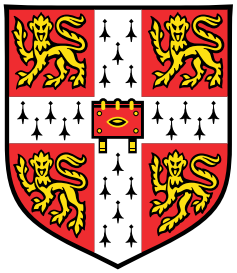
# Geographic and demographic transmission patterns of the 2009 A/H1N1 influenza pandemic in the United States



**Stephen Kissler**

Department of Applied Mathematics and Theoretical Physics
University of Cambridge

This dissertation is submitted for the degree of
*Doctor of Philosophy*

King's College October 2017

For Mark

# Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. It is not substantially the same as any that I have submitted, or is being concurrently submitted, for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my dissertation has already been submitted, or is being concurrently submitted, for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution.

Stephen Kissler
October 2017

# Acknowledgements

My first expression of thanks must go to Julia Gog. When I arrived unannounced in your office four years ago, I could not possibly have known that it would be the start of such a great adventure. For all of your well-timed nudges in the right direction, for introducing me to so many colleagues and friends, and for repeatedly entrusting me with responsibility, often more than I merited – you have my deepest appreciation and gratitude.

Regarding those aforementioned colleagues and friends, this thesis bears the imprint of many memorable conversations I've been privileged to share with Cécile Viboud, Bryan Grenfell, Ottar Bjørnstad, and Lone Simonsen. I am especially indebted to Cécile and Bryan for hosting me at the National Institutes of Health and at Princeton at various points during my PhD. Those visits are highlights of my last three years.

Back in Cambridge, I owe a great deal to Petra Klepac, Veni Karamitsou, Sophie Ip, and Maria Tang, for suffering my scrawls on the whiteboard, for critiquing my presentations, for keeping me grounded during the writeup, and for providing such a stimulating and enjoyable intellectual environment over these past years. You are the greatest research group a person could hope for. Thanks also to Stephen Eglen for checking in on me from time to time, and for assisting me through the layers of logistics associated with getting a PhD. Thanks once more to him and to Carola-Bibiane Schönlieb, for their helpful comments on my first year report. I hope I've managed to incorporate them all. And, thanks to the members of the Disease Dynamics Unit at the Department of Veterinary Medicine for taking me in as an unofficial member of their group. This thesis has benefited especially from conversations with Andrew Conlan, James Wood, Colin Russell, David Price, and Caroline Trotter.

I must also thank my thesis examiners, Matt Keeling and Steven Riley. It was a tremendous honour to share this work with two scientists I hold in such high regard. Thank you for your time, your attention, and your insight; I learned a great deal from our discussion, and have done my best to incorporate your advice. I look forward to hopefully many more conversations in the future.

I also owe a great deal to Anne Dougherty, Jim Curry, David Bortz, and Vanja Dukic, back at the University of Colorado, for showing me how to do research in the first place, and for

sustaining me during my PhD with continued advice, reference letters, and encouragement. And, I'd never be here were it not for Scot Douglass, Deborah Viles, and Joan Gabriele. Thanks to each of you, for so many things in particular, and for so much more besides than I can express.

The Gates Cambridge Trust has provided me with the financial means to study in Cambridge, but more importantly, has provided me with some of the greatest friends and fondest memories of my time here. For that I will be eternally grateful. Thanks also to King's College for their personal, academic, and financial support. And, I owe a great deal of appreciation to the fellowship at Queens' College for adopting me into their ranks – I'm beyond excited to join you formally this October.

To Fr Mark Langham, Sr Ann Swailes, Fr Philip Moller, Fr Kevin Grove, and the Revd Andrew Hammond, for providing spiritual nourishment these past years; you've ensured that my soul was shaped in tandem with my intellect. And, to my friends who have made this time in Cambridge an absolute joy, who are far too many to list, but I must at least mention Hardy Schilgen, Olly McMillan, Karly Drabot, Rachel Reckin, Dakota Spear, Willie Payne, Ilana Trumble, Jamie Gay.

Finally, to my grandmother, my parents, Mark, Katy, V, Maggie, Frank, for letting me move across an ocean to pursue this dream. Your love, prayers, and support meant, and mean, the world.

*4 October 2017*

# Abstract

This thesis describes how transmission of the 2009 A/H1N1 influenza pandemic in the United States varied geographically, with emphasis on population distribution and age structure. This is made possible by the availability of medical claims records maintained in the private sector that capture the weekly incidence of influenza-like illness in 834 US cities. First, a probabilistic method is developed to infer each city's outbreak onset time. This reveals a clear wave-like pattern of transmission originating in the south-eastern US. Then, a mechanistic mathematical model is constructed to describe the between-city transmission of the epidemic. A model selection procedure reveals that transmission to a city is modulated by its population size, surrounding population density, and possibly by students mixing in schools. Geographic variation in transmissibility is explored further by nesting a latent Gaussian process within the mechanistic transmission model, revealing a possible region of elevated transmissibility in the south-eastern US.

Then, using the mechanistic model and a probabilistic back-tracing procedure, the geographic introduction sites (the 'transmission hubs') of the outbreak are identified. The transmission hubs of the 2009 pandemic were generally mid-sized cities, contrasting with the conventional perspective that major outbreaks should start in large population centres with high international connectivity. Transmission is traced forward from these hubs to identify 'basins of infection', or regions where outbreaks can be attributed with high probability to a particular hub.

The city-level influenza data is also separated into 12 age categories. Techniques adapted from signal processing reveal that school-aged children may have been key drivers of the epidemic.

Finally, to provide a point of comparison, the procedures described above are applied to the 2003-04 and 2007-08 seasonal influenza outbreaks. Since the 2007-08 outbreak featured three antigenically distinct strains of influenza, it is possible to identify which antigenic strains may have been responsible for infecting each transmission hub. These strains are identified using a probabilistic model that is joined with the geographic transmission model, providing a link between population dynamics and molecular surveillance.

# Table of contents

# List of figures

# List of tables

# Chapter 1

# Introduction

The epidemiology of influenza escapes easy classification. At first glance, the transmission of influenza is predictable: outbreaks strike temperate regions of the world each winter without fail, and the influenza virus usually evolves so steadily that vaccines for the coming season can be produced months in advance. But upon closer inspection, this regularity gives way to inconstancy. Complex interactions between evolutionary selective pressure, changes in interpersonal contact rates, and shifts in weather patterns may cause the timing and severity of influenza outbreaks to shift unpredictably [151]. Sometimes, a novel genetic variant of the influenza virus emerges, triggering a pandemic. Pandemic outbreaks can cause infection well outside the months when seasonal outbreaks are normally observed, and can cause severe disease in age groups that are normally resilient to seasonal influenza [254]. Variation between pandemics is great as well, ranging from the 1918 pandemic, which in terms of mortality was among the worst natural disasters in human history, to the 2009 pandemic which, despite being caused by the same viral subtype as the 1918 pandemic, featured morbidity and mortality rates that were hardly worse than for normal seasonal flu [61, 232]. Apparent inconsistencies arise even at the level of individual outbreaks. The 2009 influenza pandemic spread globally far more rapidly than any previous influenza pandemic, but its transmission at the country scale, at least within the United States, proceeded in a slow, cohesive wave [91, 254].

Clearly, much uncertainty still surrounds the transmission dynamics of influenza. There is a particular need to study the spread of pandemic influenza, since pandemic influenza is often regarded as one of the foremost natural threats to human health and safety [28]. The 2009 A/H1N1pdm influenza pandemic offers a valuable opportunity to study the transmission of pandemic influenza in a contemporary setting.

This thesis makes use of a dataset of unprecedented detail that captures the incidence of weekly influenza-like illness in over 800 US cities between 2001 and 2010. A mathematical modelling analysis identifies key predictors of the between-city spread of the major autumn wave of the 2009 A/H1N1pdm influenza pandemic in the US, pinpoints the geographic locations where the outbreak first became established, and determines which age groups were most responsible for sustaining its transmission. Comparisons are drawn with two other seasonal influenza outbreaks. The transmission patterns of the 2009 A/H1N1pdm influenza pandemic challenge conventional wisdom at every turn, suggesting that extreme care must be taken when developing strategies to mitigate the spread of future pandemics.

## 1.1 Biology and epidemiology of influenza

There are three types of influenza virus capable of infecting humans, labelled types A, B, and C [44]. All three are thought to share a common ancestor, but have evolutionarily diverged to the point that genetic material may be shared within a type, but likely not between types [25]. Influenza virus types A and B cause periodic outbreaks in humans, while type C is generally not capable of sustained human-to-human transmission [44]. Type A influenza may be further classified into subtypes that specify the haemagglutinin (HA) and neuraminidase (NA) antigenic proteins that are expressed on the virus' surface. There are currently 18 known variants of HA and 11 variants of NA, each of which is assigned a number in order of its first observation [46]. Influenza virus classifications begin with the type, and then, for type A viruses, are followed by the HA and NA specification. Of the influenza A subtypes, only subtypes A/H1N1 and A/H3N2 currently cause sustained human-to-human transmission [46]. The strain responsible for the 2009 influenza pandemic is commonly labelled as type A/H1N1pdm09 or simply A/H1N1pdm, to distinguish it from other A/H1N1 subtypes.

To replicate, the influenza virus binds to a cell's surface, enters, and releases its genetic material. Copies of the virus' genome are made in the cell's nucleus, and these copies are packaged into new viral particles. Finally, the newly-formed virus particles exit the cell [25]. The genetic replication step is imperfect, so the influenza genome gradually accumulates mutations. This leads to a process called *antigenic drift*, by which the antigenic proteins expressed on the virus' surface change. Over time, this can allow the virus to escape detection by an immune system that has been challenged by an earlier form of the virus. Antigenic drift is thought to be a primary reason why human immunity to influenza wanes over time, making recurrent annual outbreaks possible [25].

The influenza A virus also evolves according to a second, more punctuated process, made possible by the fact that its genome is separated into segments. Sometimes, a single cell becomes infected by different influenza A virus subtypes. During replication, genome segments from these different strains may become packaged into a single new viral particle. This can give rise to an *antigenic shift*, in which a virus is produced with a novel combination of HA and NA surface proteins [25, 118]. If the new virus is capable of transmitting between humans, a pandemic outbreak can occur, since underlying immunity to the new strain is likely to be low [25]. Such recombination events are thought to have triggered the 1918 influenza pandemic and to have been important precursors to the 2009 A/H1N1pdm influenza pandemic [25, 118].

Influenza may be transmitted between humans through airborne droplets from coughs and sneezes at a range of up to about six feet [42, 139]. The influenza virus can also survive on surfaces, and may be transmitted via fomites [139]. Influenza infection presents clinically in humans as an upper respiratory disease, normally associated with fever, chills, and a cough [41]. Up to a third of influenza infections may remain asymptomatic [30]. Sometimes, though, influenza infection is severe enough to cause major illness and even death, often through secondary complications [41]. In the United States, influenza is thought to cause over 30,000 deaths annually, with over 90% of these in the elderly population [52]. Pandemic influenza outbreaks may feature shifts in the age groups that suffer the highest morbidity and mortality, with children and young adults often disproportionately affected [254].

At the city level, the timing of influenza outbreaks may be modulated by population size and/or surrounding population density [49, 50, 91]. A disproportionate amount of influenza transmission likely occurs in schools [33, 199, 116]. School-aged children are thought to bear the highest burden of infection during normal influenza seasons [252] and likely helped to sustain and amplify transmission of the 2009 A/H1N1pdm influenza pandemic in communities around the world [117, 180]. The geographic transmission patterns of influenza at the country and continent scales remain an area of open investigation [241]. A few studies suggest that influenza may spread in relatively coherent geographic waves at this scale [48, 91, 184].

Internationally, air travel likely contributes to the transmission of influenza [59, 96]. The rise in air travel in the late $20^{\text{th}}$ century may explain the rapid international spread of the A/H1N1pdm pandemic in the spring of 2009, in comparison to the relatively slower global spread of earlier pandemics [40]. Many of the earliest observed cases of 2009 A/H1N1pdm infection were in countries with strong air traffic links with Mexico, where the disease was first reported [132].

The drivers of the strong wintertime rhythm of seasonal influenza outbreaks in temperate regions are still not fully understood [151]. Fluctuations in temperature and absolute humidity may play a role [62, 209], as might increased indoor crowding in the winter months, and decreases in vitamin D associated with less exposure to sunlight, which may dampen the immune system [151]. Dynamic resonance theory suggests that even small shifts in these factors may be enough to sustain a seasonal pattern of influenza outbreaks, which could explain why it remains so difficult to unpick the factors that drive the seasonality of outbreaks [74]. Pandemic influenza outbreaks, however, may strike at any time of year, and often feature multiple waves of infection in quick succession [40, 183, 254].

## 1.2 Chronicle of the 2009 A/H1N1pdm influenza pandemic in the United States

In February of 2009, the town of La Gloria Mexico suffered an outbreak of an influenza-like respiratory pathogen [81]. At least one individual from La Gloria was later confirmed to have been infected by a novel strain of type A/H1N1 influenza [81]. In April 2009, two California children were found to have been infected by the same virus, confirming that the virus had begun to spread internationally [40]. Within nine weeks of the first observed cases, the virus had spread around the world, significantly more rapidly than any previous influenza pandemic [254]. On 11 June 2009, the WHO officially declared the outbreak a pandemic [40].

After an initial spike in influenza activity, incidence in the United States waned in July and August, marking the end of the first spring wave of infection [40]. Some evidence suggests that significant transmission of this first wave may have been limited to cities in the northeastern US, despite the earliest cases being detected in California and Texas [91]. Influenza activity increased again in late August, roughly coinciding with the start of the autumn school term [40, 91]. This second wave caused significant levels of infection across the US [122]. In the eastern US, the outbreak appears to have spread as a cohesive geographic wave with an epicentre in Alabama or Georgia [91]. A third, less geographically extensive wave followed in early 2010 in the southeastern US [209]. Overall, the 2009 A/H1N1pdm pandemic disproportionately affected children and young adults, both in the US and worldwide, with abnormally high rates of hospitalisation and death among those under the age of 60 [40, 254, 122].

A number of intervention strategies were implemented to mitigate the spread of the 2009 A/H1N1pdm influenza pandemic in the US. On 27 April 2009, the Centers for Disease Control and Prevention (CDC) issued a recommendation that all US citizens avoid non-essential travel to Mexico [40]. Local school closures were also reported throughout the pandemic [40]. Many individuals voluntarily adopted personal prevention measures, including washing hands more frequently, staying at home when feeling ill, and avoiding exposure to people with flu-like symptoms [225]. The CDC began developing a vaccine for the new A/H1N1pdm strain in April 2009, just six days after the first case was identified in the US [40]. The vaccine was not approved until October, however, and was initially only administered to groups deemed at high risk of infection [40]. Vaccine availability was not high enough for mass vaccination until the end of December [40]. So, while vaccination may have reduced the overall burden of infection during the pandemic, it did not have a major impact on the transmission of the spring or autumn waves of infection.

## 1.3 Key concepts in mathematical epidemiology

### 1.3.1 Early contributions to mathematical epidemiology

In 1766, Daniel Bernoulli published an essay on the increase in life expectancy that universal smallpox variolation might provide [19]. The document provides one of the earliest examples of a mathematical model used to characterise the transmission of an infectious disease [68]. The following 150 years saw significant mathematical contributions to the mitigation of disease, including Florence Nightingale's use of statistics to successfully argue for sanitary reform in Crimea [54], John Snow's famous map revealing London's Broad Street water pump as the source of a massive cholera outbreak [220], and Sir Ronald Ross's treatise on the prevention of malaria [203]. However, it is William Kermack and Anderson McKendrick who may be recognised as the founders of modern mathematical epidemiology. Their 1927 article "A contribution to the mathematical theory of epidemics" presents what are now often regarded as fundamental equations of the field [131].

### 1.3.2 Geography and age in epidemiological models

The equations developed by Kermack and McKendrick rely on strict assumptions: they require that all individuals in a population be equally susceptible to infection, that all equally contribute to onward transmission, that interpersonal contact rates be uniform, and that the infection be totally immunising [131]. Many of the developments in mathematical

epidemiology over the past ninety years may be characterised as efforts to relax these assumptions. Two key extensions to Kermack and McKendrick's original theory are the inclusion of geographic structure and age structure in epidemic models.

The importance of geography to the transmission of disease was clearly recognised even before Kermack and McKendrick's seminal article; indeed, John Snow's mapping of the 1854 cholera outbreak, sixty years prior, is seen as a foundational moment in medical geography [144]. Incorporating spatial structure into epidemic models is an obvious way to increase their relevance, since infectious diseases may only be transmitted when a susceptible host comes into physical proximity with a source of infection. However, as Lawson notes, major developments in medical geography and spatial statistics did not occur until the second half of the $20^{\text{th}}$ century [144]. In 1969, Levins made a key contribution through his introduction of the concept of a metapopulation, or a population of populations [148]. Levins was not the first biologist to consider spatially-distributed populations, but he provides the first elements of a robust mathematical framework for describing their dynamics [106]. Metapopulation theory is attractive because it allows for some consideration of spatial heterogeneity, while allowing sub-populations, or 'patches', to obey the stricter assumptions that underlie the older epidemiological theory. Though one of the theory's earliest applications was to study animal species in physically separated settings like island systems, it was soon also used to describe the spread of disease, where a pathogen spreads through a population of "island-like" hosts, who themselves may also comprise a spatially structured population [98]. Hanski and collaborators have made important contributions to metapopulation theory [103–105], and Sattenspiel *et al.* (1995) [207] provide a theoretical treatment of Kermack-McKendrick-type models on metapopulations. Metapopulation models are sometimes criticised for ignoring important aspects of the physical environment, since few populations actually obey the strict spatial clustering that the theory assumes [8]. However, since epidemiological data are usually aggregated by geographic location, metapopulation models are often the best, or only, viable option for describing the geographic transmission of disease. Furthermore, interventions are normally implemented within well-defined geographic areas (e.g. cities), making metapopulation models a natural framework for testing the potential impact of policy decisions. Metapopulation models have successfully yielded insight into the transmission of a range of diseases, suggesting that the trade-off of realism for tractability is warranted [56, 75, 91, 100, 164, 256].

Individual-based models (IBMs) provide a second framework by which geographic space may be considered in epidemic models. Rather than considering homogeneous patches of spatially-separated individuals, IBMs account for an additional level of detail by keeping

track of the movements and infectious statuses of all individuals in a population. IBMs can provide insight into how individual-level variation affects disease transmission across multiple scales, though the high level of detail incorporated into such models can sometimes be falsely mistaken for fidelity to real-world processes [200]. Keeling *et al.* (2010) [128] provide a useful comparison of epidemic dynamics on metapopulation models and on two common types of individual-based models, the commuter model and the random movement model. A version of the commuter model will be considered in §4.4.

For both metapopulation models and IBMs, the spatial range over which infection may spread may be characterised in terms of a 'kernel'. Kernels are functional forms, often only defined up to a constant of proportionality, that describe how infectivity decays with distance. For IBMs, the kernel may be more concretely interpreted as a description of the movement tendencies of individuals in the population [128]. The shape of a kernel may be inferred from pairwise movement data [127, 201], or, in the absence of such data, a parametric form for the kernel may be specified *a priori*, and parameters may be chosen so that the kernel best accounts for the observed spatial spread of a disease [48, 91, 239, 256]. Gravity and radiation kernels are the two most common parametric kernel forms used to model the geographic spread of infectious disease in humans [200, 215, 250]. These are discussed further in the introduction to Chapter 3.

In contrast to the metapopulation and IBM approaches, which partition populations into countable units, it is also possible to consider disease transmission as a continuous spatial process. This is especially common in the study of animal and plant diseases, where, for example, the spread of rabies among foxes [172] and the dispersal of pollen [181] may be seen as or approximated by diffusion processes in continuous space. There is also a rich set of statistical tools for inferring patterns from epidemic processes observed in continuous space. These include point process analysis to detect abnormal clusters of infection [70, 84] and Gaussian process regression to characterise wavefronts of infection [11, 93, 196]. Gaussian process regression is considered further in Chapter 3.

Age-structured models constitute a second key extension to Kermack and McKendrick's original theory. Host age can have a profound impact on disease transmission, both because immunity to disease may vary by age [216], and because a person's age largely dictates with whom she/he spends the most time [170]. Due to the similarities between the mathematical theories of epidemiology and demography (see [111]), age structure can sometimes be incorporated into epidemiological models in an elegant and analytically tractable way. Key contributions in this area include Dietz and Schenzle's introduction of a general model for the transmission potential of a disease in an age-structured population [69], Andreasen's

analysis of how fatal infectious diseases can regulate host population size and age structure [5], Diekmann, Heesterbeek, Roberts' development of next-generation matrix theory [66], and Klepac and Caswell's general treatment of stage-structured epidemic models [134]. Inaba (2017) [120] provides a useful overview of demographic and epidemiological models with age structure. Further background on age-structured models, especially in the context of influenza, may be found in the introduction to Chapter 5.

### 1.3.3 Turning points in the theory of infectious disease

Amidst the gradual development of increasingly realistic epidemiological models over the last century, two key ideas emerged that mark particularly important leaps in the theory of infectious diseases. These are a notion of the basic reproduction number, $R_0$, and the development of phylodynamics.

Thresholds have played a central role in the mathematical theory of infectious diseases for much of the field's history. Ross's 'mosquito theorem', for example, states that malaria control is possible by reducing the prevalence of mosquitoes below some threshold, and Kermack and McKendrick proved that epidemics may only occur when population density surpasses a baseline value [131, 203]. It was not recognised until later, however, that a central idea, the basic reproduction number, binds these diverse threshold results together. First explicitly formulated by Dietz in 1975 [67], the basic reproduction number

> "represents the number of secondary cases that one case can produce if introduced to a susceptible population." [67, 111]

The basic reproduction number provides a simple criterion for whether or not it is possible for a disease to spread: if $R_0 > 1$, an epidemic may occur, while if $R_0 < 1$, the infection will fade out of the population. Importantly, it also provides an estimate of how difficult it may be to control the spread of a disease, since the target of any vaccination strategy may be roughly summarised as an attempt to reduce $R_0$ below 1. A generalisation of the basic reproduction number, the next generation matrix (NGM), provides a natural way of characterising how a disease's infectiousness varies with different strata of the population, such as across different age groups [66]. Knowing the structure of the NGM can help develop targeted intervention strategies that reduce transmission in sectors of the population that tend to cause the most onward transmission. In Chapter 2, the basic reproduction number is roughly estimated for the autumn 2009 A/H1N1pdm epidemic in the United States, and in Chapter 5 the NGM plays a central role in the development of a method to infer the relative strengths of disease transmission between different age groups.

The second major theoretical advance came in the early 2000s with the formal incorporation of genetic theory with epidemiology, termed 'phylodynamics' [99]. The central idea is that a pathogen's genome does not simply dictate how a disease will spread, but rather that epidemiological and evolutionary processes interact. This consideration is especially important for influenza, which features complex evolution patterns that are influenced by the human immune response and by the virus' ability to infect a diverse range of host species. Since the introduction of phylodynamics, a significant body of literature has developed that considers epidemiological and genetic data in tandem [20, 82, 110, 241]. Interest in phylodynamics has been spurred by the simultaneous development of computational frameworks like BEAST that make it easier to extract information from genome sequences [73, 72] with the ever-diminishing cost of pathogen sequencing. The development of robust methods to incorporate epidemiological and genetic data remains a central priority in infectious disease research [95], and is a primary motivation for the theory developed in Chapter 4 of this thesis.

## 1.4 Mathematical contributions to the study of influenza

At certain scales, Kermack and McKendrick's original theory provides an adequate description of influenza's transmission dynamics, as Spicer *et al.* (1979, 1984) demonstrate in England and Wales and in central London [221, 222]. Using sophisticated inference techniques but few adjustments to the underlying theory, Yang *et al.* (2015) [257] use a Kermack-McKendrick-type compartmental model to estimate key transmission parameters for ten recent influenza outbreaks in the US. However, the transmission dynamics of influenza do vary across geographic space and with host age, which has spurred the development of a a range of models that account for these and other sources of heterogeneity.

Geographic mathematical models for the transmission of influenza have existed since at least the late 1960s, when Rvachev and colleagues used travel statistics from the then-USSR to describe the between-city transmission of influenza [53, 205, 221]. More recently, efforts undertaken by Viboud *et al.* (2006) [239] describe how commuting patterns affect the spread of seasonal influenza in the United States across 30 years. A model developed by Eggo *et al.* (2011) [75] for the between-city transmission of the 1918 influenza pandemic in the US and the UK laid the foundation for work by Gog *et al.* (2014) [91] and Charu *et al.* (2017) [48] who describe the between-city transmission of influenza in the United States between 2001 and 2010. City-level transmission of the 2009 A/H1N1pdm pandemic in Peru and Mexico is considered by Chowell *et al.* (2011a, 2011b) [49, 50] who, like Gog *et al.* (2014) [91] and Charu *et al.* (2017) [48], find that a city's population size and surrounding population density

may predict its outbreak timing. Geographic models for the transmission of influenza in Europe have been developed by Smieszek *et al.* (2011) [218] for Switzerland, and by Paget *et al.* (2007) [184] for the continent as a whole. Prior to the 2009 A/H1N1pdm influenza pandemic, a few studies attempted to predict the likely geographic transmission patterns of a novel influenza-like respiratory pathogen at the country scale [76, 77, 87]. These predictions generally featured hierarchical transmission, with infection reaching major cities first and then diffusing into surrounding areas. While this paradigm appears to match the transmission of some diseases such as measles [256], work by Gog *et al.* (2014) [91] suggests that the 2009 A/H1N1pdm influenza pandemic in the US did not follow this sort of hierarchical spread.

At the international scale, an early model that considers the spread of influenza via air routes is also contributed by Rvachev *et al.* (1985) [206]. Prior to the 2009 A/H1N1pdm influenza pandemic, a number of models were developed that sought to characterise the likely international transmission dynamics of a novel pandemic influenza virus via global airline networks [55, 59, 29, 114]. These generally found that, while air travel would be a key driver of the outbreak's spread, travel restrictions would do little more than delay the pandemic's inevitable progress. Retrospective studies on the role of air travel in the global dissemination of the 2009 A/H1N1pdm influenza pandemic support their findings; indeed, the pandemic did spread internationally largely along air traffic routes, and countries that imposed travel restrictions saw little delay in outbreak timing [10, 132].

Age structure is also a key component in many mathematical models of influenza transmission. An important early work by Longini *et al.* (1978) [156] evaluates possible control strategies for influenza outbreaks in an age-structured population. Castillo-Chavez *et al.* (1989) [31] introduce a model for influenza transmission that accounts for host age and cross-immunity from different strains of influenza, and demonstrate that an interplay between this cross-immunity and age-specific mortality may explain the recurrence of waves of influenza infection. Kucharski and Gog (2012a, 2012b) [135, 136] follow in this tradition to describe possible mechanisms for the development of age-varying immunity to influenza, and to describe how this variation affects the population-level transmission of influenza. Other recent approaches to incorporate age structure into mathematical models of influenza transmission centre around inferring the next-generation matrices that describe the transmission of influenza by age group, as in Nishiura *et al.* (2010) [180] and Glass *et al.* (2011) [90]. These approaches are described in greater detail in the introduction to Chapter 5. Keeling and White (2011) consider age structure and geography jointly in a mathematical analysis of

optimal vaccination campaigns, partly inspired by the outbreak of 2009 A/H1N1 pandemic influenza [130].

Recently, phylogenetic analyses have led to remarkable insights into the transmission of influenza. Phylogeography, which refers to the use of phylogenetic techniques to characterise the spatial spread of a species, has played an especially prominent important role in elucidating the transmission patterns of influenza at the international and country scales [123, 160, 177]. Russell *et al.* (2008) [204] show that international H3N2 influenza outbreaks are normally seeded from southeast Asia, while Bedford *et al.* (2015) [18] demonstrate that other strains may circulate at low levels in other parts of the world throughout the year. Nelson *et al.* (2011) demonstrate that the spring wave of the 2009 A/H1N1pdm influenza pandemic in the US may have featured multiple different strains, while the autumn wave was likely dominated by a single strain [177]. As it becomes more commonplace to sequence influenza virus genomes, phylogeographic approaches hold great promise for refining our understanding of how influenza spreads.

## 1.5 Summary of thesis

This thesis gives a detailed account of the transmission of the autumn 2009 A/H1N1pdm pandemic influenza outbreak in the United States, highlighting how geography and age affected the outbreak's trajectory, often in unexpected ways. **Chapter 2** presents the data on which the subsequent chapters rely. Special attention is given to a dataset derived from medical insurance claims records that captures the weekly incidence of influenza-like illness in over 800 US cities between 2001 and 2010. An outbreak onset detection algorithm is developed and used to infer city-level outbreak onset times from these ILI data for the autumn 2009 A/H1N1pdm epidemic in the US. An exploratory analysis of these onset times reveals a coherent geographic transmission wave that radiated across the country from a few distinct epicentres, with city-level onset times normally first detectable in children. These observations form the basis for the investigations in Chapters 3, 4, and 5.

In **Chapter 3**, a geographic transmission model for the between-city transmission of influenza in the United States is developed, following the groundwork laid by Eggo *et al.* (2011) [75] and Gog *et al.* (2014) [91]. This model is used to identify key factors that affected a city's risk of infection during the autumn 2009 A/H1N1pdm pandemic. A non-parametric extension to the model reveals how the disease's transmissibility may have varied over time and by geographic location. In particular, a region of especially high transmissibility is

identified in the southeastern US, which has already been identified as an important region for the spread of both seasonal and 2009 pandemic influenza in the US [48, 91, 209].

Then, in **Chapter 4**, the geographic transmission model is reverse-engineered to identify the establishment sites, or 'hubs', of the autumn 2009 A/H1N1pdm outbreak in the US. The most important of these lie in the southeastern US and in the central valley of California. Mapping the spread of infection onward from these hubs reveals a set of overlapping sub-outbreaks, which provide a testable hypothesis for where distinct viral strains may have circulated.

**Chapter 5** considers how the transmission strength of the 2009 A/H1N1pdm influenza pandemic in the US differed by age group. A general information-theoretic measure is adapted to identify the drivers of disease transmission from age-stratified incidence time series. Applying the measure to ILI data from the autumn of 2009 in the US reveals that school-aged children likely contributed most to sustaining transmission during the pandemic. Age-stratified ILI time series are also incorporated into the geographic transmission model from Chapter 3, revealing a slightly different picture of which age groups may have been most responsible for first sparking outbreaks in neighbouring ZIPs.

To provide a point of comparison, the above methods are applied in **Chapter 6** to ILI data from the 2003-04 and 2007-08 seasonal influenza outbreaks. Special emphasis is placed on the geographic transmission kernels for the three outbreaks, which succinctly capture key differences in the speed and wave-like character of transmission during those seasons. For the 2007-08 outbreak, a brief statistical analysis of available geo-tagged antigenic data provides a first guess of which antigenic strains may have triggered the outbreaks in each of the transmission hubs in that season. This points a way forward for combining epidemiological and genetic data to make more robust geographic inferences of how influenza spreads at the country scale.

# Chapter 2

# Data

This chapter presents the epidemiological, geographic, demographic, and antigenic datasets that underpin this thesis' findings. Of primary importance is a dataset that measures weekly influenza-like illness (ILI) incidence for twelve age groups in 884 cities across the US in 2009, covering 61.5% of all physician practices in the country. These data are presented in §2.2. An outbreak onset detection algorithm, based on the breakpoint method introduced by Charu *et al.* (2017) [48], is presented in §2.3 to help visualise the spatiotemporal patterns present in these data, and to lay essential groundwork for the transmission model developed in Chapter 3. The transmission model in Chapter 3 also relies on the population sizes, coordinates, and school term start dates for the 884 cities where ILI data are available. Datasets that provide these pieces of information are presented in sections 2.5-2.6. Finally, Chapter 6 relies in part on knowing the prevalences of distinct antigenic subtypes of influenza across the US. A dataset that captures weekly counts of laboratory-confirmed influenza cases by antigenic subtype in 10 US regions is presented in §2.7.

Sections 2.3.1 and 2.6 are adapted from "Geographic Transmission Hubs of the 2009 Influenza Pandemic in the United States" (Kissler *et al.* 2017 [133]), submitted to *Epidemics*.

## 2.1 Background

### 2.1.1 A brief history of influenza surveillance

Records of an influenza-like illness may be found as far back as the 5$^{\text{th}}$ century BCE in the writings of Hippocrates, who described a winter-time outbreak of an upper respiratory tract illness in the sixth book of *Epidemics* [186]. Though humans have likely suffered from periodic influenza outbreaks for many centuries, reliable records are sparse until *circa* 1650,

when some of the earliest confirmed references to influenza were written [195]. The first detailed records of the pathophysiology of pandemic influenza were made and shared among learned societies during the 1761-62 pandemic [230]. By the onset of the 1918 influenza pandemic, many cities had developed robust surveillance systems that made it possible to closely monitor and respond to the outbreak's spread [3]. The detailed records made during that outbreak paved the way for quantitative analysis of influenza transmission; indeed, efforts to model the transmission of the 1918 pandemic using original records continue to the present day [6, 51, 75].

Since the early 20$^{th}$ century, influenza surveillance has become increasingly centralised. In 1947, the World Health Organisation (WHO) launched the Global Influenza Programme as a central platform for sharing surveillance information and coordinating outbreak response [253]. Shortly thereafter, in 1952, the WHO launched the Global Influenza Surveillance Network (now the Global Influenza Surveillance and Response System, or GISRS) for virologic surveillance, just two decades after the influenza virus was first isolated in the laboratory [195, 253]. Currently, formal influenza surveillance is coordinated by individual countries, often according to standards specified by the WHO [248]. The rise of computer technology has supplemented these formal surveillance tracks with data from social media [141], search engines [89, 194], and electronic medical records [240, 260].

### 2.1.2 Modern influenza surveillance in the United States

Formal influenza surveillance in the United States is coordinated nationally by the Centers for Disease Control and Prevention (CDC) and locally by state and city public health services [43]. The CDC collects clinical influenza-like illness (ILI) surveillance data from a network of over 2,800 outpatient healthcare providers, and reports these data aggregated weekly for 10 geographic regions [43]. The CDC defines ILI as "fever (temperature of 100°F [37.8°C] or greater) and a cough and/or a sore throat without a known cause other than influenza" [43]. Since many non-influenza respiratory illnesses trigger ILI symptoms, the CDC also collects virologic data from approximately 100 public health and 300 clinical laboratories across the US [43]. This provides a lower-volume but more specific estimate of weekly influenza incidence than the ILI network offers, and gives information about which viral strains are circulating during a given influenza season.

The surveillance data gathered by the CDC are often used as benchmarks of accuracy [89, 214, 240]. However, the data are normally only available 1–2 weeks after the reported cases have occurred, and do not provide information about influenza activity at the state or

city levels [89]. To obtain data with better timeliness and resolution, efforts were made in the late 2000s to estimate influenza activity using Yahoo and Google search terms [89, 194]. These platforms initially only provided data for the US, though Google Flu Trends (GFT) later expanded to provide international estimates of influenza activity [94]. Despite early success, GFT mis-calculated the timing and severity of a number of outbreaks between 2009 and 2013, and was discontinued in 2015 [94, 182]. Even so, internet-based epidemic surveillance remains an active area of research [107, 211]. In addition to the above, dedicated online platforms such as FluNearYou in the US and FluSurvey in the UK gather reports of ILI symptoms from community volunteers [109, 155]. Though data from these platforms represent only a convenience sample of the population, they hold some promise for supplementing traditional ILI data streams with high-volume, real-time data [2, 190, 219].

Electronic medical claims records offer an alternate source of high-volume ILI data in the United States [240]. These records are routinely collected for insurance purposes during the majority of outpatient visits in the US. A separate claim is made for each visit, containing the physician's syndromic classification of the patient's illness [240]. Syndromes are normally classified according to the International Classification of Diseases (ICD) scheme, which is a widely-used list of possible causes of ill health and mortality [169]. Claims forms that contain syndromic classifications of influenza or influenza-related symptoms can be extracted, providing clinically-based ILI counts without requiring extra effort on the part of the clinician. Viboud *et al.* (2014) [240] demonstrate that retrospective ILI estimates from medical claims records can reliably estimate weekly influenza incidence and outbreak timing, even at fine geographic scales, with greater success than GFT. Yih *et al.* (2014) [260] demonstrate that medical claims records can provide reliable real-time influenza incidence estimates. Marsden-Haug *et al.* (2007) [161] compare electronic ICD-9 ILI estimates with laboratory-confirmed influenza incidence from US military treatment facilities, and find that ICD-9 ILI provides reliable information about temporal trends in influenza incidence. Medical claims records are not free from bias; a report by the WHO notes that "the interpretation of the data derived from [electronic medical claims-based] systems will depend heavily on the local coding practices, the external forces that influence coding decisions (such as reimbursement), and clinician understanding of the coding system" [248]. That report, however, addresses the international community; incentives and coding habits are unlikely to vary within the US as much as they might between countries. In addition, aggregating records from multiple physician practices can help to homogenise some of the variability introduced by particular practices and clinicians. Overall, electronic medical claims records offer a promising and

so-far underutilised source of fine-scale data on influenza incidence in the United States [182, 240].

### 2.1.3 Geographic divisions in the United States

For this thesis, there are two important geographic partitions of the US population: the ZIP code and the HHS region. ZIP (postal) codes are five-digit numbers assigned to groups of mailboxes by the United States Postal Service (USPS) to facilitate mail delivery. The first three digits of a ZIP code specify the processing facility through which its mail is sorted, and so can be used as a slightly coarser level of aggregation than the 5-digit ZIP code [238]. The ILI data to be considered in this thesis are tagged by the 3-digit ZIP codes of the outpatient clinics from which they are collected.

ZIP codes do not necessarily correspond to well-defined geographic regions [234, 238]. To make it possible to link ZIP codes to geographic areas, the US Census Bureau publishes ZIP Code Tabulation Areas (ZCTAs), which specify rough ZIP boundaries [234]. The US Census Bureau also publishes sets of "Gazetteer Files", which provide geographic and demographic data for ZCTAs [233]. There are 884 3-digit ZIP codes in the 48 contiguous US states for which Gazetteer data are available. Fig 2.1 depicts the population size distribution of these 3-digit ZIPs, and Fig 2.8 depicts their geographic locations (see §2.5 for more detail on how the 3-digit ZIP coordinates are obtained). The median 3-digit ZIP population size is 209,839 people.

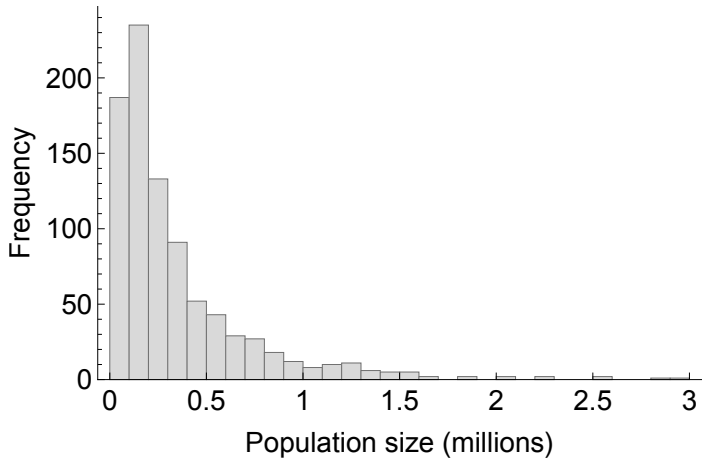



Fig. 2.1 Population size distribution for the 884 3-digit ZIPs in the US Census Bureau's Gazetteer dataset [233]. The largest 3-digit ZIP is Houston Main TX, and serves 2,983,395 people. The median 3-digit ZIP population size is 209,839 people.

HHS regions are multi-state areas specified by the US Department of Health and Human Services to act as an intermediary between state-level and national welfare-related governance [237]. There are 10 HHS regions, roughly corresponding to (1) New England, (2) New York/New Jersey, (3) the Mid Atlantic, (4) the Southeast, (5) the Great Lakes, (6) the South, (7) the Midwest, (8) the Mountain West, (9) the Southwest, and (10) the Northwest. The CDC reports much of its data, including its virologic influenza surveillance data (see §2.7), at the level of HHS regions.

Three-digit ZIPs roughly correspond to towns and cities, and HHS regions generally consist of climatically and demographically similar groups of states. However, neither of these partitions is likely to be perfectly suited for epidemiological inference. Geographic community detection is an active area of research that holds some promise for identifying more relevant ways of separating populations into discrete clusters based on movement patterns or interpersonal contacts [162, 176]. However, identifying which community structures are most relevant for the transmission of a particular disease is complicated, since disease transmission often requires specific types of contact between individuals [170], and since transmission dynamics may depend on complex interactions between interpersonal contact patterns and exogenous factors, such as temperature and humidity [62, 209]. To my knowledge, there are no suitably detailed studies on the geographic population clusters most relevant for the transmission of respiratory illnesses in the United States, or indeed in any country. The best approach appears to be to use the data at the level at which they are available, and note that models should only be used to make inferences at the geographic scale of the data with which they are parametrised.

### 2.1.4 Outbreak onset detection

Ecological invasion waves, such as those caused by a spreading pathogen, are often characterised in terms of the times at which the invading species becomes established in distinct geographic areas [92, 165, 256]. Doing so makes available a range of established modelling techniques for describing the progress of the invading species [92, 103, 75]. For infectious diseases, incidence time series are often routinely recorded as an outbreak unfolds. This motivates a need for for methods that can infer outbreak onset time, or the time of pathogen establishment, from a disease incidence time series.

There are many algorithms to identify the onset of an outbreak in real time from disease incidence time series [1, 108, 242, 248, 251]. These are mainly used as aids for making public health decisions during an outbreak. On the other hand, if the outbreak has already

passed, it may be possible to make more accurate onset time estimates by retrospectively considering the full incidence time series. There are indeed some strategies for retrospectively inferring whether an outbreak has taken place and its rough timing [112, 189], but these are not generally intended for identifying precise outbreak onset times.

At least two strategies exist for explicitly inferring outbreak onset times from retrospective ILI incidence time series [48, 91]. The first strategy, the "threshold method" from Gog *et al.* (2014) [91], defines city-level influenza outbreak onset times as the first of three consecutive weeks in which ILI incidence in a given city surpasses a sinusoidal baseline fit to the ILI incidence between flu seasons. This is similar in strategy to many of the real-time outbreak detection methods cited above. Though conceptually straightforward, defining baselines and thresholds can be difficult, and generally must be repeated for each new pathogen and geographic scale [43, 213]. The second strategy, the "breakpoint method" introduced by Charu *et al.* (2017) [48], takes a somewhat different approach. It estimates epidemic onset time as the changepoint in the slope of a bi-linear trend fit to an ILI time series in the weeks preceding the epidemic peak. This avoids the need to define a baseline and, as will be seen in §2.3, provides a natural way of characterising uncertainty in the onset estimate. The method can have difficulty identifying the onset times of outbreaks with multiple peaks, though adjustments to the method described in §2.3.1 help address this problem somewhat. Even so, it offers an alternative approach to identify epidemic onset times from noisy data while making few assumptions about the epidemic process. An evaluation of the breakpoint method and a comparison of its results with those obtained using a version of the threshold method may be found in §2.3.2.

Outbreak onset detection algorithms are intended to identify the time of outbreak establishment, but not necessarily the prior time of pathogen introduction into the population. Introduction does not guarantee establishment, especially for diseases with relatively low $R_0$, such as 2009 A/H1N1pdm influenza [129]. If establishment does follow an introduction, the time interval between the two events is stochastic. In general, geographic ecological models tend to focus on establishment because establishment is easier to detect from routinely collected data than sporadic introductions are, and because establishment is also a better indicator of a location's influence on its neighbours. Because of this, traditional metapopulation theory generally assumes that a sub-population can only contribute to the onward spread of an invading species after local establishment has occurred [103]. In the context of spatial disease transmission, the time of outbreak establishment in a sub-population can be roughly seen as the time at which that sub-population becomes infectious to its neighbours. It is therefore epidemiologically convenient to characterise the progress of a disease in terms of

its local establishment times, which can be done with the help of an onset detection algorithm. Eventually, phylogenetic data might be useful for determining outbreak establishment times. Infections in distinct small populations are likely to all have a common ancestor, and so the time of arrival of that first successful ancestor could be taken as the time of outbreak establishment. Genetic data for influenza does not currently exist at the geographic resolution necessary to make such inferences, but this would be a valuable avenue for future work.

## 2.2 Description of the ILI dataset

### 2.2.1 Data source

The ILI data to be considered in this thesis come from a convenience sample of electronic CMS-1500 medical health insurance claims forms [236] submitted by physician practices across the US between 2003 and 2009. Syndromic classifications on the forms are coded by physicians according to ICD-9 standards. Data from 354,402 practices are available for 2009, which represents approximately 61.5% of all physician practices in the US, and is thought to capture over 50% of all outpatient physician visits in that year [240]. The health records were first gathered and made available by SDI Health (now Quintiles IMS), a private healthcare data analytics company [119]. Records are anonymised and aggregated by the first three digits of the ZIP code of the practice from which they were collected, for a total of 884 locations. For the remainder of this thesis, these 3-digit ZIP codes will be referred to simply as "ZIPs". To my knowledge, this is the most geographically detailed ILI dataset ever considered for the United States. Within each ZIP, records are available as weekly aggregates across all age groups, as well as stratified into 12 age groups (<2, 2-4, 5-9, 10-14, 15-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80+). Weekly ILI counts are inferred by extracting all claims with a direct mention of influenza, or fever combined with a respiratory symptom, or febrile viral illness (ICD-9 487-488 OR [780.6 and (462 or 786.2)] OR 079.99) [240]. Cases from individuals with lower socio-economic status may be under-reported; the nature and impact of these biases are discussed further by Lee *et al.* (2017) [145].

Time series of ILI incidence are made by dividing the total number of ILI cases in each week by the total number of claims made in that week, for each ZIP. This helps adjust for variation in coverage and reporting rate by location and time. The same is done for each age-stratified subset of the data. The result is a time series of overall ILI incidence covering 496 weeks for each of the 884 ZIPs, as well as twelve age-stratified time series of the same length for each ZIP. These time series will be referred to as the IMS-ILI dataset, following

Viboud *et al.* (2014) [240]. The IMS-ILI data, aggregated across all age groups and locations, are depicted in Fig 2.2.

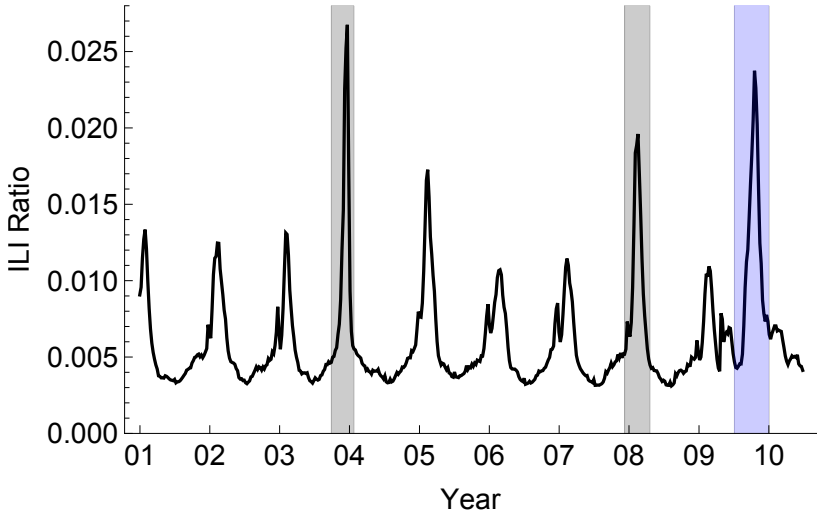



Fig. 2.2 Weekly ILI incidence in the United States, aggregated across all age groups and ZIPs, from the IMS-ILI dataset. The ILI ratio is calculated by dividing the number of recorded ILI cases by the total number medical claims submitted in each week. The blue shaded band highlights the autumn wave of the 2009 A/H1N1pdm pandemic, which is the major focus of this thesis. The 2003-04 and 2007-08 seasonal outbreaks, which will be considered in Chapter 6, are marked by the grey bands.

### 2.2.2 Data validation

The correspondence between the IMS-ILI data and reference influenza surveillance data from the CDC is described in depth in Viboud *et al.* (2014) [240]. In brief, the weekly incidence and peak timing of outbreaks in the IMS-ILI dataset both correlate highly with the weekly incidence and peak timing from CDC ILI and CDC virologic surveillance data at the regional level. The correlations remain strong when the data are stratified into four age groups. Correlations for the same metrics are also strong when the IMS-ILI data are compared with city-level ILI data from New York State. City-level correlations outside New York State could not be assessed, due to a lack of reference data. Taken together, this suggests that the IMS-ILI data may provide reliable information about epidemic timing by geographic region in the US, even when separated into age groups [240].

