual component throughout the 1944-1965 time period, and whether the biennial component peaks at the beginning or the end of the era.
Figure 3.6: **Four-rank tensor decomposition of E&W data.** The four components of the tensor decomposition for E & W urban districts. Panel A shows each of the time components, panel B shows each of the period components, and panel C shows the distribution of scores on each component by population size. Each column corresponds to a single component (e.g. the first column represents the time, period, and district-specific scores for the first component of the decomposition). We see that small districts consistently score near zero on each component. This is consistent with previous work which suggests dynamics in small districts are highly irregular and lack the seasonal signature we see in large districts.

The dominance of the annual and, in particular, the biennial signal in this data has already been documented [52]. If we were naive regarding the importance of these signals, we could investigate variation in biennial patterns regionally - including synchronicity and temporal lag. Bjornstad and Grenfell have investigated such patterns in this data set [52].

The strength of the score on all components depends largely on population size. This is consistent with previous work which has shown that dynamics in small, isolated places tend to be erratic rather than seasonal and thus do not have significant annual or biennial signals [16, 23]. Though small places consistently score near zero on each
Figure 3.7: A component by component illustration of the reconstruction of the London wavelet spectra. Each time-frequency component is multiplied by the location-specific scalar. These scalars can amplify signals or switch their signs and determine the influence of each component on the final reconstruction. The bottom row compares the original with the reconstructed spectra.
Figure 3.8: **Component reconstruction of Nowich wavelet spectra.** Each time-frequency component is multiplied by the location-specific scalar. These scalars can amplify signals or switch their signs and determine the influence of each component on the final reconstruction. If we compare this reconstruction to the London reconstruction, we can see how adjusting sign and magnitude of the scalars alters the reconstructed frequency-time spectra. The bottom row compares the reconstructed with the original spectra.
component, they generally have the largest scores on the second component relative to the others. Positive, nonzero scores on this component alone is consistent with irregular epidemics. As large districts tend to have higher magnitude scores on the other components, the influence of this irregularity component is reduced for those locations.

The third component uncovers the most variability in the dynamics of large places. Here negative values are associated with early (pre-1955) biennial signals and late (post-1955) annual signals. Positive values predict the opposite: early annual signals and late biennial signals. While all other signals tend to increase in magnitude with population size, the third component increases both in magnitude with varying sign, indicating and substantial deviation in the signal of large places. In addition to variation in the sign of the scores on this component, the magnitude of the scores is greater by nearly a factor of ten. We investigate this component further as it indicates an important dovetailing of epidemic dynamics across locations.

We examined CBRs over the entire time period as well as CBRs in the first four years of the time series. E&W experienced a post-war baby boom between 1944-1948 which peaked in 1947. We took the average of the birth rates between 1944 and 1948. We find when we average CBRs across the twenty year time period, we see little variation across locations. Baby boom CBRs have much greater variability. We see a statistically significant difference, where higher baby boom birth rates correspond to positive scores on the third component. This indicates locations with higher birth rates at the beginning of the time series have stronger annual signals early in the time series compared with their lower birth-rate counterparts (Supplement Figure C.1).

We know from mathematical models of measles transmission, that higher birth rates should lead to larger annual epidemics as a result of quicker susceptible replenishment. In concordance with these models, our decomposition shows locations with slightly lower crude birth rates would begin with more biennial seasonality and lo-
cations with crude higher birth rates would begin with more annual seasonality and settle into biennial cycles later. The baby boom CBR has been identified as a crucial bifurcation point for measles cycles using simulations [48]. This local dovetailing in dynamics as a result of variation in the CBR has not been previously illustrated.

3.5 Discussion

In this paper we explore, to our knowledge, the first application of tensor decomposition to disease surveillance data to confirm previous findings regarding (1) the dominance of annual and biennial signals in the time series across locations; (2) the deterioration in seasonality in small populations; and (3) the importance of crude birth rate in local dynamics using all 954 urban locations. We have also demonstrated the importance of baby boom births in local dynamics.