## 2.3 Inferring outbreak onset times from the IMS-ILI data

To infer outbreak onset times from the IMS-ILI data, an improved version of the breakpoint method developed in Charu *et al.* (2017) [48] is presented, and its performance is evaluated using epidemic simulations.

### 2.3.1 Breakpoint onset detection method with peak adjustment

The breakpoint method defines outbreak onset time as the changepoint in the slope of a piecewise-linear regression fit to an ILI time series in the $n$ weeks prior to and including the peak ILI incidence within some pre-specified time window. Specifically, the parameters $\{\beta_{0,i}, \beta_{1,i}, \beta_{2,i}, t_i, \sigma_i^2\}$ of the following linear model are estimated by maximum likelihood:

$$Y_{it} = \beta_{0,i} + \beta_{1,i} t + \beta_{2,i}(t - t_i)^+ + \varepsilon_{it} \tag{2.1}$$

where $Y_{it}$ is the ILI ratio for location $i$ at time $t$, and where the noise term $\varepsilon_{it}$ follows a $N(0, \sigma_i^2)$ distribution. The parameter $\beta_{0,i}$ gives the vertical-axis intercept of the first line in the trend, the parameter $\beta_{1,i}$ gives the slope of the first line, and the sum $\beta_{1,i} + \beta_{2,i}$ gives the slope of the second line. The $(t - t_i)^+$ term takes the value $t - t_i$ if the difference is positive, and zero otherwise. The time of epidemic onset in location $i$ is given by the maximum likelihood estimate of the breakpoint time $t_i$, denoted $\hat{t}_i$. To interpolate to a slightly higher degree of temporal resolution, epidemic onset times are rounded to the nearest half-week, following Gog *et al.* (2014) [91] and Charu *et al.* (2017) [48]. Uncertainty in the onset time is captured by the likelihood profile for $\hat{t}_i$. For 2009, fitting to $n = 17$ weeks prior to the peak (inclusive) provides enough ILI data points to give a robust onset estimate, while avoiding the tail end of the spring infection wave that affected a number of locations. Fig 2.3 illustrates the procedure.

For locations with very uncertain onset times (i.e. the log likelihood profile does not drop by at least 2 units), the procedure is repeated using alternative peaks, which are defined as any points in the time series prior to the true peak whose two immediately neighbouring points are lower. Whichever peak yields the narrowest onset confidence interval is chosen, with its corresponding onset. This helps to determine accurate onset times for locations that have multiple peaks. Fig 2.4 depicts the original and updated onset times for such a location in 2009. Any locations for which the onset likelihood profile does not drop by at least 2 log-likelihood units, even after this adjustment, are omitted from further analysis.

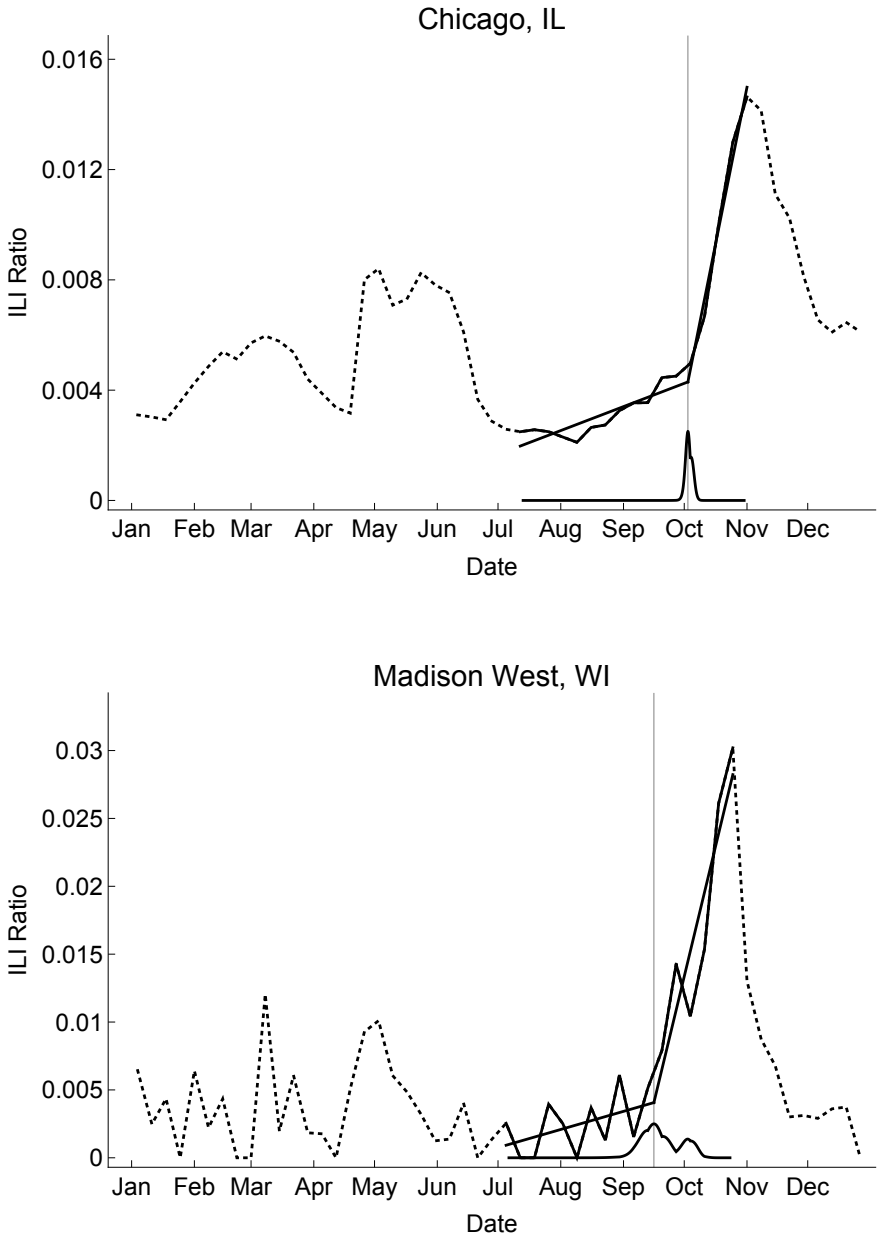



Fig. 2.3 Illustration of the breakpoint method for determining outbreak onset time from a time series of ILI incidence. The time series (dotted lines) depict the weekly ILI ratios for Chicago IL (ZIP 606) and the nearby Madison West WI (ZIP 538) in 2009. Chicago is one of the largest ZIPs in the US, with a population of 2.8 million. Madison West is one of the smallest, with a population of 58,000. A bi-linear trend is fit to the 17 weeks of the time series prior to and including the week of peak incidence. The onset date is defined as the maximum likelihood estimate of the breakpoint in the bi-linear trend, rounded to the nearest half-week. The solid curve below the time series depicts the likelihood profile (analogous to a probability density) for the breakpoint onset. Both likelihood profiles have been rescaled vertically for ease of depiction, since only their distributions matter. For Chicago, the distribution is narrow, indicating a high degree of certainty in the onset estimate. For Madison West, the distribution is wider and bimodal, where each mode corresponds to a plausible estimate for the city's outbreak onset time.

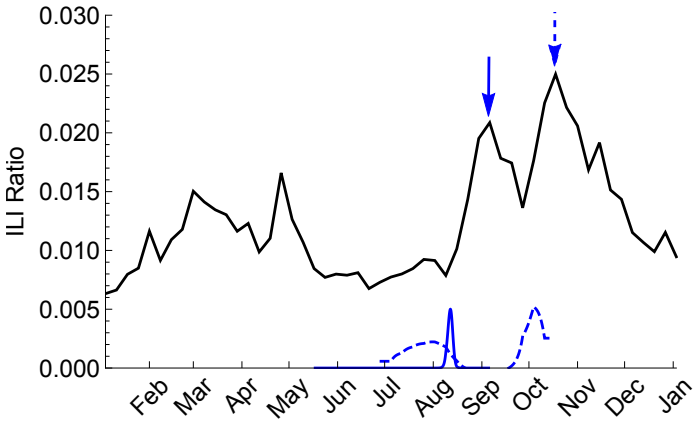



Fig. 2.4 ILI time series for Fresno CA in 2009 (black), with the original breakpoint likelihood profile (blue, dashed) obtained using the overall peak incidence (dashed arrow), and the adjusted breakpoint likelihood profile (blue, solid), obtained by using the earlier peak (solid arrow). Both likelihood profiles have been rescaled for ease of depiction, since only their distributions matter. The original likelihood profile is broad and double-peaked, and does not drop sufficiently to give a reliable onset estimate. The adjusted likelihood profile is sharp, indicating a confident onset estimate.

### 2.3.2 Evaluation of the breakpoint method

Using epidemic simulations, it is possible to check the breakpoint method's ability to accurately identify outbreak onset times. In this section, the breakpoint method is used to estimate onset times for simulated outbreaks of varying strengths and with varying degrees of autocorrelation between the incidence in consecutive weeks. Onsets are also calculated using a version of the threshold method presented in Gog *et al.* (2014). The performance of both onset detection methods improves with increasing epidemic strength. The performance of the threshold method worsens with increasing autocorrelation, while the breakpoint method is fairly robust to autocorrelation. In general, breakpoint onset estimates are both more accurate (less biased) and more precise (have lower variance) than threshold onset estimates.

Simulated epidemic time series are generated using a two-component self-exciting Poisson process model inspired by Held *et al.* (2006) [112]. This framework permits flexible modelling of stochastic epidemic processes without needing to make assumptions about the number of susceptible individuals in a population, which is difficult to estimate for influenza [257]. The incidence at discrete time $t$, $Z_t$, is defined as the sum of an endemic process $X_t$

and an epidemic process $Y_t$:

$$\begin{aligned} Z_t &= X_t + Y_t, \\ X_t &\sim \text{Poisson}(\nu_t) \\ Y_t | Z_{t-1} &\sim \text{Poisson}(\lambda_t Z_{t-1}) \end{aligned} \tag{2.2}$$

where $\nu_t$ is the mean value of the endemic process at time $t$, and $\lambda_t$ is an excitation parameter. If $\lambda_t > 1$, the epidemic will grow, and if $\lambda_t < 1$, it will decay. If the time step length is approximately equal to the disease's generation interval, then $\lambda_t$ is analogous to the reproduction number $R$.

A different approach for producing these epidemic simulations would model the endemic and epidemic components $X_t$ and $Y_t$ as independent processes. This second approach is arguably more epidemiologically accurate for influenza than Eq 2.2, if $X_t$ is interpreted as non-influenza ILI and $Y_t$ is interpreted as true influenza infections. Non-influenza infections clearly cannot directly cause influenza, but in Eq 2.2, the endemic process $X$ contributes directly to infections in the epidemic process $Y$ through the history of $Z$. The primary reason for using Eq 2.2 rather than a superposition of two independent processes is that it allows the modeller to more easily specify the time of outbreak onset/establishment, as opposed to the time of the first case introduction(s). If one were to model the endemic and epidemic processes independently, one would have to seed the epidemic process with some number of infected individuals at time $t_{seed}$. If the number of initially infected individuals is small, there is a high probability that the epidemic will fail. If the epidemic does become established, the time of establishment will not generally fall on a consistent date. If one attempts to force establishment on a particular date by introducing many infected individuals at once, this leads to an artificial deterministic jump in the number of cases at time $t_{seed}$, making it easy for an onset detection algorithm to detect the onset time, but revealing very little about the algorithm's actual performance, since sudden introductions of many infected individuals into a population are rare. A better approach seems to be to increase the infectiousness of the disease, expressed as $\lambda_t$ in Eq 2.2, at some time point $t_{start}$. This more reliably simulates a scenario in which the pathogen has been circulating at low levels for some time, and then suddenly becomes established in the population at a specified time.

In Held *et al.* (2006) [112], the endemic mean $\nu_t$ is allowed to vary in time, but does not depend on previous values of the process itself. However, in the IMS-ILI time series, there is evidence of autocorrelation in the ILI counts outside of the influenza season. This can be demonstrated in the following way: Following Gog *et al.* (2014) [91], weeks with

non-epidemic ILI are defined as those in the nationally-aggregated IMS-ILI time series (see Fig 2.2) with ILI ratio below 0.06. A sinusoid is fit to these aggregated non-epidemic ILI counts to estimate the sinusoidal phase $D$ of the out-of-season ILI. Then, a quartic function plus a new sinusoid with phase $D$ is fit to each ZIP's ILI incidence time series for the non-epidemic weeks. This trend is subtracted from the time series, providing a de-trended set of non-epidemic ILI ratios for each ZIP. A one-step autoregressive (AR1) process

$$W_{t+1} = \phi W_t + \varepsilon \tag{2.3}$$

is fit to each of these de-trended non-epidemic time series, where $W_t$ denotes the $t^{\text{th}}$ value of the de-trended time series, $\phi$ describes the level of autocorrelation, and $\varepsilon$ is an error term that follows a $N(0,\sigma^2)$ distribution. Since the de-trended non-epidemic time series are roughly stationary (i.e. not increasing or decaying over time), we should expect $\phi$ to be smaller than 1 in absolute value. Positive $\phi$ values provide evidence of positive autocorrelation, for which subsequent series values tend to be similar to their predecessors. The model parameters $\phi$ and $\sigma^2$ are fit by maximum likelihood using *Mathematica's* `TimeSeriesModelFit` function. Fig 2.5 provides a histogram of the autocorrelation parameters $\phi$ estimated from each ZIP. There is evidence of positive autoregression, with a median autoregression parameter of $\phi = 0.35$ and lower and upper quartiles of 0.24 and 0.49.

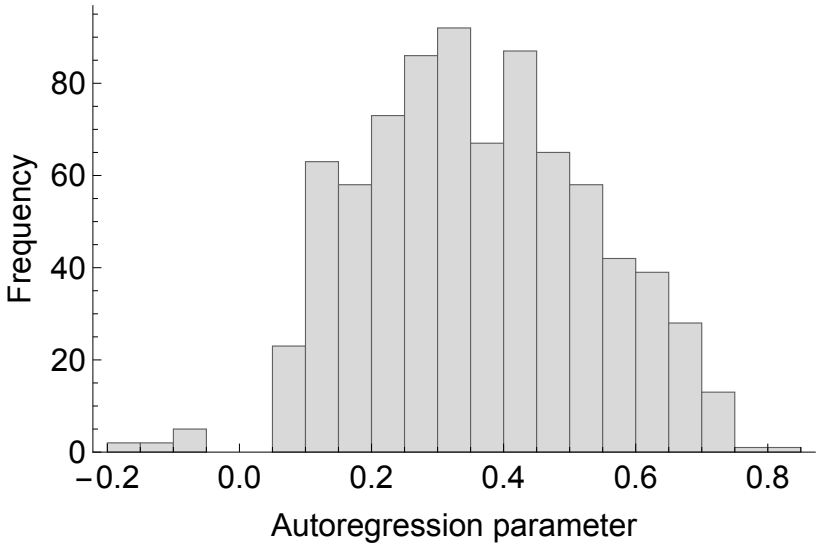



Fig. 2.5 Histogram of the ZIP-level autocorrelation parameters $\phi$ for the de-trended out-of-season ILI ratios. There is evidence of positive autocorrelation in the de-trended time series, with a median AR parameter of $\phi = 0.35$ and lower/upper quartiles of 0.24 and 0.49.

To incorporate autocorrelation into the endemic process, set

$$\nu_t = \mu_X + \phi(X_{t-1} - \mu_X) \tag{2.4}$$

where $\mu_X$ is the mean of the endemic process $X_t$, and $\phi$ is an autoregression parameter with $0 \leq \phi < 1$. Specifying the parameters in this way ensures that the endemic process has a stationary distribution with mean $\mu_X$. The variance of the endemic process is $\mu_x/(1-\phi^2)$ [101].

For the epidemic portion of the model, $Y_t$, the excitation parameter $\lambda_t$ is held fixed at some constant value $\lambda_{min} < 1$ from the start of the epidemic until a time $t_{start}$. At time $t_{start}$, $\lambda_t$ jumps to $\lambda_{max} > 1$. Then, $\lambda_t$ decreases linearly until it returns to $\lambda_{min}$ at time $t_{end}$. This yields an acute outbreak of length $t_{end} - t_{start}$.

Table 2.1 lists the outbreak simulation parameters, their interpretations, and the values they take on for the epidemic simulations. A range of $\phi$ and $\lambda_{max}$ values are tested to simulate varying levels of autoregression and epidemic strengths. Each simulation is run for a total of 52 time steps, with epidemic onset at time $t_{start} = 40$ and an epidemic duration of eight time steps. The endemic mean $\mu_X$ is set at 20. Simulations are seeded with 20 endemic-type individuals ($X_0 = 20$). For each combination of $\lambda_{max}$ and $\phi$, 1000 epidemics are simulated. Fig 2.6 depicts a representative simulation from each of these ensembles.

Table 2.1 Values and interpretations for the parameters of the epidemic simulation model Eq 2.2

| Parameter | Values | Interpretation |
|---|---|---|
| $\lambda_{min}$ | 0.5 | Epidemic excitation parameter, lower bound |
| $\lambda_{max}$ | $\{1, 1.1, 1.2, 1.3, 1.4, 1.5\}$ | Epidemic excitation parameter, upper bound |
| $\phi$ | $\{0, 0.25, 0.5, 0.75\}$ | Endemic autoregression strength |
| $\mu$ | 20 | Endemic mean number of cases |
| $X_0$ | 20 | Initial number of endemic cases |
| $Y_0$ | 0 | Initial number of epidemic cases |
| $t_{start}$ | 40 | Start time of the outbreak |
| $t_{end}$ | 48 | End time of the outbreak |

For each simulated outbreak, the onset time is calculated using both a version of the threshold method and the breakpoint method. For the threshold method, an optimal threshold value is identified for each pair of $\phi$ and $\lambda_{max}$. To do so, the mean $\mu_Z$ and standard deviation $\sigma_Z$ of the process $Z$ outside the epidemic period are estimated by simulating 10,000

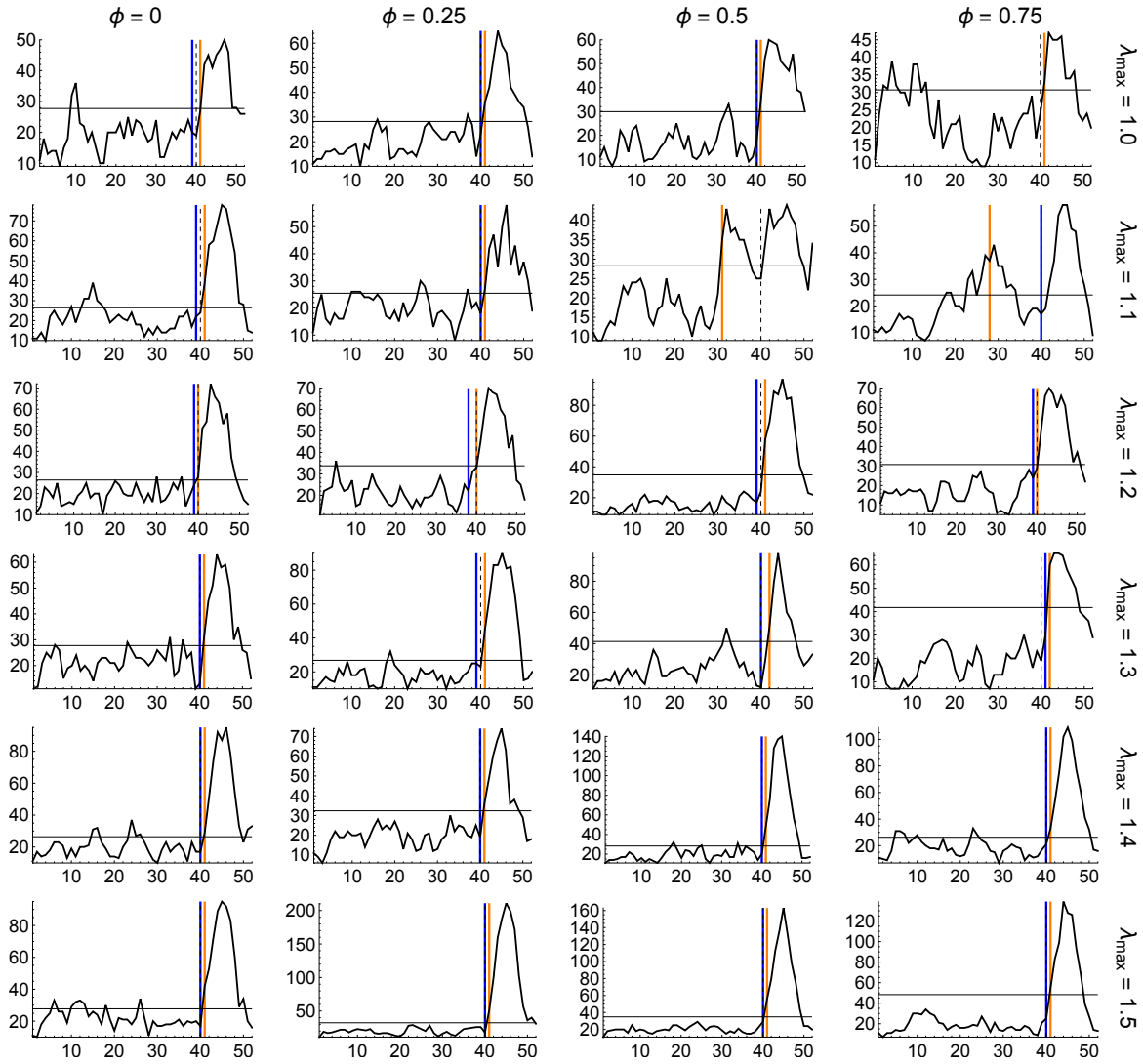



Fig. 2.6 Example outbreak simulations from model Eq 2.2 for autocorrelation parameter $\phi \in \{0, 0.25, 0.5, 0.75\}$ and epidemic strength $\lambda_{max} \in \{1, 1.1, 1.2, 1.3, 1.4, 1.5\}$. Autocorrelation increases from left to right, and outbreak strength increases from top to bottom. The vertical axes measure numbers of cases, and the horizontal axes measure time in units of weeks. The true onset time $t_{start}$ at 40 weeks is depicted by the black dashed vertical line. The estimated threshold onset time is depicted in orange, and the breakpoint onset time is in blue. The horizontal black bar depicts the optimal threshold used to calculate the threshold onsets (see Table 2.2). For two of the simulated outbreaks, those with $\phi = 0.75$ and $\lambda_{max} = 1$ and with $\phi = 0.5$ and $\lambda_{max} = 1.1$, the breakpoint method fails to detect a sufficiently certain onset time.

consecutive draws from the process $Z$ with $\lambda_t = \lambda_{min}$, and computing the empirical mean and standard deviation of these draws. This gives a 'baseline' description of the process $Z$ with epidemic forcing $\lambda_t$ is at its minimum value. Then, threshold onset times are calculated for each of the 1,000 epidemic simulations in the ensemble using thirteen possible thresholds, ranging from $\mu_Z$ to $\mu_Z + 3\sigma_Z$ in steps of size $\sigma_Z/4$. The threshold onset time is defined as the first of three consecutive weeks that lie above the specified threshold, following Gog *et al.* (2014) [91]. The threshold that yields the greatest number of onsets within $\pm$ one week of the true onset, lagged by one week (the threshold method systematically estimates onsets with a one-week lag, see Fig 2.7) is taken to be the optimal threshold for that particular combination of $\lambda_{max}$ and $\phi$. These threshold values are listed in Table 2.2 for each pair of $\phi$ and $\lambda_{max}$. The optimal thresholds lie between 1.5 and 2.5 standard deviations $\sigma_Z$ above the process mean $\mu_Z$, with higher thresholds preferred for stronger outbreaks (higher $\lambda_{max}$). Optimising the threshold in this way clearly cannot be done in a real outbreak setting, since the true onset time is not known; the intention of doing so here is to investigate which threshold values generally perform best under different epidemic scenarios, and to compare the breakpoint onset times against the best possible threshold onset times.

Table 2.2 Optimal thresholds for each combination of autoregression parameter $\phi$ and epidemic strength $\lambda_{max}$, expressed as the number of standard deviations $\sigma_Z$ above the process mean $\mu_Z$

| | $\phi = 0$ | $\phi = 0.25$ | $\phi = 0.5$ | $\phi = 0.75$ |
|---|---|---|---|---|
| $\lambda_{max}$ = 1.0 | 1.5 | 1.75 | 1.5 | 1.5 |
| $\lambda_{max}$ = 1.1 | 1.75 | 1.75 | 1.5 | 1.25 |
| $\lambda_{max}$ = 1.2 | 1.75 | 2. | 1.75 | 1.5 |
| $\lambda_{max}$ = 1.3 | 2. | 2. | 2. | 1.75 |
| $\lambda_{max}$ = 1.4 | 2.5 | 2.5 | 2.25 | 2. |
| $\lambda_{max}$ = 1.5 | 2.5 | 2.25 | 2.25 | 1.75 |

Fig 2.7 provides violin plots of the distributions of onset times inferred using the threshold and breakpoint methods for each ensemble of simulations. For all combinations of $\lambda_{max}$ and $\phi$, the most frequent threshold onset time is one week later than the true onset time, while the most frequent breakpoint onset time coincides with the true onset time. The breakpoint onset times also have lower variance than the threshold onset times in all but one scenario, with $\phi = 0.75$ and $\lambda_{max} = 1$. This suggests that breakpoint onset times are generally more precise than threshold onset times. In some cases, and particularly for low values of $\phi$ and $\lambda_{max}$, the lower variance of the breakpoint onset times may be due in part to the breakpoint

method rejecting more onsets than the threshold method. Recall that the breakpoint method rejects an onset whenever the log-likelihood profile for the onset time does not drop by at least two units, and the threshold method rejects an onset whenever there are not three consecutive weeks that lie above the threshold. Rejecting onsets should not necessarily be seen as a failure of the method; indeed, for small values of $\lambda_{max}$, the simulated outbreak is often so small that it is practically indistinguishable from the background ILI noise (see Fig 2.6). In such cases, it is likely better to reject the onset than to make a very uncertain onset estimate. As epidemic strength increases, the acceptance rates for both methods increase. The threshold method rejects more onsets as the autoregression parameter $\phi$ increases. With high autoregression ($\phi = 0.75$), the threshold method both rejects more onsets and yields onset times with higher variance than the breakpoint method. Overall, this suggests that the breakpoint method is a viable alternative to the threshold method, and generally provides more accurate onset time estimates than the threshold method, especially when background incidence is autoregressive.

## 2.4 Breakpoint outbreak onset times for the autumn wave of the 2009 A/H1N1pdm influenza pandemic in the United States

In this section, ZIP-level breakpoint outbreak onset times are calculated for the autumn wave of the 2009 A/H1N1pdm outbreak in the United States. Regressions are performed between the ZIP-level outbreak onset times and ZIP population size, latitude, longitude, and distance from the apparent epicentre of the outbreak, to summarise the outbreak's overall geographic structure. Uncertainties in onset time are mapped and regressed against ZIP population sizes, revealing no systematic relationship between population size and onset time uncertainty. A bias is revealed for the breakpoint method to preferentially estimate outbreak onset times on half-weeks. Outbreak onset times are compared with outbreak peak times, and it is shown that onset times cannot be obtained by a simple shift of the peak times. Age-stratified ZIP-level onset times are calculated for the autumn 2009 A/H1N1pdm outbreak, revealing that 10-14 year-olds tend to have the earliest observable onset times. The reproduction number $R$ at the start of the autumn wave of the 2009 A/H1N1pdm outbreak is also roughly estimated.

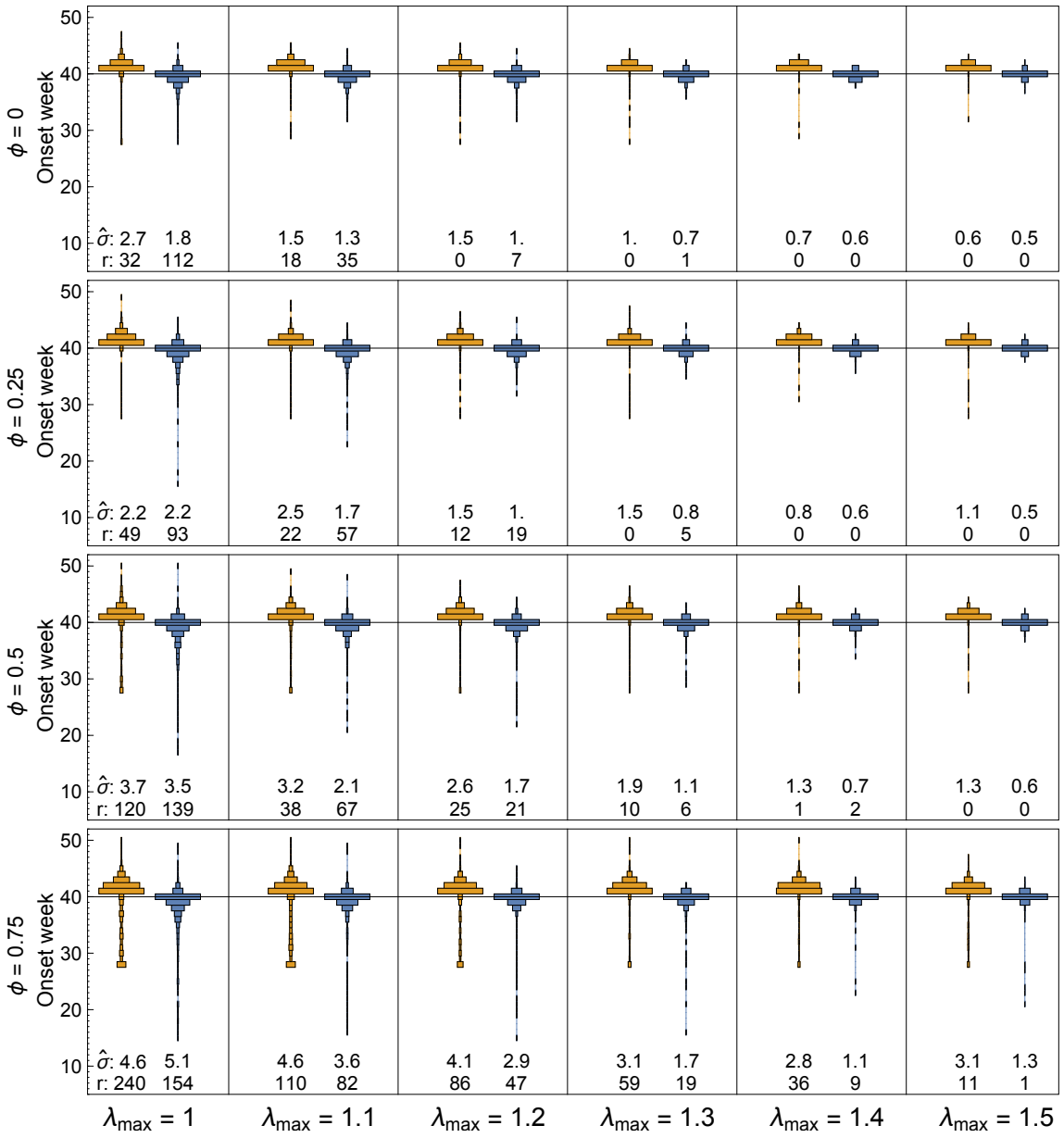



Fig. 2.7 Violin plots for the outbreak onset times estimated by the threshold (yellow) and breakpoint (blue) methods from simulated epidemics. Each column corresponds to a different epidemic strength ($\lambda_{max}$), and each row corresponds to a different level of autocorrelation ($\phi$) between consecutive incidence values. Each cell summarises the onset estimates from 1,000 simulations from model Eq 2.2. The top number at the foot of each cell gives $\hat{\sigma}$, the empirical standard deviation of the onset time distribution, with threshold on the left and breakpoint on the right. The bottom number, $r$, gives the number of onsets that were indeterminable using the method. For the threshold method, an onset is indeterminable when there are not three consecutive points in the simulation that surpass the optimal threshold. For the breakpoint method, an onset is indeterminable when the log-likelihood profile does not drop by at least two units. There is a systematic bias for the threshold to detect epidemic onset one week late. In general, the breakpoint method yields onset time estimates with lower variance than the threshold method.

### 2.4.1 Estimation of onset times

Before calculating the ZIP-level outbreak onset times for the autumn of 2009, all ZIPs with population size below 20,000 are omitted, due to the small numbers of observed ILI cases in these locations. There are 21 of these. For the remaining 863 ZIPs, the maximum ILI ratio is sought in the range between 5 July 2009 and 3 Jan 2010, depicted by the blue band in Fig 2.2. The ILI time series values for the 17 weeks prior to and including these peaks are isolated. After following the breakpoint procedure described in §2.3.1, onset estimates are obtained for 834 ZIPs. These onset times are depicted geographically in Fig 2.8. Fig 2.9 depicts the geographic locations of the 50 ZIPs that are rejected either due to their small size or to an insufficient drop in the breakpoint log-likelihood profile, as well as the ZIPs that are retained due to the alternate-peak-finding strategy.

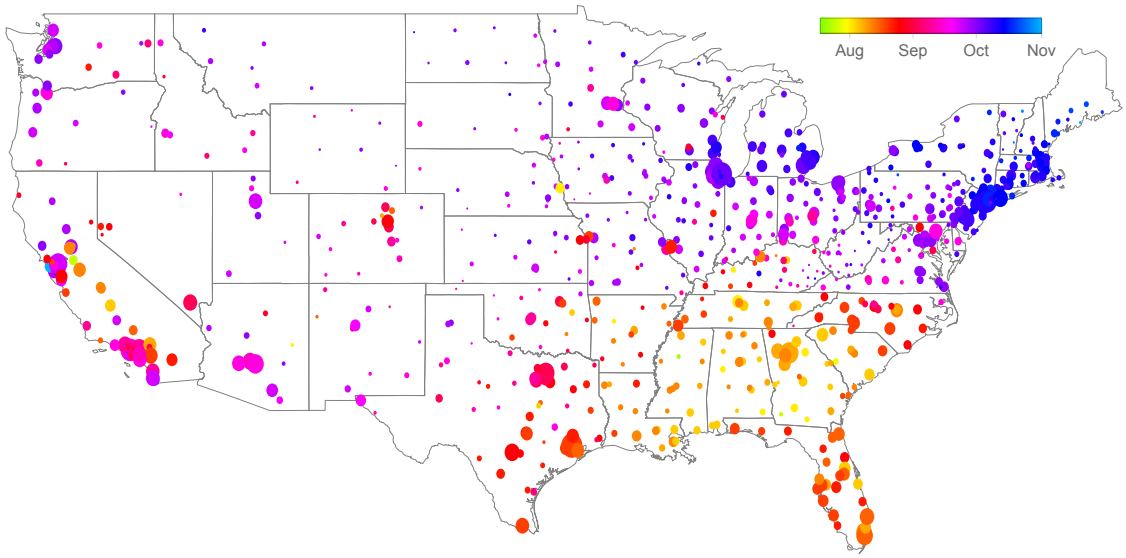



Fig. 2.8 ZIP-level outbreak onset times for 2009. Discs represent ZIPs, and disc area is proportional to the ZIP's population size. The earliest outbreaks are depicted in green/yellow, and the latest in purple/blue. A major epidemic wave appears to spread from the southeastern United States, with a possible second epidemic wave starting in the central valley of California.

The ZIP-level onset times for 2009 reveal a radial pattern of transmission, emanating from the south-eastern US. In the western half of the country, there is evidence of early transmission in the central valley of California. The epidemic lasts for about 14 weeks, with 829 (99.4%) of the 834 onsets occurring within 14 weeks of the earliest onset, in Grenada

MS. The difference between the earliest (Grenada MS) and latest (Portland ME) onset times is 19.5 weeks.

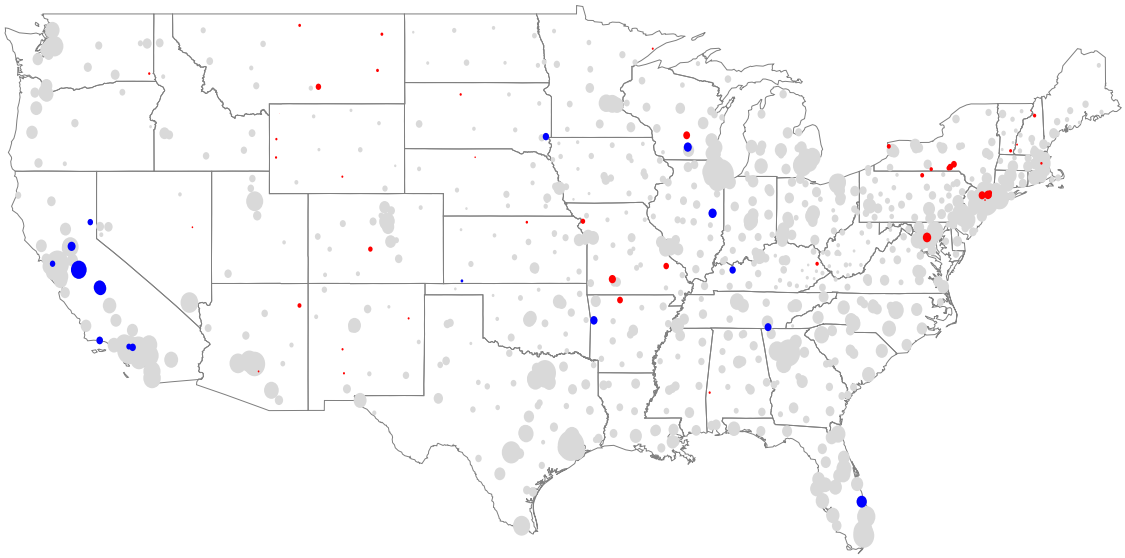


Fig. 2.9 ZIPs rejected due to low population size or high onset time uncertainty (red discs), and ZIPs retained in the analysis due to the peak-adjustment strategy (blue discs), by geographic location. ZIPs retained in the analysis with no need for peak adjustment are depicted in light grey. Disc area is proportional to the corresponding ZIP's population size. Most of the omitted ZIPs are small, and show little evidence of a geographic pattern. There are, however, two small clusters of ZIPs omitted in New York state. The relatively mild autumn outbreaks in these locations prevented a clear onset from being determined.

### 2.4.2 Investigation of the 2009 outbreak onset times and their uncertainties

Fig 2.10 depicts linear regressions between the ZIP-level autumn 2009 outbreak onset times and ZIP population size, latitude, longitude, and distance from Albany West, GA. Regressions are performed for all ZIPs (black) and also for only the ZIPs in the eastern US (red), which are here defined as the ZIPs lying in HHS regions 1-5, following Gog *et al.* (2014) [91]. Separating out the ZIPs in the eastern US helps isolate trends associated with the major pandemic wave in the eastern half of the country.

For both the eastern US and for all US ZIPs, there is a significant negative correlation ($p < 10^{-4}$) between outbreak onset times and ZIP population size. That is, more populous ZIPs tend to have earlier onsets. While this may suggest that influenza tends to arrive in large

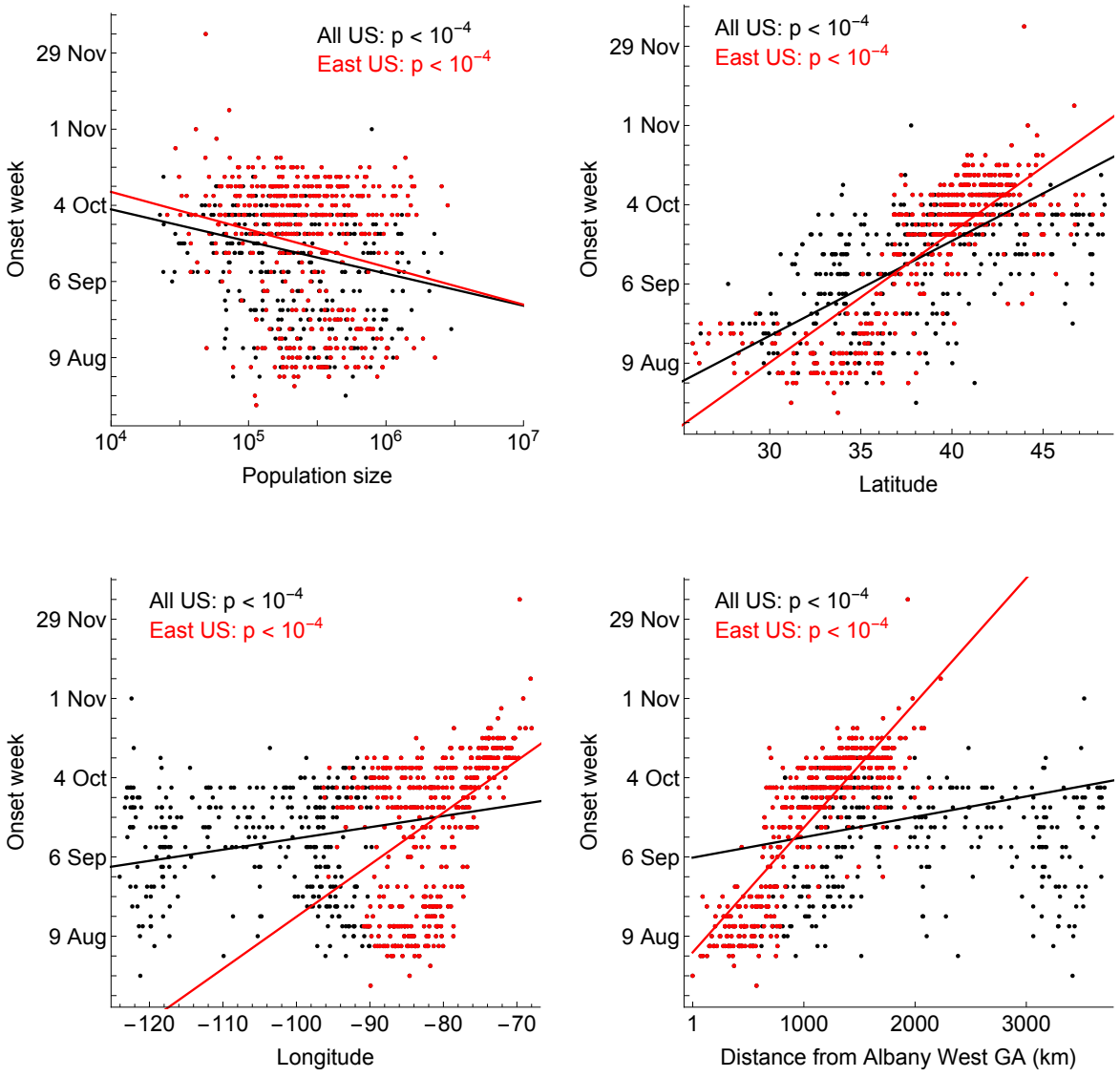



Fig. 2.10 Regressions between epidemic onset times and ZIP population size (top left), latitude (top right), longitude (bottom left), and distance from Albany West GA (bottom right). Red points correspond to ZIPs in the eastern US (HHS regions 1-5), and black points correspond to ZIPs in the western US (HHS regions 6-10). Regressions are performed for the full US (black lines) and for the eastern US (red lines). All correlations are significant ($p < 10^{-4}$). There are negative correlations between onset time and population size, suggesting that larger ZIPs tend to have earlier onsets. All other correlations are positive, indicating that the outbreak spread in a general north-easterly path, and may have had an epicentre near Albany West GA.

cities first, the trend may also be explained by the fact that there are more large ZIPs in the southern US than in the northern US. This can be demonstrated by dividing the US along 39.15 degrees of latitude, which splits the country into 418 northern ZIPs and 416 southern ones. The southern half of the US has 112 ZIPs with more than 500,000 people vs. 74 in the north, and has 34 ZIPs with more than 1,000,000 people vs. 23 in the north. Conversely, there are just 60 ZIPs with fewer than 100,000 people in the southern half of the US, and 87 in the northern US. Since the major epidemic wave in the autumn of 2009 travelled from south to north, larger ZIPs would tend to have earlier onsets simply because they lay earlier in the epidemic's path.

To demonstrate this idea more rigorously, the epidemic onset times may be regressed simultaneously on latitude and population size. That is, the coefficients of the linear model

$$y_i = \beta_0 + \beta_1 lat_i + \beta_2 pop_i \tag{2.5}$$

may be estimated using linear least squares. Here, $y_i$ represents ZIP $i$'s outbreak onset time, $lat_i$ is the ZIP's latitude, and $pop$ is the base-10 logarithm of the ZIP's population size. For both the eastern US and for all US ZIPs, the coefficients $\beta_0$ and $\beta_1$ (for the intercept and latitude) are significant ($p < 10^{-4}$), while the coefficient $\beta_2$ (for population size) is insignificant at confidence level $\alpha = 0.05$ ($p = 0.080$ for the eastern US and $p = 0.087$ for the full US). This indicates that, when accounting for latitude, ZIP population size offers little additional explanation of the variation in outbreak onset time.

There are positive correlations between epidemic onset times and both latitude and longitude, suggesting that the epidemic wave tended to travel in a north-easterly path. The slopes of the onset vs. latitude regression lines are similar for the eastern and the full US, suggesting a consistent pattern of south-to-north spread across the country. The slope of the onset vs. longitude regression line for the eastern US is perceptibly steeper than the slope for the full US, indicating that the west-to-east spread of the epidemic was strongest for the major pandemic wave in the east. The onset vs. longitude scatter also shows that very few locations in the central US had early onsets, possibly indicating that transmission started near the coasts and spread inward.

Finally, there are positive correlations between both the eastern and full-US outbreak onset times and distance from Albany West GA. Albany West is chosen as the possible epicentre of the outbreak because the main epidemic wave appears to emanate from Alabama or Georgia (see Fig 2.8), and Albany West has the earliest onset date in those states, on 26 July. The correlation is more significant for the eastern ZIPs than for the full US, again

pointing to a strong radial transmission pattern in the eastern half of the US emanating from the southeast. Epidemic epicentres for 2009 are considered more thoroughly in Chapter 4.

Next, we check for geographic patterns in onset uncertainty. Onset uncertainty is measured here as the width of the log-likelihood profile for the onset time $t_j$ (see Fig 2.3) at 1.92 log-likelihood units below the maximum value, which corresponds to a rough 95% confidence interval. These widths are depicted geographically in Fig 2.11. There are patches of high uncertainty in the vicinity of Los Angeles CA, San Francisco CA, and New York City NY. There is also a vertical band of ZIPs with elevated onset uncertainty in the central US and a patch in Kentucky. For the ZIPs in New York City, initial rises in ILI incidence were generally less pronounced than in the rest of the country, making onset detection more difficult. The weaker New York outbreaks could be due to local immunity from an earlier pandemic wave that struck the city in the spring of 2009 [91]. In California, many of the ILI time series have multiple peaks, possibly due to multiple epidemic waves passing through the state (see for example the time series for Fresno CA in Fig 2.4), which generally makes detecting onset times more difficult. The locations with high onset uncertainty in the central US lie along a boundary where population density drops sharply. Different epidemiological dynamics on the western vs. the eastern side of this boundary may have exposed these cities to multiple transmission waves, making onset estimates less precise. It is unclear what may have caused the wide onset uncertainties in Kentucky, though it appears that Kentucky lies in a region where the major epidemic wave in the eastern US slowed briefly (see §3.2.2). If outbreaks in that region were for some reason comparatively less severe than in the rest of the country, this would account for both the slowing and the elevated onset uncertainty.

Fig 2.12 provides a scatter of onset uncertainty vs. ZIP population size. There is no significant correlation ($p = 0.16$). The total range of uncertainties is high, at just over 10 weeks. However, 95% of ZIPS have uncertainties below 4 weeks and 79% of ZIPs have uncertainties below 2 weeks, so most onset estimates are fairly confident.

When interpolating onset times to the nearest half-week, the breakpoint method tends to place outbreak onset times on half weeks rather than full weeks. Fig 2.13 depicts the number of city-level outbreak outbreak onset times in each half week during the autumn of 2009. The line is jagged, with locally more onsets on half weeks than on full weeks. Interestingly, the threshold method has a similar bias in the other direction; though somewhat less consistent, there is a tendency for the threshold method presented by Gog *et al.* (2014) [91] to place onset times in 2009 on full weeks rather than on half-weeks (see Fig 2.14). One way to circumvent this bias would be to simply ignore the step of rounding to half-weeks, but the additional detail provided by interpolation arguably justifies introducing the bias. It may be

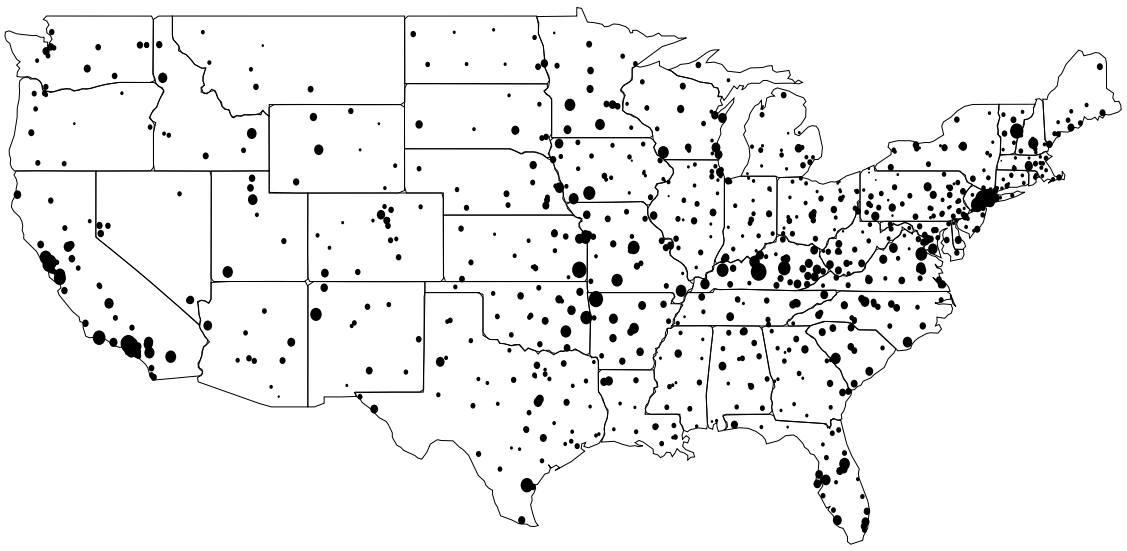


Fig. 2.11 Epidemic onset uncertainty by geographic location. Disc area is proportional to the onset uncertainty in the corresponding ZIP. Onset uncertainty is measured as the width of the breakpoint log likelihood profile (see Fig 2.3) at 1.92 log-likelihood units below the peak (same as the widths depicted in Fig 2.12). Onset uncertainty is largest in Los Angeles CA, San Francisco CA, New York City NY, in Kentucky, and along a vertical band through the central US.

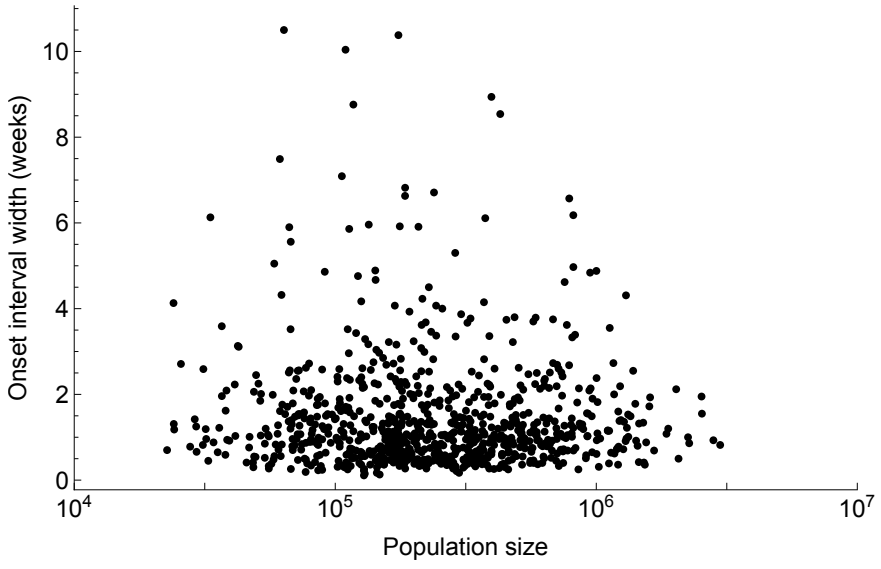



Fig. 2.12 Scatter of onset uncertainty vs. ZIP population size. Onset uncertainty is defined as the width of the breakpoint log-likelihood profile (see Fig 2.3) at 1.92 log-likelihood units below the maximum. This gives an approximate 95% confidence interval for the onset. There is no correlation ($p = 0.16$).

possible to adjust the breakpoint method to reduce this bias, perhaps by performing some sort of smoothing on the ILI time series before fitting the breakpoint regression, but for now I leave this for future work, and proceed using the breakpoint method as developed above in §2.3.1. This bias reveals itself again in Chapter 3 when we estimate temporal variation in the transmissibility of the autumn 2009 pandemic wave. The implications of the bias will be discussed further there.

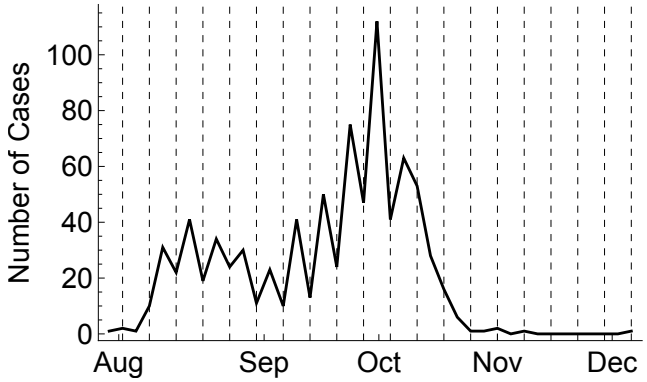



Fig. 2.13 Number of ZIP-level outbreak onsets per half-week in the autumn of 2009, as estimated by the breakpoint method. The vertical dashed lines mark full weeks. There is a bias for the breakpoint method to put epidemic onset times on half-weeks.

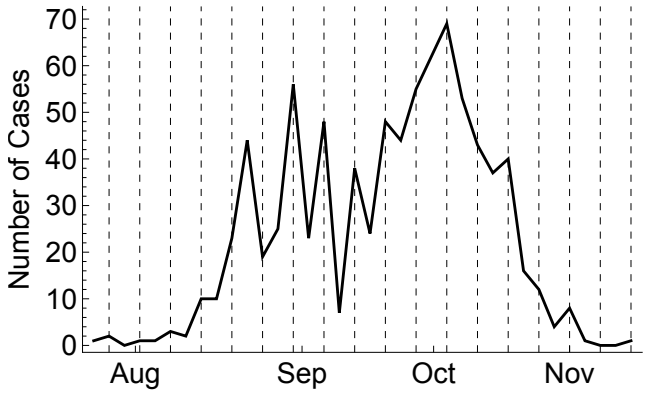



Fig. 2.14 Number of ZIP-level outbreak onsets per half-week in the autumn of 2009, as estimated by the threshold method presented by Gog *et al.* (2014) [91]. The vertical dashed lines mark full weeks. There is a moderate bias for the threshold method to put epidemic onset times on full weeks, though there are some exceptions to this, as in the large half-week spike in late August.

Finally, we compare outbreak onset times with outbreak peak times, to check whether there is a simple relationship between the two. Fig 2.15 depicts the difference between

the maximum-likelihood outbreak onset time and the outbreak peak time in each ZIP. The onset-to-peak time varies geographically, with ZIPs in a band separating the southeastern and northeastern US, as well as in coastal California and in New York City, tending to have longer intervals between onset and peak times. The long onset-to-peak intervals in the horizontal band across the eastern US roughly matches with a region where the major epidemic wave in the eastern US slowed briefly (see §3.2.2). The long onset-to-peak intervals in New York may be due to underlying immunity from the spring wave of infection in that city. The long onset-to-peak intervals in California may be due to multiple epidemic waves affecting that state, which could cause especially long intervals between the onset of the first wave and the peak of the most severe wave. Fig 2.16 depicts the distribution of times between outbreak onset and peak. In most cases, the peak time follows the onset time by 3-4 weeks. The distribution is wide however, ranging from 0.5 to 12 weeks, suggesting that onset times cannot be obtained from a simple time shift of the peak times. In Fig 2.17, the difference between peak time and onset time is regressed against ZIP population size. There is a slight but significant positive trend ($p < 0.01$), suggesting that larger ZIPs tend to have longer spans of time between outbreak onset and peak. Overall, this suggests that peak time tends to roughly follow a few weeks behind outbreak onset time, with a longer time interval for larger ZIPs. However, there does not appear to be a reliable way to infer epidemic onset times from peak times.

### 2.4.3 A rough calculation of $R$ at the start of the autumn 2009 pandemic wave

Using the IMS-ILI data, it is possible to make a rough estimate of the reproduction number $R$ at the start of the autumn wave of the 2009 A/H1N1pdm pandemic in the US. Following Wallinga and Lipsitch (2007) [243], the value of $R$ at the start of an outbreak may be estimated from the initial exponential rate of increase in incidence $r$ and the disease's mean generation interval $T_c$. If the disease is assumed to follow traditional SIR dynamics described by a system of ordinary differential equations, then the generation interval must follow an exponential distribution [243], and an estimate for $R$ is given by

$$R_{exp} = 1 + rT_c. \tag{2.6}$$

On the other hand, an upper bound estimate for $R$ may be obtained by assuming that all secondary infections occur exactly $T_c$ time units after the onset of the infection that caused them. In this case, the generation interval distribution is a delta function centred at $T_c$ [243].

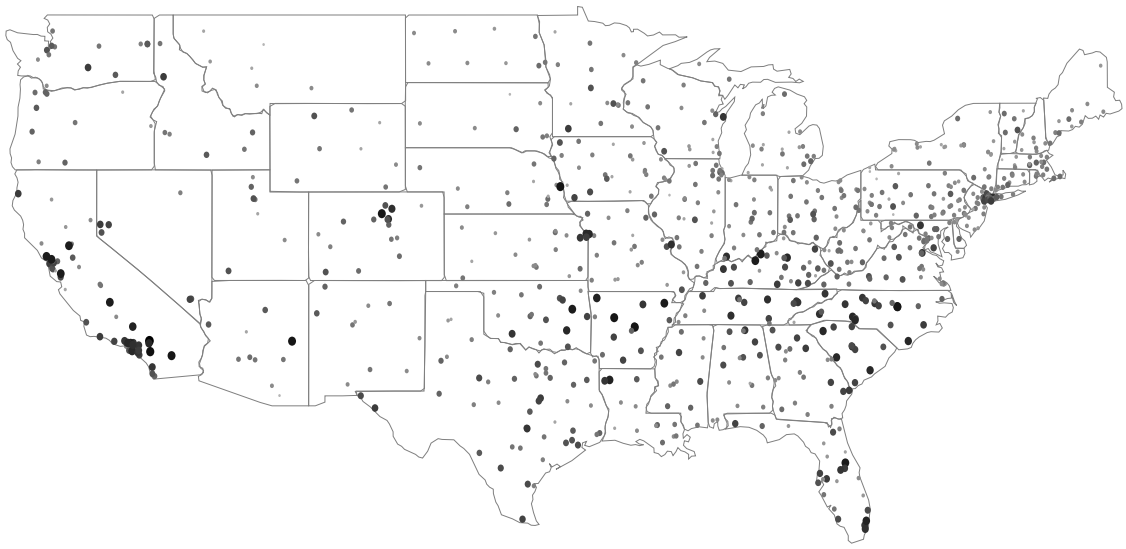


Fig. 2.15 Difference between epidemic peak time and epidemic onset time by geographic location. Disc area is proportional to the size of the difference. The largest time intervals between onset and peak are found in California and in the mid-Atlantic states.

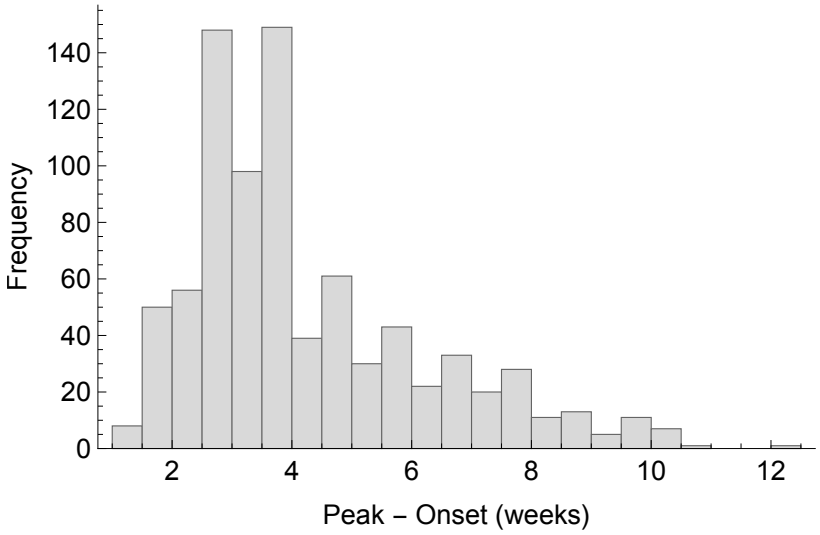



Fig. 2.16 Histogram of the difference between epidemic peak time and epidemic onset time in 2009. The distribution is wide, suggesting that epidemic peak times cannot be used as a simple proxy for epidemic onset times.

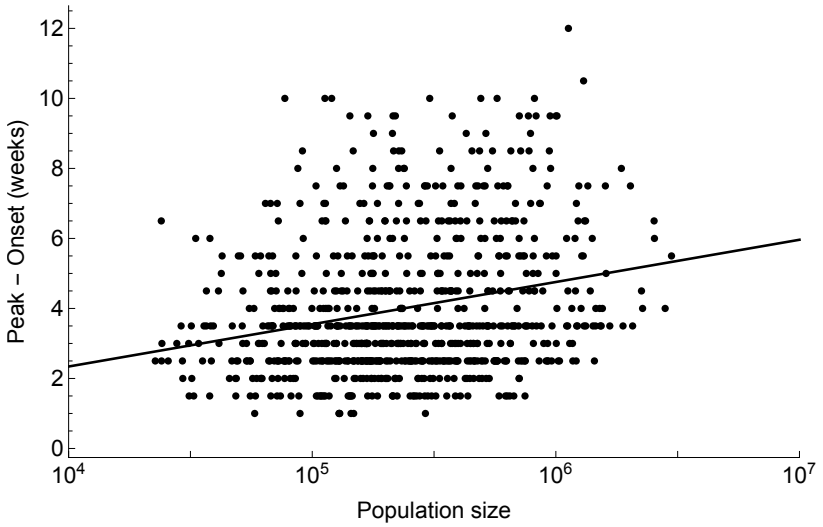



Fig. 2.17 Scatter of the difference between epidemic peak and epidemic onset time vs. ZIP population size. There is a significant trend ($p < 0.01$), with larger ZIPs tending to have longer intervals between onset and peak.

An estimate for this upper bound is given by

$$R_{max} = e^{rT_c}. \tag{2.7}$$

To estimate $R$ at the start of the autumn 2009 A/H1N1pdm pandemic wave, a linear regression is fit to the logarithm of each ZIP's ILI time series from the onset week to either four weeks past the onset, or the epidemic peak, whichever comes first. The slope of this line provides an estimate of $r$ for each ZIP. There are many estimates of the generation interval for 2009 A/H1N1pdm influenza, 17 of which are listed in a review by Boelle *et al.* (2011) [23]. Six of these estimates are made using data from the US. Table 2.3 provides the median $R_{exp}$ and $R_{max}$ across all ZIPs calculated using these six generation interval estimates and the initial rates of increase $r$. All lie near Yang *et al.*'s (2015) [258] estimate of $R = 1.24$ at the beginning of the autumn wave of the 2009 pandemic in the US. There is no clear geographic pattern in the $R$ estimates.

### 2.4.4 Age-stratified autumn 2009 onset times

Outbreak onset times can also be calculated by age group in 2009. To do so, epidemic peaks are sought in the age-stratified IMS-ILI data within the same timespan as for the age-aggregated case, from 5 July 2009 – 3 Jan 2010, and the ILI ratios for the 17 weeks

Table 2.3 Median $R_{exp}$ and $R_{max}$ (Eq 2.6 and 2.7) across all ZIPs using five different estimates of the mean generation interval $T_c$ for 2009 A/H1N1pdm influenza in the US

| $T_c$ | Source of $T_c$ estimate | $R_{exp}$ (IQR) | $R_{max}$ (IQR) |
|---|---|---|---|
| 2.6 | Cauchemez *et al.* [32], White *et al.* [247] | 1.17 (1.13, 1.22) | 1.18 (1.14, 1.24) |
| 2.7 | Lessler *et al.* [147] | 1.18 (1.14, 1.23) | 1.19 (1.15, 1.25) |
| 3.0 | France *et al.* [80] | 1.20 (1.15, 1.25) | 1.22 (1.16, 1.28) |
| 3.4 | Morgan *et al.* [167] | 1.22 (1.17, 1.28) | 1.25 (1.19, 1.33) |
| 4.4 | Yang *et al.* [259] | 1.29 (1.22, 1.37) | 1.33 (1.25, 1.44) |

prior to and including the peak are isolated. Onset times are estimated using the breakpoint method. Fig 2.18 depicts these outbreak onset times by age group. The radial transmission pattern observed in the age-aggregated onsets (Fig 2.8) is apparent in the age-stratified onsets as well, though it becomes more difficult to detect in the oldest two age groups, which tend to have very noisy ILI time series and therefore many undetectable outbreak onset times.

Previous studies suggest that school-aged children tend to suffer the highest burden of influenza infection during outbreaks [81, 179, 245, 252, 259]. This may lead to earlier detectable onsets on those age groups as well. Fig 2.19 depicts the the number of ZIPs for which each age group has the earliest onset. If two age groups have tied outbreak onset times, both are counted as having the earliest onset in that ZIP. The 10-14 year-old age group tends to have the earliest detectable outbreak onset times. More generally, infants and school-aged children (ages 0-19) tend to have earlier onset times than adults and the elderly (ages 20+). Chapter 5 presents a more thorough analysis of the relative roles of age groups during the spread of the 2009 A/H1N1pdm influenza pandemic in the US.

## 2.5 Geographic data

The US Census Bureau's Gazetteer files provide coordinates and population sizes for ZCTAs, which roughly correspond to five-digit ZIP codes [233]. Since the US census is only conducted every ten years, the 2009 population sizes included in the Gazetteer files are intercensal estimates made by the US Census Bureau that account for estimated births, deaths, and migrations [235]. To identify coordinates for the 3-digit ZIP codes, the population-weighted centre of mass is calculated for all 5-digit ZIPs within a 3-digit ZIP code. The distribution of 3-digit ZIP population sizes is depicted in Fig 2.1, and the coordinates and

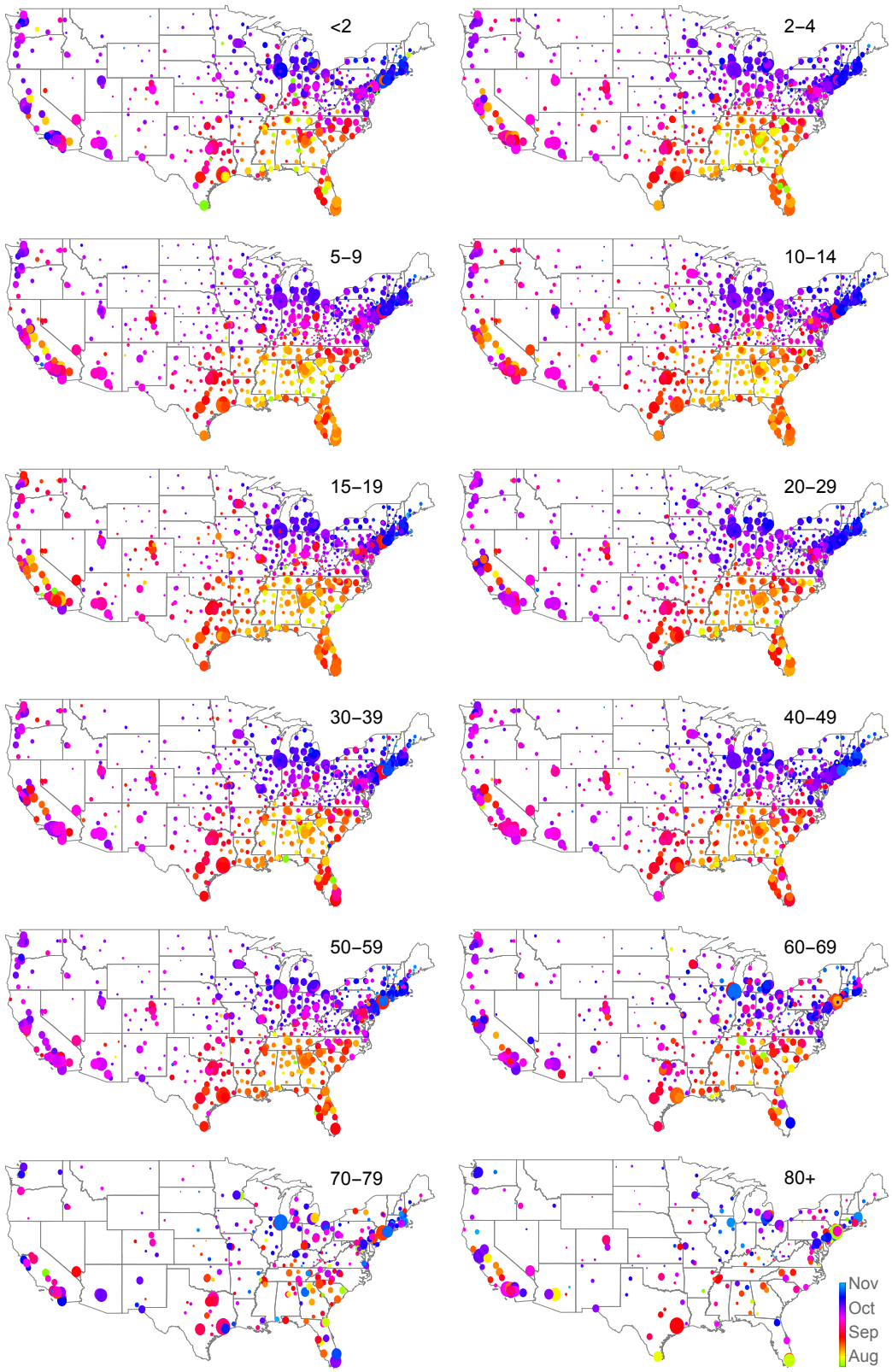



Fig. 2.18 ZIP-level outbreak onset times for the autumn 2009 pandemic wave, stratified by age group. Disc area is proportional to the ZIP's population size. ZIPs are only depicted where where a reliable onset estimate could be obtained for the age group. The radial pattern is visible in all age groups, though the time series for the two oldest age groups are too noisy to obtain many reliable onset estimates.

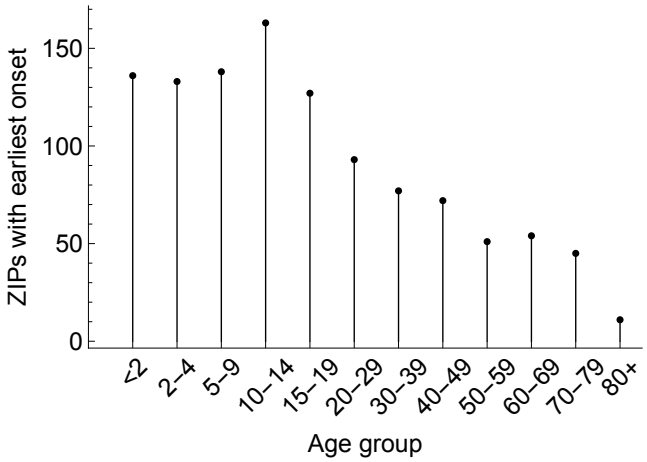



Fig. 2.19 Frequency with which each age group has the earliest onset in a given ZIP. If two age groups have a simultaneous onset in a ZIP, the frequency count is increased by 1 for both. Children (0-19) tend to have the earliest onsets. The 10-14 year-old age group has the earliest onset time most frequently, in 163 of the 884 ZIPs.

populations sizes are depicted geographically in Fig 2.8, with the autumn 2009 ZIP-level outbreak onset times.

## 2.6 Schools data

Data on school start dates in the autumn of 2009 are available at the state level from Chao *et al.* (2010) [47]. In Alabama, Florida, Georgia, Mississippi, and South Carolina, the five states near the apparent epicentre of the autumn 2009 outbreak in the eastern US, school start dates are available at the finer district level, also from [47]. Most ZIPs contain multiple school districts, so ZIP-level school start dates are defined to be the median of all district start dates within that ZIP. The model fits in later chapters do not change appreciably when the mean start date or earliest start date are used instead.

Fig 2.20 depicts ZIP-level median school start dates in 2009 by geographic location. Schools open earliest in the southeastern US and latest in the northeastern US, roughly anticipating the trajectory of the autumn 2009 pandemic wave. However, the pandemic wave proceeded much more slowly than the opening of schools; all schools were open within a span of about six weeks, while the epidemic wave took over 14 weeks to spread from the southern US through Maine. Fig 2.21 depicts a linear regression of epidemic onset times on school start dates. The trend is positive and significant, indicating an association between school start dates and epidemic onset times. However, again, an increasing lag between

school start dates and outbreak onset times is apparent from the regression, suggesting that outbreak onsets cannot be explained simply by the opening of schools.

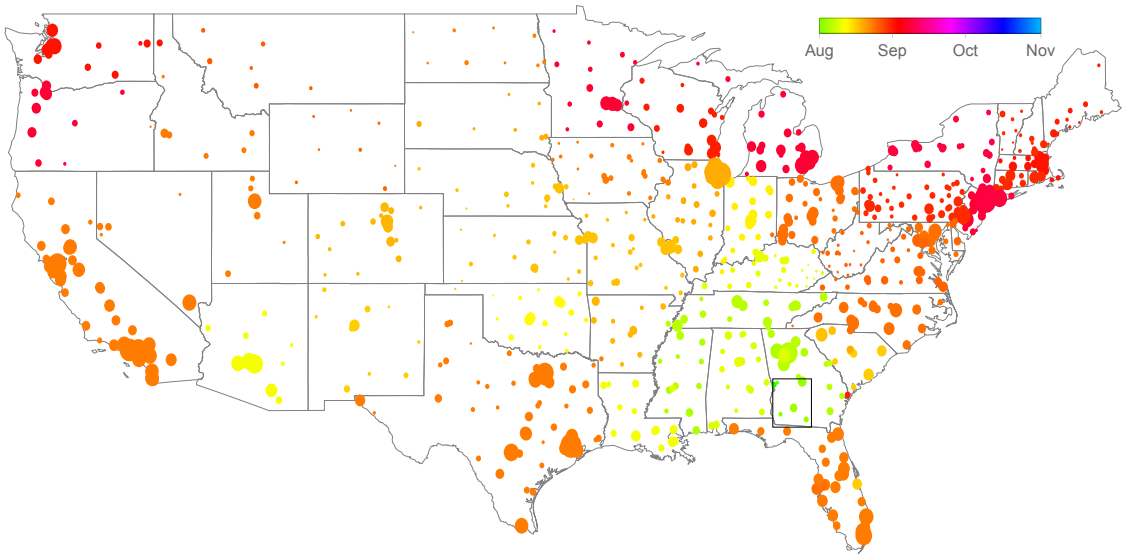



Fig. 2.20 Median school start date by geographic location. Discs represent ZIPs, and disc area is proportional to population size. The colour scale matches the one in Fig 2.8, to allow comparison between school start dates and outbreak onset times in 2009. Schools tend to open earliest in the southeast and latest in the northeast, in a radial pattern that roughly anticipates the spread of the 2009 pandemic. However, the pandemic wave spread much more slowly than the opening of schools; in Maine, there is roughly a two-month lag between the school start date and the state-averaged outbreak onset time. A cluster of six ZIPs in southern Georgia (boxed) had the country's earliest school start dates in 2009, on or before 6 Aug 2009.

## 2.7 Antigenic data

The CDC publishes weekly virologic data for circulating influenza strains through its FluView portal, with records going back to 1997 [45]. These data consist of the weekly number of laboratory-confirmed observations of four antigenic subtypes of influenza (A/H1N1pdm, A/H1, A/H3, and B) for each HHS region. The data are collected from approximately 100 public health laboratories and 300 clinical laboratories across the US [43]. Clinical laboratories are generally only equipped to identify the type of influenza virus (A or B), but not the subtype, so some of the CDC virologic samples are tagged as 'A/Unsubtyped' [43]. The number of laboratory-observed observations of each subtype by week and HHS region

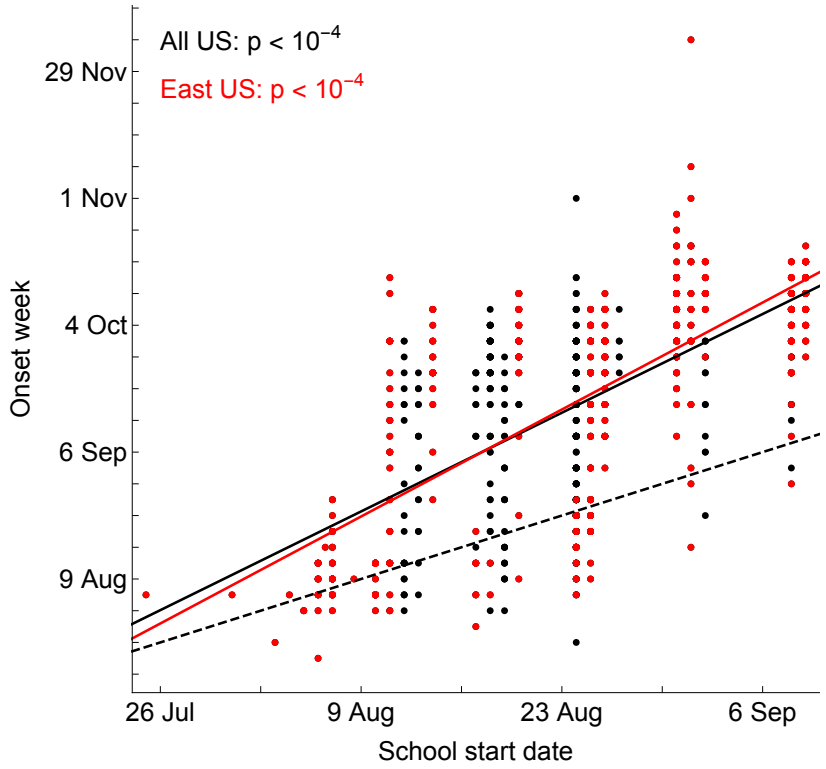



Fig. 2.21 Scatter of outbreak onset times vs. school start dates. Red points correspond to ZIPs in the eastern US (HHS regions 1-5), and black points correspond to ZIPs in the western US (HHS regions 6-10). The black regression line is fit to all points, and the red regression line is fit just to those points corresponding to ZIPs in HHS regions 1-5. There is a significant positive correlation between outbreak onset times and school start dates ($p < 10^{-4}$) both for the full US and for the eastern ZIPs. However, the difference between onset times and school start dates increases as the epidemic progresses. The dashed line depicts equivalence between outbreak onset times and school start dates. Early in the outbreak, the regression trend lines lie close to this equivalence line, suggesting that outbreak onsets and school start dates were closely associated at the start of the autumn 2009 pandemic wave. As the outbreak progressed, however, the discrepancy between onset times and school start dates increased.

from 2001-2010 are depicted in Fig 2.22. In 2009, transmission was dominated by subtype A/H1N1pdm09. Because the outbreak was dominated by a single antigenic subtype, there are no detectable geographic patterns in antigenic prevalence in 2009. Virologically-based geographic analyses of 2009 pandemic influenza transmission therefore must generally rely on more detailed genetic information [160, 177]. Investigation of influenza genomic data lies beyond the scope of this thesis, but remains an important area for future research [241]. For other influenza seasons, such as the 2007-08 seasonal outbreak, multiple antigenic subtypes co-circulated, making geographic inferences possible using antigenic data. This will be considered further in Chapter 6.

## 2.8 Discussion

This chapter's primary focus is on a dataset that captures weekly influenza-like illness (ILI) incidence for 12 age groups in 884 locations across the US between 2001 and 2009. The data are available from medical claims records maintained in the private sector. To my knowledge, they provide the finest geographic resolution of ILI data ever considered for studying the spatial spread of influenza in the United States. Viboud *et al.* (2014) [240], Gog *et al.* (2014) [91], and Charu *et al.* (2017) [48] use the same ILI data aggregated to a coarser geographic resolution, with the number of geographic locations ranging between 200 and 400. The only other study of which I am aware that considers a comparable number of locations in the US is provided by Yang *et al.* (2015) [258], who estimate epidemiological parameters for 10 influenza seasons using ILI data from 115 US cities. A number of surprising epidemiological characteristics of the 2009 A/H1N1pdm influenza pandemic in the United States have already been pointed out in other studies, including the unusually early timing of the autumn wave [122], unusually high rates of morbidity and mortality among young adults [122, 198], and an unusually slow and coherent geographic wave of transmission that seems to have been seeded, unexpectedly, in the southeastern US [91]. The fine-scale data considered in this thesis provide an opportunity to study this unconventional outbreak in close detail.

A number of other studies characterise the spread of influenza at the country scale using spatially-resolved ILI data. In addition to the studies just mentioned by Gog *et al.* (2014) [91], Charu *et al.* (2017) [48], and Yang *et al.* (2015) [258], Chowell *et al.* (2011a) [50] study the geographic transmission of influenza in Peru across 134 provinces using a combination of ILI and laboratory-confirmed data. Chowell *et al.* (2011b) [49] use a combination of ILI and mortality data to describe the geographic transmission of the 2009 A/H1N1pdm pandemic in Mexico. Smieszek *et al.* [218] use sentinel ILI data from Switzerland to model

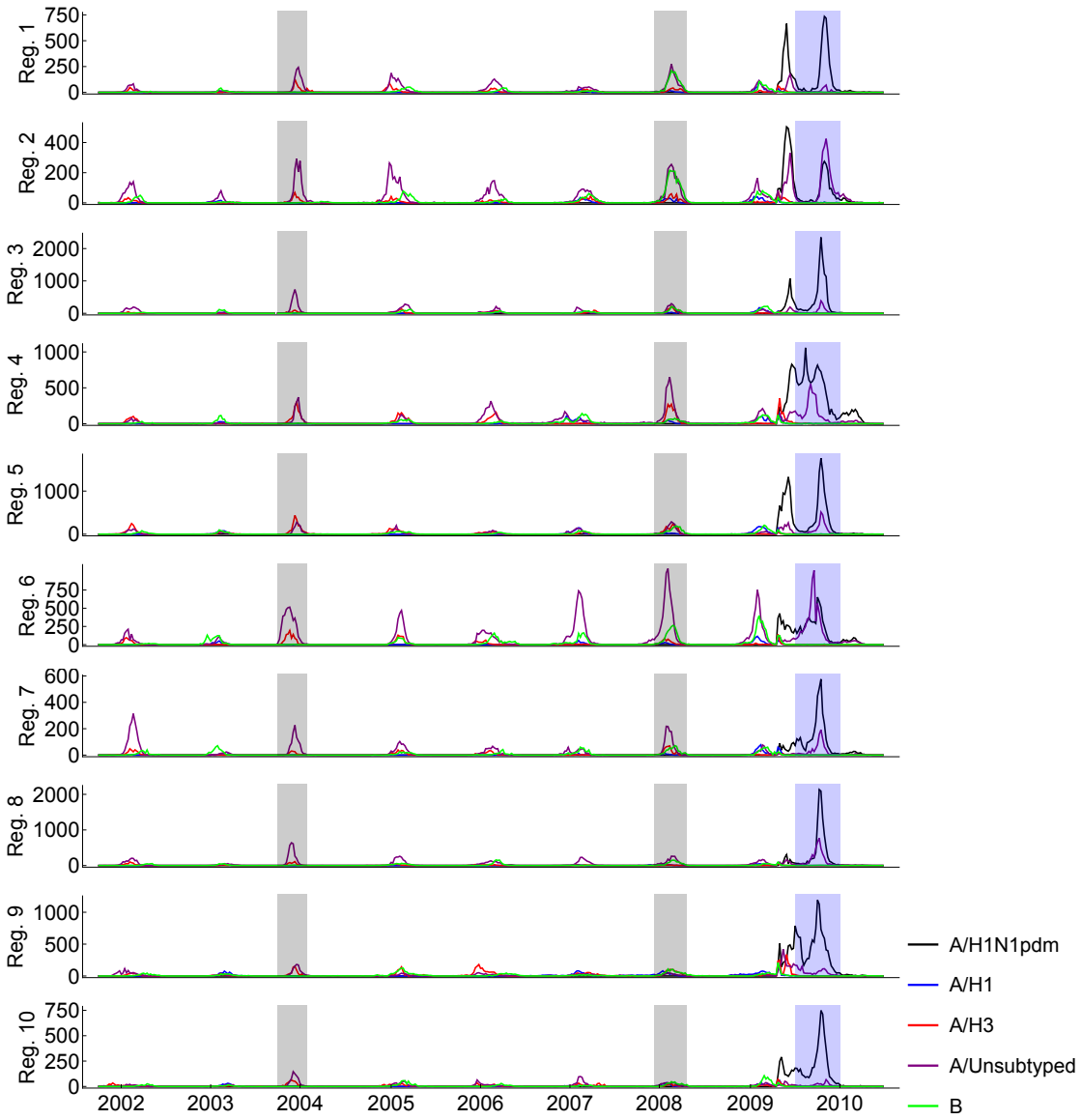



Fig. 2.22 Weekly number of laboratory-confirmed cases of each antigenic subtype collected by the CDC between 2001 and 2010. The blue shaded region corresponds to the autumn wave of the 2009 A/H1N1pdm influenza pandemic, and the grey shaded areas correspond to the 2003-04 and 2007-08 seasonal outbreaks, to be discussed in Chapter 6. The autumn wave of the 2009 pandemic was dominated by the A/H1N1pdm strain.

the geographic transmission of the 2003-04 influenza outbreak in that country. Paget *et al.* (2007) [184] use sentinel ILI data and virological data aggregated to the country level to describe the transmission of eight seasonal influenza outbreaks across Europe. All of these find significant differences in outbreak timing by geographic location, motivating the study of the geographic transmission of influenza outbreaks using ILI data.

Electronic medical claims records offer a promising source of disease surveillance data, especially in the context of influenza [240, 260]. In particular, they appear to provide more finely-resolved information about the geography and age structure of influenza outbreaks than traditional surveillance methods can, while improving upon the accuracy of social media- and search query-based ILI estimates [182, 240]. However, the electronic medical record data stream still carries a number of limitations. As mentioned in §2.1.2, differing local incentives and coding practices can reduce the reliability of medical claims data [248]. Conflicting incentives can be problematic at the overall health-system level, too; in the United States, the ownership of health-related data in the private sector can drive up the cost of access, making electronic medical records, while easy to collect in theory, sometimes very difficult to obtain. In addition, privacy concerns rightly place a limit on the resolution with which medical claims data can be reported, so that some degree of aggregation will always be necessary. Not all aggregation strategies are the same, however, and more research is required to understand how to achieve aggregation with a proper balance between privacy, ease of coding, epidemiological relevance, and pertinence for intervention strategies.

Aside from the particular difficulties associated with electronic medical claims data, ILI itself is an imperfect measurement of influenza incidence. ILI incidence is normally reported as a proportion of physician visits due to influenza-like illnesses, which may not correspond to the *per capita* incidence of ILI [258]. This makes it difficult to estimate population-level influenza intensity from ILI data. Indeed, Viboud *et al.* (2014) [240] find that, while outbreak timing in the IMS-ILI data correlates highly with outbreak timing in the CDC's ILI and virologic data, correlations in outbreak intensity between the datasets, measured as total additional ILI after subtracting out a sinusoidal baseline, are significantly lower. This provides a rationale for focusing on outbreak timing when considering the IMS-ILI data, and motivates this chapter's focus on developing a robust algorithm to detect epidemic onset times from ILI incidence time series.

The breakpoint method, originally introduced by Charu *et al.* (2017) [48] and presented in updated form here, appears to offer a robust means of identifying outbreak onset times from noisy, potentially autocorrelated time series of ILI incidence. A major advantage of the breakpoint method is that it avoids a need to define baseline and threshold levels of ILI

activity, which generally must be done in an *ad hoc* way [248]. The adjustments to the breakpoint method introduced in this chapter include (1) fitting the breakpoint regression to a fixed number of time series points for all locations, to ensure that onset uncertainties are comparable between locations, (2) introducing a strategy to measure onset uncertainty using the likelihood profile of the breakpoint estimate, (3) using that onset uncertainty estimate as a criterion for accepting or rejecting a time series from analysis, rather than simply rejecting the 20% of locations with the smallest differences between maximum and minimum ILI intensity, and (4) introducing a strategy for identifying accurate onset times from time series with multiple incidence peaks. This chapter also presents the first systematic evaluation of the breakpoint method's performance. According to this analysis, the breakpoint method performs best when the ILI time series has a clear, sudden rise in incidence at the outbreak onset time, though its performance is still good in noisier settings. Most ZIPs in the US have a sharp increase in ILI activity at the beginning of the autumn wave of the 2009 A/H1N1pdm pandemic, so the breakpoint method should give fairly accurate onset estimates for that epidemic. The breakpoint method can also generally detect epidemic onset times with higher precision and accuracy than an optimised threshold method, especially when autocorrelation between subsequent incidence values is high. Interpolating the breakpoint onset time estimates to the nearest half-week introduces a bias, with relatively more onsets estimated to occur on half weeks than on whole weeks. However, the overall trajectory of the autumn wave of the 2009 A/H1N1pdm influenza pandemic that becomes visible when mapping the breakpoint onsets (Fig 2.8) matches well with the patterns observed by Gog *et al.* (2014) [91] and Charu *et al.* (2017) [48], the only other studies of which I am aware that provide detailed pictures of the geographic transmission of the autumn 2009 pandemic wave in the US.

The breakpoint method was not tested for its ability to detect outbreak onset times in real time for an ongoing outbreak. Since the breakpoint method relies on the full epidemic time series prior to and including the peak, it is possible that when presented with less data, such as before the epidemic has peaked, a threshold-based method might perform equally well or better. The breakpoint method might be prone to detecting frequent spurious onsets due to stochastic rises in ILI that yield 'false' peaks. Also, to use the breakpoint method in real time, one would have to decide upon an acceptable width for the onset likelihood profile, below which an onset would be identified. This seems to simply push the task of defining a threshold to a higher level of abstraction, which may not ultimately be helpful. It is therefore unlikely that the breakpoint method will contribute to real-time outbreak onset detection, but further work might still be warranted in this area.

Statistical analysis of the breakpoint outbreak onset times offers insight into the geographic transmission of the autumn 2009 A/H1N1pdm influenza pandemic wave in the US. In general, and in agreement with findings from Gog *et al.* (2014) [91] and Charu *et al.* (2017) [48], the outbreak featured a major geographic transmission wave that spread from the southeastern US. Wave-like geographic transmission patterns for influenza at the continent and country scales have been reported in a few of other studies as well: Paget *et al.* (2007) [184], for example, find evidence of west-to-east and south-to-north spread of influenza across Europe in four seasons between 1999 and 2007, and Smieszek *et al.* (2011) [218] report a north-easterly spread of influenza across Switzerland in 2003. As may be expected by the relatively high incidence of influenza in children [81, 179, 245], outbreak onset times from the autumn of 2009 in the US are generally first detectable in school-aged children, with 10-19 year-olds leading the estimated outbreak onset times in more ZIPs than any other age group. At the full US scale, it appears that the overall geographic transmission pattern of the autumn 2009 pandemic wave was fairly consistent between age groups, with all age groups showing evidence of a transmission wave spreading from the southeastern US. The geography and timing of the start of the pandemic wave may be associated with the relatively early start of the school term in the southeastern US. This would agree with a number of previous studies that have identified schools as key sites of transmission during influenza outbreaks [117, 199, 255]. However, the opening of schools cannot fully explain the onward spread of the autumn 2009 pandemic wave, since the transmission wave lagged well behind the start of the school term in many ZIPs, especially in the northern US. The themes uncovered by these statistical analyses constitute major areas of focus for the rest of this thesis: Chapter 3 examines potential contributors to the geographic transmission of the autumn 2009 A/H1N1pdm outbreak in the US, including the role of schools; Chapter 4 considers the geographic establishment sites of the outbreak; and Chapter 5 takes a closer look at how different age groups may have contributed to both sparking and sustaining transmission.

## 2.9 Summary

This chapter presents the key datasets that underpin this thesis' findings. Special focus is placed on a set of geo-tagged influenza-like illness (ILI) data obtained from electronic medical claims records. This dataset captures weekly ILI incidence in 884 3-digit ZIP (postal) codes across the US between 2001 and 2009. An onset detection algorithm is presented, and its performance is evaluated using epidemic simulations. Epidemic onset

times are calculated for the autumn 2009 pandemic wave in the US, revealing a radial pattern of geographic transmission from a likely epicentre in the southeastern US. This overall geographic transmission pattern remains visible when the ILI data are stratified into 12 age groups. An estimate of the reproduction number $R$ at the start of the autumn wave of the 2009 A/H1N1pdm pandemic aligns closely previous estimates. Additional datasets are introduced that provide the coordinates, population sizes, and school start dates for the 884 3-digit ZIPs covered by the ILI dataset. While the start of the autumn wave of the 2009 A/H1N1pdm outbreak in the US roughly coincided with the opening of schools in the southeast, the discrepancy between school start dates and local epidemic onset times increased as the epidemic spread, suggesting that schools alone cannot explain the geographic transmission of the outbreak.

# Chapter 3

# A geographic model of between-city influenza transmission

In this chapter, a mechanistic mathematical model is presented to describe the between-city geographic transmission of influenza in the United States. The model is fit to city-level outbreak onset times from the autumn wave of the 2009 A/H1N1pdm influenza outbreak in the US. Analysis of the best-fit model reveals that a given city's outbreak onset time was likely influenced by its population size, the surrounding population density, its geographic distance from infected locations, and possibly the mixing of children in schools within the city. To flexibly account for additional unobserved factors that may have influenced transmission, the model's transmissibility parameter is allowed to vary in time and space according to a Gaussian process. The optimal Gaussian process fits reveal a probable region of increased transmissibility in the southeastern United States at the beginning of the autumn wave of the 2009 A/H1N1pdm pandemic.

## 3.1 Background

### 3.1.1 The gravity model

Metapopulation models provide a useful framework for describing the transmission of disease between discrete geographic sub-populations (see §1.3.2). Modelling disease dynamics on metapopulations requires specifying how members of the host species move between sub-populations, or patches. A key challenge therefore consists in modelling how an individual migrating from a given patch chooses a destination. Ideally, this is done with empirical mobility data. For humans, this information can sometimes be obtained from survey-based

commuting data or mobile phone geo-tags [239, 246]. In many cases, however, such data are unavailable or lacks sufficient detail. In the United States, privacy laws prohibit telecommunications companies from sharing mobile phone locations with third parties. Commuting data, which are available from the US Census Bureau, could be used instead. But, if the bulk of transmission occurs between unemployed individuals, such as children, the relevance of commuting data may be limited [91, 239].

The gravity model provides a simple and well-tested alternative for describing the movements of individuals between metapopulation patches. Introduced by Zipf in 1946 [261], the model states that the frequency of trips between two sub-populations is related to the product of the sub-populations' sizes and to the distance between them. In one common form, the gravity model describes the amount of movement $m$ between two sub-populations $i$ and $j$ as

$$m_{i,j} = cN_i^{\alpha} N_j^{\beta} \kappa(d_{i,j}) \tag{3.1}$$

where $N_i$ and $N_j$ are the population sizes of locations $i$ and $j$, $\kappa$ is a decaying function of the distance $d_{i,j}$ between the two locations (power and exponential curves are common choices), $\alpha$ and $\beta$ modulate the relative importances of the "donor" and "recipient" population sizes respectively, and $c$ is a scaling factor. When $\alpha = \beta = 1$ and $\kappa(d_{i,j}) = 1/d_{i,j}^2$, the equation has the same form as Newton's law of gravitation, giving the model its name.