Tensor decomposition shows promise in its ability to distill multidimensional data into lower dimensional components. Though tensors have been used in many fields of research, their applications to epidemiological data are still under-explored. Here we use tensor decomposition to reveal heterogeneities in time, space and frequency. Since many of the differences illustrated here are well documented in previous studies of this data, we are confident in the ability of tensor decomposition to uncover the dominant trends in the time series.

Here we have demonstrated the utility of tensor decomposition for infectious disease data, more broadly this method is applicable to any spatiotemporal cyclical phenomena in ecology or population science. Though we have only touched on all its applications here, this method can (1) reconstruct original signals without additional noise; (2) concisely summarize dominant trends and heterogeneities in an otherwise unwieldy data set; (3) identify the appropriate frequencies at which to evaluate phase differences and synchrony and lag. An obvious extension of this analysis would be to
use tensor decomposition on the phase or phase-differenced matrices for these locations in order to concisely summarize spatiotemporal dissipation of epidemics [52].
Appendix A

Chapter One Supplemental Information

A.1 TSIR

As a robustness check, we profiled each location across all fitting options available in the tsiR package. Specifically, we fit tsiR using every reasonable combination of regression type for susceptible reconstruction, switching the axes for cumulative births and cases, as well as the GLM family used to estimate parameters. We then select the model with the best likelihood score and use those parameter estimates. This produces slightly different parameter estimates for each location, but the comparative results are consistent. In general we observe a slightly attenuated correlation with population size for our aggregate measures, this is consistent with our belief that there may be a slight bias in small populations. This will be examined more thoroughly in upcoming work. When we examine compare the district pairs, the shape of the curve in Figure 3 which shows the difference of pairs by population difference is virtually the same even when the effect of population size is attenuated. Regression results (discussed later in the supplement) demonstrating the relative importance of
population size also hold. As the results did not change, we opt to present the simpler fitting options in order to be consistent with previous analyses.

Figure A.1: Estimates of the basic reproductive number by population. Here we see some evidence of potential downward bias in the small populations. Though this trend holds even when parameters from optimized regression selections are used. Future work will examine the potential contributions of modeling bias, but we expect they are quite small.

### A.2 Pair selection

Urban and rural pairs were selected by virtue of having the same name. Each represents an urban district and a corresponding rural area. We then confirm that each rural district neighbors the urban district with the same name. In total we have 179 districts of various sizes in a variety of locations.

<table>
<thead>
<tr>
<th>Distincts</th>
<th>Med Pop</th>
<th>Mean Pop</th>
<th>Min Pop</th>
<th>Max Pop</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban</td>
<td>11411</td>
<td>23472</td>
<td>720</td>
<td>245236</td>
<td>32047</td>
</tr>
<tr>
<td>Rural</td>
<td>16552</td>
<td>19433</td>
<td>3060</td>
<td>8608</td>
<td>12600</td>
</tr>
</tbody>
</table>

Table A.1: Paired districts: descriptive statistics
Table A.2: Loadings for All Principal Components