Zipf originally developed the model, naming it the "P1 P2 / D hypothesis", as a heuristic description of the between-city movement of individuals in the United States [261]. Later, the model was set on sounder theoretical footing, most notably by Wilson (1970) [250] and Batty and Sikdar (1982) [14–17]. Both theoretical treatments use an entropy maximization approach to derive the model. The model in its basic form (Eq 3.1) has some clear inconsistencies: it is possible for the flux between a very small and a very large population to exceed the size of the small population, and doubling the sizes of two populations quadruples the movement between them, eventually turning all (or more than all) of the inhabitants of the cities into commuters. To circumvent these problems, Wilson's and Batty and Sikdar's formulations introduce unique balancing coefficients for each city. In practice, however, the simpler version of the gravity model is normally used. As long as the model is parametrised for the same metapopulation on which it is used, and the sub-populations do not grow or shrink significantly, the apparent inconsistencies can be safely ignored. Some recent examples that use the gravity model to describe the geographic transmission of disease in humans are given in [91, 164, 231, 239, 256].

There is at least one important alternative to the gravity model. The radiation model, introduced by Simini *et al.* (2012) [215], is intended to describe how workers commute between cities. The model is founded on two basic assumptions: that the number of jobs available in a city is related to the city's population size, and that an individual will always choose the closest job to her/his home that offers higher benefits than the best job available in her/his own city. The model can be formulated in terms of the equations that describe particle radiation and absorption. The model is parameter-free, which may be seen as both an advantage and a disadvantage: in the complete absence of data, the model can still be used, but when data are available, the model cannot adapt even if the fit is poor. A radiation model was tested as an alternative to the gravity model kernel discussed below. However, simulations from that model were incapable of reproducing the wave-like structure observed during the autumn 2009 A/H1N1pdm pandemic wave in the US, due to too-frequent long-distance jumps of infection between major population centres. All attempts to reproduce the qualitative behaviour of the geographic spread of the autumn 2009 A/H1N1pdm pandemic wave using the radiation model were unsuccessful.

### 3.1.2 Survival analysis

Colonisation and extinction are the two key processes that underlie ecological dynamics on metapopulations [106]. In traditional metapopulation theory, they are discrete binary events: a given patch is either extinct or is colonised, and the switch from one state to another happens instantaneously. A common assumption states that patches may only transition once, from colonised to extinct or vice-versa. This assumption roughly holds true when modelling non-recurrent ecological invasion waves, as are often observed for acute infectious diseases, for which colonisation corresponds to infection. Survival analysis, which is the statistical theory that describes the expected waiting time for sudden, permanent events to occur, offers a natural mathematical framework for studying such systems.

In survival analysis, the time at which an irreversible event occurs is a random variable with distribution uniquely specified by a hazard function $h$. The hazard function expresses the time-varying rate at which the event occurs, conditional upon the event not occurring prior to that time. Normally, time is taken to be a continuous variable. For epidemiological modelling, however, observations are often binned into regular time intervals, making it more natural to treat time as a discrete variable. To make the standard survival-analytic framework amenable to epidemiological data, then, it is necessary to discretise time. In particular, we

seek an expression for the probability of onset occurrence at a given discrete time $t_j$, in terms of some underlying continuous process. The following derivation is adapted from [202]:

Consider a continuous random variable $T$, which describes the occurrence time of an irreversible event. Its probability density is given by $f(t)$, and it has cumulative density function

$$F(t) = \Pr(T < t) = \int_0^t f(x)dx. \tag{3.2}$$

Define the survival function

$$S(t) = \Pr(T \geq t) = 1 - F(t) = \int_t^\infty f(x)dx \tag{3.3}$$

as the probability that the event does not occur before time $t$.

The hazard function

$$h(t) = \lim_{dt \to 0} \frac{\Pr(t \leq T < t + dt | T \geq t)}{dt} = \frac{f(t)}{S(t)} \tag{3.4}$$

is the instantaneous occurrence rate of the event, conditional on the event not occurring before time $t$. This is the central object of study in survival analysis. From Eq 3.3, it can be seen that $S'(t) = -f(t)$, and so

$$h(t) = -\frac{d}{dt} \log S(t). \tag{3.5}$$

Integrating gives an expression for the survival function $S$, and thus the overall distribution of the event's occurrence time, in terms of the hazard function:

$$S(t) = \exp\Big[-\int_0^t h(x)dx\Big]. \tag{3.6}$$

Next, consider a set of time intervals $t_1, t_2, \ldots$, where $t_i$ is the time interval $[\tau_{i-1}, \tau_i)$. Define the discrete random variable $\hat{T}$, which describes the (discrete) occurrence time of an event. Its probability mass is given by

$$\hat{f}_j = \Pr(\hat{T} = t_j). \tag{3.7}$$

The discrete hazard function is defined as

$$\hat{h}_j = \Pr(\hat{T} = t_j | \hat{T} \geq t_j). \tag{3.8}$$

Note that

$$\hat{h}_j = 1 - \Pr(\hat{T} > t_j | \hat{T} > t_{j-1}) \tag{3.9}$$

$$= 1 - \exp\Big[-\int_{\tau_{j-1}}^{\tau_j} h(t)dt\Big] \tag{3.10}$$

which follows from Eq 3.6, where $\tau_{j-1}$ is allowed to take the lower bound of the survival function integral since survival has been guaranteed up to time $\tau_{j-1}$. Now, define the "force function" [1]

$$\lambda_j = \int_{\tau_{j-1}}^{\tau_j} h(t)dt. \tag{3.11}$$

A high cumulative continuous hazard $h$ over a given time interval translates into a high value for the force function, so that $\lambda_j$ may be interpreted the amount of "force" driving the occurrence of the event in time step $t_j$.

To conclude, the hazard of occurrence of an irreversible event at discrete time $t_j$ may be expressed in terms of the force function $\lambda_j$:

$$\hat{h}_j = \Pr(\hat{T} = t_j | \hat{T} \geq t_j) = 1 - e^{-\lambda_j}, \tag{3.12}$$

where $\lambda_j$ is the integrated continuous hazard function $h$ over the bounds of the time interval $t_j$ (Eq 3.11). This provides a way of expressing the discrete-time probability distribution of event occurrence in terms of the continuous-time hazard function $h$. Specifying either $h$ or $\lambda$ is sufficient to specify the discrete hazard function $\hat{h}$. Eggo *et al.* (2011) [75] and Gog *et al.* (2014) [91] define expressions for the force of infection $\lambda$ on cities during the the 1918 and 2009 influenza pandemics, and use relation 3.12 to construct the probability of observing some full set of outbreak onset times across multiple cities. A derivation of this probability expression may be found in §3.2.2.

### 3.1.3 Gaussian processes

Parametric models, which characterise data using functions of finite collections of parameters, are useful when an underlying mechanism for the process that generated the data is hypothesised or known. When the mechanism is unclear, non-parametric models, for which the number of parameters grows indefinitely with the amount of available data, can

[1] In traditional survival analysis texts, the hazard is denoted $\lambda$, rather than $h$. In epidemiological literature, however, the symbol $\lambda$ often corresponds to the force of infection, which is why $\lambda$ has been reserved for the force function here.

help to identify meaningful patterns in the data. Rasmussen and Williams (2006) [196], p.166, provide a helpful discussion of parametric and non-parametric models. A useful type of non-parametric model is the Gaussian process. A Gaussian process is a flexible random surface that can be regressed onto data. Gaussian processes are infinite-dimensional generalisations of multivariate normal distributions, and have many of the advantages of analytic tractability that accompany normal distributions.

Formally, a Gaussian process is defined as a collection of random variables, any finite number of which follow a multivariate normal distribution [196]. Following [196], a Gaussian process is fully specified by its mean function $m(x)$ and covariance function $k(x,x')$, and is denoted

$$f(x) \sim GP(m(x),k(x,x')). \tag{3.13}$$

The Gaussian process at any finite collection of points $\boldsymbol{x} = (x_1,x_2,\ldots,x_n)$ follows the multivariate normal distribution

$$MVN((m(x_1),m(x_2),\ldots,m(x_n))^T, \begin{bmatrix} k(x_1,x_1) & k(x_1,x_2) & \ldots & k(x_1,x_n) \\ k(x_2,x_1) & k(x_2,x_2) & \ldots & k(x_2,x_n) \\ \vdots & & \ddots & \vdots \\ k(x_n,x_1) & k(x_n,x_2) & \ldots & k(x_n,x_n) \end{bmatrix}).$$

The covariance function $k$ must be positive semidefinite and symmetric in its arguments, analogous to the constraints on the covariance matrix of a multivariate normal distribution [196]. Different choices of covariance function yield processes that differ in smoothness and how rapidly they vary with distance. The squared exponential (SE) covariance function is a simple yet versatile example of covariance function. It is defined as

$$k_{SE}(d) = \exp\Big(-\frac{d^2}{2l^2}\Big) \tag{3.14}$$

where $d$ is the distance $|x-x'|$ between any two input points $x$ and $x'$, and $l$ defines the characteristic length scale of the process, which is roughly the distance that must be travelled for the function value to change 'significantly'. Gaussian processes with a SE covariance function are smooth, with mean-square derivatives of all orders [196]. This may be contrasted with the exponential covariance function, which has form

$$k_E(d) = \exp\Big(-\frac{d}{l}\Big). \tag{3.15}$$

The exponential covariance function is jagged (see Fig 3.1), without guaranteed mean-square differentiability of any order.

The SE covariance function is a special case of both the rational quadratic (RQ) and the Matérn covariance functions. The RQ covariance function has form

$$k_{RQ}(d) = \left(1 + \frac{d^2}{2\alpha l^2}\right)^{-\alpha} \tag{3.16}$$

which approaches the SE covariance function as $\alpha \to \infty$. The RQ covariance function may be interpreted as an infinite sum of SE covariance functions with different length scales, and also has mean-squared differentiability of all orders, regardless of $\alpha$ [196]. The Matérn covariance function has form

$$k_{\text{Matérn}}(d) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}d}{l}\right)^{\nu} K_{\nu}\left(\frac{\sqrt{2\nu}d}{l}\right) \tag{3.17}$$

with $\nu$ and $l$ both positive and $K_{\nu}$ a modified Bessel function. The argument $\nu$ is a smoothness parameter, such that processes with the Matérn covariance function are $\lfloor \nu \rfloor$-times mean-square differentiable [196]. The SE covariance function is obtained when $\nu \to \infty$. The exponential covariance function is obtained when $\nu = 1/2$. Figure 3.1 depicts draws from Gaussian processes with the squared exponential, exponential, and RQ covariance functions. When developing a Gaussian process model, a covariance function is normally chosen *a priori* by the modeller to match the anticipated 'character' of the underlying process to be described, such as its smoothness and rough rate of variation.

### Gaussian process regression

In practice, one often wishes to fit a Gaussian process $f(\cdot)$ with covariance function $k(\cdot,\cdot)$ to $n$ observed data points $\boldsymbol{y}$ taken at locations $\boldsymbol{x}$. For example, one might wish to model how temperature varies across a geographic region, given readings $\boldsymbol{y}$ from a set of $n$ weather stations at coordinates $\boldsymbol{x}$. One way to approach this problem is to model the observations as a Gaussian process $f(\boldsymbol{x})$ with additional noise:

$$\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\varepsilon} \tag{3.18}$$

where $\boldsymbol{\varepsilon}$ is a vector of the measurement errors at each site. If the measurement errors are assumed to follow independent and identically distributed draws from a normal distribution with mean 0 and variance $\sigma_n^2$, the joint distribution of the observations $\boldsymbol{y}$ and a set of $m$

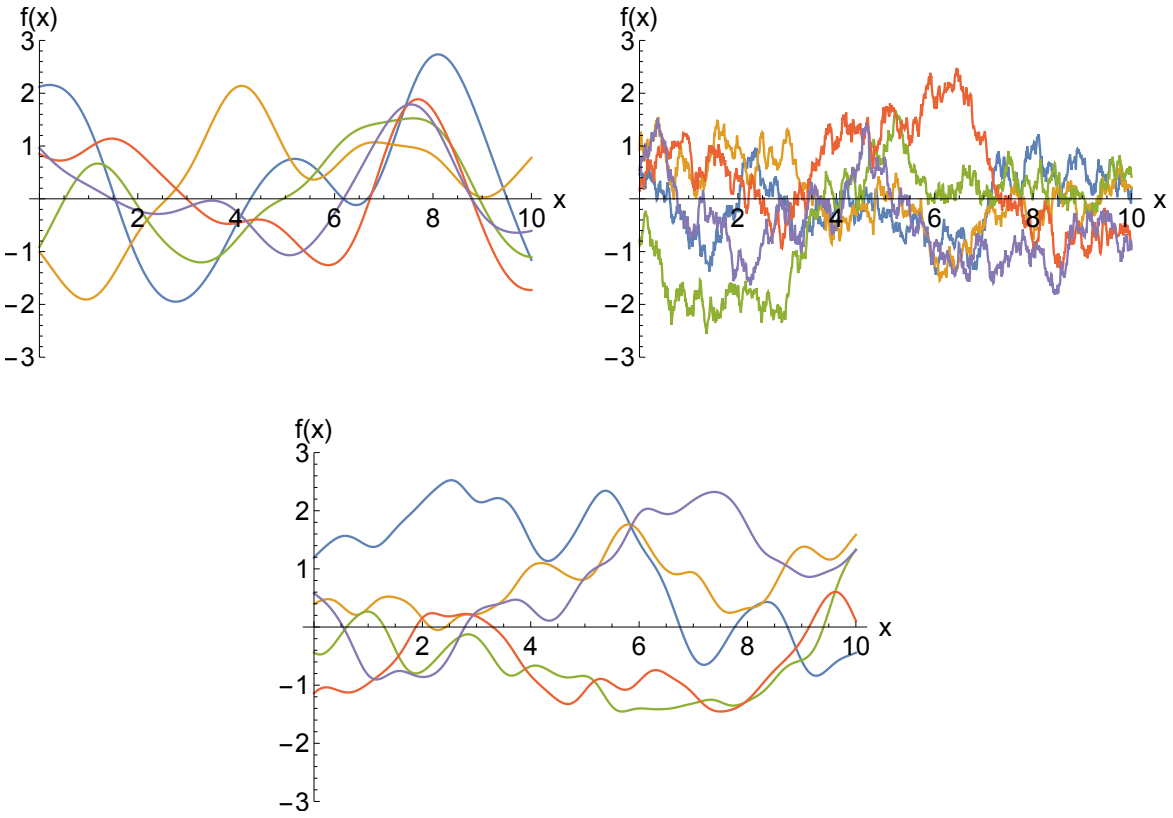



Fig. 3.1 Five draws each from Gaussian processes with the squared exponential (SE) covariance function (Eq 3.14; top left plot), exponential covariance function (Eq 3.15; top right plot), and rational quadratic (RQ) covariance function (Eq 3.16; bottom plot) with $\alpha = \frac{1}{2}$. The length scale $l$ is fixed at 1 for all three processes, and the process mean $m(x)$ (see Eq 3.13) is fixed at 0. The processes differ primarily in smoothness: the squared exponential and RQ covariance functions yield smooth processes with mean-square differentiability of all orders, while the exponential covariance function yields a jagged process without guaranteed mean-square differentiability of any order. Choice of covariance function is normally made *a priori* by the modeller, depending on the anticipated qualitative behaviour of the underlying process to be described.

unobserved temperatures $\mathbf{f}_*$ at locations $\boldsymbol{x}_*$ is

$$\begin{bmatrix} \boldsymbol{y} \\ \mathbf{f}_* \end{bmatrix} \sim MVN(\mathbf{0}, \begin{bmatrix} K(X,X)+\sigma_n^2 I & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}) \tag{3.19}$$

where $K(X,X)$ is the $n \times n$ covariance matrix obtained by evaluating the covariance function for each pair of observation points $\boldsymbol{x}$, $K(X,X_*)$ is the $n \times m$ matrix of covariances between the observation points $\boldsymbol{x}$ and the prediction points $\boldsymbol{x}_*$, $K(X_*,X_*)$ is the $m \times m$ covariance matrix of the prediction points $\boldsymbol{x}_*$ with themselves, and $K(X_*,X) = K(X,X_*)^T$.

The posterior distribution of $\mathbf{f}_*$, conditional on the observed data points $\boldsymbol{y}$, is given in [196]:

$$\mathbf{f}_*|X,\boldsymbol{y},X_* \sim MVN(\bar{\mathbf{f}}_*, \text{cov}(\mathbf{f}_*)) \qquad \text{where} \tag{3.20}$$

$$\bar{\mathbf{f}}_* = K(X_*,X)[K(X,X)+\sigma_n^2 I]^{-1}\boldsymbol{y}, \tag{3.21}$$

$$\text{cov}(\mathbf{f}_*) = K(X_*,X_*) - K(X_*,X)[K(X,X)+\sigma_n^2 I]^{-1}K(X,X_*). \tag{3.22}$$

So, the posterior mean and variance of the process can be evaluated at any set of points $\boldsymbol{x}_*$ from just the covariance function and the observed data. This is the simplest case of Gaussian process regression.

Often, however, the training data $\boldsymbol{y}$ are not observed directly. Instead, some process that depends on $\boldsymbol{y}$ is observed, and the posterior process $\mathbf{f}_*$ must be inferred indirectly. This may be accomplished using a link function $\Phi$ that connects some set of direct observations $\boldsymbol{t}$ with the process $f(\cdot)$. In one common case, the link function defines the probability with which a particular event $\boldsymbol{t}$ is observed:

$$P(T=\boldsymbol{t}) = \Phi(\Omega,\boldsymbol{y},\boldsymbol{t}) \qquad \text{where} \tag{3.23}$$

$$\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\varepsilon} \qquad \text{and} \tag{3.24}$$

$$f(\boldsymbol{x}) \sim GP(0,k(x,x')). \tag{3.25}$$

Here, $T$ is a random variable that describes the observation process, and $\Omega$ is a finite list of additional model parameters. In this formulation, $\Phi$ is a semiparametric model, which means that it is composed of both parametric ($\Omega$) and nonparametric ($\boldsymbol{y}$) elements. To estimate $\boldsymbol{y}$ and $\Omega$, one can consider the likelihood function

$$L(\boldsymbol{y},\Omega|\boldsymbol{t}) = \Phi(\Omega,\boldsymbol{y},\boldsymbol{t}). \tag{3.26}$$

If it is sufficiently easy to evaluate $\Phi$, posterior estimates of both $\boldsymbol{y}$ and $\Omega$ may be obtained using a Metropolis Hastings algorithm. First, a prior distribution for the parameter $\Omega$ must be specified. The prior for $\boldsymbol{y}$ is given by the Gaussian process, Eq 3.13. Proposals for $\boldsymbol{y}$ and $\Omega$ are drawn from their respective prior distributions, and the likelihood (3.26) is evaluated. The proposals are accepted or rejected with probability proportional to the likelihood ratio between the current and most recently accepted proposals, eventually yielding a good estimate of the posterior densities of $\boldsymbol{y}$ and $\Omega$.

In addition to the detailed background on Gaussian processes and Gaussian process regression provided by Rasmussen and Williams in [196], Gelfand *et al.* (2003) [85] introduce a theoretical framework for fitting generalised linear models with coefficients that vary spatially or temporally according to a Gaussian process. Building upon the related work of Banerjee *et al.* (2003) and Banerjee and Gelfand (2006) [11], Goldstein *et al.* (2015) [93] identify local trends in the speed and direction of the spread of an invasion wave of the gypsy moth *Lymantria dispar* in the north-eastern United States.

## 3.2 Model definition

To describe the geographic transmission of the 2009 A/H1N1pdm outbreak in the United States, two related models are presented. They differ only in the parameter that describes the transmissibility of the disease: in the first, transmissibility is constant across locations and time, while in the second, it is allowed to vary by location and time according to a Gaussian process.

### 3.2.1 Motivating the model structure

A metapopulation model is a natural choice for describing the geographic transmission of a human disease at the continent scale. Humans tend to cluster into relatively well-defined communities, such as cities and towns, and epidemiological data is often collected at the same scale. For this analysis, data are available for 834 3-digit ZIP (postal) codes in the United States (see Chapter 2), each of which roughly corresponds to a city. These ZIPs are treated as the sub-populations of the metapopulation model.

The model is constructed to explain the time of epidemic onset in each ZIP as a function of relevant predictors. Following Gog *et al.* (2014) [91], the model tests for effects from population size and density, the onset of the autumn school term, and proximity to infected ZIPs. It is assumed that each ZIP transitions exactly once from uninfected to infected, since

the model only seeks to explain the initial invasion of the autumn 2009 pandemic wave. Survival analysis offers a mathematically natural way of addressing this problem. From this perspective, specifying the transmission model consists in defining a force function (see §3.1.2) in terms of relevant predictors, which in turn specifies the probability distribution of each city's outbreak onset time via the discrete hazard function, Eq 3.12.

In both of the transmission models presented in this chapter, the force of infection $\lambda$ consists of two summed parts. The first part captures the baseline risk of importing infection from far away (either from abroad or from some distant location within the country), while the second describes the risk of importing infection from close neighbours. This form is chosen in an attempt to reproduce the outbreak dynamics of the autumn wave of the 2009 A/H1N1pdm outbreak in the US, described in §2.4, consisting of a few long-distance introductions followed by short-distance wave-like spread.

### 3.2.2 The fundamental transmission model

The force of infection on location $i$ is given by the equation

$$\lambda_i(t) = \beta_0 + (\beta_d + I_a \beta_{ds}) N_i^{\mu} \frac{\sum_{j \in \Lambda_t} n_{j,t}^{\theta} N_j^{\nu} \kappa(d_{i,j})}{[\sum_{j \neq i} N_j^{\nu} \kappa(d_{i,j})]^{\varepsilon}} \tag{3.27}$$

where $\lambda_i(t)$ is the force of infection on location $i$ at (discrete) time $t$. This force of infection may be interpreted as the cumulative continuous hazard of infection over half-week $i$ (see Eq 3.11). The input $I_a$ is an indicator function that is 1 if the school term in location $i$ begins in week $t+a$, and 0 otherwise. Possible lags $a \in \{0, 0.5, 1, 1.5, 2\}$ are considered, to test for heightened transmission due to increased social mixing between children up to two weeks before the start of the school term. The summation index $\Lambda_t$ is the set of all locations with epidemic onset prior to time $t$. The input $n_{j,t}$ is the observed ILI ratio (see §2.2) in location $j$ at time $t$, normalised by the mean ILI ratio in location $j$ from June 2009 through June 2010, and fixed at 0 for all $t$ prior to location $j$'s epidemic onset time. Models that incorporated the ILI ratio into $n_{j,t}$ prior to location $j$'s onset time were also tested, but these generally performed worse in terms of AIC than those that fixed $n_{j,t}$ at 0 prior to the onset in location $j$. This jump in ILI essentially characterises outbreak establishment as a sudden event, rather than a gradual process. This perspective is defensible in the context of a highly infectious pathogen, for which the sudden presence of even a small amount of infection in a neighbouring city leads to a large increase in the force of infection. The ILI ratio on half-week values is taken to be the geometric mean of the ILI ratios on the full weeks just

before and just after the half-week, following Gog *et al.* (2014) [91]. Inputs $N_i$ and $N_j$ are the population sizes of the recipient and donor populations $i$ and $j$, respectively, normalised by the mean ZIP population size. The parameter $\beta_0$ is the background force of infection due to long-distance seeding of infection, assumed constant for all locations, following [91]; $\beta_d$ is the transmissive strength of the disease; $\beta_{ds}$ is a boost in transmissive strength due to schools being in session in location $i$; $\mu$ and $\nu$ are gravity model exponents on the population sizes of recipient location $i$ and donor location $j$, respectively; $\theta$ modulates the importance of the epidemic time series; $\varepsilon$ modulates population density dependence; and $\kappa(d_{i,j})$ is a decreasing function that describes how epidemiological connectivity between cities decays with great-circle distance $d_{i,j}$. The model is based on the most parsimonious model from Gog *et al.* (2014) [91], though the donor population size $N_j$ is re-incorporated, the epidemic time series term $n_{j,t}^{\theta}$ is introduced, and a more general distance kernel is used (see Eq 3.28). Table 3.1 summarises the parameters' interpretations and gives their possible ranges.

The distance kernel has form

$$\kappa(d_{i,j}) = \left(1 + \frac{d_{i,j}}{\rho\gamma}\right)^{-\gamma} \tag{3.28}$$

where $d_{i,j}$ is the distance between locations $i$ and $j$, $\rho$ is the kernel's distance scale, and $\gamma$ is the kernel's rate of decay. Note that

$$\lim_{\gamma\to\infty} \kappa(d_{i,j}) = e^{-d_{i,j}/\rho},$$

following a definition of the exponential function. In other words, the parameter $\rho$ defines the exponential curve that the kernel $\kappa(d_{i,j})$ approaches as $\gamma \to \infty$ (see Fig 3.2). This way, the kernel incorporates both power-law decay and exponential decay, the two most common kernels for gravity models. Roughly speaking, $\rho$ adjusts the kernel's rate of decay at short distances, and $\gamma$ adjusts the kernel's rate of decay at long distances. This is apparent by examining the curves depicted in Fig 3.2.

The interpretations of $\rho$ and $\gamma$ become clearer when the kernel is seen as a custom function designed to feature a set of desired properties. The first such property might be that the kernel should describe power-law decay in the force of infection with distance, since human movements are so often cited to follow power laws. So, one might propose

$$\kappa(d_{i,j}) = d_{i,j}^{-\gamma}.$$

This kernel has a singularity at $d_{i,j} = 0$, however. To avoid this, one might propose the kernel

$$\kappa(d_{i,j}) = (1 + d_{i,j})^{-\gamma}$$

which shifts the kernel horizontally so that it intersects the vertical axis at $\kappa(0) = 1$. Next, it might be desirable to scale the distance $d_{i,j}$ by some value $\sigma$ to provide greater flexibility in the rate of decay. Such a kernel would have form

$$\kappa(d_{i,j}) = \left(1 + \frac{d_{i,j}}{\sigma}\right)^{-\gamma}.$$

The parameters $\sigma$ and $\gamma$ interact: increasing $\gamma$ increases the steepness of the curve, while increasing $\sigma$ makes the curve more shallow. These adjustments are different in character, however. Changes in $\sigma$ squeeze or stretch the curve horizontally, which affects the shape of the kernel most at shorter distances. On the other hand, $\gamma$ specifies the rate at which the kernel decays, which affects the kernel's shape most significantly at larger distances. If both $\gamma$ and $\sigma$ increase together, the parameters roughly offset each other, but the kernel changes subtly in form: it approaches an exponential. This limiting exponential function has a decay rate that depends on how quickly $\gamma$ increases compared to $\sigma$. To untangle this relationship, $\sigma$ may be factored into $\rho\gamma$, where $\rho$ captures this difference in the rate of increase in $\sigma$ vs $\gamma$. So, $\rho$ explicitly gives the decay rate of the limiting exponential function as $\gamma \to \infty$. This yields Eq 3.28.

### Parameter estimation

Parameter values for the fundamental transmission model, Eq 3.27, may be estimated using a maximum likelihood scheme. To accomplish this, we here derive an expression for the probability of observing some full set of outbreak onset times $T = \{T_1, \ldots, T_n\}$, where $T_i$ is the outbreak onset time for city $i$, in terms of $\Theta = \{\beta_0, \beta_d, \beta_{ds}, \mu, \nu, \rho, \gamma, \varepsilon\}$, the set of model parameters.

Consider a time step $t$ for which the set of all previously infected locations, $\Lambda_t$, is known. Recall from §3.1.2 that $1 - \exp(-\lambda_i(t))$ is the probability that location $i$ becomes infected at time $t$, given that it has not been infected prior to time $t$, where $\lambda_i(t)$ is the force of infection on location $i$ at time $t$. It follows that $\exp(-\lambda_i(t))$ is the probability that an uninfected location $i$ remains uninfected at discrete time $t$. Let $\Psi_t$ be the (random) set of locations that become infected at time $t$, so that $\Lambda_{t+1} = \Psi_t \cup \Lambda_t$, with $\Psi_t \cap \Lambda_t = \emptyset$. Also let $\bar{\Lambda}_t$ be the complement of $\Lambda_t$, or the set of locations that remain uninfected at time $t-1$. Then, the

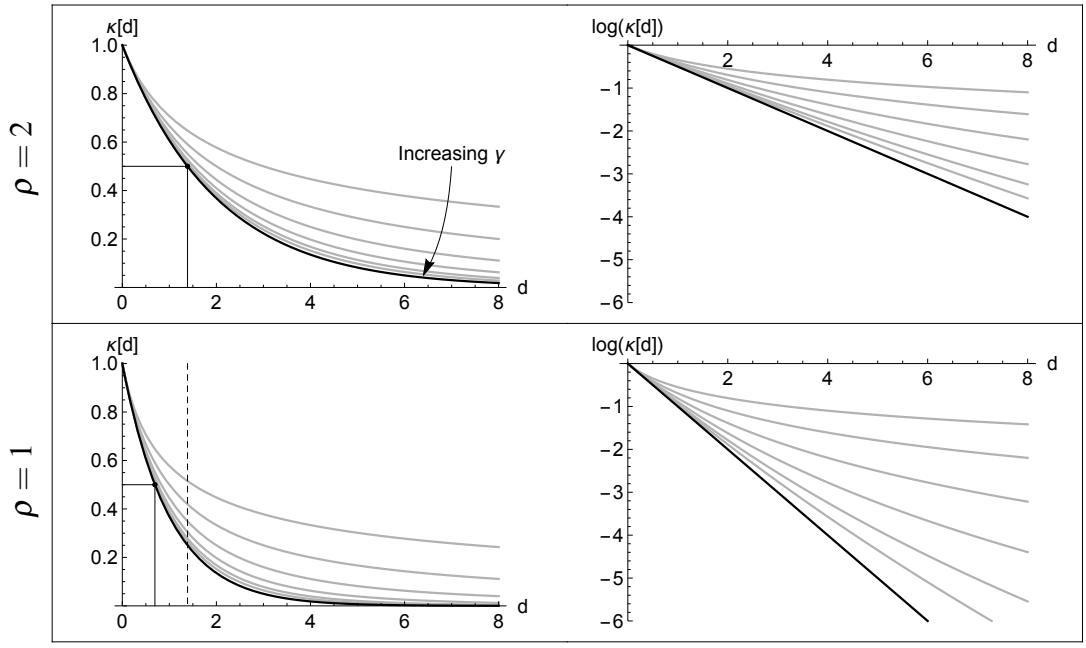



Fig. 3.2 Plots of the distance kernel $\kappa(d_{i,j})$ for $\rho = 2$ (top pane) and $\rho = 1$ (bottom pane), with $\gamma = \{0.5, 1, 2, 4, 8, 16\}$ (grey curves). The solid black curves represent the case $\gamma \to \infty$, for which the kernel decays exponentially. As $\gamma$ increases, the kernel approaches this limiting exponential form. The kernel with $\rho = 1$ decays twice as fast as the kernel with $\rho = 2$; in both plots, the point where $\kappa(d_{i,j}) = 0.5$ is marked, which occurs for $\rho = 1$ at exactly half the distance it takes for the $\rho = 2$ kernel to decay by the same amount. The plots on the right-hand side depict the logged kernels for the same $\rho$ and $\gamma$ values. These demonstrate that at long distances, the kernel with finite $\gamma$ will always decay more slowly than its exponential ($\gamma \to \infty$) counterpart: the logged power kernel (finite $\gamma$) remains always convex, while the logged exponential kernel (infinite $\gamma$) is linear. This is why the power kernel is often characterised as having 'thick tails' compared to its exponential counterpart.

Table 3.1 Geographic transmission model parameters, possible ranges, and interpretations

| Parameter | Range | Interpretation |
|---|---|---|
| $\beta_0$ | $[0,\infty)$ | Force from external seeding (importations from abroad or distant within-country) |
| $\beta_d$ | $[0,\infty)$ | Transmissibility |
| $\beta_{ds}$ | $[0,\infty)$ | Boost in transmissibility from schools |
| $\mu$ | $[0,1]$ | Importance of recipient population size |
| $\nu$ | $[0,1]$ | Importance of donor population size |
| $\rho$ | $(0,\infty)$ | Characteristic transmission distance |
| $\gamma$ | $(0,\infty)$ | Transmission kernel decay rate |
| $\varepsilon$ | $[0,1]$ | Population density dependence (at high values, the fraction of neighbours infected matters more than total neighbours infected) |
| $\theta$ | $[0,1]$ | Importance of true neighbouring number infected (i.e. importance of neighbouring epidemic time series) |

probability that the set $\Psi_t$ of locations becomes infected at time $t$ is

$$P(\Psi_t|\Lambda_t,\Theta) = \prod_{i\in\Psi_t}(1-e^{-\lambda_i(t)}) \prod_{i\in\bar{\Lambda}_{t+1}} e^{-\lambda_i(t)} \prod_{i\in\Lambda_t} 1 \tag{3.29}$$

$$= \prod_{i\in\Psi_t}(1-e^{-\lambda_i(t)}) \prod_{i\in\bar{\Lambda}_{t+1}} e^{-\lambda_i(t)} \tag{3.30}$$

which may be interpreted as a product of the probabilities that the locations in $\Psi_t$ become infected at time $t$, multiplied by a product of the probabilities that all locations not in $\Psi_t$ that were previously uninfected remain uninfected at time $t$, multiplied by a product of the probabilities that each location that has already been infected prior to time $t$ remains infected at time $t$ (which is 1 for this model, since locations can never become uninfected after having been infected). Then, the probability of observing some full set of epidemic onset times $T$ is

a product of Eq 3.30 over each time step in the epidemic; that is,

$$P(T|\Theta) = \prod_{t=1}^{\max(T)} P(\Psi_t|\Lambda_t) \tag{3.31}$$

$$= \prod_{t=1}^{\max(T)} \Big[ \prod_{i:T_i=t} (1-e^{-\lambda_i(t)}) \prod_{i:T_i>t} e^{-\lambda_i(t)} \Big] \tag{3.32}$$

$$= \Big( \prod_{t=1}^{\max(T)} \prod_{i:T_i=t} (1-e^{-\lambda_i(t)}) \Big) \Big( \prod_{t=1}^{\max(T)} \prod_{i:T_i>t} e^{-\lambda_i(t)} \Big) \tag{3.33}$$

$$= \Big( \prod_{i=1}^{n} (1-e^{-\lambda_i(t)}) \Big) \Big( \prod_{i=1}^{n} \prod_{t<T_i} e^{-\lambda_i(t)} \Big) \tag{3.34}$$

$$= \prod_{i=1}^{n} \Big[ (1-e^{-\lambda_i(T_i)}) \prod_{t=1}^{T_i-1} e^{-\lambda_i(t)} \Big] \tag{3.35}$$

where (3.32) follows from substituting Eq 3.30, (3.33) follows from regrouping terms, (3.34) follows from rearranging the products, and (3.35) follows from recombining the terms. The initial set of infected locations, $\Lambda_1$, is the empty set. The upper product limit $n$ in Eqs 3.34 and 3.35 is the total number of locations.

Eq 3.35 may be interpreted as the likelihood of the model parameters $\Theta$ given the observed outbreak onset times. Taking the logarithm gives the log-likelihood:

$$\ell(\Theta;T) = \sum_{i=1}^{n} \Big( \log(1-e^{-\lambda_i(T_i)}) - \sum_{t=1}^{T_i-1} \lambda_i(t) \Big). \tag{3.36}$$

The parameter values $\Theta$ that maximise (3.36), given outbreak onset times $T$ estimated by the breakpoint method (see §2.3), are calculated using the Nelder-Mead simplex algorithm, as implemented in MATLAB's `fminsearch()` function.

### Model selection

Holding certain parameters in Eq 3.27 at null values (see Table 3.2) yields a set of simpler nested models. These models can be compared against one another in terms of their goodness of fit and complexity using an information criterion such as the Akaike information criterion (AIC) [4]. The AIC is defined as

$$\mathrm{AIC} = 2k - 2\log(L)$$

where $k$ is the number of parameters in the model and $L$ is the model's likelihood. Smaller AIC values correspond to better models. To identify the "best" model, maximum likelihood parameter values are estimated for the full model (Eq 3.27) and for all possible nested models, yielding 480 fits in total. An AIC value is calculated for each model, and the model with the best (lowest) AIC is identified. Table 3.3 provides the AIC values for a set of 20 representative nested models. While including donor population size $N_j^\nu$ in the model never yields an improvement in AIC, including an effect from schools ($\beta_{ds}$) is preferred for some classes of nested models. Specifically, for models in which the explicit ILI time series is not included ($\theta = 0$), models that include $\beta_{ds}$ with one-week-advanced school start dates ($a = 1$) are preferred by AIC. So, even though the best overall model does not include an influence from schools, there is some evidence that mixing among schoolchildren – perhaps up to a week before the start of the school term – may have influenced transmission of the outbreak.

Table 3.2 Null values of geographic transmission model parameters and interpretations

| Parameter | Null value | Interpretation |
|---|---|---|
| $\beta_{ds}$ | 0 | No effect from schools opening |
| $\mu$ | 0 | No effect from recipient population size |
| $\nu$ | 0 | No effect from donor population size |
| $\gamma$ | $\infty$ | Distance kernel is exponential |
| $\varepsilon$ | 0 | No population density dependence |
| | 1 | Full population density dependence |
| $\theta$ | 0 | No effect from neighbouring number infected |

*Null values for $\beta_0$, $\beta_d$, and $\rho$ are not considered. Without $\beta_0$, there is no seeding, so the outbreak can never begin. Without $\beta_d$, all parameters except $\beta_0$ are also practically removed. Setting $\rho = 0$ makes local transmission impossible, while $\rho \to \infty$ makes trips to all locations equally probable, making the model no longer a geographic one.

The best model in terms of AIC has form

$$\lambda_i(t) = \beta_0 + \beta_d N_i^\mu \frac{\sum_{j\in\Lambda_t} n_{j,t}^\theta \kappa(d_{i,j})}{\sum_{j\neq i} \kappa(d_{i,j})}. \tag{3.37}$$

That is, the optimal model includes effects from all parameters except school onsets ($\beta_{ds}$) and donor population size ($\nu$). The population density parameter $\varepsilon$ estimated as identically 1. The maximum-likelihood parameter values for this model are listed in Table 3.4, and their profile likelihoods are depicted in Fig 3.3.

The absence of a parameter that captures the effect of the autumn school term in Eq 3.37 contrasts with the findings of Gog *et al.* (2014) [91]. Incorporating the neighbouring

Table 3.3 AIC values and significant parameters for 20 of the "best" nested models

| I-curve param | Kernel param | $a$ | Best $\Delta$AIC | $\beta_0$ | $\beta_d$ | $\beta_{ds}$ | $\mu$ | $\nu$ | $\rho$ | $\varepsilon$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $\theta = 0$ | $\gamma \to \infty$ | 0 | 28.5 | • | • | | • | | • | • |
| | | 0.5 | 28.5 | • | • | | • | | • | • |
| | | 1 | 25.5 | • | • | • | • | | • | • |
| | | 1.5 | 26.8 | • | • | • | • | | • | • |
| | | 2 | 28.5 | • | • | | • | | • | • |
| | $\gamma$ free | 0 | 23.4 | • | • | | • | | • | • |
| | | .5 | 23.4 | • | • | | • | | • | • |
| | | 1 | 21.0 | • | • | • | • | | • | • |
| | | 1.5 | 21.6 | • | • | • | • | | • | • |
| | | 2 | 23.4 | • | • | | • | | • | • |
| $\theta$ free | $\gamma \to \infty$ | 0 | 2.8 | • | • | | • | | • | • |
| | | .5 | 2.8 | • | • | | • | | • | • |
| | | 1 | 2.8 | • | • | | • | | • | • |
| | | 1.5 | 2.8 | • | • | | • | | • | • |
| | | 2 | 2.8 | • | • | | • | | • | • |
| | $\gamma$ free | 0 | **0** | • | • | | • | | • | • |
| | | .5 | **0** | • | • | | • | | • | • |
| | | 1 | **0** | • | • | | • | | • | • |
| | | 1.5 | **0** | • | • | | • | | • | • |
| | | 2 | **0** | • | • | | • | | • | • |

To generate this table, all 480 models generated by setting all possible combinations of the parameters to null values, and considering all possible school onset lags up to two weeks, were fit to data. The AIC values reported here are for all combinations of $\theta$ (fixed at 0 or free), $\gamma$ (fixed at $\infty$ or free), and $a \in \{0, 0.5, 1, 1.5, 2\}$. For each of these combinations, the best model in terms of AIC was identified. The bullets in the columns on the right-hand side of the table indicate which parameters are included in those best models. The parameter $\nu$ is never included (it is always fit to effectively 0), and the parameter $\beta_{ds}$ is only included for the $\theta = 0$ models, at $a = 1$ and $a = 1.5$. Note that lower AIC values correspond to "better" models. The raw AIC score for the best model is 4275.6.

outbreak ILI intensity, $n_{j,t}$, which was not done in Gog *et al.* (2014) [91], can account for this difference. Specifically, the neighbouring outbreak intensity $n_{j,t}$ seems to capture most of the relevant information encoded in the school opening term $I_a$, rendering $\beta_{ds}$ redundant. Indeed, in a subset of ZIPs – especially those in the southeast – a rise in ILI intensity is associated with the start of the autumn school term. Fig 3.4 depicts the ZIPs with outbreak onset time within one week of the start of the autumn school term. While school term start dates are closely linked with outbreak onset times in the southeast, the lag between school start dates and outbreak onset times increases over the course of the epidemic, with some outbreaks in the northeast trailing the start of the school term by two months or more (see Fig 2.21). The ILI intensity term $n_{j,t}^{\theta}$ may therefore capture much of the relevant school-term information at the start of the outbreak, and, unlike the school term, remain a relevant predictor of onset times as the epidemic unfolds, leading to an ultimate rejection of the parameter $\beta_{ds}$ in the most parsimonious model, Eq 3.37.

Table 3.4 Maximum likelihood parameter values for the best transmission model, Eq 3.37 (see also Fig 3.3)

| Parameter | Estimated value (95% CI) | Units |
|---|---|---|
| $\beta_0$ | 0.00043 (0.00015, 0.00087) | $(\Delta t)^{-1}$ |
| $\beta_d$ | 0.61 (0.53, 0.70) | $(\Delta t)^{-1}(km)^{1-\varepsilon}$ |
| $\mu$ | 0.32 (0.24, 0.40) | none |
| $\rho$ | 66 (48, 96) | $km$ |
| $\gamma$ | 8.9 (5.5, 74) | none |
| $\varepsilon$ | 1.0 (fixed) | none |
| $\theta$ | 0.56 (0.35, 0.77) | none |

### Global parameter sensitivities to onset uncertainty

The parameters of the transmission model, Eq 3.27, are fit using two stages of maximum likelihood inference. The first stage specifies the outbreak onset times (see §2.3). The second stage fits the model parameters to these estimated onsets, as described in this chapter. The 95% confidence intervals given in Table 3.4 report uncertainty from the second stage of maximum likelihood only. However, uncertainty from the first stage (the onset calculations) also propagates to the parameters.

It is possible to compare the uncertainties introduced by each stage of inference. Fig 3.5 depicts the MLE parameter values for the best transmission model, Eq 3.37, as pairwise scatters (small black points) for 1,000 sets of re-sampled onset times. The new onset times are calculated by independently drawing a new onset time from each location's onset likelihood

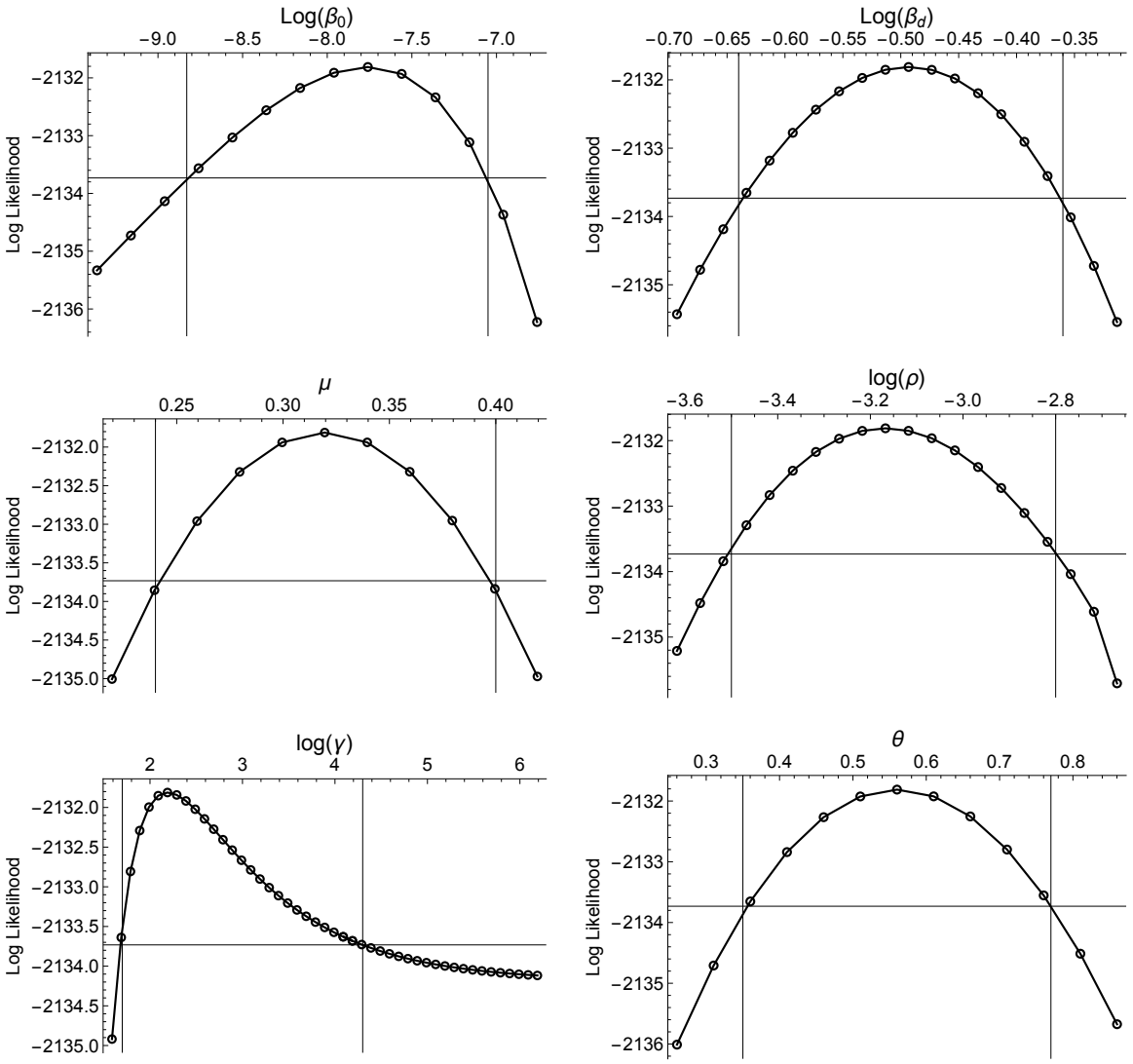



Fig. 3.3 Profile likelihood curves for the six free parameters of the best transmission model, Eq 3.37. The parameter values in Table 3.4 correspond to the maximum values of these curves, and the confidence intervals are specified by the intersection points between the curves and the horizontal line at 1.92 log-likelihood units below the maximum log-likelihood.

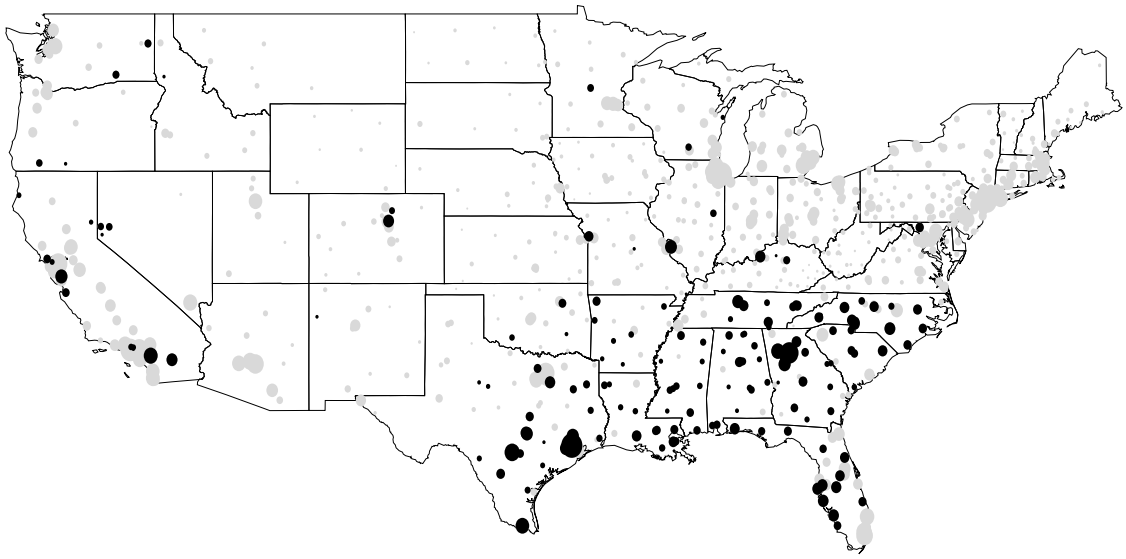


Fig. 3.4 Map depicting ZIPs where the median start date of the autumn school term is within one week of the outbreak onset time (black discs). Grey discs represent ZIPs where the discrepancy between outbreak onset time and school start date is greater than one week. Outbreak onset times are closely associated with the start of the autumn school term in the southeastern US.

distribution (see §2.3). To provide a point of comparison, the larger red points depict the maximum likelihood model parameters estimated using the maximum likelihood onsets. These are the same parameter values that are listed in Table 3.4. The red ellipses depict the 95% confidence regions for the MLE parameters using MLE onsets. These are analogous to the confidence intervals reported in Table 3.4. These ellipses mark the intersection of the log-likelihood surface for the two parameters depicted, maximizing over the other parameters, with the plane at 1.92 log-likelihood units below the maximum log-likelihood. Note that the ellipses are calculated without using any input from the re-sampled onsets. Even so, the scatters generated by re-sampling onsets generally lie acceptably within the ellipses. This suggests that the "stage-2" (model) uncertainty reported in Table 3.4 already captures much of the additional uncertainty introduced through onset uncertainty. This is just a rough test of the relationship between onset uncertainty and model uncertainty; one should note in particular that each small black point plotted in Fig 3.5 would have its own 95% confidence interval. However, this test does provide some evidence that ignoring onset uncertainty does not lead us to egregiously under-estimate the model parameter uncertainties. With this in mind, we proceed using the transmission model parameters estimated using the maximum likelihood onset times, listed in Table 3.4.

### Simulations

To verify that the transmission model can reproduce the wave-like geographic transmission pattern observed during the autumn wave of the 2009 A/H1N1pdm outbreak in the US, epidemics may be simulated from the transmission model. Two such simulations are depicted in Fig 3.6. The epidemics are seeded in Albany West GA, Augusta Northeast SC, Grenada MS, and Stockton North CA, the four locations with outbreak onset times in the first week of the true epidemic. It is assumed that no further seeding takes place, so $\beta_0$ is fixed at 0 for the rest of each simulation. The best transmission model with $\theta = 0$ is used for these simulations, since simulating from a model with $\theta > 0$ would require specifying an underlying model to generate the within-ZIP ILI incidence time series. The parameter values for the simulation model are listed in the caption to Fig 3.6. They lie within the confidence intervals of the overall best model's parameter values (see Table 3.4), suggesting that the simulations from this simpler model should still give valuable insight into the optimal model's behaviour.

The simulations broadly capture the wave-like spread of the true outbreak, and last for a similar duration ($\sim$14 weeks). Systematic differences between the simulated epidemics and the true one are discussed and addressed further in the following subsection.

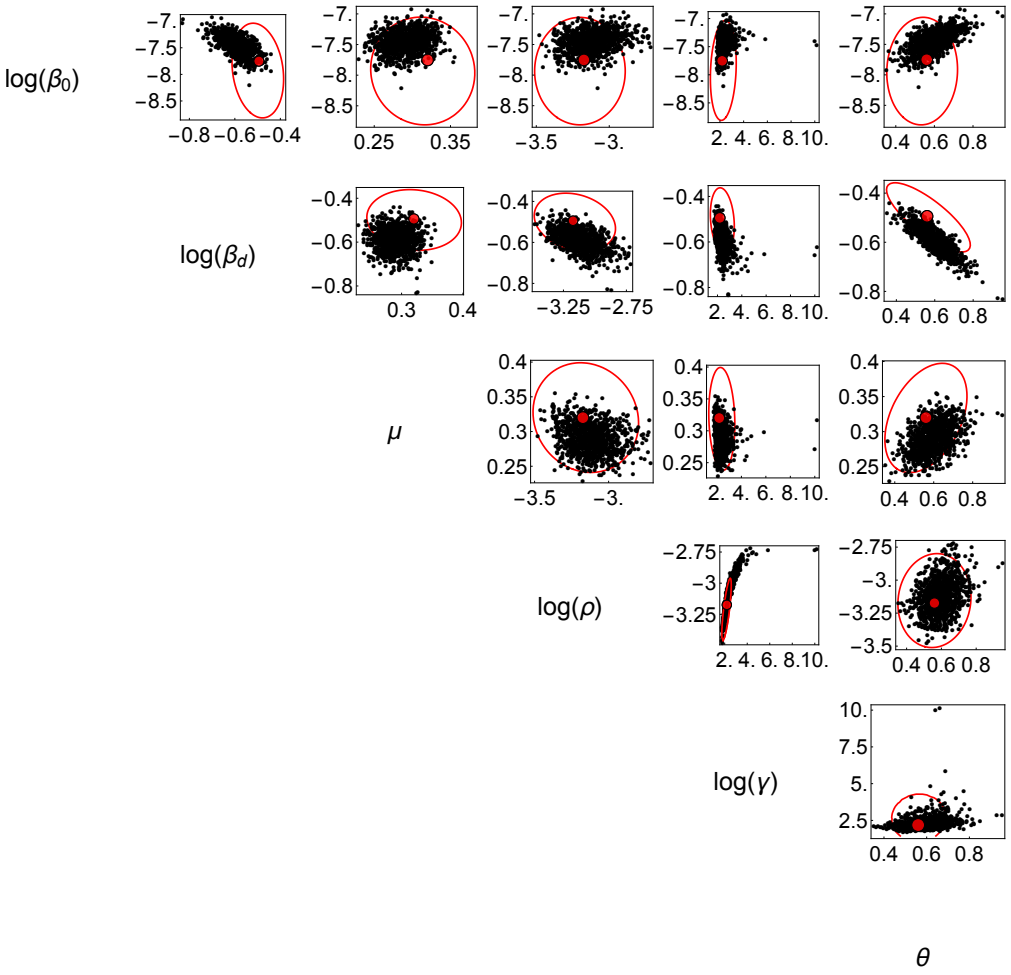



Fig. 3.5 Scatters of MLE parameter values of the most parsimonious transmission model, Eq 3.37, using re-sampled epidemic onset times (black points). The red points mark the MLE parameter values using the MLE onset times. The red ellipses represent the intersection of the log-likelihood surface of the most parsimonious transmission model's parameters with the plane at 1.92 log-likelihood units below the model's maximum log-likelihood, giving an approximate 95% confidence region for the parameters. These ellipses are analogous to the confidence intervals reported in Table 3.4. The ellipses are calculated with no input from the scatters of black points. Even so, the scatters lie largely within the ellipses, indicating that the uncertainty in epidemic onset times is already captured acceptably by the confidence intervals reported in Table 3.4.

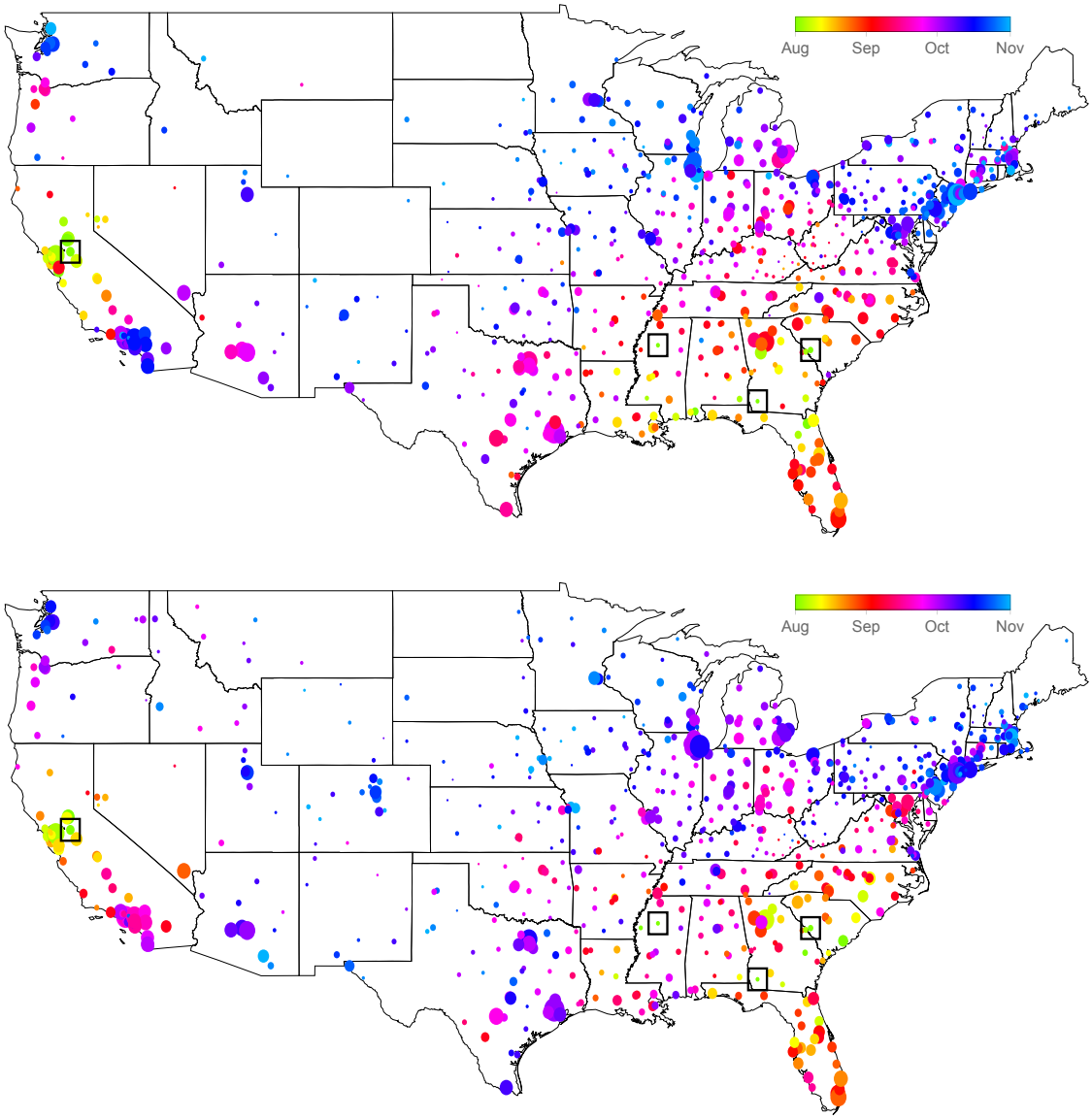



Fig. 3.6 Two simulations from the most parsimonious transmission model with $\theta = 0$, seeded in the four locations with epidemic onset in the first week of the true epidemic (boxed). Comparing with the onsets of the true epidemic (Fig 2.8), the simulated epidemics have similar radial spread patterns and overall durations. Parameters for these simulations are $\beta_d = 0.57$, $\mu = 0.25$, $\nu = 0$, $\rho = 62$, $\gamma = 7.8$, $\varepsilon = 1$, $\theta = 0$.

### Outbreak onset time residuals

In addition to checking output from simulations, the model's quality can be checked by examining the differences between the observed epidemic onset times and the onset times predicted by the model. For a given location $i$, the expected onset time is

$$E[T_i] = \sum_{k=1}^{\infty} k \cdot P(T_i = k) \tag{3.38}$$

$$= \sum_{k=1}^{\infty} k(1 - e^{-\lambda_i(k)}) \prod_{t=1}^{k-1} e^{-\lambda_i(t)} \tag{3.39}$$

$$= \prod_{t=1}^{0} e^{-\lambda_i(t)} - \prod_{t=1}^{1} e^{-\lambda_i(t)} + 2\prod_{t=1}^{1} e^{-\lambda_i(t)} - 2\prod_{t=1}^{2} e^{-\lambda_i(t)} + 3\prod_{t=1}^{2} e^{-\lambda_i(t)} - \dots \tag{3.40}$$

Since $P(T_i = \tau) = (1 - e^{-\lambda_i(\tau)}) \prod_{t=1}^{\tau-1} e^{-\lambda_i(t)}$ and $P(T_i = 1) = (1 - e^{-\lambda_i(1)})$, it makes sense to define $\prod_{t=1}^{0} e^{-\lambda_i(t)} = 1$. Substituting this into (3.40) and simplifying gives

$$E[T_i] = \sum_{k=1}^{\infty} \prod_{t=1}^{k-1} e^{-\lambda_i(t)}. \tag{3.41}$$

Now, let $t_f$ denote the onset time of the outbreak with the latest onset. By noting that $\lambda_i(t)$ remains constant for all $t > t_f$, the infinite sum can be evaluated exactly:

$$E[T_i] = \Big[\sum_{k=1}^{t_f} \prod_{t=1}^{k-1} e^{-\lambda_i(t)}\Big] + \Big[\sum_{k=t_f+1}^{\infty} \prod_{t=1}^{k-1} e^{-\lambda_i(t)}\Big] \tag{3.42}$$

$$= \Big[\sum_{k=1}^{t_f} \prod_{t=1}^{k-1} e^{-\lambda_i(t)}\Big] + \Big[(\prod_{t=1}^{t_f} e^{-\lambda_i(t)})(1 + e^{-\lambda_i(t_f+1)} + (e^{-\lambda_i(t_f+1)})^2 + \dots)\Big] \tag{3.43}$$

$$= \Big[\sum_{k=1}^{t_f} \prod_{t=1}^{k-1} e^{-\lambda_i(t)}\Big] + \Big[(\prod_{t=1}^{t_f} e^{-\lambda_i(t)}) \frac{1}{1 - e^{-\lambda_i(t_f+1)}}\Big]. \tag{3.44}$$

Fig 3.7 depicts the difference between the observed and expected epidemic onset time in each ZIP. There is a clear band in Missouri, Kentucky, and Virginia where the true outbreaks begin significantly later than the model expects. A discrepancy is also visible when plotting the true vs. expected cumulative number of outbreaks (Fig 3.8). The true epidemic gets off to a faster start than expected, but progresses at a slower rate, before increasing in rate in late September, after which the true and expected cumulative number of outbreaks match relatively well. These observations suggest that there is some variation in transmissibility in space and/or time for which the model does not yet fully account.

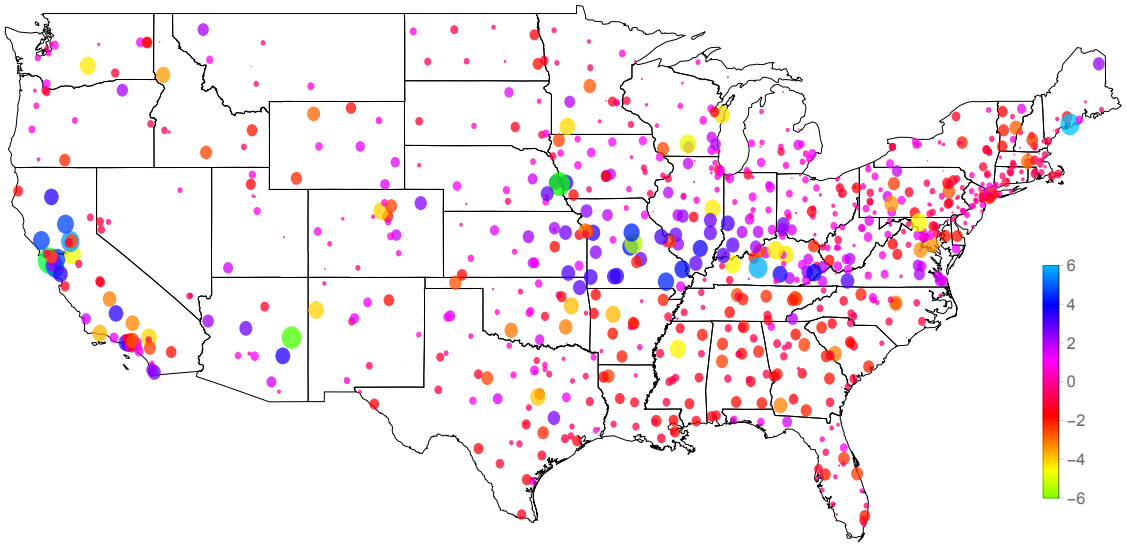



Fig. 3.7 Difference in weeks between the observed and expected outbreak onset time in each ZIP. The area of each disc is proportional to the magnitude of the difference. Blue/purple discs correspond to ZIPs where the true epidemic onset time is later than the expected onset time, and green/red discs correspond to ZIPs where the opposite is true. There is a band of locations in Missouri, Kentucky, and Virginia, and another patch near San Francisco CA, where the model consistently predicts earlier epidemic onset times than actually occurred. The model compensates by predicting slightly later-than-accurate epidemic onset times in much of the rest of the country.

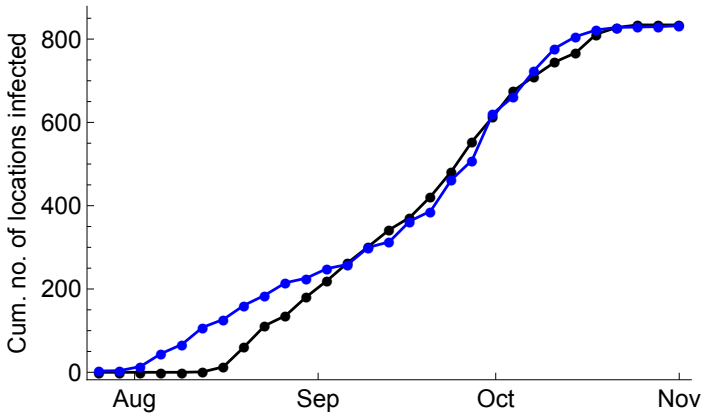



Fig. 3.8 Expected (black) and observed (blue) cumulative number of locations infected over time. The expected onsets are calculated from Eq 3.44. The true epidemic starts more quickly than expected by the transmission model, and has an increase in rate in late September that is not reflected in the expected onsets.

### 3.2.3 The transmissibility surface transmission model

The basic transmission model Eq 3.27 accounts for short-range predictors of transmission, and treats long-distance jumps as a stochastic process with constant rate across all locations. However, it overlooks mid-scale variation in transmission strength. To adjust for this variation, one can allow the transmissibility $\beta_d$ to depend on time and geographic location. However, fitting an independent value for the transmissibility at each time and location would require fitting far more parameters than there are available data points and, even if it were possible, would yield an over-fit model from which underlying patterns would be difficult to identify.

These issues may be avoided by imposing a correlation structure on the transmissibility values. In particular, a time-specific adjustment $\xi_t^T$ and a space-specific adjustment $\xi_i^S$ are incorporated into the model's transmissibility term, yielding a new model of form

$$\lambda_i(t) = \beta_0 + \beta_d \,\mathrm{Exp}[\xi_t^T + \xi_i^S] N_i^{\mu} \frac{\sum_{j\in\Lambda_t} n_{j,t}^{\theta}\, \kappa(d_{i,j})}{\sum_{j\neq i} \kappa(d_{i,j})} \tag{3.45}$$

where

$$\boldsymbol{\xi}^T = f^T(\boldsymbol{t}) + \boldsymbol{\varepsilon}^T, \qquad \boldsymbol{\xi}^S = f^S(\boldsymbol{x}) + \boldsymbol{\varepsilon}^S$$

and

$$f^T \sim GP(\mu^T(\cdot), k^T(\cdot,\cdot)), \qquad f^S \sim GP(\mu^S(\cdot), k^S(\cdot,\cdot)).$$

That is, the temporal and spatial adjustments to transmissibility, $\boldsymbol{\xi}^T = (\xi_1^T, \ldots, \xi_{tmax}^T)$ and $\boldsymbol{\xi}^S = (\xi_1^S, \ldots, \xi_n^S)$, are described by the Gaussian processes $f^T$ and $f^S$. Here, $\boldsymbol{t}$ is an array of the time intervals during which the epidemic occurs, and $\boldsymbol{x}$ is an array of the ZIP coordinates. The vectors $\boldsymbol{\varepsilon}^T$ and $\boldsymbol{\varepsilon}^S$ are additive noise terms, where each element $\varepsilon_t^T$ and $\varepsilon_i^S$ follows an independent, identically distributed Normal distribution with mean 0 and variance $\sigma_T^2$ and $\sigma_S^2$, respectively; that is,

$$\varepsilon_t^T \sim N(0, \sigma_T^2) \quad \text{and} \tag{3.46}$$

$$\varepsilon_i^S \sim N(0, \sigma_S^2) \tag{3.47}$$

for all $t$ and $i$. The Gaussian processes $f^T$ and $f^S$ are defined by the temporal and spatial mean functions $\mu^T$ and $\mu^S$ and the temporal and spatial covariance functions $k^T$ and $k^S$. The transmissibility terms $\xi_t^T$ and $\xi_i^S$ enter the transmission model via an exponential function to ensure that the full transmissibility term remains positive.

Specifying the model in this way allows the modeller to impose a correlation relationship between the temporal and spatial transmissibility values respectively, preventing the problems

associated with overfitting. The new terms may be interpreted as a "global" transmissibility adjustment $\boldsymbol{\xi}^T$ that varies over time equally across all locations, plus a "local" transmissibility adjustment $\boldsymbol{\xi}^S$ that accounts for additional spatial variation. Fixing $\mu^T(t) = \mu^S(x) \equiv 0$ and $k^T(t,t') = k^S(x,x') \equiv 1$ for all $t,x,t',x'$ yields the original best transmission model, Eq 3.37. In this case, the adjustments $\boldsymbol{\xi}^T$ and $\boldsymbol{\xi}^S$ are identically equal to zero.

Other choices of covariance function allow for more flexible transmissibility surfaces. The task of choosing a prior form for the covariance function can be simplified by considering what sort of predictor the Gaussian process might be replacing. Broad-scale weather changes and geographic and temporal differences in human behaviour might all explain the additional variation in transmissibility. It is reasonable to assume that these predictors might vary somewhat smoothly in space and time. A squared-exponential kernel, then, for its smoothness and analytical simplicity, might be a natural choice. One might seek a temporal process with length scale of approximately one month (eight half-weeks), since a shorter length scale would risk reintroducing the problem of overfitting, while a larger one might not be flexible enough to detect important structure. For the spatial process, a length scale of approximately $3\rho$ is reasonable. At this distance, the distance kernel $\kappa$ has dropped by 95%, so the Gaussian process would capture mid-scale variability, picking up where the distance kernel leaves off. According to the MLE parameter values from Table 3.4, a distance of $3\rho$ is approximately 200 km.

For the following model fits, $\boldsymbol{\xi}^T$ is assumed to follow a Gaussian process prior with mean function $\mu^T = 0$ and SE covariance function $k^T$ with length scale $l = 8$ half-weeks. The spatial transmissibility adjustment $\boldsymbol{\xi}^S$ is assumed to follow a Gaussian process prior with mean function $\mu^S = 0$ and squared-exponential covariance function $k^S$ with length scale $l = 200$ km. Spatial length scales of $l = 100$ km and $l = 500$ km were also tested, as well as a rational quadratic covariance function with length scale $l = 200$ km.

### Parameter estimation

Posterior distributions for $\boldsymbol{\xi}^T$ and $\boldsymbol{\xi}^S$ are estimated using a Metropolis Hastings algorithm. In each iteration of the algorithm, proposed values for a random subset of $\boldsymbol{\xi}^T$ are drawn from a multivariate normal distribution specified by Eq 3.20-3.22, where $\boldsymbol{x}_*$ are the proposal time points and the $\boldsymbol{x}$ are the other time points, with values $\boldsymbol{y}$. The model's likelihood is evaluated with these new proposed values. The proposal is accepted with probability proportional to the ratio of the new likelihood to the most recently accepted likelihood (or, for the first iteration, the original model's likelihood, with no $\boldsymbol{\xi}^T$ or $\boldsymbol{\xi}^S$). This is repeated for mutually exclusive random subsets of $\boldsymbol{\xi}^T$ until a new value has been proposed and either accepted or rejected

for each element of $\boldsymbol{\xi}^T$. The same procedure is applied to update $\boldsymbol{\xi}^S$. This constitutes a single iteration of the algorithm. The algorithm is run four times for 10,000 iterations each. Following Gelman *et al.* (2013) [86], the first 5,000 iterations of each run are discarded to avoid effects from the burn-in period. To assess convergence, the Gelman-Rubin statistic is calculated for each location and each half-week [86]. The Gelman-Rubin statistic is a ratio of the between-chain to the within-chain variance that approaches 1 from above as the number of iterations approaches infinity. The above procedure yields a Gelman-Rubin statistic of below 2 for all chains, and below 1.1 for all $\boldsymbol{\xi}^S$ chains and 37 of the 40 $\boldsymbol{\xi}^T$ chains. This suggests that the chains have converged. The ordered Gelman-Rubin statistics for the $\boldsymbol{\xi}^S$ and $\boldsymbol{\xi}^T$ chains are depicted in Fig 3.9.

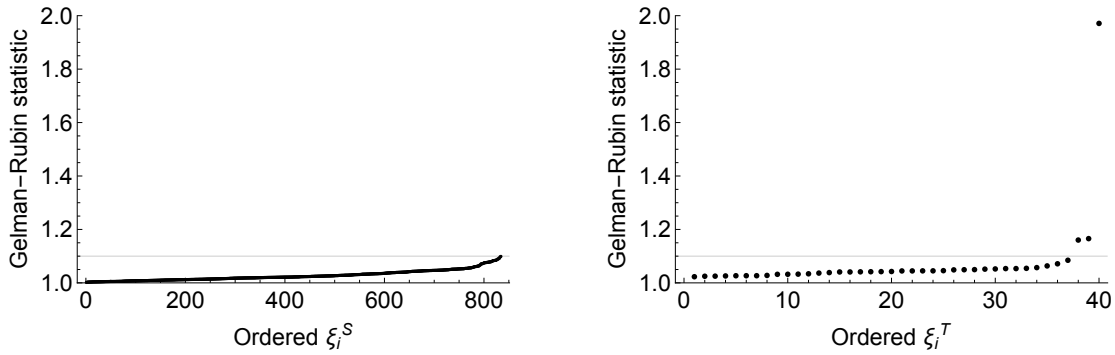



Fig. 3.9 Gelman-Rubin statistics for the spatial ($\boldsymbol{\xi}^S$) and temporal ($\boldsymbol{\xi}^T$) transmissibility surface Markov chains, produced using the Metropolis Hastings algorithm described in §3.2.3. The Gelman-Rubin statistic for all spatial and temporal chains is below 2, and is below 1.1 for all spatial chains and all but three of the temporal chains. This provides good evidence that the chains have converged.

The last 5,000 iterations for each of the four runs are combined, yielding 20,000 posterior estimate draws of $\boldsymbol{\xi}^T$ and $\boldsymbol{\xi}^S$. Fig 3.10 depicts the mean exponentiated $\boldsymbol{\xi}^T$ values with $\pm 2$ standard deviations. The exponentiated $\boldsymbol{\xi}^T$ may be interpreted as a temporally-varying multiplicative factor for the transmissibility term $\beta_d$. There is evidence of a drop in transmissibility in August that rises again from September to mid-October, before dropping again to average by the end of the outbreak. The curve is jagged, reflecting the breakpoint onset detection method's tendency to place epidemic onset times preferentially on half weeks (see §2.3.2); there are local peaks in $\boldsymbol{\xi}^T$ on half weeks and dips on whole weeks. Despite this small-scale bias, there is still a clear trend in the large-scale structure. Fig 3.11 depicts the mean exponentiated $\boldsymbol{\xi}^S$ values geographically, which may be interpreted as a spatially-varying multiplicative factor for $\beta_d$. There is evidence of higher-than-average transmissibility in

the southeast and lower-than-average transmissibility in the mid-Atlantic region where the epidemic wave slowed.

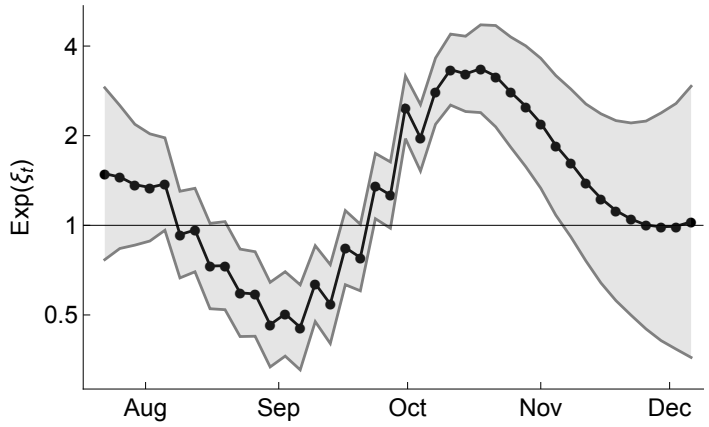



Fig. 3.10 Mean exponentiated temporal transmissibility values, Exp[$\boldsymbol{\xi}^{\boldsymbol{T}}$] (black line), with $\pm 2$ standard deviations (grey band). Values above 1 indicate higher-than-average transmissibility, and values lower than 1 indicate lower-than-average transmissibility. The temporal transmissibility adjustment begins slightly above 1, before dipping in mid-August through the beginning of September. It then rises until mid-October, and finally decreases again back to 1 at the end of the epidemic. The locally jagged pattern is an artefact of the breakpoint onset detection method's preference to place epidemic onsets near half-week values (see §2.3.2); this makes transmission appear to be stronger on half weeks vs. full weeks.

For comparison, the transmissibility adjustments $\boldsymbol{\xi}^T$ and $\boldsymbol{\xi}^S$ are re-estimated using (1) a SE covariance function with spatial length scale of 100 km and temporal length scale of 8 half-weeks, (2) a SE covariance function with spatial length scale of 500 km and temporal length scale of 8 half weeks, and (3) a RQ covariance function with spatial length scale of 200 km and temporal length scale of 8 half weeks. The temporal length scales in all scenarios is kept at 8 half-weeks since a shorter length scale risks over-fitting, while a longer length scale would make the process too inflexible to show significant variation over the 14-week outbreak. The posterior estimates of Exp[$\boldsymbol{\xi}^T$] under each scenario are depicted in Fig 3.12, and the posterior estimates of Exp[$\boldsymbol{\xi}^T$] under each scenario are depicted in Fig 3.13. The overall shape of Exp[$\boldsymbol{\xi}^T$] under all three scenarios is similar to the shape obtained using a SE covariance function with spatial length scale of 200 km and temporal length scale of 8 half weeks (Fig 3.10). All have a dip in transmissibility in September and a rise in transmissibility in late October. This is perhaps unsurprising, since the temporal length scale for all three scenarios is the same, but it does show that inference of the temporal process is fairly robust to changes in the spatial length scale and the covariance function. The

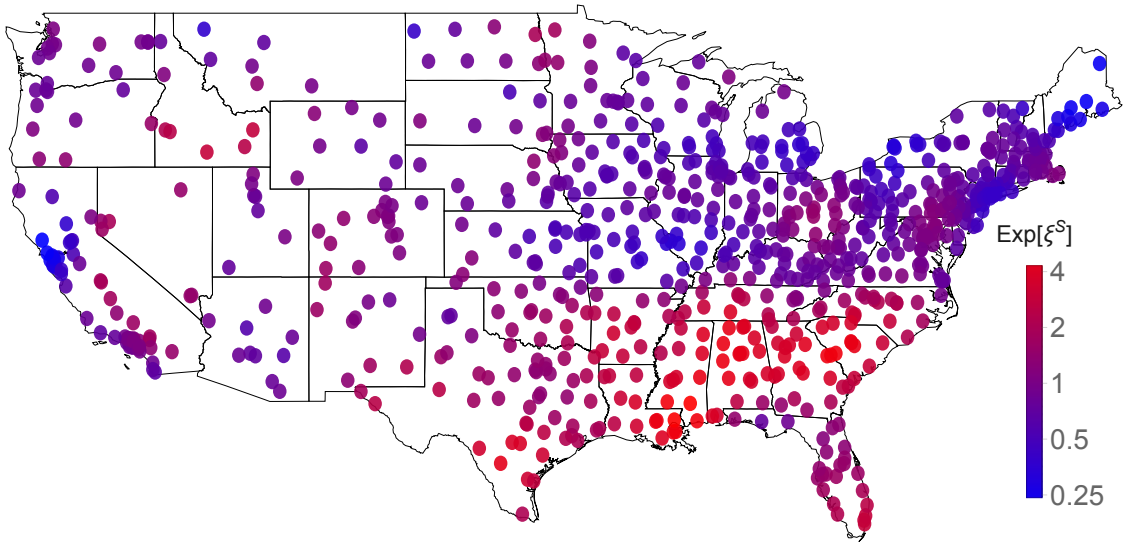



Fig. 3.11 Map of the mean exponentiated geographic transmissibility values, Exp[$\boldsymbol{\xi}^{\boldsymbol{S}}$]. Discs represent ZIPs, and are coloured according to the ZIP's estimated mean Exp[$\xi_i^S$] value. Values higher than 1 (red) indicate higher-than-average transmissibility, and values lower than 1 (blue) indicate lower-than-average transmissibility. The transmissibility adjustment is highest in the southeast, while there is a clear band of low transmissibility in Missouri, Illinois, and Kentucky, where the epidemic wave slowed. There is also evidence of higher-than-average transmissibility in the central valley of California, where a second epidemic wave appears to have been sparked.

spatial transmissibility surfaces depicted in Fig 3.13 also reveal roughly similar patterns as the transmissibility surface in Fig 3.11, with elevated transmissibility in the southeastern US. Under the first scenario, with spatial length scale of 100 km, the transmissibility surface is more locally variable than the surfaces produced using longer spatial length scales. This is to be expected, since a length scale of 100 km yields a very flexible transmissibility surface. The transmissibility surface produced under the second scenario, with spatial length scale of 500 km, is less locally variable than the transmissibility surfaces produced using shorter length scales. The overall range of transmissibility values is also smaller, with $\text{Exp}[\boldsymbol{\xi}^T]$ ranging from 0.3 to 3.5, rather than from about 0.2 to over 6 for the SE scenarios with shorter length scale. This reduced variability is due to the surface's higher rigidity. The transmissibility surface produced under the third scenario, with spatial length scale of 200 km and a RQ covariance function, is similar to the transmissibility surface produced using a SE covariance function and the same spatial length scale (Fig 3.11). This suggests that the estimated transmissibility surface is somewhat robust to the choice of covariance function, though it should be noted that both covariance functions belong to the same general class, yielding a surface with infinite mean-square differentiability everywhere (see §3.1.3).

We proceed using the mean posterior estimates for $\boldsymbol{\xi}^T$ and $\boldsymbol{\xi}^S$ obtained using a SE covariance function with spatial length scale of 200 km and temporal length scale of 8 half weeks. Fig 3.14 depicts the difference between the actual and expected outbreak onset times by location using the new model, Eq 3.45, with these posterior mean estimates substituted in. Comparing with Fig 3.7 indicates that including $\boldsymbol{\xi}^T$ and $\boldsymbol{\xi}^S$ resolves many of the systematic discrepancies. Fig 3.15 depicts the expected and true cumulative number of locations infected over time under the new model. The shapes of the curves now match closely. The gap between the curves points to a remaining model mis-specification. In general, there are always a few more true outbreaks than expected, likely due to mid-range jumps of infection that the model cannot reliably predict. There appear to be about 50 such jumps at any given time; adding 50 to the expected cumulative onset curve makes the two curves match almost perfectly, except at the very beginning and very end of the epidemic. The discrepancy may be due to a mis-specification in the shape of the transmission kernel, and suggests that exploring more flexible kernel forms may be warranted.

### 3.2.4 Further exploration of the Gaussian process fitting procedure

To check whether the above inferences of temporal and spatial variation in the transmissibility of the autumn 2009 A/H1N1pdm outbreak are reliable, epidemics can be simulated using the

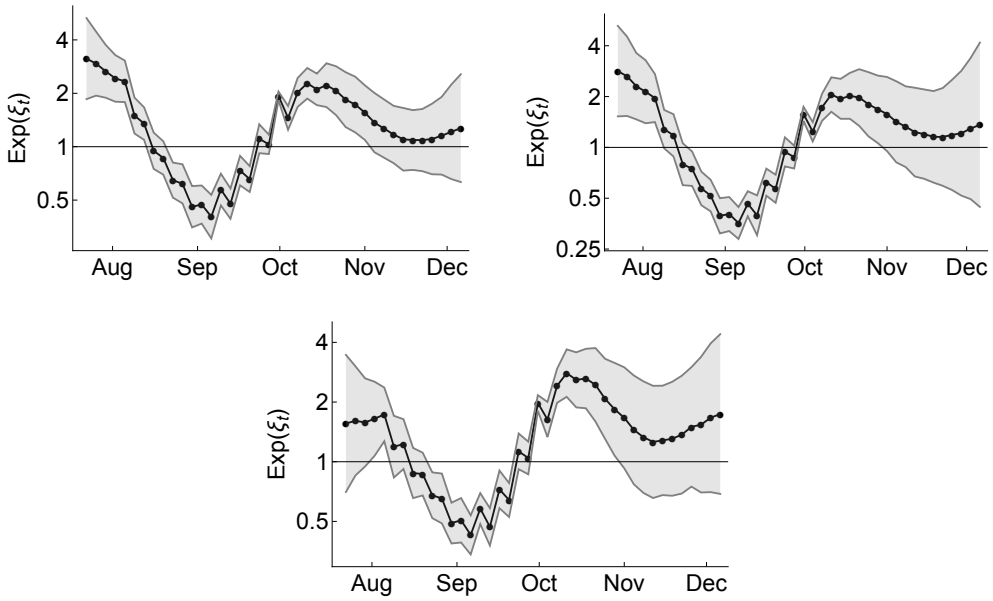



Fig. 3.12 Mean exponentiated temporal transmissibility values, Exp[$\boldsymbol{\xi}^{\boldsymbol{T}}$] (black lines), with $\pm 2$ standard deviations (grey bands), using a SE covariance function with spatial length scale of $l = 100$ km and temporal length scale $l = 8$ half-weeks (upper left), a SE covariance function with spatial length scale $l = 500km$ and temporal length scale $l = 8$ half-weeks (upper right), and a rational quadratic covariance function with spatial length scale $l = 200$ km and temporal length scale $l = 8$ half-weeks (bottom). The exponentiated transmissibility adjustment Exp[$\xi^T$] may be interpreted as a multiplicative factor for the transmissibility term $\beta_d$ (see Eq 3.45) that varies over time. Values above 1 indicate higher-than-average transmissibility, and values lower than 1 indicate lower-than-average transmissibility. All three plots resemble the temporal transmissibility surface fit using a SE kernel with spatial length scale $l = 200$ km and temporal length scale $l = 8$ half-weeks (Fig 3.10).

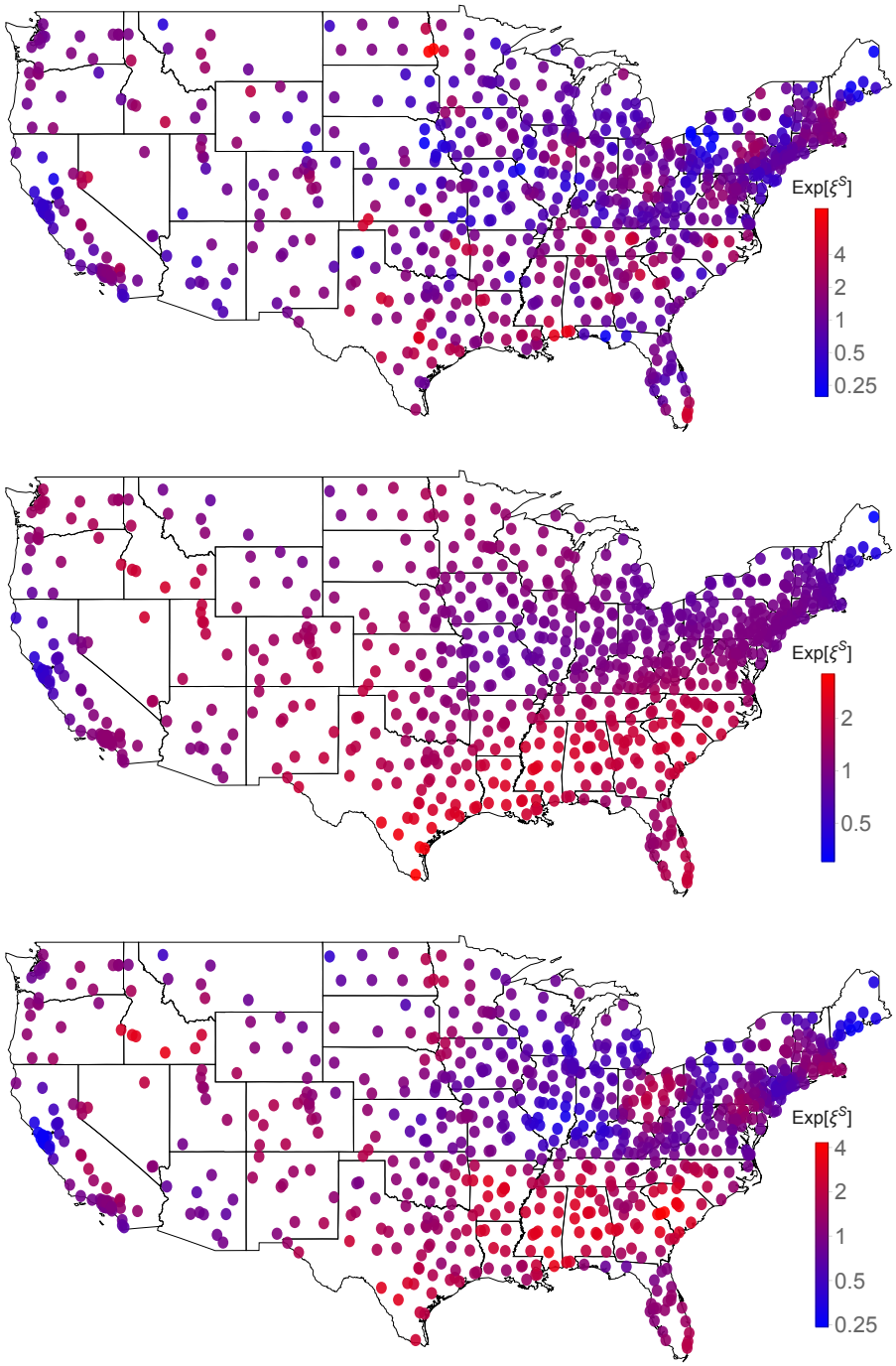



Fig. 3.13 Maps of the mean exponentiated geographic transmissibility values, Exp[$\boldsymbol{\xi}^{S}$], using a SE covariance function with spatial length scale $l = 100$ km (top), $l = 500$ km (middle), and a rational quadratic covariance function with length scale $l = 200$ km (bottom). In all three scenarios, the temporal characteristic length scale is 8 half-weeks. The exponentiated transmissibility adjustment Exp[$\xi^{S}$] may be interpreted as a multiplicative factor for the transmissibility term $\beta_d$ (see Eq 3.45) that varies across space. A SE covariance function with characteristic distance of $l = 100$ km yields a patchy spatial transmissibility surface, though there is still a high concentration of ZIPs in the southeast with elevated transmissibility. A SE covariance function with characteristic distance of $l = 500$ km yields a surface with relatively less variation than the shorter distance scales. A RQ covariance function with $l = 200$ yields a spatial transmissibility surface that closely resembles the one estimated using the SE kernel with the same length scale, depicted in Fig 3.11.

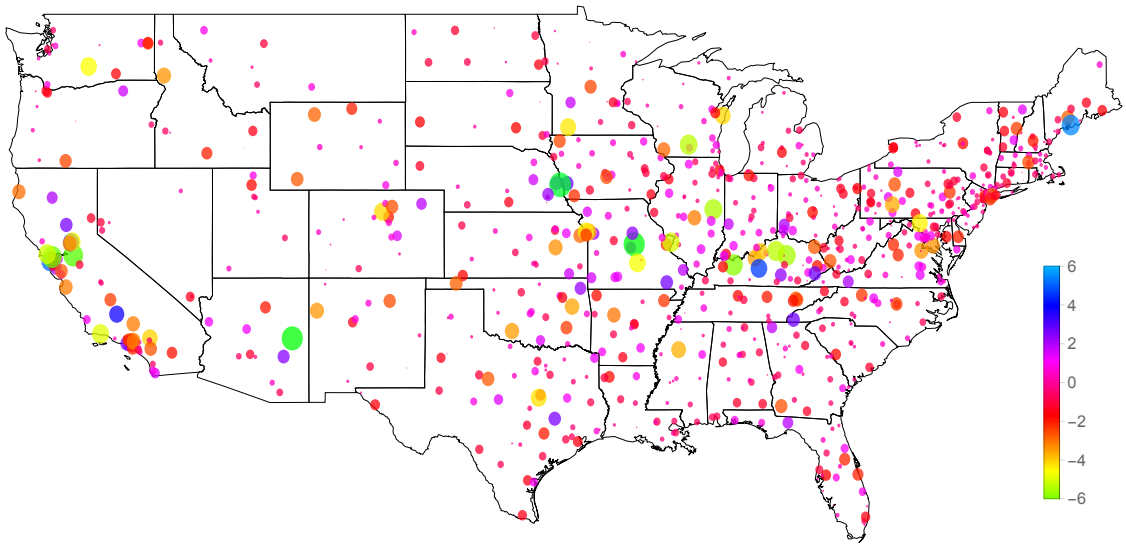



Fig. 3.14 Difference in weeks between the observed and expected epidemic onset time in each ZIP under the transmissibility-adjusted model, Eq 3.45. The area of each disc is proportional to the magnitude of the difference. Blue/purple discs correspond to ZIPs where the true epidemic onset time is later than the expected onset time, and green/red discs correspond to ZIPs where the opposite is true. The band of locations in Missouri, Kentucky, and Virginia with later-than-expected onsets in Fig 3.7 is no longer as apparent.

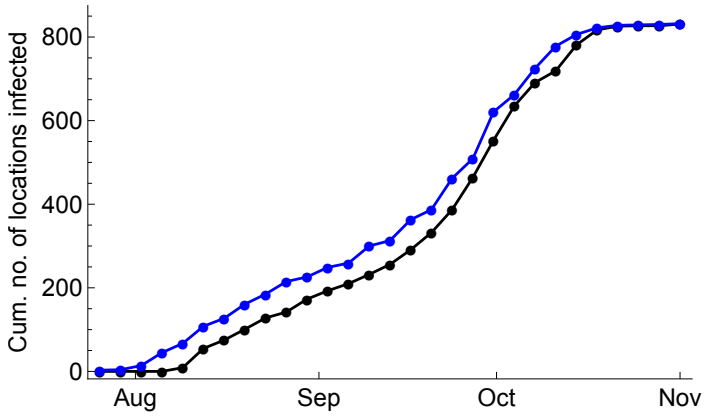



Fig. 3.15 Expected (black) and observed (blue) cumulative number of locations infected over time under the transmissibility-adjusted model, Eq 3.45. The shapes of the two curves match closely. The lag between the expected and observed cumulative number of onsets may be due to mid-range jumps for which the transmission model still does not fully account. If this is true, then there were approximately 50 of these mid-range outbreaks at any given time during the outbreak; adding 50 to the expected cumulative number of outbreaks causes the two curves to match almost perfectly, except at the very beginning and very end of the epidemic.

geographic transmission model Eq 3.37 with known transmissibility $\beta_d$, and the transmissibility surfaces can be estimated using the synthetic onsets. In this section, two scenarios are considered. First, the transmissibility $\beta_d$ is held constant across all locations and for all time, to ensure that the Gaussian process fits do not yield spurious patterns. Second, the transmissibility $\beta_d$ is increased in the southeastern US (HHS regions 4 and 6) to check that the Gaussian process fits can correctly identify authentic differences in transmissibility. In both scenarios, $\beta_d$ is held constant across time.

Outbreaks are simulated from the best mechanistic transmission model Eq 3.37 with $\beta_{ds}$, $\nu$, and $\theta$ all equal to zero. As for the simulations presented in §3.2.2, $\theta$ is held at zero to avoid having to simulate full epidemic curves in each ZIP. Other parameter values are fixed at $\mu = 0.25$, $\rho = 62$km, and $\gamma = 7.8$, equal to the values used to produce Fig 3.6. Also following §3.2.2, epidemics are seeded in the four locations with onset in the first week of the true outbreak, and $\beta_0$ is fixed at 0, so that no additional seeding occurs. For the constant-transmissibility scenario, $\beta_d$ is fixed at 0.61 for all locations throughout the epidemic. For the spatially-varying transmissibility scenario, $\beta_d$ is fixed at 1.64 in the 244 ZIPs in HHS regions 4 and 6, and at $1.64/4 = 0.41$ in the 590 ZIPs in the rest of the country. This makes the mean $\beta_d$ across all locations equal to 0.61, and makes $\beta_d$ in the southeast four times higher than it is in the rest of the country.