<table>
<thead>
<tr>
<th>variable</th>
<th>PC 1</th>
<th>PC 2</th>
<th>PC 3</th>
<th>PC 4</th>
<th>PC 5</th>
<th>PC 6</th>
<th>PC 7</th>
<th>PC 8</th>
<th>PC 9</th>
<th>PC 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>coupling</td>
<td>0.488</td>
<td>-0.069</td>
<td>0.165</td>
<td>-0.067</td>
<td>-0.075</td>
<td>0.121</td>
<td>0.170</td>
<td>-0.037</td>
<td>0.520</td>
<td>0.637</td>
</tr>
<tr>
<td>susceptibles</td>
<td>-0.307</td>
<td>-0.372</td>
<td>-0.350</td>
<td>-0.210</td>
<td>-0.047</td>
<td>0.057</td>
<td>0.169</td>
<td>0.692</td>
<td>0.296</td>
<td>0</td>
</tr>
<tr>
<td>growth rates</td>
<td>0.025</td>
<td>0.539</td>
<td>-0.305</td>
<td>0.106</td>
<td>0.109</td>
<td>0.720</td>
<td>-0.214</td>
<td>0.157</td>
<td>0.051</td>
<td>0.030</td>
</tr>
<tr>
<td>fadeout length</td>
<td>-0.489</td>
<td>-0.012</td>
<td>0.114</td>
<td>0.212</td>
<td>0.133</td>
<td>0.026</td>
<td>0.061</td>
<td>0.070</td>
<td>-0.419</td>
<td>0.707</td>
</tr>
<tr>
<td>birth rates</td>
<td>0.207</td>
<td>-0.110</td>
<td>-0.233</td>
<td>0.742</td>
<td>-0.554</td>
<td>-0.063</td>
<td>0.031</td>
<td>0.140</td>
<td>-0.093</td>
<td>-0.009</td>
</tr>
<tr>
<td>β variation</td>
<td>0.254</td>
<td>-0.453</td>
<td>0.294</td>
<td>-0.125</td>
<td>-0.066</td>
<td>0.257</td>
<td>-0.651</td>
<td>0.233</td>
<td>-0.281</td>
<td>0.028</td>
</tr>
<tr>
<td>β</td>
<td>-0.336</td>
<td>0.291</td>
<td>0.369</td>
<td>0.235</td>
<td>-0.056</td>
<td>-0.294</td>
<td>-0.477</td>
<td>0.185</td>
<td>0.509</td>
<td>-0.011</td>
</tr>
<tr>
<td>birth variation</td>
<td>0.283</td>
<td>-0.134</td>
<td>-0.282</td>
<td>0.353</td>
<td>0.765</td>
<td>-0.272</td>
<td>-0.155</td>
<td>0.110</td>
<td>0.047</td>
<td>0.030</td>
</tr>
<tr>
<td>number fadeouts</td>
<td>0.218</td>
<td>0.378</td>
<td>-0.390</td>
<td>-0.385</td>
<td>-0.228</td>
<td>-0.479</td>
<td>-0.274</td>
<td>0.173</td>
<td>-0.228</td>
<td>0.267</td>
</tr>
<tr>
<td>$R_0$</td>
<td>0.284</td>
<td>0.322</td>
<td>0.486</td>
<td>0.036</td>
<td>0.106</td>
<td>-0.058</td>
<td>0.376</td>
<td>0.582</td>
<td>-0.250</td>
<td>-0.144</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>SD</th>
<th>Prop</th>
<th>Cum</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>35.5</td>
<td>0.356</td>
<td>0.356</td>
<td>25.8</td>
<td>0.188</td>
<td>0.544</td>
<td>21.7</td>
<td>0.133</td>
<td>0.677</td>
<td>18.5</td>
</tr>
<tr>
<td></td>
<td>18.5</td>
<td>0.097</td>
<td>0.068</td>
<td>15.5</td>
<td>0.068</td>
<td>0.774</td>
<td>15.5</td>
<td>0.097</td>
<td>0.842</td>
<td>13.8</td>
</tr>
<tr>
<td></td>
<td>13.8</td>
<td>0.068</td>
<td>0.054</td>
<td>12.3</td>
<td>0.054</td>
<td>0.896</td>
<td>12.3</td>
<td>0.068</td>
<td>0.940</td>
<td>11.1</td>
</tr>
<tr>
<td></td>
<td>12.3</td>
<td>0.054</td>
<td>0.043</td>
<td>11.1</td>
<td>0.043</td>
<td>0.975</td>
<td>11.1</td>
<td>0.068</td>
<td>0.992</td>
<td>7.8</td>
</tr>
<tr>
<td></td>
<td>11.1</td>
<td>0.043</td>
<td>0.035</td>
<td>7.8</td>
<td>0.017</td>
<td></td>
<td>7.8</td>
<td>0.068</td>
<td></td>
<td>5.3</td>
</tr>
<tr>
<td></td>
<td>7.8</td>
<td>0.017</td>
<td>0.017</td>
<td>5.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## A.3 PCA