Three epidemics are simulated under each scenario. For each simulation, posterior Gaussian process estimates of $\boldsymbol{\xi}^T$ and $\boldsymbol{\xi}^S$ are generated using the procedures described in §3.2.3, using a SE covariance function with spatial length scale $l = 200$ km and temporal length scale $l = 8$ half-weeks. Figs 3.16-3.17 depict the mean posterior $\boldsymbol{\xi}^T$ and $\boldsymbol{\xi}^S$ for the constant $\beta_d$ scenario, and Fig 3.18-3.19 depict the mean $\boldsymbol{\xi}^T$ and $\boldsymbol{\xi}^S$ for the scenario with elevated $\beta_d$ in the southeast. The fits for the constant $\beta_d$ scenario in Figs 3.16-3.17 show no substantial variation in transmissibility over time or space, as expected. For the second scenario with elevated transmissibility in the southeast, the temporal transmissibility surfaces $\boldsymbol{\xi}^T$ are significantly elevated at the start of the outbreak, and decrease near the fourth week of the simulation. This is because three of the four outbreak seeds are in the southeastern US (see Fig 3.6), so the early epidemic is associated with high transmissibility in all three cases. The spatial transmissibility surfaces in Fig 3.17 also reveal elevated transmissibility in the southeastern US. Interestingly, the transmissibility surfaces here have smaller ranges than the transmissibility surfaces obtained using the true onset times (see Figs 3.10 and 3.11). Despite the true transmissibility in the southeast being four times higher than in the rest of the country, the posterior mean estimates for $\text{Exp}[\boldsymbol{\xi}^S]$ in that region are just over 1.25. This underestimate of the true elevation in transmissibility may be due in part to rigidity in the

posterior process imposed by the correlation structure, and may also be exacerbated by the smooth process having difficulty fitting to the sharp, sudden increase in transmissibility in the southeast. If it is true that the Gaussian process fits generally underestimate the overall variation in transmissibility, then the true transmissibility elevation in the southeastern US during the autumn 2009 A/H1N1pdm pandemic wave may have been very pronounced, in order to produce posterior mean $\boldsymbol{\xi}^S$ estimates well over 4.

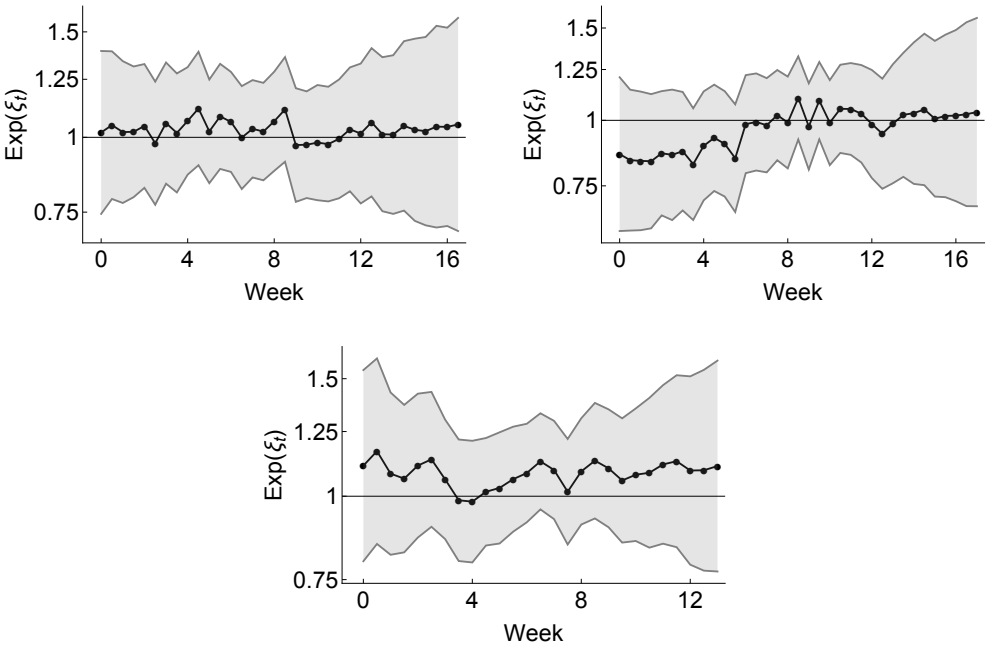



Fig. 3.16 Mean exponentiated temporal transmissibility values, Exp[$\boldsymbol{\xi}^T$] (black lines), with $\pm 2$ standard deviations (grey bands), for three epidemic simulations using the transmission model Eq 3.37 with parameter values $\beta_0 = 0$, $\beta_d = 0.61$, $\mu = 0.25$, $\rho = 62$km, $\gamma = 7.8$, and $\theta = 0$. The outbreaks are seeded in the four locations with earliest onset in during the true autumn wave of the 2009 A/H1N1pdm outbreak in the US (see Fig 3.6). The uncertainty bands all overlap with the horizontal line at Exp[$\boldsymbol{\xi}^T$] = 1, correctly identifying no substantial variation in transmissibility over time.

## 3.3 Discussion

In this chapter, a mechanistic transmission model is presented that describes the outbreak onset times in 834 3-digit ZIP codes in the United States during the 2009 A/H1N1pdm influenza pandemic. A model selection procedure indicates that the recipient ZIP's population size, its surrounding population density, the distance to neighbouring infected ZIPs, and

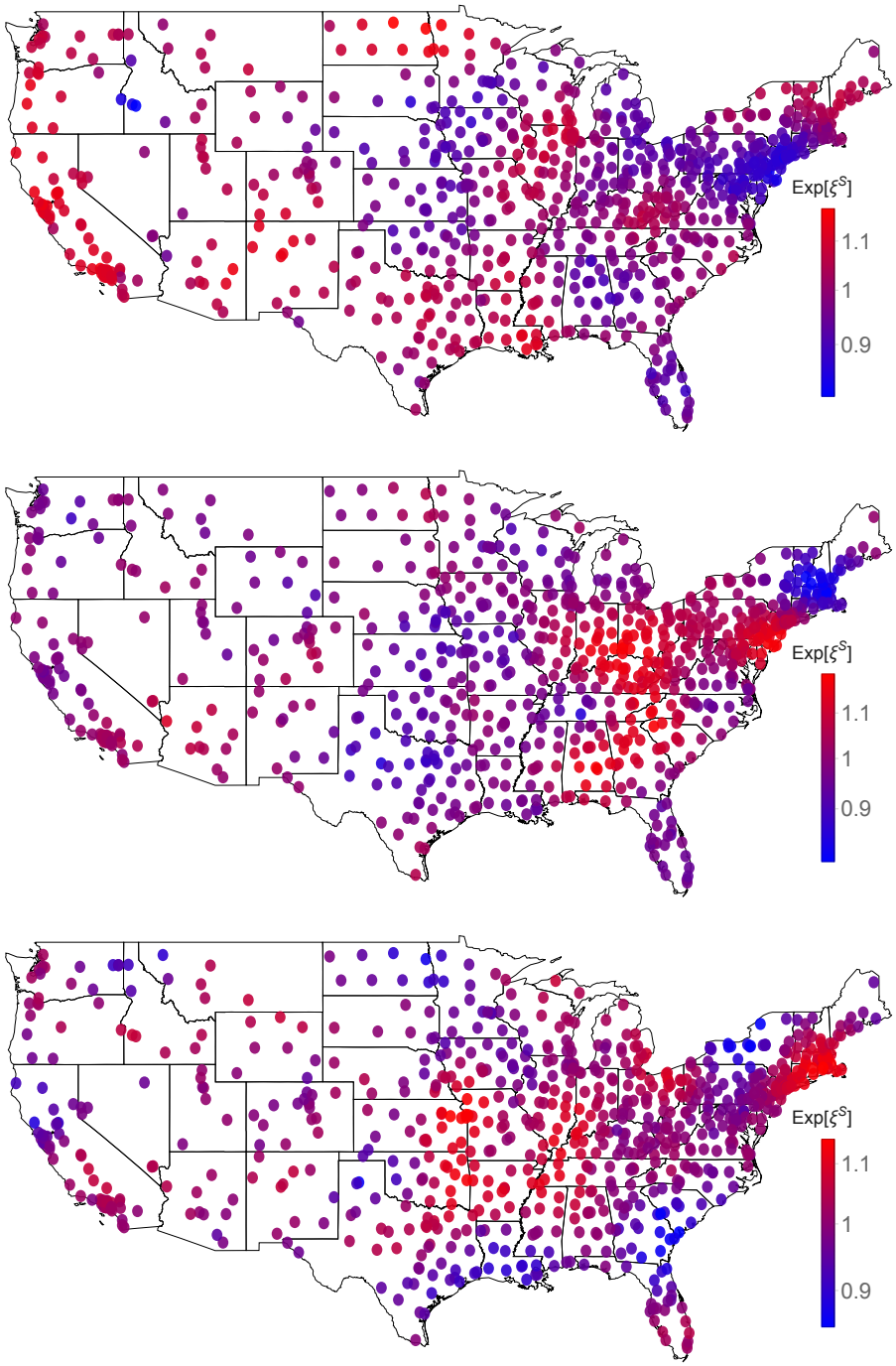



Fig. 3.17 Maps of the mean exponentiated geographic transmissibility values, Exp[$\boldsymbol{\xi}^S$] for three epidemic simulations using the transmission model Eq 3.37 with parameter values $\beta_0 = 0$, $\beta_d = 0.61$, $\mu = 0.25$, $\rho = 62$km, $\gamma = 7.8$, and $\theta = 0$. The outbreaks are seeded in the four locations with earliest onset in during the true autumn wave of the 2009 A/H1N1pdm outbreak in the US (see Fig 3.6). Though the surfaces show some spatial variation, the patches of high/low transmissibility are not consistent across the three simulations. The range of posterior mean Exp[$\boldsymbol{\xi}^S$] estimates is small, between 0.8 and 1.2 in all three cases, correctly identifying virtually no substantial spatial variation in transmissibility.

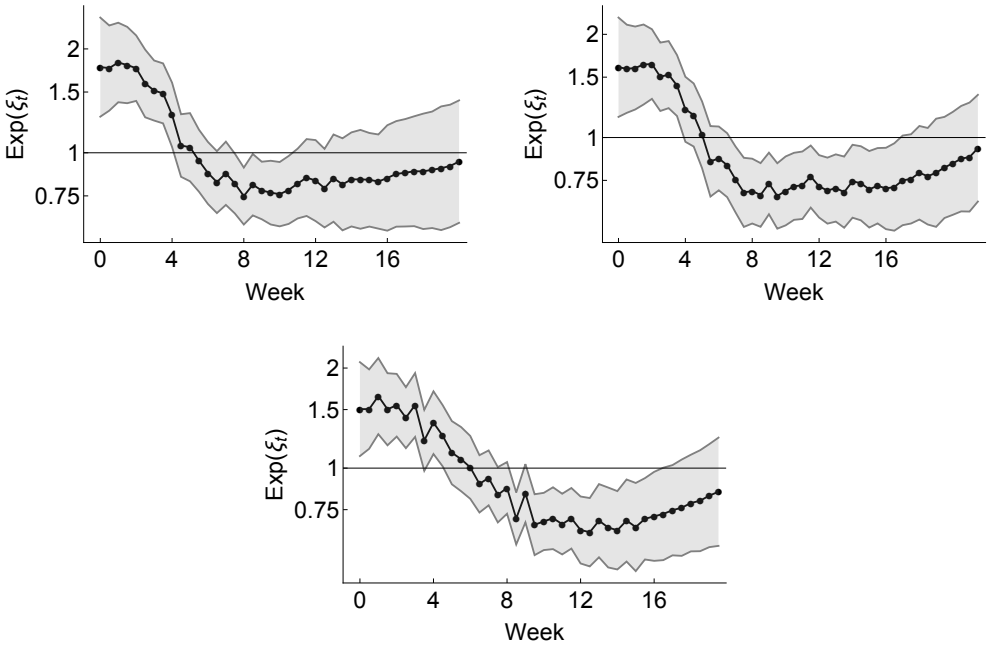



Fig. 3.18 Mean exponentiated temporal transmissibility values, Exp[$\boldsymbol{\xi}^{\boldsymbol{T}}$] (black lines), with $\pm 2$ standard deviations (grey bands), for three epidemic simulations using the transmission model Eq 3.37 with parameter values $\beta_0 = 0$, $\mu = 0.25$, $\rho = 62$km, $\gamma = 7.8$, and $\theta = 0$. The transmissibility parameter $\beta_d$ is four times higher in the southeast (HHS regions 4 and 6) than in the rest of the country, at a value of 1.64 vs. 0.41. The mean $\beta_d$ across all locations is 0.61. The outbreaks are seeded in the four locations with earliest onset in during the true autumn wave of the 2009 A/H1N1pdm outbreak in the US (see Fig 3.6). The posterior mean Exp[$\boldsymbol{\xi}^{\boldsymbol{T}}$] is elevated at the start of the outbreak, reflecting the fact that the epidemics generally begin in the southeast, where transmission strength is high. The transmissibility estimates decrease below 1 after the epidemic has passed out of the southeast into areas with below-average transmissibility, around week 4.

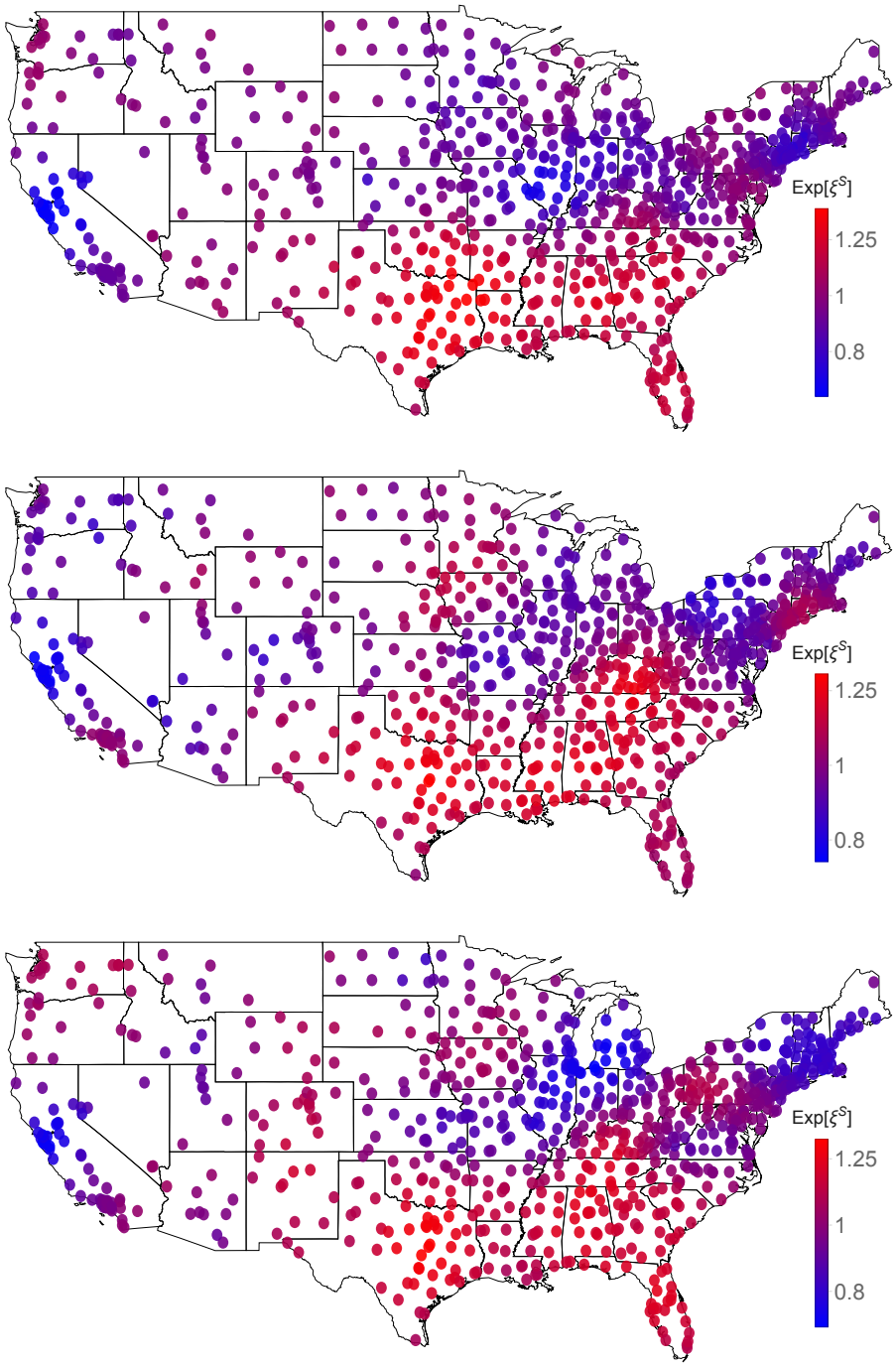



Fig. 3.19 Maps of the mean exponentiated geographic transmissibility values, Exp[$\boldsymbol{\xi}^{S}$] for three epidemic simulations using the transmission model Eq 3.37 with parameter values $\beta_0 = 0$, $\mu = 0.25$, $\rho = 62$km, $\gamma = 7.8$, and $\theta = 0$. The transmissibility parameter $\beta_d$ is four times higher in the southeast (HHS regions 4 and 6) than in the rest of the country, at a value of 1.64 vs. 0.41. The mean $\beta_d$ across all locations is 0.61. The outbreaks are seeded in the four locations with earliest onset in during the true autumn wave of the 2009 A/H1N1pdm outbreak in the US (see Fig 3.6). The posterior mean Exp[$\boldsymbol{\xi}^{S}$] is elevated in the southeast in all three cases, correctly identifying the imposed geographic variation in transmissibility, though underestimating the magnitude of the difference.

the intensity of the outbreaks in those neighbouring ZIPs all predict a given ZIP's outbreak onset time. There is also some evidence that mixing between children a week before the start of the autumn school term may have contributed to transmission, though this effect is only identified in sub-models that ignore the intensity of the outbreaks in neighbouring ZIPs ($\theta = 0$).

The best transmission model's rejection of an effect from the start of the autumn school term deserves further commentary, since on the surface it seems to contradict a range of other findings that schools act as catalysts for influenza transmission [91, 117, 199, 255]. Indeed, rises in ILI intensity corresponded closely with the start of the autumn school term in ZIPs in the southeastern US, where the major transmission wave of the autumn 2009 epidemic originated. As the outbreak progressed, however, epidemic onset times lagged increasingly behind the start date of the autumn school term, while infection in neighbouring ZIPs remained a good predictor of epidemic onset time. This ultimately leads the best transmission model to include an effect from the neighbouring ILI intensity and to reject an effect from the start of the autumn school term. It is possible that heightened interpersonal mixing between children in schools made cities receptive to epidemic "sparks", but that other factors were more important for specifying a city's precise outbreak onset time. The relative roles of different age groups in the transmission of the autumn 2009 A/H1N1pdm pandemic in the US is analysed in greater depth in Chapter 5.

In previous studies, power-law and exponential distance kernels have been used to describe the geographic transmission range of the 2009 A/H1N1 influenza pandemic in the United States [48, 91, 133, 239]. The hybrid kernel considered here (Eq 3.28) is chosen to help identify which type of kernel is preferred. The best transmission model in terms of AIC uses a power-law kernel, with finite power-law decay $\gamma$, but the MLE estimate for $\gamma$ is high, so that the optimal kernel is in practice very nearly exponential. In general, power kernels have thicker tails than exponential kernels, and so geographic epidemic processes that spread according to power law kernels tend to have relatively more long-range jumps than ones that spread according to exponential kernels. The high MLE value for $\gamma$ identified here suggests that long-range jumps of infection occurred during the 2009 outbreak in the US, but were rare.

There are many other types of distance kernels that might be tested. An especially rich set of kernels may be found in the literature on pollen dispersal [7]. It would be interesting to see whether any of these kernels yield a better model fit, and in particular whether they clear up the discrepancy between the predicted and true cumulative number of onsets over time depicted in Fig 3.15. Alternatively, a more rigorous approach for choosing a kernel would

involve tracking the movements of individuals over a period of time, as in Read *et al.* (2014) [197], and either parametrising a functional form or fitting a non-parametric curve to the distribution of trip distances. In the latter case, a cubic spline might be fit to the distribution of trips between ZIPs, and vertical and horizontal stretch parameters would be estimated as part of the transmission model. Unfortunately, to my knowledge, movement data from the US with sufficient detail to make such inferences do not exist.

The parameter $\theta$, which extends the transmission model presented in Gog *et al.* (2014) [91], has a few possible interpretations. According to the MLE value for $\theta$, the force of infection in a given ZIP is approximately related to the square root ($\theta \approx 0.5$) of the normalised number of infected individuals in the neighbouring ZIPs, $n_{j,t}$. Setting $\theta < 1$ vertically squashes the neighbouring ILI curve relative to the $\theta = 1$ case, so that the peak contributes relatively less and the trough contributes relatively more to the force of infection (see Fig 3.20). This could be due in part to variable behaviour over the course of the outbreak; for example, it is possible that healthcare providers are more vigilant for influenza-like illness near the peak of an outbreak, and individuals are more likely to seek healthcare. This would artificially inflate the number of observed cases near the peak of the outbreak. Setting $\theta < 1$ could help offset this effect. Alternatively, it is possible that infected individuals may be less likely to travel during the peak of an outbreak than when baseline influenza levels are low. This would somewhat reduce the risk of infection to neighbouring cities during the peak.

The value of $\theta$ may also be influenced by the basic reproduction number, $R_0$, of the disease. Assuming the within-city infection process may be modelled as a Poisson branching process, the risk of sparking an outbreak in a susceptible city with $k$ independent introductions of infection is

$$P(infect) = 1 - \left(\frac{1}{R_0}\right)^k \tag{3.48}$$

as shown in Keeling and Rohani (2011) [129]. Alternatively, according to the transmission model presented in this chapter, the probability that a ZIP $i$ becomes infected in a given time step $t$ may be expressed as

$$P(infect) = 1 - \text{Exp}(-\lambda_i(t)). \tag{3.49}$$

If we consider the force of infection from just one city $j$, and interpret the force of infection $\lambda$ as a function of $\theta$ (holding all other parameters constant), we may approximate the force of infection on city $i$ from city $j$ as $\lambda_{i,j}(t) \approx c_1 n_{j,t}^{\theta}$, where $c_1$ is a constant that captures all

other parts of the transmission model besides $n_{j,t}$. Eq 3.49 may then be re-written as

$$P(infect) = 1 - \text{Exp}(-c_1 n_{j,t}^{\theta}). \tag{3.50}$$

If we assume that the number of introductions of infection, $k$, to city $i$ from city $j$ is proportional to the number of people who are infected in city $j$ (that is, $k \propto n_{j,t}$), then we may express Eq 3.50 as

$$P(infect) = 1 - \text{Exp}(-c_2 k^{\theta}) \tag{3.51}$$

where the $c_2$ is another scaling constant. Now we may compare Eq 3.48 and Eq 3.51, both of which express the probability of infection as a function of the number of introductions $k$ of infection. The two equations have fundamentally different forms, but both describe monotonically increasing functions that are equal to 0 when $k = 0$, and approach 1 as $k$ increases (see Fig 3.21). It is possible that the value of $\theta$ is chosen so that the graphical form of Eq 3.51 matches the form of the 'true' probability of infection, modelled by 3.48, as closely as possible. This suggests that $\theta$ may be linked to the basic reproduction number $R_0$ of the disease.

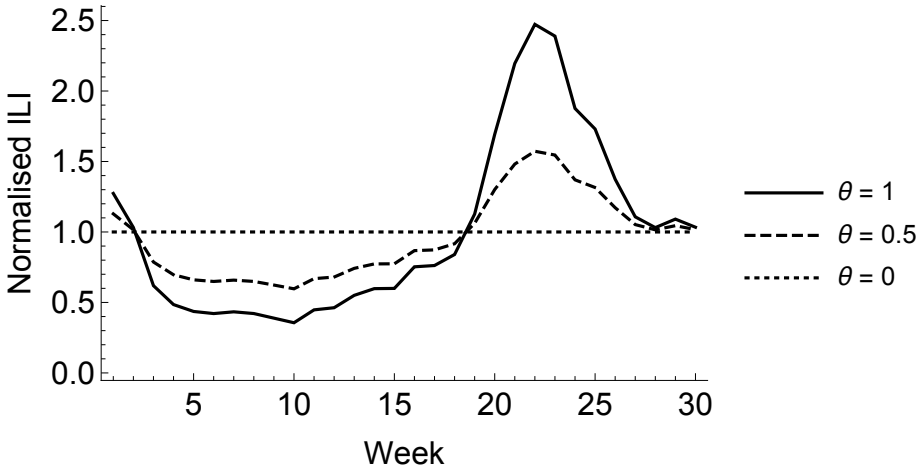



Fig. 3.20 Normalised ILI time series for ZIP 606 (Chicago IL) from the last 30 weeks of 2009, taken to various powers $\theta$. Values of $\theta < 1$ dampen the relative intensity of the peak compared with the rest of the time series. The normalised time series approaches a flat line as $\theta$ approaches 0 from above.

The parameters for the mechanistic transmission model are fit using a maximum likelihood strategy, as in [91]. Sensitivity of the model parameter values to uncertainty in the outbreak onset times is assessed by re-drawing onset times and re-fitting the model

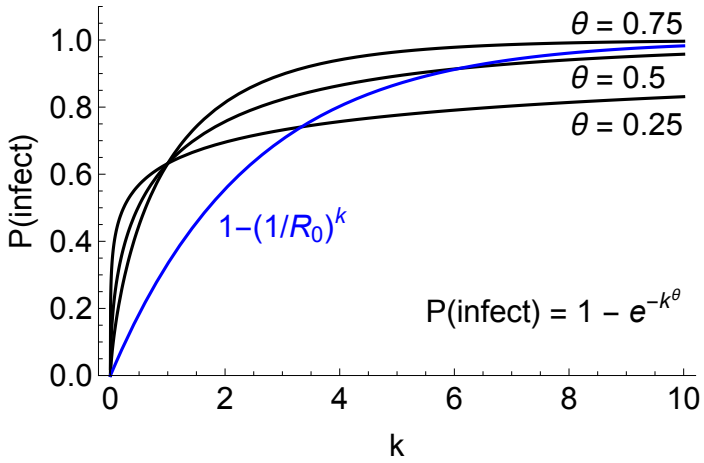



Fig. 3.21 Probability of infection as a function of the number of introductions $k$ of infection into a susceptible population, as described by two different models. The blue line corresponds to the assumption that the within-city transmission of disease follows a Poisson branching process, yielding Eq 3.48. The black lines correspond to the assumption that the probability of infection may be derived from the force of infection exerted by a neighbouring city, leading to Eq 3.51. Both ways of describing the probability of infection give monotonically increasing functional forms. The value of $\theta$ could be chosen so that the probability of triggering an outbreak as described by the transmission model (black) most closely matches the underlying 'truth' (blue), which is dictated by the basic reproduction number $R_0$.

parameters. Another option would be to define a full likelihood function for the model parameters that takes into account the onset time uncertainties, and then to fit the mechanistic transmission model parameters using a MCMC scheme, as in [75]. This would offer a more unified approach for incorporating the onset time uncertainties into the full model, but is also more computationally intensive, and so for now is left for future work. The parameter estimates and the optimal model structures resulting from the maximum likelihood and the MCMC schemes should be compared, to check that the model fitting procedure does not lead to conflicting qualitative interpretations of the 2009 pandemic's transmission.

Another potentially useful model extension would be to incorporate a full stochastic SIR model into the mechanistic transmission model. In this scenario, the model parameters would be fit to the entire ILI curve, rather than just the outbreak onset times. Due to the high degree of noise and heterogeneity between the ILI curves for different ZIPs, it is likely that an adaptive fitting algorithm, such as the particle filtering method employed by Yang *et al.* (2015) [257], would be needed. However, this would involve not just a significant increase in methodological complexity, but it would also require a fundamentally more complex underlying model: for example, the model would have to account for geographic differences in ILI reporting rates to fully explain the observed differences in ILI intensity between locations. Furthermore, Viboud *et al.* (2014) [240] find that estimates of epidemic intensity from the IMS-ILI data may not be fully reliable, so it may be difficult to defend epidemiological conclusions based on an SIR model fit explicitly to the IMS-ILI curves. Developing and fitting such a model remains a potentially fruitful area for future work, but lies beyond the scope of this thesis.

Simulations verify that the fundamental geographic transmission model, Eq 3.37, can reproduce the overall epidemic timing and wave-like trajectory observed during the autumn wave of the 2009 A/H1N1pdm outbreak in the United States. However, the fundamental model does not fully account for variation in the geographic structure of the outbreak at the regional scale. Plotting the difference between the true ZIP-level outbreak onset times and the outbreak onset times predicted by the fundamental transmission model reveals a clear horizontal band in the eastern US where the epidemic wave appears to have slowed. To account for this variation, a temporal and a spatial correction parameter, $\boldsymbol{\xi}^T$ and $\boldsymbol{\xi}^S$, are introduced into the model. These parameters are each constrained to follow a Gaussian process. Fitting these parameters to the data reveals a dip in transmissibility in early September and a peak between October and November, as well as elevated transmissibility in the southeast and decreased transmissibility in the north, with an apparent dividing line in southern Missouri, Illinois, and Kentucky. While it remains unclear what causes this variation, it is possible that

geographic differences in weather-related factors, such as temperature and humidity, may be responsible [151, 209]. The temporal variation could also be attributed to changes in human behaviour: the dip in transmissibility might reflect an abundance of caution in the initial phases of the epidemic that subsided as people realised that the disease was not as severe as initially feared. Poletti *et al.* (2011) [193] attribute a sudden increase in weekly incidence of A/H1N1pdm influenza in mid-October of 2009 in Italy, similar to the one observed in the United States, to changes in human behaviour associated with a reduced perception of risk. An interesting area for future study would be to seek factors that co-vary with the posterior $\boldsymbol{\xi}^T$ and $\boldsymbol{\xi}^S$ through regression studies and incorporate these into the transmission model, to see if a better fit might be obtained.

The temporal and spatial posterior transmissibility surfaces $\boldsymbol{\xi}^T$ and $\boldsymbol{\xi}^S$ should not be interpreted in isolation from one another. The second set of simulation studies in §3.2.4, with elevated transmissibility in the southeast, demonstrate that spatial variation in transmissibility can make it appear that there is also temporal variation in transmissibility. The $\boldsymbol{\xi}^T$ and $\boldsymbol{\xi}^S$ are best interpreted as joint windows into the same underlying process. For the autumn wave of the 2009 A/H1N1pdm outbreak in the US, transmissibility appears to have been high during the early transmission of the outbreak in the southeastern US, decreased in the mid-autumn as the outbreak spread from that region, and then rose again to an average value as the epidemic traversed the rest of the country in the late autumn. These differences may be due to either spatial or temporal changes in transmissibility, or both. Further simulations are warranted to determine the extent to which explicitly spatial or temporal differences in transmissibility are distinguishable using the methods presented above. These might involve simulating outbreaks spreading on a hypothetical metapopulation under various transmissibility scenarios, including fixed spatial/varying temporal transmissibility, fixed spatial/varying temporal transmissibility, and varying spatial/varying temporal transmissibility. The posterior transmissibility surfaces would be estimated from the simulated outbreak onset times using the methods described in this chapter, and these would be compared with the true variation in transmissibility used to generate the simulations. It is possible that the ability to distinguish between spatial and temporal variation in transmissibility will increase with the number of independent introduction sites, since, for example, a universal slowdown in all sub-outbreaks would be best explained by temporal variation in transmissibility, while local slowdowns would be best explained by spatial variation in transmissibility. Ideally, one would estimate an independent transmissibility value for each location at each time point, rather than locking the temporal variation in transmissibility across all locations, as is done in the methods presented in this chapter. In such a scenario, one might assume that the transmissibility at

each location is correlated with its neighbours and also with its own transmissibility at the previous time step(s). While theoretically possible, this more general case requires vastly more transmissibility values to be estimated, making the overall calculation computationally prohibitive.

Simulating outbreak onset times from the geographic transmission model and re-fitting the Gaussian process transmissibility surfaces helps to reveal the relationship between the transmissibility surface fits and the true underlying epidemic dynamics. While the Gaussian process fits correctly identify no spatiotemporal differences in transmissibility when $\beta_d$ is held constant constant across time and space, the Gaussian process fits underestimate the variation in transmissibility for the simulations with elevated $\beta_d$ in the southeast (Figs 3.18-3.19) by a factor of over 2. If this underestimation extends to the Gaussian process fits on the true data, then it would appear that transmissibility in the southeastern US during the spread of the autumn 2009 pandemic wave was comparatively explosive, up to eight times higher than in many other parts of the country. However, there are a few reasons why the true elevation in transmissibility in the southeast may not have been quite so severe. First, the smooth Gaussian processes generated using SE and RQ covariance functions are not well-suited to identify sudden jumps in an observed process, such as the sudden jump in transmissibility built into the epidemic simulations [196]. This unrealistic ridge could contribute to the underestimation observed in the simulation study. Second, it appears that the temporal process artificially absorbs some of the spatial variation in transmissibility in the simulations, reducing the observed spatial variance in the Gaussian process fits. Third, the transmission model itself may provide an incomplete description of how local epidemic establishment occurs. As it stands, the transmission model only accounts for outbreaks that are caused by active seeding of infection from neighbours or from far-away reservoirs of infection. Alternatively, the A/H1N1pdm influenza virus may have been circulating at low levels in the late summer of 2009, and some combination of schools opening in the southeast, favourable meteorological conditions, and/or changes in interpersonal contact behaviour could have triggered many outbreaks in the southeast nearly simultaneously. This would lead the transmission model to infer extremely high transmissibility, to account for the sudden onset of many nearby outbreaks. More research is needed to better understand whether and to what extent influenza virus circulates in populations between seasons, and also to identify exogenous factors that may trigger the sudden establishment of infection in a population.

## 3.4 Summary

A mechanistic geographic transmission model is developed to explain the epidemic onset times in 834 3-digit ZIP codes across the United States during the autumn wave of the 2009 A/H1N1pdm influenza pandemic. Population size, population density, the distance to neighbouring ZIPs, and the epidemic intensity in those same ZIPs predict outbreak onset times. There is some evidence that mixing among children a week before the start of the autumn school term also facilitated transmission. To account for regional variation in transmissibility, time- and location-specific adjustments are built into the model following Gaussian process priors. Fitting these adjustment factors to data reveals that overall transmissibility dipped in early September and rose again between October and November, that transmissibility in the southeast was higher than average, and transmissibility in the north was lower than average.

# Chapter 4

# Transmission hubs of the autumn wave of the 2009 A/H1N1pdm influenza pandemic in the United States

A central goal in epidemiology is to identify the geographic sites where an epidemic first becomes established. These are normally places to which infection is introduced from somewhere outside the population, sparking a chain of outbreaks within the population. We refer to establishment sites of this type as 'hubs', since they are locations through which infection passes from outside the population to within it. In this chapter, a general mathematical strategy is presented to identify the transmission hubs of an outbreak that occurs on a metapopulation. Using this strategy, the transmission hubs of the 2009 A/H1N1pdm influenza pandemic in the United States are identified, and the total onward transmission triggered by each hub is mapped. Most of the ZIP-level outbreaks in the US can be traced back to three hubs, in Albany West GA, Grenada MS, and Stockton CA. Onward transmission from these and an additional six hubs accounts for 90% of the observed ZIP-level outbreaks in the autumn of 2009. Mapping the onward transmission triggered by each hub yields a hypothesis of where distinct viral strains may have circulated.

Sections 4.2–4.5 are adapted from "Geographic Transmission Hubs of the 2009 Influenza Pandemic in the United States" (Kissler *et al.* 2017 [133]), submitted to *Epidemics*.

## 4.1 Background

### 4.1.1 Terms for various epidemiological hotspots

Identifying 'hotspots' of disease transmission is a key part of describing the transmission history of an epidemic. Hotspots come in a variety of forms: examples include individual people who pass infection on to many neighbours, geographic locations responsible for infecting large regions of a country, individuals or locations in which genetic recombination frequently occurs yielding new viral strains, or locations where contact between humans and animals leads to host species crossover. Each type of hotspot contributes to transmission in a distinct way, but overlapping nomenclature can lead to confusion between the concepts. Here, a variety of hotspots are named and described to distinguish them from the central theme of this chapter, the transmission hub.

*Superspreaders* are either individuals or geographic locations that spread disease to many immediate neighbours [188]. The term has existed since at least 1973, when it was conjectured that individual superspreaders might contribute to the transmission of influenza [79]. Lloyd-Smith *et al.* (2005) [154] provide a rigorous definition of "super-spreading events", and demonstrate that individual-level variability in disease transmission rates (i.e. the presence of superspreaders) leads to more frequent epidemic extinction, but also to more explosive outbreaks. Importantly, superspreaders need not be index cases of a disease. Rather, a disease may be circulating at low levels within a population, reach a superspreader, and then spread rapidly.

*Sources*, on the other hand, are locations in which novel genetic strains of a pathogen emerge, or sometimes where animal-to-human transmission first occurs; often, these coincide [60, 140, 204, 241]. This term is also, perhaps misleadingly, sometimes used to refer to reservoirs of infection, like schools and workplaces [102].

In contrast, a *hub* is a location through which infection circulating outside a population passes into the population. Hubs are sites that (1) receive a long-distance jump of infection that (2) sparks significant onward transmission within the population. These too are sometimes called 'sources' [258] – but we avoid that term here. Taken together, outbreaks generally emerge from a source, enter into distinct populations via hubs, and spread explosively once they reach superspreaders (see Fig 4.1).

Various strategies exist for identifying specific superspreaders [212, 226] and for testing whether superspreaders may have significantly contributed to the transmission of an outbreak [83, 154]. Similarly, there are techniques to identify epidemic sources, using both epidemiological and genetic data [95, 140, 204, 241]. A few strategies have also been proposed to

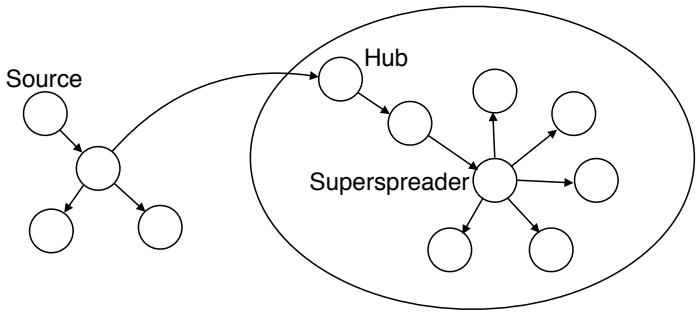



Fig. 4.1 Schematic diagram depicting the relationship between a source, a superspreader, and a hub. Circles could represent cities or individual people, the ellipse could represent a country or a community of interest, and arrows represent disease transmission. In general, a disease first emerges in a source, enters into a population of interest via a hub, and spreads explosively through transmission by superspreaders.

identify transmission hubs. LeGrand *et al.* (2009) [146] develop a back-calculation method to identify the probable location of an aerosol pathogen release, motivated by a hypothetical bioterrorism event. The algorithm assumes diffusive spread of the pathogen, and can only identify a single source of infection. Levy *et al.* (2011) [150] re-trace an outbreak of Chagas disease in Peru, also using a diffusion-based transmission model. Their strategy allows for the identification of multiple hubs, but the number of hubs must be specified before carrying out the calculation. Yang *et al.* (2015) [258] identify transmission hubs of the 2014-2015 Ebola virus outbreak in Sierra Leone using a gravity-based transmission model and an MCMC Kalman filtering method. There remains a need for a mathematical strategy that identifies multiple hubs automatically, using a mechanistic model, without relying on the convergence of MCMC algorithms that can take large amounts of time and computing power when dealing with big datasets.

In the particular case of influenza, it is generally accepted that many new viral strains originate in East and Southeast Asia [204], and that superspreaders may contribute to the transmission of disease [138, 191]. Charu *et al.* (2017) [48] propose an empirical hub-finding technique for cities in the United States where outbreak onset times are available. Under that method, the distance to the nearest previously-infected city is recorded for each city at its outbreak onset time, and hubs are defined as the locations with distances in the top 1% of this distribution. To my knowledge, no other study has attempted to identify the transmission hubs for influenza at the country scale.

### 4.1.2 Epidemiology and phylogeography

Phylogeography refers to the inference of an organism's geographic spread using geo-tagged genetic sequences and the statistical tools of genetic phylogeny [115]. Like traditional epidemiology, the central aim of pathogen phylogeography is to reconstruct the transmission history of an outbreak. Combining epidemiological and genetic data offers a powerful way to improve inferences of an epidemic's geographic transmission history. However, there remains a need for robust theoretical frameworks that synthesise these two disparate data sources [95].

One way to link epidemiological data with genetic data in a geographic setting is to use phylogeographic methods to infer transmission chains, and then to refine these inferences using epidemiological data [63, 64, 142, 166]. Since inferring phylogenies is computationally intensive, these approaches have to my knowledge only been applied to relatively small outbreaks, consisting of up to 150 simulated cases, or up to about 50 real cases. Alternatively, it is possible take the reverse approach, and use epidemiological data to generate hypotheses for the geographic patterns that should be visible in genetic data. This strategy is less common, probably because epidemiological datasets with sufficient spatial resolution to reconstruct detailed geographic transmission routes are rare. When such data are available, however, it should in theory be possible to use a mathematical model to reconstruct how an observed outbreak spread from a set of possible introduction sites. Using this reconstruction, geographic regions could be delineated where infection was triggered predominately by onward transmission from a particular introduction site. According to the ecological founder effect, a high prevalence of genetically-related pathogens should be found within each of these regions, due to the pathogens having a common ancestor, while between-region genetic variance should be higher. The region structure can therefore be tested and refined using genetic clustering analyses. This chapter lays a foundation for such an approach, by presenting a method to infer introduction sites and the extent of onward transmission of an outbreak using a metapopulation transmission model and known outbreak onset times in each sub-population. The method is demonstrated using city-level outbreak onset times from the autumn wave of the 2009 A/H1N1pdm influenza pandemic in the United States.

## 4.2 Mathematical framework

Given a mathematical model that describes outbreak onset times in distinct metapopulation patches as an additive force of infection from each previously-infected patch, a transmission

network may be constructed that describes all possible routes of transmission. Reversing this transmission network in a particular way yields a Markov chain that can be used to trace the epidemic to its most likely sites of introduction. This section describes this strategy in detail.

### 4.2.1 Characterising the forward transmission network

Suppose an epidemic occurs on $n$ metapopulation patches, or 'locations'. Each location's outbreak onset time is observed. Without loss of generality, let the locations be assigned indices $i = 1, \ldots, n$ in order of their outbreak onset time, from earliest to latest. Consider a function $\lambda_i$ that characterises the force of infection on location $i$ at its outbreak onset time as the sum of the forces exerted by all previously-infected locations, plus some background risk of seeding from outside the metapopulation. That is,

$$\lambda_i = \beta_0 + \sum_{j=1}^{i-1} \lambda_{i,j} \tag{4.1}$$

where $\beta_0$ is the force from external seeding, and $\lambda_{i,j}$ is the force exerted on location $i$ by location $j$. The first outbreak ($i = 1$) could only have been triggered by external seeding, so define $\sum_{j=1}^{0} \lambda_{1,j} = 0$.

The partial forces of infection $\lambda_{i,j}$ can be visualized as a transmission network, as depicted in the left-hand diagram in Fig 4.2. Locations are represented by nodes (circles), connected with arrows that indicate possible transmission pathways. In addition, $n$ 'seeding states' (clouds) are introduced, each of which exerts a force of $\beta_0$ on a single location. Summing the forces of all of the arrows going into a node $i$ yields Eq 4.1.

### 4.2.2 Reversing the infection process

To identify epidemic hubs, transmission chains are traced probabilistically back to likely points of introduction. This is done by reversing the direction of the transmission network and noting that, with the proper normalisations, the resulting 'reverse transmission network' represents a Markov chain for which the probability of transitioning from state $i$ to state $j$ is equivalent to the probability that location $i$ was infected by 'parent' location $j$. The aim is to trace each outbreak to a most probable parent, then to a most probable parent's parent, and so on, until the outbreak is ultimately traced back to a first ancestor, where the infection was introduced into the system – that is, a hub.

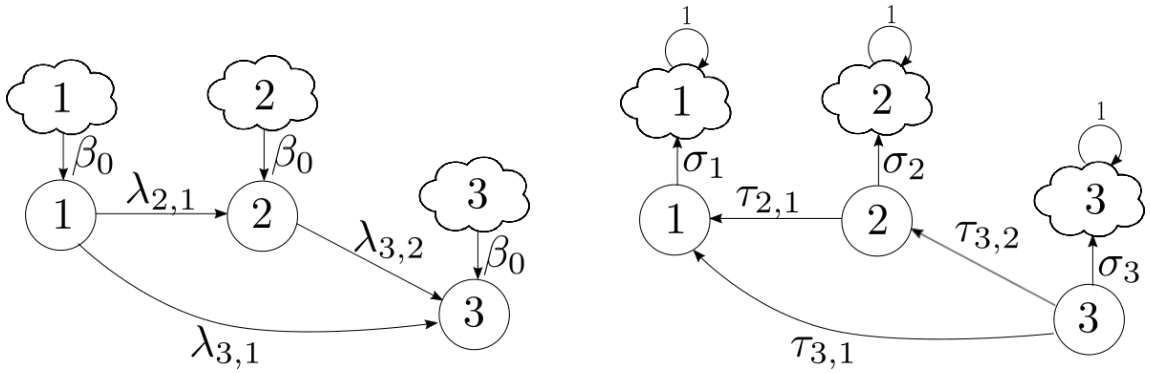



Fig. 4.2 Forward transmission network (left) and reverse transmission network (right) for an idealised outbreak taking place on three locations. Circles represent real locations, and clouds represent 'seeding states' – conceptual reservoirs of infection that contribute infective force from outside the population. In reality, it is more natural to think of a single external reservoir of infection that contributes a constant force $\beta_0$ on all locations; however, artificially separating the reservoirs is mathematically convenient. In this example, the outbreak begins in location 1, then infects location 2, and finally infects location 3, in three subsequent time steps. In the left-hand diagram, arrows denote possible transmission paths, and arrow labels give the partial forces of infection. In the right-hand diagram, arrows point towards possible 'parent' outbreaks, and arrow labels give the probability that the location at the tip of the arrow directly sparked the outbreak in the location at the tail of the arrow. Definitions of the arrow weights are given in §4.2.2. In this simplified setting, location 1 would be a hub, since the outbreaks in locations 2 and 3 can be traced back to the seeding state attached to location 1 with high probability.

Mathematically, this back-stepping procedure is done by taking powers of the reverse transmission network's transition matrix. The $i, j^{th}$ entry of the transition matrix gives the probability that the outbreak in location $i$ was immediately triggered by location $j$ (i.e. that location $j$ was a 'parent' to the outbreak in location $i$). The $i, j^{th}$ entry of the squared transition matrix gives the probability that the outbreak in location $i$ was triggered by location $j$ via any one intervening location (i.e., that location $j$ was a 'grandparent' to the outbreak in location $i$). As these powers approach infinity, each outbreak is traced back to a most likely 'seeding state'. In the limit, the $i, j^{\text{th}}$ entry of the $p^{\text{th}}$ power of the transition matrix gives the probability that the outbreak in location $i$ was initially triggered by a seeding event in location $j$.

To illustrate the procedure, refer again to the idealised outbreak depicted in Fig 4.2. Reversing the arrows in the left-hand diagram gives the reverse transmission network (right-hand diagram), where each arrow now points towards a possible contributor of infection. The transition probabilities are denoted

$$\tau_{ij} = P(\text{transmission from } j \text{ to } i) = \frac{\lambda_{ij}}{\lambda_i}$$

and

$$\sigma_i = P(\text{external seeding in } i) = \frac{\beta_0}{\lambda_i}.$$

The $\tau_{ij}$ represent the probability that the outbreak in location $i$ came from parent location $j$, and the $\sigma_i$ represent the probability that the outbreak in location $i$ was due to a seeding event.

Define $\boldsymbol{\tau}_{n\times n}$ to be the matrix whose $i, j^{th}$ entry is $\tau_{ij}$. Note that $\tau_{ij} = 0$ for all $j \geq i$, so $\boldsymbol{\tau}$ is strictly lower triangular. Also define $\boldsymbol{\sigma}_{n\times n}$ to be the matrix with $\sigma_1, \sigma_2, \ldots, \sigma_n$ along the diagonal and with zeros elsewhere. The transition matrix $\mathbf{M}_{2n\times 2n}$ that describes the reverse transmission network can be written using these matrices:

$$\mathbf{M} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \boldsymbol{\sigma} & \boldsymbol{\tau} \end{pmatrix}.$$

The first $n$ elements of the state space of $\mathbf{M}$ correspond to the seeding states (clouds in Fig 4.2), and the remaining $n$ elements correspond to the real locations. Entry $\mathbf{M}_{ij}$ is the probability that 'parent' location $j$ directly sparked location $i$'s outbreak (or, equivalently, the probability that the reverse transmission process transitions from state $i$ to state $j$). The identity matrix in the upper left block indicates that the seeding states are ultimate sources of infection; they can only transition to themselves. Similarly, the $\mathbf{0}$ matrix in the upper right

block indicates that transmission cannot occur from a real location to a seeding state. The $\boldsymbol{\sigma}$ matrix in the lower left block captures the probability of a seeding event in each real location. The $\boldsymbol{\tau}$ matrix in the lower right captures the transmission probabilities between real locations. Note that, as required, the row sums of $\mathbf{M}$ all equal 1.

The $p^{th}$ power of $\mathbf{M}$ contains the probabilities of transitioning between any two nodes via $p-1$ intermediate steps. Finding the ultimate ancestor of each location's outbreak, then, requires calculating $\lim_{p\to\infty}\mathbf{M}^p \equiv \mathbf{M}^\infty$. Since $\boldsymbol{\tau}$ is strictly lower triangular (has zeros along its diagonal), $\boldsymbol{\tau}^m = \mathbf{0}$ for $m \geq n+1$. Thus, $\mathbf{M}^m = \mathbf{M}^\infty$ for $m \geq n+1$, yielding

$$\mathbf{M}^\infty = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ (\mathbf{I}+\boldsymbol{\tau}+\boldsymbol{\tau}^2+\ldots)\boldsymbol{\sigma} & \mathbf{0} \end{pmatrix}.$$

Element $(\mathbf{M}^\infty)_{i,j}$ gives the probability that state $j$ was the ultimate source of the outbreak in location $i$. The identity matrix in the upper left block indicates that seeding states are sources unto themselves; this is so by definition. Each real location's ultimate source is a seeding state, since the real→real transitions in the lower-right block of the matrix all go to zero. The lower-left block of $\mathbf{M}^\infty$ contains the values of greatest interest. Denote this block $\mathbf{P}_{n\times n} \equiv (\mathbf{I}+\boldsymbol{\tau}+\boldsymbol{\tau}^2+\ldots)\boldsymbol{\sigma}$. The entries $\mathbf{P}_{i,j}$ are the probabilities that external seeding in location $j$ ultimately led to an outbreak in location $i$. The row sums $\sum_j \mathbf{P}_{ij}$ equal 1 for all $j$. The column sums of $\mathbf{P}$, denoted $C_j = \sum_i \mathbf{P}_{ij}$, can be interpreted as the expected number of outbreaks triggered by seeding in location $j$. Hubs are locations with high $\sigma$ and high $C$ values – that is, locations where external seeding probably triggered an outbreak, which then led to significant onward spread.

This result may also be derived in a more compact, though perhaps less intuitively satisfying, way. We seek a matrix $\mathbf{P}$ such that entry $\mathbf{P}_{i,j}$ is the probability that seeding in location $j$ caused the outbreak in location $i$, via any number of intermediate steps. It is helpful to introduce notation for the following events:

$\mathring{j} \Rightarrow i$ : "Seeding in location $j$ ultimately caused the outbreak in location $i$"

$j \to i$ : "The outbreak in location $j$ immediately caused the outbreak in location $i$"

So, $\mathbf{P}_{ij} = \Pr(\mathring{j} \Rightarrow i)$. Note that $\Pr(\mathring{j} \Rightarrow j) = \sigma_j$. For $i \neq j$, the law of total probability states that

$$\Pr(\mathring{j} \Rightarrow i) = \sum_k \Pr(\mathring{j} \Rightarrow k)\Pr(k \to i).$$

The probabilities $\Pr(k \to i)$ are the $i,k^{th}$ entries of the transition matrix $\tau$. Taken together, this implies that

$$\mathbf{P} = \boldsymbol{\sigma} + \boldsymbol{\tau}\mathbf{P}. \tag{4.2}$$

Solving for $\mathbf{P}$ gives

$$\mathbf{P} = (\mathbf{I} - \boldsymbol{\tau})^{-1}\boldsymbol{\sigma} = (\mathbf{I} + \boldsymbol{\tau} + \boldsymbol{\tau}^2 + \dots)\boldsymbol{\sigma}. \tag{4.3}$$

Since $\boldsymbol{\tau}$ is strictly lower triangular, $\boldsymbol{\tau}^p = \mathbf{0}$ for $p > n$, and so the infinite sum is guaranteed to converge. Eq 4.3 matches the earlier result for $\mathbf{P}$.

## 4.3 Hubs of the 2009 A/H1N1pdm influenza pandemic in the United States

### 4.3.1 The transmission model revisited

The transmission model given in Eq 3.27 expresses the force of infection on a location $i$ as a sum of the forces exerted by all potential contributors. The methods developed in §4.2 may therefore be applied to identify transmission hubs of the autumn wave of the 2009 A/H1N1pdm influenza pandemic in the United States. For the following, we define forces of infection using the best transmissibility-adjusted transmission model developed in Chapter 3, Eq 3.45. Results are the same when using the best model without transmissibility adjustments (Eq 3.37), as well as the best exponential-kernel model (Eq 3.37 with $\gamma \to \infty$) with or without an effect from 1-week-advanced school start dates.

In particular, define $\lambda_{ij}$, the force of infection on location $i$ from location $j$ at $i$'s time of onset, as

$$\lambda_{i,j} = \begin{cases} \beta_d \operatorname{Exp}[\xi_t^T + \xi_i^S] N_i^{\mu} \dfrac{n_{j,t}^{\theta}\kappa(d_{i,j})}{\sum_{k \neq i}\kappa(d_{i,k})} & \text{if } j < i \\ 0 & \text{otherwise} \end{cases}$$

where $\kappa(d_{i,j})\left(1 + \frac{d_{i,j}}{\rho\gamma}\right)^{-\gamma}$. Recall that under the most parsimonious transmission model, $\beta_{ds} = \nu = 0$ and $\varepsilon = 1$ (see Eq 3.27 and Table 3.3). The total force of infection on location $i$ at its time of onset can be written as the sum of these partial forces:

$$\lambda_i = \beta_0 + \sum_{j=1}^{n} \lambda_{ij}.$$

This is equivalent to the transmission model given in Eq 3.37, evaluated at location $i$'s outbreak onset time.

### 4.3.2 Calculating hubs

Using these $\lambda_{i,j}$, the transition probabilities $\tau$ and $\sigma$ of the reverse transmission network are calculated, giving the transition matrix $\mathbf{M}$. Taking powers of $\mathbf{M}$ until the lower-right block becomes a zero matrix yields sub-matrix $\mathbf{P}$, which gives the probability that seeding in any one location triggered the outbreak in any other. We define hubs as locations with $\sigma > 0.3$ and $C > 2$ (recall that $C_j = \sum_i \mathbf{P}_{i,j}$ is the effective number of locations infected by seeding in location $j$). That is, hubs are locations in which an outbreak was caused by external seeding with at least 30% probability, and that went on to infect at least two other locations. These cutoff values for $\sigma$ and $C$ correspond to natural gaps in the $\sigma$- and $C$-values estimated for the ZIPs in 2009 (see Figs 4.3-4.6). The cutoff values are liberal, in the sense that they err on the side of identifying locations that may have had near-negligible influence on the onward geographic transmission of the outbreak, rather than risk excluding any possible hubs. Indeed, a seeding event that triggered just $C = 2$ onward outbreaks would have had very little practical effect on this set of over 800 cities. There is an even lower natural cutoff for $\sigma$, around $\sigma = 0.12$ (see Fig 4.3), that could be used, but would arguably be too lax of a criterion, admitting as hubs cities where an external seeding event was very unlikely.

Despite these fairly relaxed criteria for defining transmission hubs, just nine ZIPs surpass both cutoffs. These are listed in Table 4.1. Reducing the $\sigma$ cutoff to the lower value to 0.12 introduces just one additional transmission hub, in Frederick MD, with $\sigma = 0.14$ and $C = 11.2$. Seeding in the three most influential hubs is expected to have triggered 617 of the 834 observed ZIP-level outbreaks. Overall, seeding in the nine hubs is estimated to account for 90% (758.8) of the observed ZIP-level outbreaks in the US in the autumn of 2009 through onward geographic transmission. To visualise the geographic influence of hub $j$, matrix element $\mathbf{P}_{i,j}$ is translated into a colour intensity and mapped for all locations $i$. These 'basins of infection' are depicted in Fig 4.7.

### 4.3.3 Accounting for onset uncertainty

Each ZIP's outbreak onset time is associated with some uncertainty, which could affect the placement of the hubs. Re-drawing outbreak onset times from the breakpoint likelihood profiles (see §2.3) and re-calculating the transmission hubs generally yields a set of locations either identical to or geographically close to the hubs listed in Table 4.1. So, the hubs

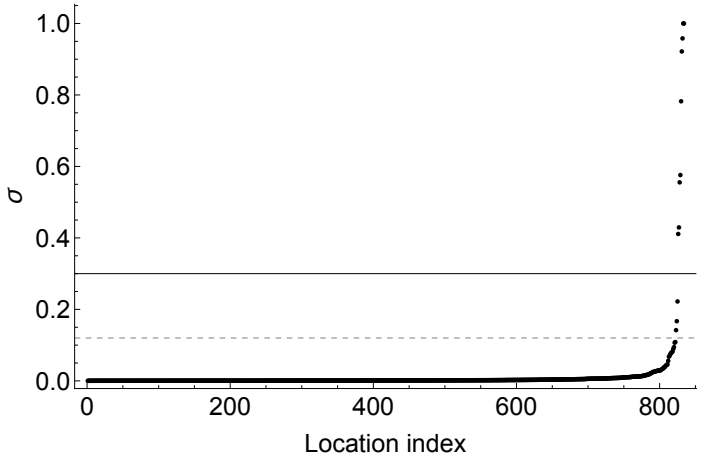



Fig. 4.3 Seeding probability $\sigma$ for all 834 ZIPs, ordered by magnitude of $\sigma$. Most locations have $\sigma$-values below 0.1, which translates into a less than 10% estimated chance that external seeding caused the location's outbreak, according to the transmission model Eq 3.45. A few locations, however, have much higher $\sigma$-values. A natural gap occurs around $\sigma = 0.3$ (solid horizontal line), between the ninth- and tenth-highest $\sigma$-values. We adopt this as the $\sigma$ cutoff value for defining transmission hubs of the autumn wave of the 2009 A/H1N1pdm pandemic. Another cutoff arises around $\sigma = 0.12$ (dashed horizontal line), between the $12^{\text{th}}$- and $13^{\text{th}}$-highest $\sigma$-values. The geographic locations of all ZIPs with $\sigma > 0.12$ are depicted in Fig 4.5.

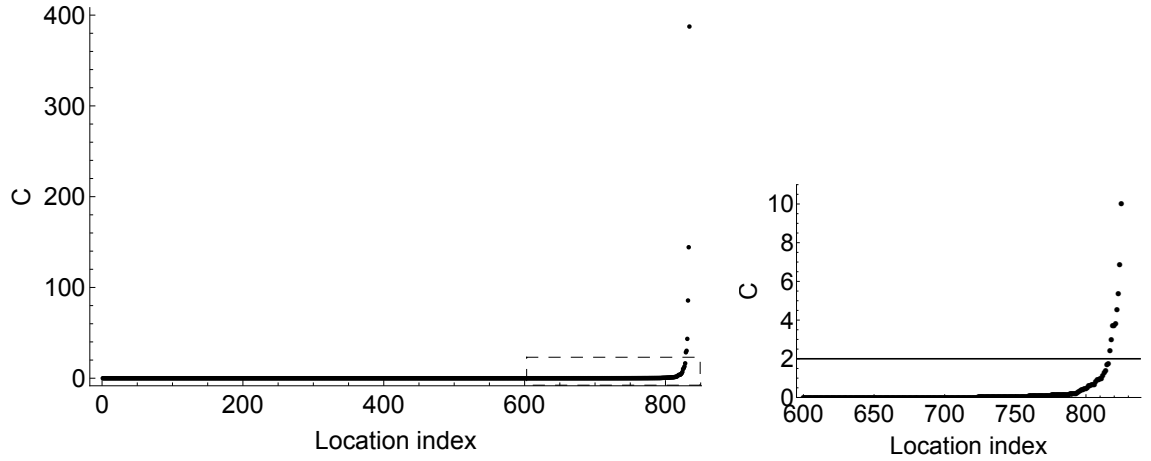



Fig. 4.4 Effective number of locations infected ($C$) for all ZIPs, ordered by magnitude of $C$. The left-hand plot depicts the $C$ values for all ZIPs, and the right-hand plot provides detail of the section in the dashed box. A natural cutoff arises around $C = 2$, marked by the horizontal black line in the left-hand plot, which 12 ZIPs surpass. The names and geographic locations of these ZIPs are depicted in Fig 4.6.

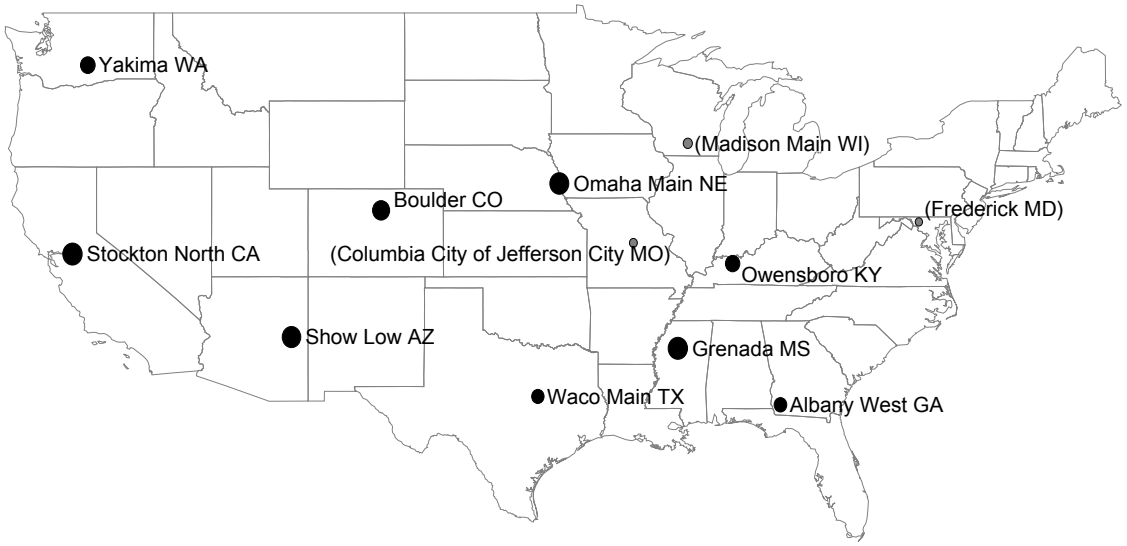



Fig. 4.5 Map depicting the seeding probabilities $\sigma$ by geographic location for all ZIPs with $\sigma > 0.12$. Disc area is proportional to the magnitude of $\sigma$. ZIPs with $\sigma > 0.3$ are in black and labelled without parentheses. ZIPs with $0.12 < \sigma < 0.3$ are in grey, and their names are in parentheses. These potential sites of external seeding are scattered roughly evenly across the country.

Table 4.1 Transmission hubs of the autumn wave of the 2009 A/H1N1pdm influenza pandemic in the United States

| Name | ZIP | Pop. size | $C$ | $\sigma$ | Onset date |
|---|---|---|---|---|---|
| Grenada, MS | 389 | 113,782 | 387.4 | 1.00 | 23 Jul |
| Albany West, GA | 398 | 111,263 | 144.3 | 0.43 | 26 Jul |
| Stockton North, CA | 952 | 508,759 | 85.6 | 1.00 | 26 Jul |
| Omaha Main, NE | 681 | 573,828 | 43.4 | 0.96 | 2 Aug |
| Owensboro, KY | 423 | 167,975 | 30.2 | 0.56 | 2 Aug |
| Boulder, CO | 803 | 112,702 | 28.5 | 0.78 | 9 Aug |
| Show Low, AZ | 859 | 77,189 | 16.4 | 0.92 | 2 Aug |
| Yakima, WA | 989 | 275,599 | 13.0 | 0.81 | 23 Aug |
| Waco, TX | 767 | 171,493 | 10.0 | 0.58 | 6 Aug |

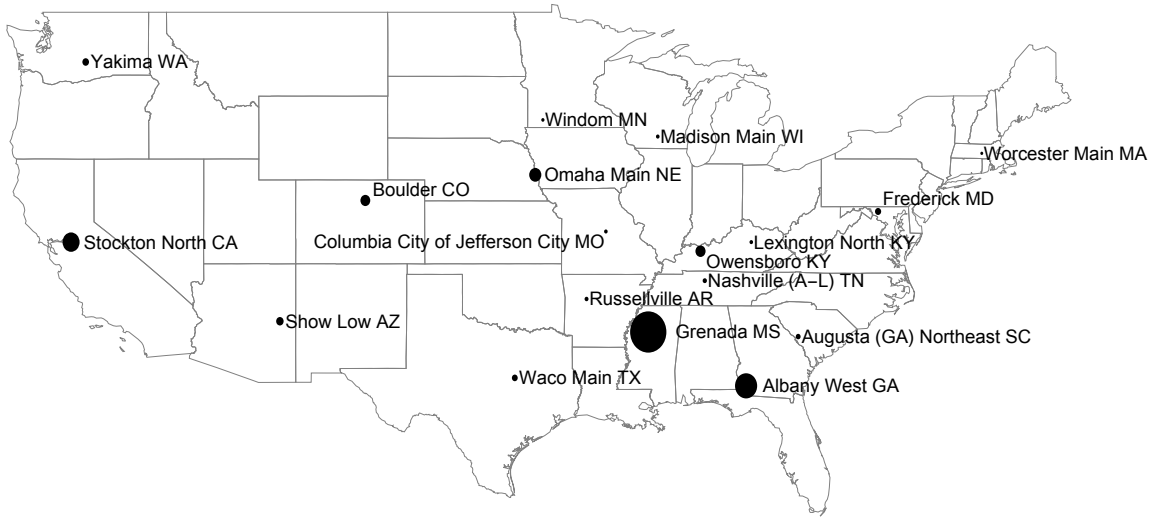



Fig. 4.6 Map depicting the effective number of locations infected $C$ by geographic location for all ZIPs with $C > 2$. Disc area is proportional to the magnitude of $C$. Stockton North CA, Grenada MS, and Albany West GA have the highest $C$-values, indicating that seeding in these three ZIPs contributed significantly to the onward geographic transmission of the 2009 A/H1N1pdm influenza outbreak in the United States. Relatively smaller contributors are scattered throughout the rest of the country.

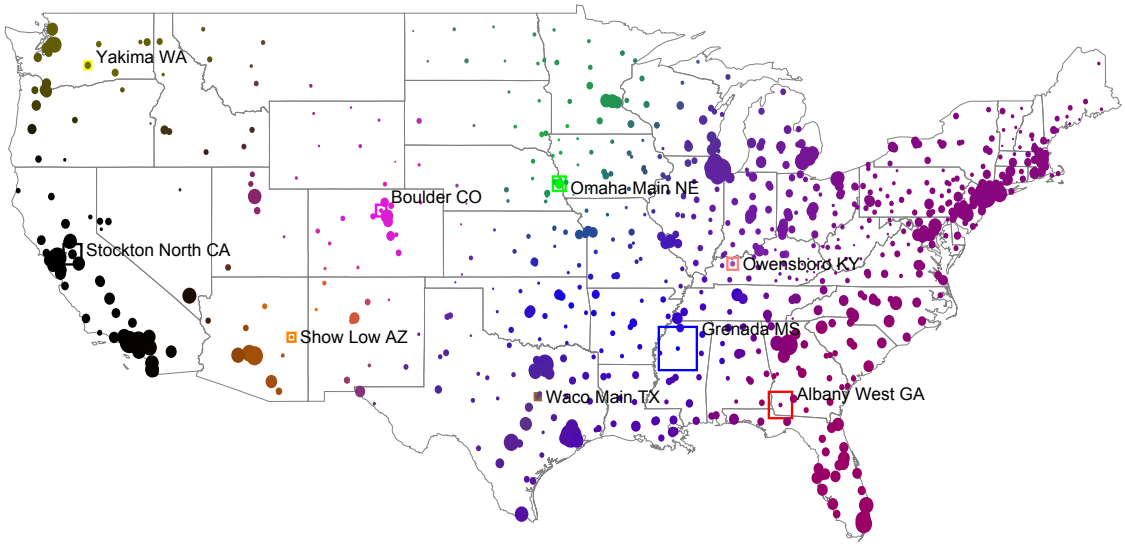



Fig. 4.7 Basins of infection for the nine hubs listed in Table 4.1. Hubs are outlined with boxes. Box area is proportional to the number of outbreaks that seeding in the hub triggered through gravity-driven onward transmission. Each hub $j$ is assigned a colour (the colour of the surrounding box), and then all locations $i$ are coloured with intensity proportional to the probability $\mathbf{P}_{ij}$ that hub $j$ sparked its outbreak (see §4.2). The prevailing black in California indicates that outbreaks in that state can be chiefly attributed to the hub in Stockton CA. The purple in the eastern US indicates mixing from Grenada MS (blue) and Albany West GA (red).

identified above may be interpreted as representatives of larger geographic regions in which an introduction of infection likely took place, sparking onward chains of short-distance transmission.

To determine the precise effect of this onset uncertainty on the hubs, 250 new sets of outbreak onset times were drawn from the onset likelihood profiles, and hubs were re-calculated. Fig 4.8 depicts the average $C$ for each ZIP after the 250 onset re-samples, in ascending order. There remains a cutoff around $C = 12$, which 16 ZIPs surpass. Fig 4.9 depicts the average $C$ for each ZIP geographically. To identify ZIPs that triggered infections in similar geographic areas, the columns of $\bar{\mathbf{P}}$, the element-wise average $\mathbf{P}$ across all 250 re-samples, are normalised to each sum to 1. Then, for each of the ZIPs that infected at least 12 others on average, the Kullback-Leibler (KL) divergence is calculated between the column corresponding to that ZIP and every other column. The columns of ZIPs that infected similar geographic regions have low KL divergence. The grey lines in Fig 4.9 depict these divergences, where thickness is proportional to the inverse of the divergence. There tends to be a sharp break between divergences of less than 0.1 and divergences greater than 0.1. Lines connecting ZIPs with KL divergence less than 0.1 are given colours. This yields seven connected clusters of ZIPs. These clusters are estimates of the larger regions that the transmission hubs listed in Table 4.1 represent. There are two large connected components in the southeast, surrounding Grenada MS and Albany West GA. Each of these consists of 9 and 22 ZIPs, respectively. Boulder CO has a connected component consisting of 3 ZIPs. Yakima WA, Stockton North CA, and Show Low AZ have connected components consisting of 2 ZIPs each. Omaha NE is isolated.

Only one of the seven connected components, the one surrounding New York City PO Box/Unique (1), NY, does not include a hub listed in Table 4.1. This could be taken as evidence that there was a possibly significant seeding event in New York City, breaking the trend of hubs in small- to mid-sized cities. However, this ZIP has a relatively noisy ILI time series and a correspondingly wide onset confidence interval. Because of this, the ZIP is sometimes assigned a significantly earlier onset than its neighbours. The average $C$ for this ZIP across the 250 re-sampled sets of onset times is 21, indicating that seeding in this ZIP may have triggered about 21 downstream outbreaks. There are exactly 21 ZIPs within a 50-km radius of this one, which suggests that whatever influence this hub may have had was limited to a small geographic area, and hence did not significantly impact the geographic transmission of the outbreak in the rest of the US.

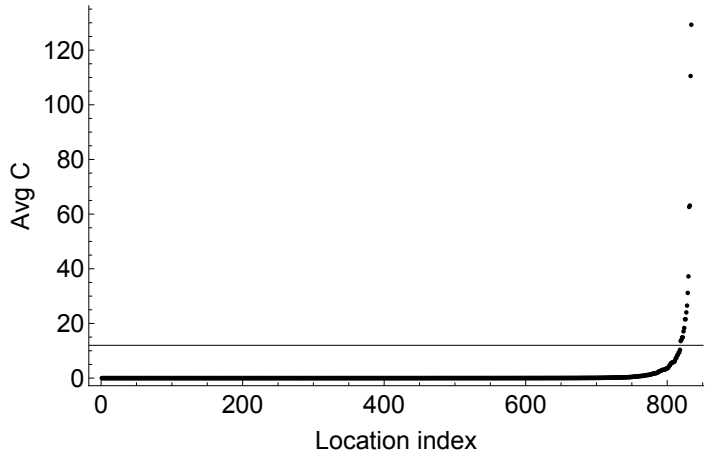



Fig. 4.8 Average effective number of locations infected ($\bar{C}$) for all ZIPs under 250 sets of re-sampled onsets, ordered by magnitude of $C$. The horizontal line depicts the cutoff at $C = 12$, which 16 ZIPs surpass.

### 4.3.4 Re-calculating the transmission hubs with the true hubs missing

According to the methods developed above, the transmission hubs of the 2009 A/H1N1pdm influenza pandemic in the United States were small- to mid-sized cities (see Table 4.1). Though this conclusion runs against the conventional wisdom that outbreak epicentres ought to be in large, well-connected cities, there are a few epidemiological arguments for why these findings are plausible (see §4.5). However, one possible explanation for the counter-intuitive distribution of hubs is that (1) there are far more small- to mid-sized ZIPs than there are large ZIPs, and (2) smaller ZIPs tend to have noisier epidemic times series than larger ZIPs. As a result, by chance, smaller ZIPs may sometimes appear to pre-empt the outbreak onsets in their larger neighbours, even if the larger ZIP were the true hub. This would cause our method to erroneously identify the smaller city as a hub.

One way to roughly test whether this is the case is to omit the transmission hubs identified above, in §4.3.2, from the set of ZIPs and to re-calculate the hubs. If the true establishment sites of the autumn 2009 A/H1N1pdm outbreak in the US were in fact in large cities, and the small-city hub assignments are simply one-off errors, then this exercise should correctly reveal the large cities as the transmission hubs. If, however, a set of nearby smaller cities are again identified as hubs, then either the small-city observational noise is pronounced enough to cause a second layer of mistaken hub assignments, or there is reason to believe that the transmission hubs of the autumn 2009 A/H1N1pdm pandemic wave in the US may truly have been in small- to mid-sized cities.

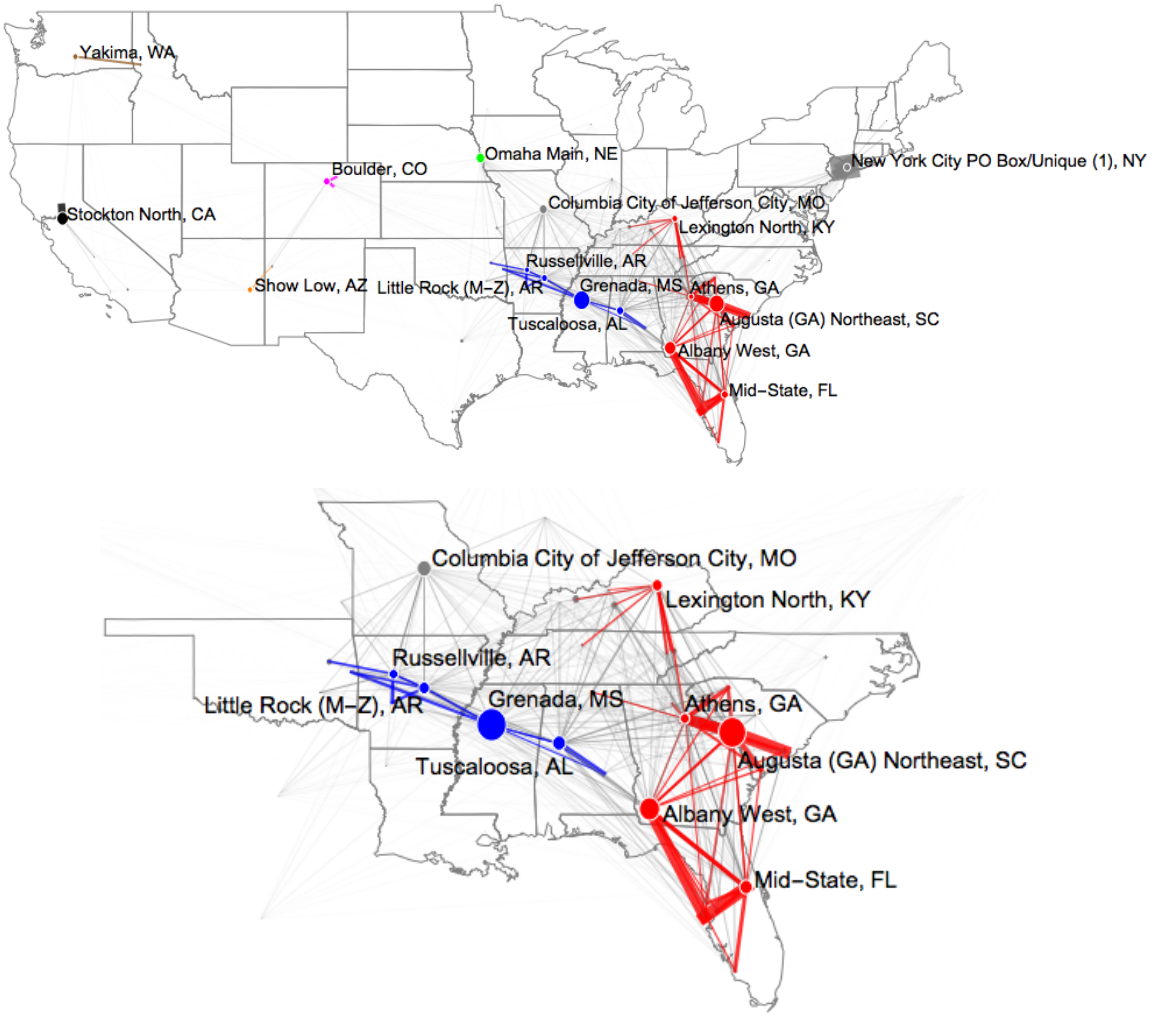



Fig. 4.9 Map of clusters of ZIPs that triggered outbreaks in similar geographic areas. The upper map depicts the clusters for the entire US, and the lower map provides detail for the souteastern US. Disc area is proportional to the ZIP's average $C$ across 250 re-sampled sets of onsets. ZIPs that triggered more than 12 downstream outbreaks on average are labelled. ZIPs responsible for infecting similar geographic areas are connected with grey lines, where the thickness of the line is proportional to the number of shared downstream outbreaks. Similarity is measured by normalising each column of matrix $\bar{\mathbf{P}}$ to sum to 1, and then calculating the Kullback-Leibler (KL) divergence between those distributions. For each ZIP that triggered more than 12 downstream outbreaks on average, coloured lines are drawn to all other ZIPs where the calculated KL divergence is less than 0.1 (that is, to ZIPs that infected very similar geographic areas). This forms seven connected components. These clusters are estimates of the larger regions that the hubs listed in Table 4.1 represent, in which an introduction of infection likely occurred, triggering significant onward short-distance transmission.

Fig 4.10 depicts the seeding probabilities $\sigma$ by geographic location, and Fig 4.11 depicts $C$ by geographic location, after omitting the nine hubs identified in §4.3. Nine more ZIPs are identified with $\sigma > 0.3$ and $C > 2$; these are labelled in Fig 4.11. Once again, these new hubs are all small- to mid-sized cities. Importantly, these new hubs lie geographically near the hubs identified previously, which would not be expected if noise alone were responsible for classifying a ZIP as a hub; if noise alone were responsible, the geographic placement of the new hubs should have little correlation with the placement of the previous ones. There remains a high concentration of new hubs in the southeast, providing further support that the southeast was an important region for epidemic establishment in the US in the autumn of 2009. Though only a rough test, this exercise indicates that a more nuanced explanation than one-off observational noise is likely required to account for why the observed set of hubs were not in major cities. A more thorough test of the hub-identification method using individual-based epidemic simulations is presented in the next section, §4.4.

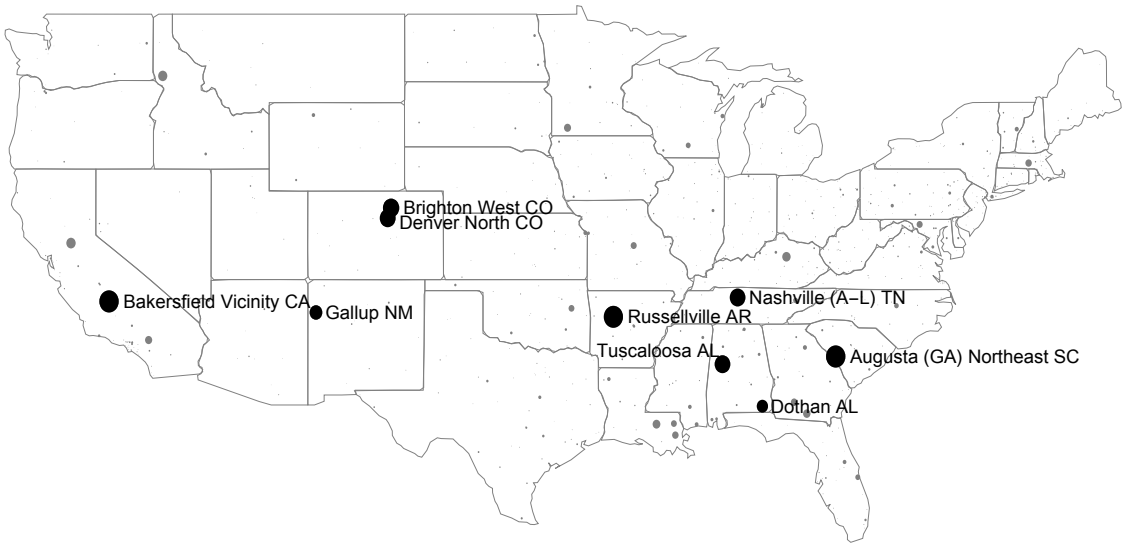



Fig. 4.10 Seeding probability $\sigma$ by geographic location, with the transmission hubs identified in §4.3.2 and listed in Table 4.1 removed from the set of ZIPs. Discs represent ZIPs, and disc area is proportional to $\sigma$. Locations with $\sigma > 0.3$ are labelled. In general, the ZIPs with high $\sigma$ depicted here lie close to a previously-identified transmission hub, but are still not major cities.

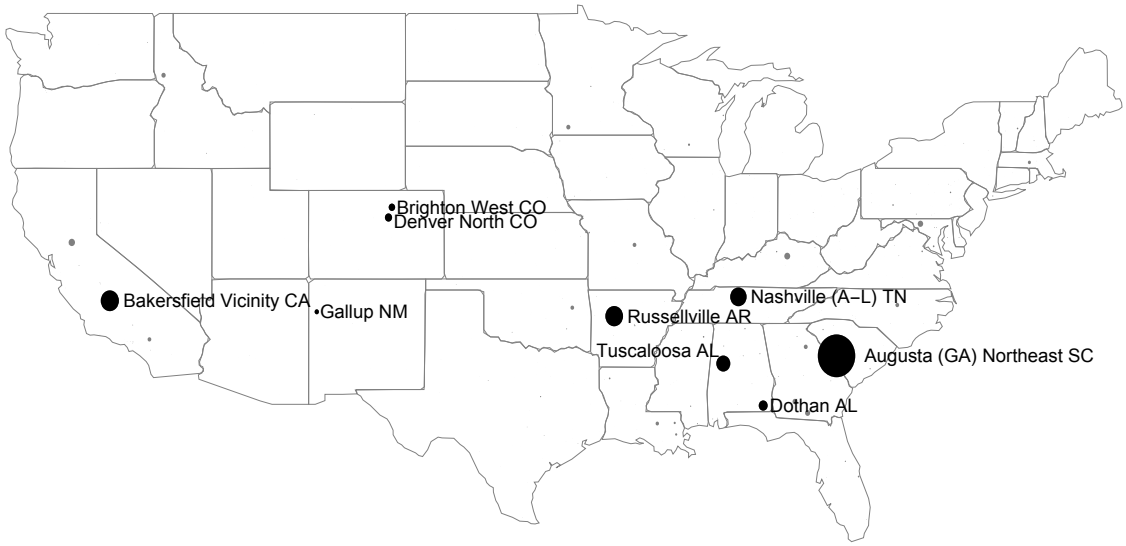



Fig. 4.11 Effective number of outbreaks $C$ triggered by geographic location, with the transmission hubs identified in §4.3.2 and listed in Table 4.1 removed from the set of ZIPs. Discs represent ZIPs, and disc area is proportional to $C$. Locations with $C > 2$ are labelled. These happen to be the same cities labelled in Fig 4.10 with $\sigma > 0.3$. So, the labelled cities may be interpreted as the new transmission hubs that are identified when the true transmission hubs are removed from the analysis. Once again, probable epidemic establishment sites are identified in the central valley of California, in Colorado, in Arizona/New Mexico, and in the southeast, but not in major cities.

## 4.4 Simulation-based validation of methods

To more thoroughly test the accuracy of the hub-identification method developed in this chapter, we turn to individual-based epidemic simulations on metapopulations. Here, epidemics are seeded in a large city surrounded by smaller ones, allowed to propagate, and then the transmission hubs are calculated. The introduction site is identified accurately around 80% of the time. It is therefore unlikely that, in the real autumn 2009 outbreak, many small hubs would be identified if in fact the outbreak were triggered in large cities. This section describes the epidemic simulation strategy in detail, and provides more information about the hub detection method's accuracy.

### 4.4.1 Overview of the epidemic simulation methods

The simulation model described here is a version of the commuter model presented in Keeling *et al.* (2010) [128]. Epidemics are simulated on metapopulations consisting of either 16 or 25 cities with random coordinates. Using random coordinates allows for a general exploration of how city placement affects the accuracy of hub identification; simulations using actual ZIP population sizes and coordinates are considered in §4.4.6. City population sizes are chosen to match the distribution of ZIP population sizes in the US. Within-city infection and recovery are governed by SIR dynamics, and modelled using the Gillespie algorithm with additional noise. Infection travels between cities via commuting individuals, who spend one-third of their day in a 'workplace' city and two-thirds in a 'residence' city. Epidemic onset times are calculated from the simulated time series using the breakpoint method (§2.3). Hubs are identified using these epidemic onset times according to the methods developed in §4.2.

### 4.4.2 Specifying city coordinates and population sizes

Each simulation takes place on $\mathscr{N} = 16$ or 25 cities, which just surpass the median number of ZIPs (13 and 22) that surround the real hubs within a radius of $3\rho$ and $4\rho$ respectively, where $\rho = 66$km is the maximum likelihood length scale of the distance kernel from the geographic transmission model developed in Chapter 3 (see Table 3.4). Considering a radius of 3–4 times the characteristic distance $\rho$ from the true hub should be sufficient, since beyond a distance of $4\rho$ cities are virtually epidemiologically uncoupled, according to the transmission model developed in Chapter 3. At a distance of $3\rho$, there is an approximately 95% drop in force of infection, according to Eq 3.37. At a distance of $4\rho$, the drop is over 98%. Early outbreaks that occur at a distance of more than 3-4$\rho$ from a true transmission hub are therefore likely

to be identified as separate seeding events, rather than erroneously replacing the hub. Also, simulations that incorporate more than about 25 cities quickly become computationally infeasible.

Cities are uniform-randomly assigned coordinates within a square of width $w$. Population sizes are assigned by randomly drawing $\mathcal{N}$ values from from the distribution of true ZIP population sizes in the US. The populations of the simulated cities are divided by 500 before simulating. This makes the noise of the simulated epidemics match or exceed the noise of the true epidemics (see Fig 4.12), and ensures that the the population sizes are small enough for stochasticity to matter. Introducing a division factor may be justified in part by noting that the effective population size of a city should be smaller than the true population size, since in reality individuals are not in constant contact with every other member in their city (i.e. real populations are not well-mixed).

### 4.4.3 Commuting

The simulated disease spreads between cities via commuting individuals. A proportion $f_c$ of each city's inhabitants commutes to other cities each day. Workplaces for these commuters are assigned according to the gravity model with an exponentially-decaying distance kernel. For each city $i$, potential workplaces $j$ are given relative weights

$$w_j = N_j^{\mu} e^{-d_{ij}/\rho}.$$

The commuters are assigned in proportion to these weights. For example, if city 1 has 100 commuters, and cities 2, 3, and 4 have relative weights of $w_j = 15$, 30, and 5, respectively (and there are no other cities in the simulation), 30 commuters would be assigned to city 1, 60 to city 2, and 10 to city 3.

Commuters are the only mobile individuals in the simulation, and they are only permitted to travel to and from their workplace. Between the hours of 5pm and 9am (i.e. overnight), the commuters reside in their home city. At 9am, the commuters instantaneously move to their workplaces, and remain there until 5pm, when they instantaneously return home.

### 4.4.4 Model running

The disease is simulated according to the Gillespie algorithm, assuming that each city's population is well-mixed, and that each individual's times of infection and recovery are random variables that follow exponential distributions.

At the start of the simulation, one city is chosen as the epidemic introduction site. An outbreak is seeded in that city by randomly choosing ten individuals and updating their status to 'infected'. The outbreak begins at midnight. Transition rates between susceptible ($S$), infected ($I$), and recovered ($R$) classes for individuals within a city are specified by the SIR model:

| Transition | Rate |
|---|---|
| $S \to I$ | $\beta \frac{SI}{N}$ |
| $I \to R$ | $\gamma I$ |

In the rate definitions, $N$ is the city's population size, which between 5pm and 9am is the city's total number of residents, and between 9am and 5pm is the number of non-commuting residents plus the total number of commuters to the city. $S$ and $I$ are the numbers of susceptible and infected individuals in the city, respectively. The parameter $\beta$ specifies the infectiousness of the pathogen, and $\gamma$ is the recovery rate.

### 4.4.5 Binning and noise

In conventional disease incidence data, the precise times of infection and recovery for individuals are not recorded. Instead, new infections are aggregated over fixed units of time. To represent this, the simulated model output is gathered into bins that contain the number of new infections in each city over consecutive spans of 7 days.

Furthermore, ILI data reflect other respiratory illnesses in addition to influenza. To account for this, additional ILI cases are added into each week's incidence according to a Poisson distribution with mean equal to a fraction $f_n$ of the city's population size.

### 4.4.6 Parameters and model validation

The model parameters used for the simulations are listed in Table 4.2. Epidemics are simulated on either $\mathscr{N} = 16$ or 25 cities. The gravity model population size exponent $\mu$ is set at 0.32, its maximum-likelihood value from Chapter 3 (see Table 3.4). Only $\rho = 1$ is considered, since adjusting this parameter is equivalent to placing the city coordinates in a larger or smaller square. The weekly background noise $f_n$ is fixed at 1% of the population size, which yields epidemic time series that resemble the observed time series across all population sizes (see Fig 4.12), and surpasses the background ILI rate of 0.6% used to specify the 'low ILI threshold' in Gog *et al.* (2014) [91]. The recovery rate $\gamma$ is fixed at 1/168, with units of 1/hour, which corresponds to an individual's infection lasting for approximately

one week (168 hours). The transmission parameter $\beta$ is fixed at 1/112, also with units of 1/hour, which gives a within-city basic reproduction number $R_0 = \beta/\gamma = 1.5$. The width of the simulation window is $3\rho = 3$ for the 16-city simulations and $4\rho = 4$ for the 25-city simulations. The fraction of commuters $f_c$ in each city is set at 10%. Ideally, $f_c$ would be chosen such that the overall epidemic length roughly matches the amount of time it took the true outbreak to spread through an area of about $3\rho - 4\rho$ (about 200–266) square kilometres, which is generally around 5 weeks. At $f_c = 10\%$, epidemics are somewhat longer than in reality, on the order of 10-15 weeks between the earliest and latest onsets. Values of $f_c = 5\%$ and $f_c = 20\%$ were also considered. Setting $f_c = 5\%$ leads to long epidemics, with many simulations yielding overall epidemic lengths of more than 20 weeks. Setting $f_c = 20\%$, on the other hand, leads to rapid epidemics with no evidence of wave-like structure. Setting $f_c = 10\%$ appears to give the best compromise between epidemic length and realistic epidemic structure, so results are reported for that value.

Table 4.2 Simulation model parameters

| Parameter | Meaning | Value(s) |
|---|---|---|
| $\mathcal{N}$ | Number of cities | $16, 25$ |
| $\mu$ | Gravity model population effect | 0.32 |
| $\rho$ | Transmission kernel length scale | 1 |
| $f_c$ | Fraction of commuters in each city | 0.10 |
| $f_n$ | Background ILI noise parameter | 0.01 |
| $\beta$ | Disease transmission rate | $1/112$ |
| $\gamma$ | Disease transmission rate | $1/168$ |
| $w$ | Width of simulation window | $3, 4$ |

In each simulation, the introduction site is set as the city with the largest population size. After simulating the epidemic, outbreak onset times are calculated for each city using the breakpoint method. The hubs are then identified according to the procedure described in §4.2. The force of infection expression is taken to be

$$\lambda_i(t) = \beta_0 + \beta_d N_i^{\mu} \frac{\sum_{j \in \Lambda_t} e^{-d_{i,j}/\rho}}{\sum_{j \neq i} e^{-d_{i,j}/\rho}} \tag{4.4}$$

where parameters $\rho$ and $\mu$ are set as their simulation values (see Table 4.2). The seeding parameter $\beta_0$ is set at $e^{-10}$ to match the order of magnitude of the true $\beta_0$ (see Table 3.4). The transmission parameter $\beta_d$ is trivially set at 1; since the hubs calculation relies exclusively on ratios of the force of infection, the parameter essentially cancels out.

In total, 20 epidemics were simulated on each of 25 random ensembles of 16 cities, and 10 epidemics were simulated on each of 10 random ensembles with 25 cities. Also, 15 epidemics were simulated on the true geographic positions of the ZIPs in Arizona/New Mexico with the true hub in Phoenix AZ, and 15 epidemics were also simulated on ZIPs in Alabama/Georgia with the true hub in Atlanta GA. Fig 4.13 depicts four of the simulated epidemics on 25 cities. The epidemic spreads from deepest to lightest red. Fig 4.14 depicts two simulated epidemics in Arizona/New Mexico with seed in Phoenix AZ. Fig 4.15 depicts two simulated epidemics in Alabama/Georgia with seed in Atlanta GA.

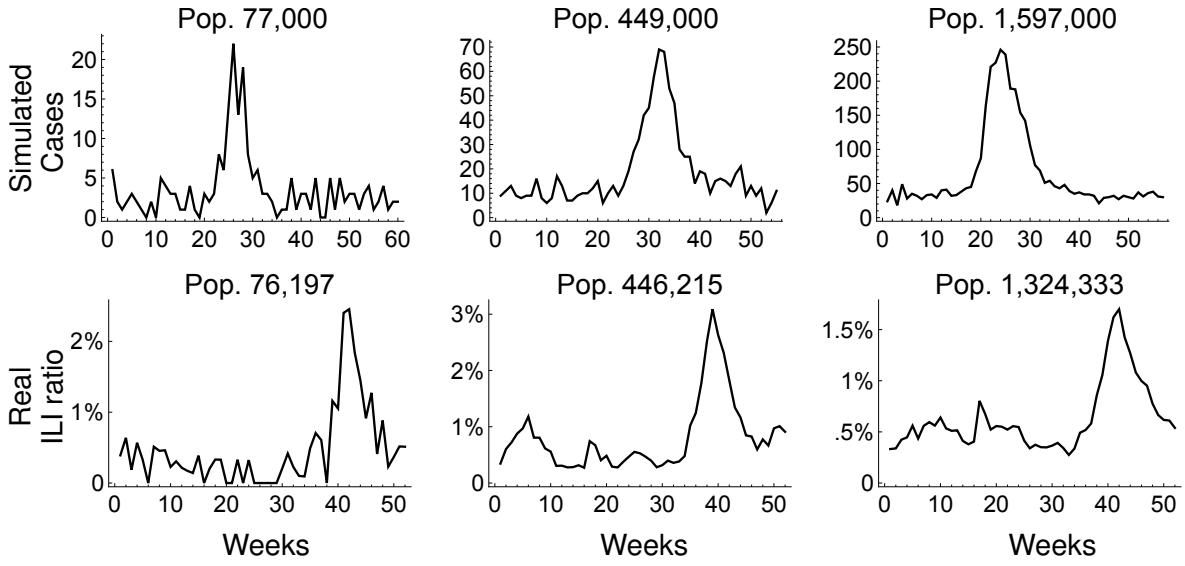



Fig. 4.12 Simulated (top row) and true (bottom row) ILI time series for a range of population sizes. The simulations in the top row are generated from the simulation model described in §4.4 with parameter values given in Table 4.2. The time series in the bottom row are from selected cities in the IMS-ILI dataset.