## A.4 Pair Regression

We tested the association of urban and rural designation as well as population size using linear regression. As the shape of the association with population size is demonstrably non-linear (Figure 1.3), we allowed for different intercepts and slopes for small (log population $\leq 4.15$) and large populations (log population $\geq 4.15$). This threshold was the most generous in terms of finding an effect of urban/rural designation and maximizing the adjusted r-squared of the regression. Details regarding the selection of this threshold and result sensitivity are included in the supplement. Even with the
Figure A.2: Additional figures demonstrating (A) the difference in size and number of epidemics between urban and rural districts and (B) the consistency in the number of large outbreaks ($\geq 14$ weeks). Another example of an urban and rural pair which demonstrate the subtle difference even when population sizes are comparable.
most generous threshold, the association of urban/rural designation is statistically significant only for small locations, with the effect size representing approximately one tenth that of log population. In other words, the marginal difference between urban and rural locations is comparable to a .1% change in population size. We believe this does not constitute a substantive difference. Even still this represents only a difference on the first principal component which explains only 39% of the total variance.

Since the data is locally linear with population size, we split the data above and below around $10^4$. We test sensitivity to this threshold by moving the threshold from $10^4$ to $10^{4.5}$ by .01 and rerunning the regression. Results of this analysis are included in the appendix. The statistical significance of the coefficient on urban is sensitive to the high/low threshold. When we select the threshold which maximizes the adjusted r-squared, the coefficient on urban status is significant for small places but not for large places is and is comparable to one tenth that on log population. In other words, for small places a shift from rural to urban is equivalent to a .1% increase in population size. Therefore, though the results are statistically significant, the size of the association is inconsequential. We further test these results by comparing with a regression on the spectral density results. The difference between urban and rural districts on their components is of a comparable size as that of the qualitative measures, however, the statistical significance is diminished.
Figure A.3: Size and significance of the coefficient urban/rural dummy when different thresholds are used for large/small populations. These are the coefficients calculated for the epidemic parameter data. The shaded region demonstrates the area where the adjusted R-squared was highest for the regression. Red points indicate statistical significance at the .05 level. We see some statistical significance at thresholds above .415, but the magnitude of these coefficients is consistently near 0.
Appendix B

Chapter Two Supplemental Information

B.1 Simulation Details

The SIR model can be adjusted to include the possibility of returning to the susceptible class post-infection if the disease is not immunizing, or being placed into a vaccinated class if vaccines are in use. The model assumes susceptible individuals move to the infected class at a rate determined by both the number of susceptible and the number of infected individuals and a contact rate ($\beta$). Individuals then recover at a rate ($\gamma$) which is determined by the life course of the infection. If individuals are guaranteed lifelong immunity as a result of the infection (as is the case with measles, mumps, varicella, and many other childhood diseases) they never return to the susceptible class post-infection. In this case new susceptibles are provided via births.
B.1.1 Alpha

We vary the $\alpha$ value we use because the theoretical model of transmission in continuous time would have $\alpha = 1$ but fitting and simulating epidemics in discrete time has been most successful with $\alpha < 1$. In particular, fitting London the TSIR model to the London epidemic dynamics has produced an $\alpha$ value of .97. Some epidemiologists have suggested this adjustment is necessary when moving from continuous to discrete time to keep simulated dynamics from becoming more explosive than the continuous time dynamics we measure. Other epidemiologists have suggested $\alpha$ can be interpreted as an adjustment for heterogeneous mixing patterns. In order to deal with both interpretations of $\alpha$ we simulate with a range of values from 0.95-1.

B.1.2 Beta

We assume transmission has a roughly sine shape, consistent with what has been estimated for large cities in the E&W data set. We fix the mean $\beta$ such that it produces an $R_0$ of 15, 20, or 25 as appropriate and adjust the shape such that it demonstrates the same sine fluctuation we see in the data while keeping the mean $R_0$ the same. Recall that $\bar{R}_0 = \overline{\beta} \ast S_0 = \overline{\beta} \ast N$ so for each population $N$, we want:

$$\overline{\beta} = \frac{R_0}{N} \quad (B.1)$$

Where $\overline{\beta}$ is the mean transmission rate over the course of the year.

B.2 Estimating With Fixed Alpha

Estimating with and without fixing alpha. These figures demonstrate the advantages of fixing alpha even when the value chosen is not equal to the true value. Panel A Demonstrates the $R_0$ bias grouped by population size when $\alpha$ is not fixed. In some
cases we observe large bias in $R_0$ when $\alpha$ is not fixed. Panel B demonstrates how bias in the estimate of $\alpha$ is related to bias in the estimate of $R_0$. In particular, for lower values of $\alpha$ used in the simulations, we see TSIR is prone to vastly underestimating $\alpha$ and compensating for this misestimation by overestimating $\beta$ and thus inflating estimates of $R_0$. Further, we see that biases in $R_0$ tend to be small when bias in $\alpha$ is small. Panel C shows the bias in $R_0$ when $\alpha$ is fixed, compared to Panel A we see a drastic reduction in bias across populations. Panel D shows bias in $R_0$ when $\alpha$ is fixed in the estimation procedure, grouped by the true $\alpha$ value. We there is a possibility for bias even when we fix $\alpha$ to the value used in the simulations and that the bias is slightly worse for locations where we select an $\alpha$ value which is not the true value. However, across values the bias is substantially less than when $\alpha$ is not fixed.

**B.3 Estimating Transmission Shape**

Though the focus of this paper is predicting and correcting for possible biases in parameter estimates across scales, another key contribution of the TSIR model is the recovery of the seasonal shape of transmission. In the E&W data, this seasonal shape is largely driven by the school calendar, [22] in other contexts the shape may be driven by seasonal migration patterns.[45, 42] We provide a brief discussion of the ability of the TSIR model to recover the seasonal shape of transmission across simulations in the supplement (Figure [B.2]).

Seasonal shape recovery from simulations. We regress the estimates of $\beta_t$ on the true $\beta_t$ to estimate how well TSIR captured the original shape of the seasonal transmission. We fit the shape of seasonal is best recovered in populations above 100,000 and/or in cases where the disease goes extinct for less than 20% of the observation period. Though the seasonal shape of transmission is generally not captured for small
Figure B.1: Estimating with and without fixing alpha. These figures demonstrate the advantages of fixing alpha even when the value chosen is not equal to the true value. Panel A Demonstrates the $R_0$ bias grouped by population size when $\alpha$ is not fixed. In some cases we observe large bias in $R_0$ when $\alpha$ is not fixed. Panel B demonstrates how bias in the estimate of $\alpha$ is related to bias in the estimate of $R_0$. In particular, for lower values of $\alpha$ used in the simulations, we see TSIR is prone to vastly underestimating $\alpha$ and compensating for this misestimation by overestimating $\beta$ and thus inflating estimates of $R_0$. Further, we see that biases in $R_0$ tend to be small when bias in $\alpha$ is small. Panel C shows the bias in $R_0$ when $\alpha$ is fixed, compared to Panel A we see a drastic reduction in bias across populations. Panel D shows bias in $R_0$ when $\alpha$ is fixed in the estimation procedure, grouped by the true $\alpha$ value. We there is a possibility for bias even when we fix $\alpha$ to the value used in the simulations and that the bias is slightly worse for locations where we select an $\alpha$ value which is not the true value. However, across values the bias is substantially less than when $\alpha$ is not fixed.

stochastic locations, we know from our main findings that TSIR is often still able to recover the mean transmission rate, $\bar{\beta}$. 
Figure B.2: **Seasonal shape recovery from simulations.** We regress the estimates of $\beta_t$ on the true $\beta_t$ to estimate how well TSIR captured the original shape of the seasonal transmission. We fit the shape of seasonal is best recovered in populations above 100,000 and/or in cases where the disease goes extinct for less than 20% of the observation period. Though the seasonal shape of transmission is generally not captured for small stochastic locations, we know from our main findings that TSIR is often still able to recover the mean transmission rate, $\bar{\beta}$.

### B.4 Time-Series Measures

1. **Seasonality:** For these estimates we tally the number of cases in each biweek across the observation period. We then check how well this tally approximates the typical sinusoidal measles transmission shape.

   (a) **Standard Deviation of Seasonality:** The standard deviation of cases across biweeks. This provides a measure of how different each biweek’s caseload is.

   (b) **Seasonality R-Squared:** The r-squared from regressing the seasonal case tally on a sinusoidal shape.

   (c) **Coefficient of Variation in Seasonality:** the standard deviation of the seasonal tally divided by the mean number of cases across weeks.
2. Case Burden and Persistence:

(a) Coefficient of Variation in Annual Cases: We bin the total number of cases for each year and divide the standard deviation across years by the average annual cases.

(b) Number of Fadeouts: The number of times the disease goes extinct.

(c) Proportion Fadeouts: The number of biweeks without cases divided by the total number of biweeks.

3. Signal Processing: The purpose of these measures is to test various properties of the incidence data from a signal perspective. For many of these measures we use Fourier analysis and calculate the periodogram of the spectral density using daniell smoothers. This gives us an estimate of the dominant periods in the signal.

(a) Red Noise Test: Red noise (or guassian noise) is a signal created from a random walk across frequencies. This tends to produce a lower frequency signal than white noise. We compare the power spectrum of the incidence data using fourier analysis to the power spectrum of red noise and apply a $\chi^2$ test for difference. We use the p-values from this test to characterize how similar the incidence data is to red noise [109].

(b) Coefficient of Variation in Top Periods: We calculate the spectral density and select the top ten most powerful frequencies. We then calculate the coefficient of variation across these frequencies.

(c) Average Power: We take find the average period of the ten strongest periods from the spectral density.

(d) Biennial to Noise Ratio: We divide the power of the biennial period by the average of low period (one week to one month) periods.
(e) Annual to Noise Ratio: We divide the power of the annual period by the average of low period (one week to one month) periods.

(f) Biennial to Annual Ratio: We divide the power of the biennial period by the power of the annual period.

4. Growth Rates: For these we isolate the beginning of each epidemic and examine the growth rates along a four biweek window. We approximate growth rates by adding unity to the data, taking the log and then taking first differences. We believe growth rates will have a relationship to estimated transmission and a lot of variation or a large gulf between minimum and maximum observed values may impact the precision of our estimates.

(a) Standard Deviation of Growth Rates: We take the standard deviation of the growth rates and average across each epidemic.

(b) Growth Mean-Max: We take the average of the max growth rate across each window.

(c) Min Growth Rate: We take the minimum of all values across each window and epidemic.

(d) Max Growth Rate: We take the maximum value across each window and epidemic.

(e) Mean Growth Rate: We take the mean of all values.
Figure B.3: Coverage of time series measures across simulations and data, plotted against population (part one)
Figure B.4: Coverage of time series measures across simulations and data, plotted against population (part two)
Figure B.5: Coverage of time series measures across simulations and data, plotted against population (part three)
Appendix C

Chapter Three Supplemental Information

C.1 Crude Birth Rates and the Third Component

C.2 Simulation Raw and Reconstructed Wavelet Spectra
Figure C.1: Crude Birth Rates for locations above the critical community size in England and Wales grouped by positive and negative scores on the third component of the tensor decomposition. We see that locations with higher crude birth rates are more likely to have a positive score on this component, indicating larger annual epidemics early in the time series.

C.3 Sample Location Raw and Reconstructed Wavelet Spectra
Figure C.2: Averaged and reconstructed power spectra for simulated data. We see a large amount of variation as a result of stochasticity within each group. However we do see constant annual power in each location. We see the biennial signal becomes stronger over time in the third (variable birth rate) group. Additionally we see more long-term cycles in the first (low birth rate) group. We see the reconstructed power spectra are able to pick up much of this variation.
Figure C.3: Reconstructed and original power spectra for (A) London, (B) Manchester, (C) Norwich, and (D) Leeds. We can see that the four components are able to capture variation in the timing of annual and biennial periodicities. We also see that much of the noise and idiosyncrasies of individual time series is reduced. In other words, the reconstructions pull out the dominant signals and enable a comparison across these patterns.
Bibliography