### 4.4.7 Simulation results

Table 4.3 gives the accuracy of the hub detection method for all simulated ensembles. In general, the correct hub is identified around 75-80% of the time. Around 5-10% of the time, one or more incorrect hubs are identified alongside the correct hub. The correct hub fails to be identified about 20-25% of the time.

It is more common for cities with small population sizes to be incorrectly identified as a hub. Fig 4.16 depicts the distribution of population sizes of the cities that are incorrectly identified as hubs across all 500 simulations with $\mathscr{N} = 16$ cities and 100 simulations with

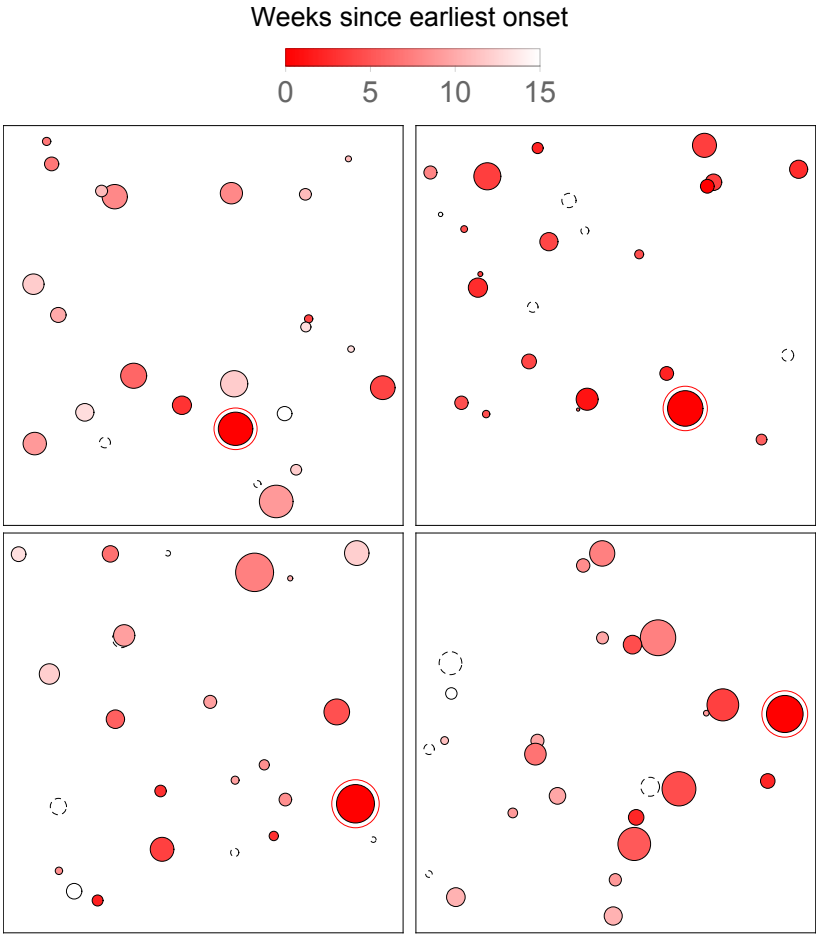



Fig. 4.13 Four simulated epidemics on 25 cities, using the simulation procedure described in §4.4. Discs represent cities, and disc area is proportional to the city's population size. Disc colour represents the outbreak onset time in each city as estimated by the breakpoint method (see §2.3.1). Cities with dashed outlines had indeterminable onset times. The city circled in red is the introduction site, where ten infected individuals are introduced at the beginning of the simulation. In each instance, the largest city is chosen as the introduction site.

Table 4.3 Hub identification accuracies for all simulated ensembles.

| $\mathscr{N}$ | Number of simulations | Times correct hub identified | Times correct hub identified with another hub | Times correct hub not identified |
|---|---|---|---|---|
| 16 | 500 | 385 (77%) | 55 (11%) | 115 (23%) |
| 25 | 100 | 83 (83%) | 7 (7%) | 17 (17%) |
| AZ/NM | 15 | 12 (80%) | 1 (7%) | 3 (20%) |
| AL/GA | 15 | 11 (73%) | 0 (0%) | 4 (27%) |

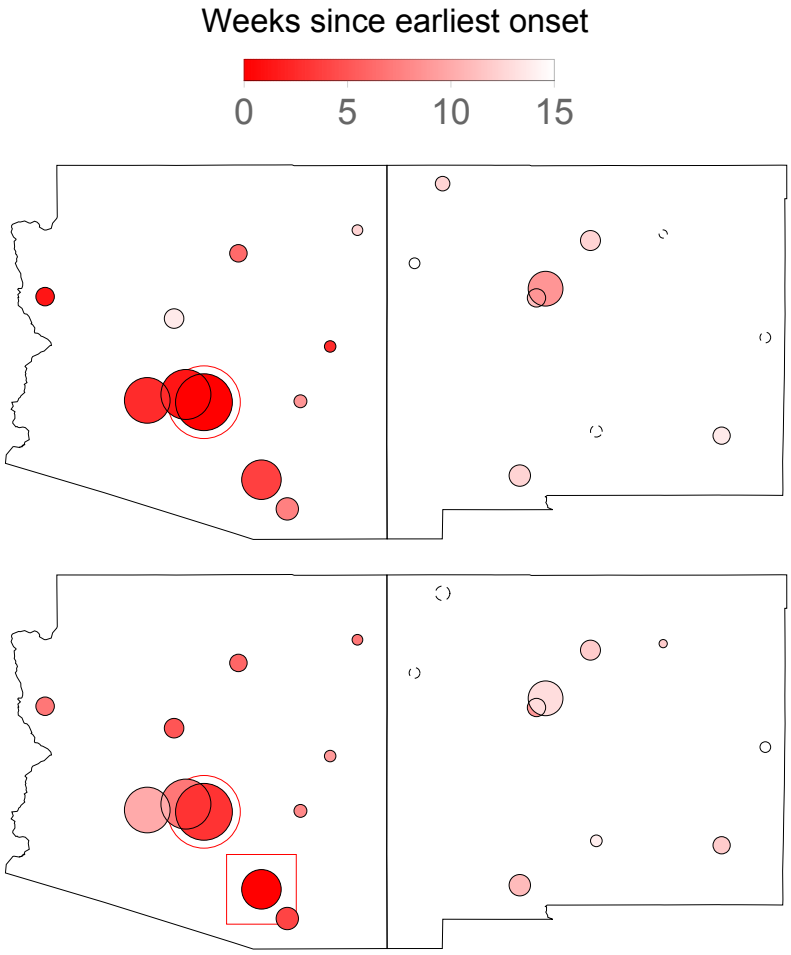



Fig. 4.14 Two simulated outbreaks in Arizona and New Mexico. Discs represent ZIPs, and disc area is proportional to population size. The epidemic is seeded in Phoenix (surrounded by a thin red circle). The epidemic progresses from deeper to lighter red. The time scale on the legend is in weeks. In the upper simulation, the hub identification scheme succeeds in identifying Phoenix as the hub. In the lower simulation, a different hub, Tucson Main AZ (boxed), is identified incorrectly.

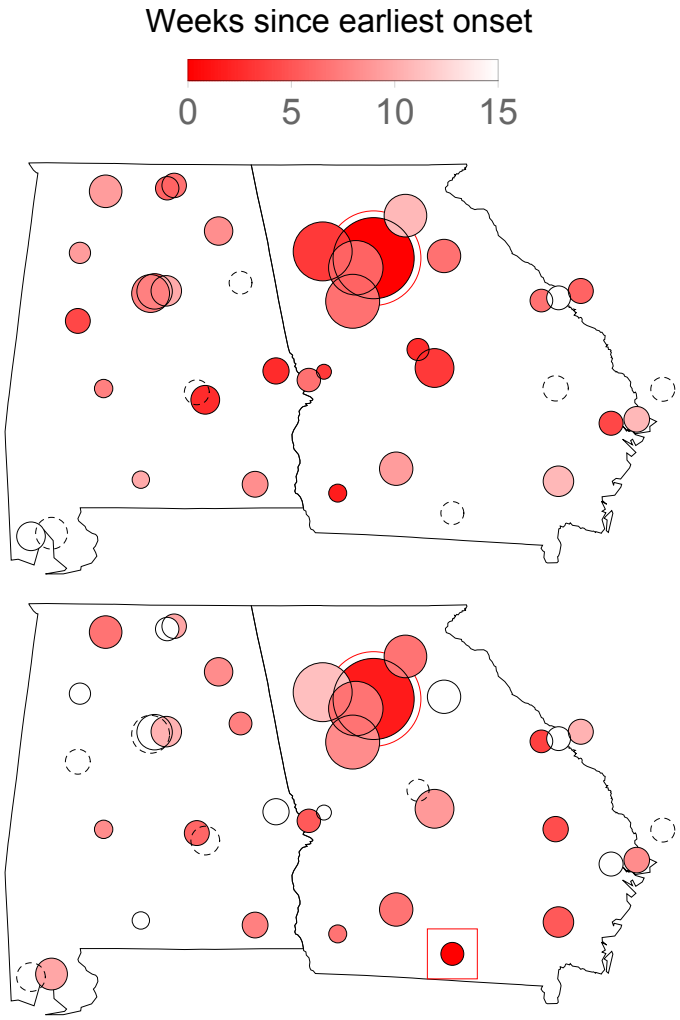



Fig. 4.15 Two simulated outbreaks in Alabama and Georgia. Discs represent ZIPs, and disc area is proportional to population size. The epidemic is seeded in Atlanta (surrounded by a thin red circle). The epidemic progresses from deeper to lighter red. In the upper simulation, the hub identification scheme succeeds in identifying Atlanta as the hub. In the lower simulation, a different hub, Valdosta GA (boxed), is identified incorrectly.

$\mathscr{N} = 25$ cities. This could be due in part to the smaller cities having noisier time series, and thereby more frequently appearing to pre-empt the outbreak onset in the true hub. The overall population distribution for all ZIPs (see Fig 2.1) roughly follows the same shape as the population distribution for the mis-identified ZIPs in Fig 4.16, raising the question of whether the incorrect hubs could simply represent a random sample of ZIPs. However, a bootstrapping test suggests that this is not the case. To test whether the 204 incorrectly identified hubs might represent a simple random sample of ZIPs, 10,000 unbiased random samples of size 204 were drawn with replacement from the set of ZIP population sizes, and the sample mean, sample standard deviation, and sample skewness were calculated. The sample 95% confidence intervals for the bootstrapped mean, standard deviation, and skewness values were (294229, 398331), (290205, 466937), and (1.64, 3.48), respectively. The sample mean, standard deviation, and skewness of the population sizes of the mis-identified ZIPs are 349282, 288363, and 1.41, respectively. So, while the sample mean of the mis-identified ZIP population sizes lies within the bootstrapped confidence interval for the sample mean, the standard deviation and skewness of the mis-identified ZIP population sizes lie outside of the bootstrapped confidence intervals. The skewness of the mis-identified ZIP population sizes is below the lower bound of the bootstrapped confidence interval, which is evidence that smaller ZIPs are disproportionately mis-identified as hubs, since the presence of larger ZIPs in such a distribution would tend to increase skewness. So, there does appear to be a bias for smaller ZIPs to be mistakenly identified as hubs.

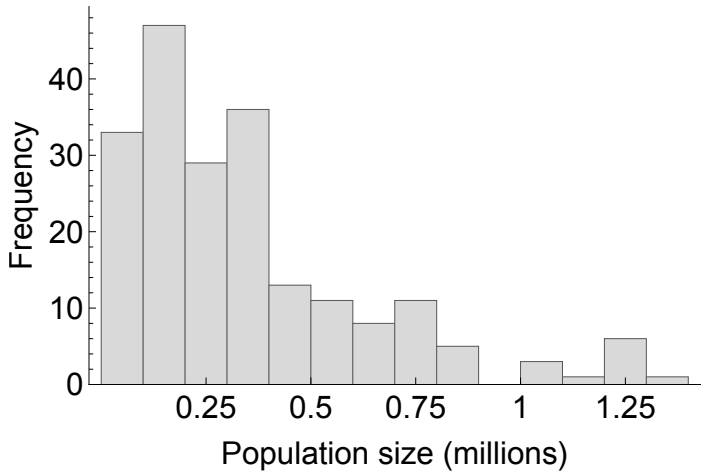



Fig. 4.16 Histogram of the population sizes of cities that are incorrectly identified as hubs for the combined 500 simulations on $\mathscr{N} = 16$ cities and the 100 simulations on $\mathscr{N} = 25$ cities. Small cities are more often mistakenly identified as hubs than large ones.

However, the results from the epidemic simulations suggest that it is unlikely that stochastic effects can explain why *all* of the observed hubs are in small- to mid-sized ZIPs. Even if we assume that the probability of mis-specifying a single transmission hub is 25%, then the probability of mis-specifying three or more hubs is is about 1%, if the probabilities of success are assumed to be independent for each hub. So, if the true hubs of the autumn wave of the 2009 A/H1N1pdm influenza pandemic in the US were in large cities, this simulation study suggests that it is unlikely that the hub identification scheme would consistently mistake them for small- to mid-sized cities.

### 4.4.8 Comments on model formulation

While the simulation model presented in this section provides some support for the accuracy of the hub identification method, there are a number of reasons why this may not be the best way of validating the method. First, it should be noted that a transmission hub and an introduction site are not the same thing. A transmission hub is a site of outbreak establishment, which may not coincide with the location to which infection was first introduced. In the simulation model developed in this section, we attempt to force establishment by introducing ten infected individuals at once into an index city. However, with a basic reproduction number of 1.5, the outbreak still has a 1.7% chance ($1/1.5^{10}$) of not becoming established in this index city. Even so, establishment may occur elsewhere, especially if some of the initially-infected individuals are commuters. In this scenario, a different city would be recognised as the transmission hub – and rightly so, because it would be the first city in which the epidemic became established. However, under the evaluation criteria presented in this section, this would be identified as a failure of the method. The probability of this sort of error is low, but it still likely leads to some false deflation in the apparent accuracy of the hub identification method.

Second, for the simulated epidemics, an exponential kernel is used to assign workers to work places, and the same kernel is used to describe the decay in the the force of infection with distance. However, there is no guarantee that the two should match. Indeed, Keeling *et al.* (2010) [128] show that epidemic dynamics from models that assume continuous decay in transmission strength (like the mechanistic transmission model developed in Chapter 3) can differ significantly from the dynamics of models based on random individual movements (like the individual-based model presented in this section), even when both are governed by similar kernels. It is unclear how a kernel that describes individual-level commuting patterns might translate into a kernel that describes metapopulation-level epidemic establishment.

Understanding this link is key for properly evaluating the hub identification method using individual-based simulations. Unfortunately, such an in-depth investigation lies outside the scope of this thesis. It remains an important area for future work.

## 4.5 Discussion

This chapter presents a method for identifying the geographic transmission hubs of an outbreak on a metapopulation, for which outbreak onset times for the sub-populations are known, and for which the force of infection on each sub-population can be separated into a additive contributions from all potential immediate sources of infection. The method is used to identify the transmission hubs of the autumn wave of the 2009 A/H1N1pdm influenza pandemic in the United States. These were generally small- to mid-sized cities, with two of the three most important hubs in the southeastern US, and the third in the central valley of California. This Discussion begins with a few comments on the epidemiological assumptions nested into the hub-finding scheme. Then, since the hubs that are identified for the autumn 2009 pandemic wave are somewhat unexpected, there follows a discussion of possible epidemiological justifications for the set of inferred hubs. The section concludes with a brief discussion on how the methods presented in this chapter open new possibilities for synthesising genetic and epidemiological data to improve the geographic reconstruction of outbreaks.

### 4.5.1 Epidemiological interpretation of the hub-finding procedure

There are two possible interpretations of the transmission network depicted in Fig 4.2. Under the first, the outbreak in each city $i$ is sparked by one and only one introduction from a nearby city or from external seeding. The quantities $\tau_{i,j}$ and $\sigma_i$ respectively give the probabilities that this index case came from location $j$ or from external seeding. Under the second interpretation, the outbreak in each city $i$ is caused by the sum total of the infective force exerted by multiple introductions from all possible 'parent' locations. Here, the quantities $\tau_{i,j}$ and $\sigma_i$ give the relative contribution of each predecessor. The second interpretation arguably is the most realistic: the relatively low basic reproduction number of influenza suggests that most within-city outbreaks are probably caused by multiple introductions from neighbouring cities, and are partially sustained by a continued influx of infection. Under either scenario, however, the hub identification method is equally valid, since it traces transmission through all possible chains, weighted by their probability, to a most likely introduction site.

### 4.5.2 Accounting for the set of observed transmission hubs

The nine identified hubs listed in Table 4.1 are all small- to mid-sized cities, not the major population centres that conventional wisdom would predict as gateways of infection to the United States. Three of them – Grenada MS, Albany West GA, and Stockton CA – were likely responsible for sparking the majority of infections, triggering an estimated three-quarters (617) of the 834 observed outbreaks through gravity-driven onward transmission.

These unexpected results demand further discussion. While it is likely that air travel played an important role in disseminating the 2009 A/H1N1pdm virus both internationally and within the US during the early spring wave [55, 59], these results indicate that other critical ingredients are needed to explain the spatial introduction patterns of the autumn wave of the pandemic in the continental US. The simulation study in §4.4 casts doubt on the possibility that the observed distribution of hubs is due to a bias in data or methods, which might estimate earlier epidemic onset times in smaller ZIPs than in larger ones by chance. In addition, there is little evidence that small locations in the true data set have systematically earlier or more uncertain epidemic onsets than large ones. Fig 2.12 depicts epidemic onset uncertainty vs. ZIP population size, and finds no significant relationship. The upper-left scatter in Fig 2.10 depicts epidemic onset week vs. population size, and reveals that larger ZIPs tend to have earlier onsets than smaller ones, not the reverse. Furthermore, when comparing the epidemic onset times of the 45 smallest ZIPs (those with population size below 50,000) with the epidemic onset time of the nearest ZIP with more than 100,000 people, there are 12 pairs where the smaller ZIP onset precedes the larger, 27 where the larger ZIP onset precedes the smaller, and 6 where the two onsets coincide. This again suggests that onsets in smaller ZIPs do not systematically precede onsets in larger ZIPs. Finally, for simulated outbreaks, the breakpoint method reliably estimates epidemic onset time, even when the epidemic time series is noisy (see §2.3).

Before examining possible explanations for the geographic distribution of the hubs, it is worth mentioning that epidemic establishment can be a highly stochastic process. According to classic probability theory of epidemic spread [9, 26], a chain of transmission can stochastically break early in an epidemic, even if the basic reproductive number ($R_0$) is above 1. If the chain is not broken during this early phase, a major epidemic is predicted to unfold roughly deterministically, infecting a large fraction of all susceptible individuals. For directly-transmitted immunizing infections, the probability of stochastic fadeout is approximately $(1/R_0)^x$, where $x$ is the number of initial infected individuals. With a reported $R_0$ of 1.6 for A/H1N1pdm [257], this would predict a 62% probability of early fade-out for a single long-distance introduction, and a 39% probability of early fade-out even with two

long-distance introductions into the same community. This highly random filter between introduction and robust establishment may help explain the curious spatial invasion pattern of the autumn wave of the 2009 A/H1N1pdm outbreak in the US.

Also, identifying small ZIPs as hubs is not necessarily a cause for surprise, since even large cities are sometimes partitioned into small ZIPs. For instance, there are 12 ZIPs in New York City with fewer than 600,000 people; if one of these had been a hub, there would be no need to question the conventional wisdom of epidemic establishment in highly-connected locations. The real counter-intuitive result is that the hubs do not geographically coincide with major cities. However, the highly-dispersed geographic distribution of the US population makes the observed set of hubs more plausible. To calculate the probability of observing hubs in minor cities, it is necessary to shift attention away from ZIPs, which do not generally reflect an epidemiologically or socially relevant partition of the US population. The 2010 US Census' definition of an incorporated place [233] corresponds more directly to the common notion of a city. There are 26 incorporated places in the US with population size greater than 600,000 (just over the size of the largest hub). These cities account for only 15.5% of the total US population. So, if a person is chosen at random to spark an outbreak in his/her home town, there is a 84.5% chance that this person will not be from a major city. There is a 22% chance that nine consecutively-chosen random people are not from major cities – and thus a 22% chance of observing a distribution of hubs similar to the one observed here, assuming all individuals had equal probability of seeding an outbreak. There is therefore no evidence to reject the hypothesis that an unbiased stochastic process was responsible for the long-range jumps of A/H1N1pdm in the US in the autumn of 2009.

Nevertheless, previous immunity, the start of the autumn school term, and meteorological effects may have tipped the balance further toward outbreak establishment in these smaller hubs. An early wave of A/H1N1pdm influenza struck some major US cities, including New York and Chicago, between April and June of 2009, and may have conferred some immunity on those cities' populations. A brief calculation shows that this underlying immunity could have doubled the number of importation events needed to trigger an outbreak. To begin, note that the probability of outbreak establishment in a population after $k$ introductions of a disease with basic reproduction number $R_0$ is approximately [129]

$$p = 1 - \frac{1}{R_0^k} \tag{4.5}$$

which may be solved for $k$:

$$k = \frac{-\log(1-p)}{\log(R_0)}. \tag{4.6}$$

Now, imagine that a previous wave of infection has dropped the reproduction number to some value $R$. The number of infections needed to spark an outbreak in this partially immune population with the same probability as before is

$$k^* = \frac{-\log(1-p)}{\log(R)}. \tag{4.7}$$

The proportional increase in the number of cases needed to spark an outbreak with probability $p$ in the partially immune population is obtained by dividing Eq 4.7 by Eq 4.6, giving

$$\frac{k^*}{k} = \frac{\log(R_0)}{\log(R)}. \tag{4.8}$$

From Eq 4.8, if the spring infection wave in New York City and Chicago decreased the reproduction number roughly from 1.6 to 1.2, then it would have required $k^*/k = \log(1.6)/\log(1.2) = 2.6$ times more introductions on average to spark an outbreak in these cities than in a fully susceptible city. This may have prevented outbreak establishment during the autumn of 2009 in cities that suffered an earlier spring wave of infection.

Elsewhere, the start of the autumn school term may have increased the likelihood of pathogen establishment. Chao *et al.* (2010) [47] provide evidence of this at the state level. If schools in smaller towns went into session before those in nearby large cities, it would help explain why the smaller towns acted as hubs. Currently, this can only be investigated in Alabama, Florida, Georgia, Mississippi, and South Carolina, the five states where school start dates in 2009 are available at the district level. School start dates elsewhere are only available as a state average, from [47]. There are two hubs in these five southern states, in Grenada MS and Albany West GA. Both hubs are relatively close to Atlanta GA, a major urban centre and international transit hub. The median school start date in Grenada MS precedes the median school start date in Atlanta by about one half week, and the median school start date in Albany West precedes the median school start date in Atlanta by about one week. Perhaps more convincingly, a cluster of six ZIPs surrounding and including Albany West, and excluding Atlanta, had the country's earliest school start dates (see Fig 2.20). Though the difference in school term timing between this cluster and Atlanta is slight, between one-half and one week, this could help explain why Albany, rather than Atlanta, was an epicentre of transmission for the eastern half of the US. More detailed data are needed to determine whether similar differences in school term timing are associated with hubs in other states.

Finally, meteorological factors such as humidity may have influenced the geography of the hubs. Ambient absolute humidity has been linked to the survival and subsequent

transmissibility of the influenza virus [151, 209]. Indeed, Shaman *et al.* (2011) [209] correctly predicted a third pandemic wave in the southeastern US based on a spatiotemporal model of the effective reproductive number $R_E$ driven by absolute humidity. The results presented in this chapter show that the southeast also played a crucial role in the spread of the second (autumn) pandemic wave, since the two most influential hubs lie in that region. This warrants further investigation of meteorological effects that may have predisposed the southeast to outbreak establishment in 2009.

It is unfortunately impossible to identify or assess the importance of international hubs using the IMS-ILI dataset. This may especially affect inferences for the southwestern United States, since a major A/H1N1pdm outbreak was also occurring in the central and northern states of Mexico during the autumn of 2009 [49]. For example, the influenza activity in southern California, which is currently traced with high probability to the hub in Stockton (see Fig 4.7), might be explained better by some unobserved hub just across the US-Mexico border. This issue highlights the need for fine-scale influenza incidence data that can be compared across national boundaries.

### 4.5.3 Linking with genetic data

Geographic incidence data make it possible to identify pathogen establishment sites, as in this study and in [258]. A complementary approach for inferring establishment sites uses genetic data instead, as in [160] and [158]. Linking epidemiological and virological observations has proven difficult for human influenza [241], but the methods presented here may help bridge the gap by providing a spatially-detailed, testable hypothesis of the mixing patterns one might expect to see in spatially-referenced sequence data. Testing the observed patterns could proceed in two steps. First, geolocated A/H1N1pdm viral sequences from the US in 2009 would need to be gathered from databases such as FluDB[175] and GenBank[174]. These would be subdivided into regions according to the basins depicted in Figure 4.7; for example, a California group, an Idaho-Oregon-Washington group, an Arizona group, a Colorado-Utah group, an Iowa-Minnesota-Nebraska group, and an Eastern States group. A clustering analysis, similar to the one performed in [177], would reveal whether significant differences exist in the viral sequences between these regions. Importantly, Nelson *et al.* (2011)[177] conclude that similar viral strains caused the outbreaks in New York City NY, Milwaukee WI, and Houston TX, the only three cities studied in that article, in the autumn of 2009. In Figure 4.7, these three cities have similar hues, so there is already some agreement between this chapter's results and existing phylogeographic analysis of the pandemic. Second,

depending on the resolution of the available genetic data, the probabilities of pathogen jumps between regions or between cities could be calculated according to phylogenetic similarity and gathered into a Markov transition matrix. Hubs could then be identified using the step-tracing method presented in this article. These hubs could be cross-checked with the nine hubs identified in Table 4.1. Combining the data streams in this way would shed more light on the true transmission network of the 2009 pandemic, improving in turn our ability to develop effective and efficient interventions for future outbreaks.

## 4.6 Summary

A general method is introduced to identify the geographic transmission hubs of an epidemic on a metapopulation for which outbreak onset times are modelled as a function of the force of infection, expressed as the sum of independent contributions from various potential contributors of infection. Using this method, the transmission hubs of the autumn wave of the 2009 A/H1N1pdm influenza pandemic in the United States are identified. There are nine, most of which are small- to mid-sized cities. A simulation study shows that error due to noise is insufficient to explain the counter-intuitive set of transmission hubs. Instead, it is likely that epidemic establishment was governed by a highly stochastic process, possibly influenced by previous immunity, weather, and the start of the autumn school term.

# Chapter 5

# Age-specific transmission of the 2009 A/H1N1pdm influenza pandemic in the United States

The transmission dynamics of infectious diseases in humans can vary widely by the age of the infected host. Identifying which age groups are most responsible for sustaining overall transmission can improve strategies for outbreak control and prevention. Existing methods to infer the relative roles of different age groups in epidemic transmission, however, can normally only accommodate a few age classes, and require data that are highly-specific for the disease being studied. In this chapter, symbolic transfer entropy (STE), a concept adapted from the signal processing literature, is presented as a measure that can be used to identify the most transmissive age groups in an epidemic when data are noisy and split into many age groups. STE provides a relative ranking of which age groups dominate transmission, rather than a reconstruction of the explicit between-age-group transmission matrix. Simulation studies establish that STE can identify which age groups dominate transmission, even when there are systematic differences in reporting rates between the age groups. Then, the pairwise STE is calculated between time series of influenza-like illness for 12 age groups in 834 US cities during the autumn of 2009. Elevated STE from 5-19 year-olds to most other age groups indicates that school-aged children were likely the most important transmitters of infection within cities during the autumn wave of the 2009 pandemic in the US. The pairwise STE is also calculated between age groups in cities between which infection likely spread, as identified by the geographic transmission model presented in Chapter 3. These estimates suggest that school-aged children may also have contributed disproportionately to the short-distance geographic transmission of the outbreak. The results may be partially

confounded by higher rates of physician-seeking behaviour in children compared to adults, but it is unlikely that differences in reporting rates can fully explain the observed differences in STE. Finally, as an alternative test of age-specific geographic transmission of the 2009 influenza pandemic in the US, the age-stratified ILI data are explicitly incorporated into the transmission model developed in Chapter 3. A model selection procedure indicates that between-city transmission is best predicted by the combined ILI intensity in young infants ($<$2 years), young school-aged children (5-9 years), and young adults (20-29 years). This suggests that transmission from school-aged children alone, though dominant overall, still may not fully explain the between-city spread of the pandemic.

## 5.1 Background

Age is a predictor of both susceptibility to and transmissibility of many diseases, including influenza [143, 168, 187, 217, 245]. This section provides an overview of how host age affects the transmission of influenza, and describes the modelling approaches that have contributed to this knowledge. The benefits and limitations of these approaches are discussed, particularly with respect to inference on high-volume syndromic epidemiological data. Calculating symbolic transfer entropy (STE) is introduced as a candidate method for identifying patterns of transmission between age groups using this type of data.

### 5.1.1 Age as a key characteristic in disease transmission

A person's age largely dictates her/his interpersonal contact patterns. The POLYMOD study [170], one of the most complete sources of information on age-specific interpersonal contacts, provides strong evidence that people tend to interact most frequently with others close to their own age. This trend holds true within British primary schools as well, where mixing is highly age-specific [57].

For diseases that spread via casual contact, age is therefore a key predictor of an individual's risk of infection. However, care must be taken when using contact data to infer disease transmission patterns. Survey-based contact data are normally based on some criterion, such as time spent together, number of words spoken, or a physical touch [57, 71, 113, 170]. If a different type of contact is a better predictor of infection, the relevance of the contact data is diminished. Furthermore, age-related contact rates vary geographically, so parametrising a model for one geographic location using a different location's contact data may yield unrealistic results [170]. Finally, the way in which contact information is incorporated into

models of disease transmission must be chosen with care. Kucharski *et al.* (2014) for example find that an individual's risk of infection is influenced more by her/his age group's average mixing patterns than by the individual's own number of contacts [137].

With these caveats in mind, however, it is possible to infer which age groups may suffer the greatest burden of disease during an emerging outbreak. A mathematical model using the POLYMOD contact data shows that 5-19 year-olds are expected to bear the highest burden of infections during an outbreak of an emerging pathogen spread by casual contact [171]. For influenza, and particularly 2009 A/H1N1 pandemic influenza, a range of studies indicate that school-aged children suffer a relatively higher risk of infection compared to other age groups [81, 179, 245, 252, 259]. Findings by Brownstein *et al.* (2005) contrast somewhat with these studies, using hospital data to suggest that preschool-aged children may in fact be the first age group to become infected in seasonal influenza outbreaks, and may also be the first to transmit the disease to other age groups [27].

School-aged children may also be the main drivers of transmission during influenza epidemics. This notion is supported by a range of studies from different countries on different outbreaks [91, 170, 218, 245]. In light of this, many have suggested that preferentially vaccinating children may be the most effective way to disrupt the spread of an emerging influenza outbreak [157, 173, 245], and there is some empirical evidence that suggests that such a strategy would be effective [199].

The relationship between host age and influenza transmission may differ between pandemic and seasonal outbreaks [163]. Pandemic outbreaks are often marked by shifts in the age groups that suffer the highest morbidity and mortality. The 1918 and 2009 influenza pandemics both featured shifts in morbidity and mortality from the very young and very old towards young adults [126, 183, 217]. This may be partially due to underlying immunity in older individuals who have been exposed to a wider array of strains, some of which may resemble the pandemic strain more closely than any strains to which the younger age groups have been exposed [198]. The relationship between morbidity/mortality and onward transmission remains unclear, however. Reichert *et al.* (2012) [198] find that immunity from previous infection may protect against severe morbidity during a pandemic, but may not prevent infection – and therefore may also not prevent transmission. It is therefore not valid to assume that the age groups that suffer the worst health consequences are necessarily the ones most responsible for transmission.

Differences in healthcare-seeking behaviour between age groups can confound ILI-based estimates of which age groups dominate transmission. For example, if children seek healthcare for ILI more frequently than adults, this could lead to artificially high

estimates of ILI in children, and thus to artificially high estimates of children's contribution to overall transmission. Methods that aim to identify age-related differences in transmission should account for known differences in healthcare-seeking behaviour between age groups. Biggerstaff *et al.* (2012) [22] estimate that 40% of adults (defined as individuals over 18 years) and 56% of children with ILI sought healthcare during the 2009 influenza pandemic in the United States. These estimates remained fairly consistent during the following 2010-11 influenza season, when 45% of adults and 57% of children with ILI sought healthcare [21]. In a 2002-03 study of ILI in University of Minnesota students, 29.3% of students with ILI sought healthcare, which is substantially lower than the healthcare-seeking rates reported by Biggerstaff *et al.* for adults during the 2009 pandemic and 2010-2011 seasonal outbreak [21, 22, 178]. This difference could be due in part to a lower level of perceived personal risk from ILI before vs. after the 2009 pandemic, and also to the different age categories used in the two studies; Biggerstaff *et al.* (2012) [22] report healthcare-seeking rates for adults as a whole, while Nichol *et al.* (2005) [178] focus just on university-aged students. Overall, the evidence suggests that children with ILI seek healthcare more frequently than adults with ILI do.

### 5.1.2 Incorporating age into epidemiological models

Failing to incorporate demographic structure into epidemiological models can yield incorrect estimates of key epidemiological parameters. Nishiura *et al.* (2010), for example, find that models without age structure overestimated the reproduction number of the 2009 A/H1N1pdm influenza pandemic in Japan [180]. Also, epidemiological processes and demographic processes can sometimes interact, making it important to include age structure into epidemic models. Recurring outbreaks of the measles, for example, have been attributed to replenishment of the susceptible population as children are born [58]. Klepac and Caswell (2011) [134] provide a theoretical framework for these sorts of models. For the study of single outbreaks of acute infections such as those caused by influenza, however, these long-term interactions can normally be safely ignored; it is enough to account for the differences in risk of infection due to differences in contact patterns [171, 245] and, in some cases, previous exposure to the pathogen [78, 217].

The reproduction number, $R$, is perhaps the most commonly estimated quantity for characterising the transmissibility of a disease. For structured populations, however, it is natural to work with the next-generation matrix, which is a generalisation of the reproduction number [65, 66]. The structure of the next-generation matrix can reveal which sub-populations are

most responsible for sustaining the transmission of a disease. Just as the reproduction number captures the expected number of secondary infections caused by a single infected individual, the $i, j^{th}$ element of the next-generation matrix gives the expected number of secondary infections in class $i$ caused by a single infected individual in class $j$. The reproduction number is the dominant eigenvalue of the next-generation matrix [66]. Diekmann *et al.* (2010) [66] provide a recipe for calculating the next-generation matrix from compartmental epidemiological models, but epidemiological reasoning alone is often enough to construct the matrix, if the relative transmission rates between sub-populations are known.

Multiple studies use epidemiological models to infer which age groups drive the transmission of influenza. Nishiura *et al.* (2009) [179] parametrise a next-generation matrix for the 2009 A/H1N1pdm influenza outbreak in Japan, and find that children would have been capable of sustaining the outbreak amongst themselves, while the outbreak would have died out in a population of only adults. Glass *et al.* (2011) [90] extend a method introduced by Wallinga and Teunis (2004) [244] to estimate a time-varying next-generation matrix for the same Japanese outbreak [90]. Worby *et al.* (2015) use a simple ratio of pre- and post-peak influenza cases in five age groups to demonstrate that 5-19 year-olds were primarily responsible for transmission of the 2009 A/H1N1pdm pandemic in certain US states [252]. The methods presented in each of these studies are limited by the type of data they require and the amount of detail they can accommodate. The strategies proposed by Nishiura *et al.* (2009) [179] and Glass *et al.* (2011) [90] make strong assumptions about the structure of the next-generation matrix, which become increasingly unrealistic as the number of age classes grows. The method introduced by Worby *et al.* (2015) requires data that is highly specific for influenza, so ILI data is unsuitable. For syndromic ILI data with as many as 12 age classes, a different strategy is required.

### 5.1.3 Transfer entropy

To identify the epidemiological interactions between multiple age groups from noisy data governed by complicated dynamics, we seek a measure of stochastic 'driving' that makes as few assumptions as possible about the processes' underlying dynamics. Transfer entropy (TE) provides such a measure. Introduced by Schreiber (2000) [208], the TE gives the amount of information the past states of one stochastic process provide about the transition probabilities of another. It is sometimes referred to as the amount of information "transferred" from one process to another [208]. If $I$ and $J$ are discrete-state and discrete-time random processes such that $i_t$ and $j_t$ are the states of processes $I$ and $J$ at time $t$, then the TE from

process $J$ to process $I$ is defined as

$$T_{J\to I} = \sum_{\Omega_I,\Omega_J} p(i_{t+1}, i_t^{(k)}, j_t^{(l)}) \log\Big(\frac{p(i_{t+1}|i_t^{(k)}, j_t^{(l)})}{p(i_{t+1}|i_t^{(k)})}\Big) \tag{5.1}$$

where $i_t^{(k)}$ is shorthand notation for $(i_t, \ldots, i_{t-k+1})$, and similarly $j_t^{(l)} = (j_t, \ldots, j_{t-l+1})$. The logarithm has base 2, so that the transfer entropy is measured in bits. The sum is over all possible combinations of states $(i_{t+1}, i_t^{(k)}, j_t^{(l)})$, where $i_{t+1}, i_t^{(k)} \in \Omega_I$ and $j_t^{(l)} \in \Omega_J$, and $\Omega_I$ and $\Omega_j$ are the state spaces for processes $I$ and $J$. Eq 5.1 is a Kullback-Leibler divergence that measures how much process $I$ deviates from the generalised Markov property $p(i_{t+1}|i_t, \ldots, i_1) = p(i_{t+1}|i_t^{(k)})$, given the last $l$ states of process $J$. In practice, the histories are often fixed at length 1 ($k = l = 1$) [208]. Note that the TE from a Markov process to itself, $T_{I\to I}$, is zero.

The TE is related to mutual information. The mutual information for two random variables $I$ and $J$ is defined as

$$M_{IJ} = \sum p(i,j) \log\Big(\frac{p(i,j)}{p(i)p(j)}\Big) \tag{5.2}$$

[208] where the sum is over all possible states $i$ and $j$ of the variables $I$ and $J$. This is another Kullback-Leibler divergence that measures the deviation of the joint process $p(i,j)$ from the assumption that the processes $I$ and $J$ are independent. Unlike the TE, mutual information is symmetric; that is, it measures the probabilistic dependence between two processes, but cannot determine the direction of information transfer between them, if there is any [208]. Measuring the delayed mutual information between two processes is one way to introduce asymmetry. This takes a step toward inferring whether one process influences another, by measuring shared information between the present state of one process and the past states of another [208]. While the lagged mutual information describes how one process' history predicts the static probabilities of another, the TE measures how one process' history influences the transition probabilities of another. Because of this, the TE is less likely to be confounded by a shared input signal, and is a better measure of stochastic 'driving' [208]. Section 2 of Kaiser and Schreiber (2002) [124] provides a detailed description of the differences between TE and mutual information.

There are at least two other strategies for detecting causal-type interactions between time series. Granger causality [97] measures how knowledge of one autoregressive stochastic process improves predictions of another. Granger causality is a special case of TE, when the stochastic processes are jointly Gaussian-distributed [12]. The TE is thus better suited

than Granger causality for making inferences on more general, possibly nonlinear, processes, though this comes at the expense of requiring more data and having no clear way to test statistical significance [12]. Convergent cross mapping (CCM) [228], on the other hand, was developed to detect causal relationships in stochastic systems with underlying deterministic structure. CCM relies on Takens' theorem [229] to reconstruct candidate manifolds of the underlying dynamical system using lagged observations from two time series. Causality is inferred if nearby points on one reconstructed manifold consistently map to nearby points on the other reconstructed manifold. CCM has been used to provide evidence that temperature and absolute humidity fluctuations drive the timing of global seasonal influenza outbreaks [62], though some controversy surrounds these findings [13, 227]. Nevertheless, it would be interesting to see whether CCM can reveal asymmetric epidemiological interactions between age groups, and to compare its findings with those identified using TE. Lungarella *et al.* (2007) [159] provide more detail on the relationships between various methods that infer causal relationships from time series data.

As a brief aside, we prefer to avoid the term 'causality' and instead speak of processes influencing, or sometimes 'driving', one another. Despite its frequent use in the literature, causality is a philosophically fraught term, and the methods discussed above only detect limited types of causality. Some argue that it is impossible for these methods to detect any true notion of causality, since identifying causality requires perturbing the alleged causal source, or otherwise intervening in the system [152]. Regardless of the vocabulary used, these methods have successfully detected meaningful relationships between real-world stochastic processes [24, 125, 185, 223, 224, 228].

The TE is limited by only applying to discrete-state random processes. In practice, one often wishes to estimate the transfer of information between random processes with continuous or near-continuous state spaces. A few strategies have been introduced to make TE amenable to these sorts of processes, including coarse-graining the data into bins with adaptive widths, and reconstructing the processes' probability densities using non-parametric kernel estimators [124]. Both methods demand many modelling choices that may be difficult to defend, and can lead to widely divergent results [124].

Staniek and Lehnertz (2008) [223] introduce a more robust strategy for calculating information transfer between time series processes that have continuous- or near-continuous state spaces. Motivated by the insight that the relative amplitudes of subsequent observations from these sorts of processes may provide enough information to identify interactions between them, they propose symbolising the time series based on ordered $m$-tuples of observations. This creates an 'alphabet' of symbols that describe the qualitative structure of the time series

while greatly reducing the state space. First, the modeller chooses the number of consecutive points $m$ that make up each symbol. For $m = 2$, there are only two possible symbols: $i_{t-1} < i_t$ and $i_{t-1} > i_t$. For $m = 3$, there are six possible symbols, which we label $A$ through $F$:

$$A: \ i_{t-2} < i_{t-1} < i_t \tag{5.3}$$

$$B: \ i_{t-2} < i_t < i_{t-1} \tag{5.4}$$

$$C: \ i_{t-1} < i_{t-2} < i_t \tag{5.5}$$

$$D: \ i_{t-1} < i_t < i_{t-2} \tag{5.6}$$

$$E: \ i_t < i_{t-2} < i_{t-1} \tag{5.7}$$

$$F: \ i_t < i_{t-1} < i_{t-2} \tag{5.8}$$

These symbols are depicted in Fig 5.1. In general, for a given $m$, there are $m!$ possible symbols. For a process $I$ with $n$ observations, a symbol is assigned to each state $(i_m, i_{m+1}, \ldots, i_n)$. Fig 5.2 depicts how a given time series would be symbolised for $m = 2$ and $m = 3$. Doing so yields a new process $\hat{I}$ with observations $(\hat{i}_m, \hat{i}_{m+1}, \ldots, \hat{i}_n)$. The same may be done for process $J$. The symbolic transfer entropy (STE) is then defined as

$$T^S_{J\to I} = \sum p(\hat{i}_{t+1}, \hat{i}_t, \hat{j}_t) \log\Big(\frac{p(\hat{i}_{t+1}|\hat{i}_t, \hat{j}_t)}{p(\hat{i}_{t+1}|\hat{i}_t)}\Big) \tag{5.9}$$

where the sum is over all possible symbols for states $\hat{i}_{t+1}$, $\hat{i}_t$, and $\hat{j}_t$.

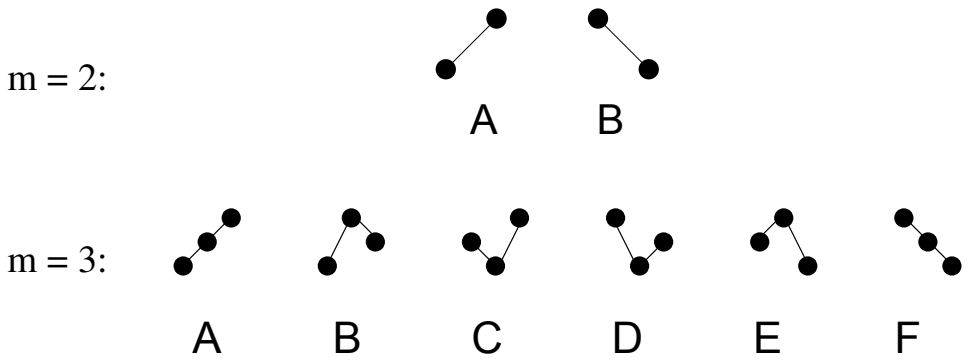



Fig. 5.1 List of the two possible symbols for $m = 2$ (top) and the six possible symbols for $m = 3$ (bottom). The six $m = 3$ symbols are specified by Eqs 5.3-5.8. The ordering of the symbols is arbitrary, but will remain consistent throughout this chapter.

In practice, the joint and conditional probabilities in Eq 5.9 are estimated using the relative frequencies of the symbols in the observed dataset. For example, to calculate the

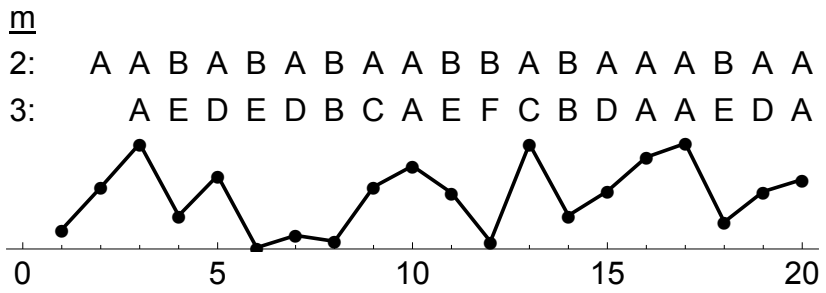



Fig. 5.2 An example time series and its symbolisation with $m = 2$ and $m = 3$. For $m = 2$, $A$ represents an increase and $B$ represents a decrease. The symbols corresponding to the letters for $m = 3$ are defined in Eq 5.3-5.8 and depicted in Fig 5.1.

joint probability $p(\hat{i}_{t+1} = A, \hat{i}_t = B, \hat{j}_t = C)$ for symbols $A$, $B$, and $C$, one counts the number of times a consecutive $B$ and then $A$ are observed in the $I$ process, with a $C$ in the $J$ process simultaneous with the $B$ in the $I$ process. Dividing this count by the total length of the process, $n - m + 1$, gives an estimate of the joint probability.

If multiple realisations of the process are available, the probabilities may be estimated by the relative frequencies of the symbols across all realisations. The process need not have reached a stationary distribution, as long as enough realisations are available [124]. Fig 5.3 depicts how this is done. Consider $k$ realisations from two possibly related stochastic processes $I$ and $J$. Each realisation consists of a time series of length $n$. So, there are $k$ pairs of length-$n$ time series available for estimating the STE between variables $I$ and $J$. In Fig 5.3, $k = 3$ and $n = 17$. To estimate the STE, each of the time series is symbolised using symbols of length $m$. This yields $k$ pairs of symbol strings, where each string within a pair has length $n_S = n - (m - 1)$. In Fig 5.3, $m = 2$, so each symbol string has length $n_S = 17 - (2 - 1) = 16$.

Next, the probabilities that make up the STE sum, Eq 5.9, are estimated using the relative frequencies of the symbols in the symbol strings. For example, $P(\hat{i}_{t+1} = B, \hat{i}_t = A, \hat{j}_t = A)$ is calculated as the number of times a consecutive $A$ then $B$ are observed in the $I$ process, with an $A$ in the $J$ process concurrent with the $A$ in the $I$ process. It is helpful to interpret this as the number of times the pattern

$$\begin{matrix} A & B \\ A & \end{matrix}$$

is observed when the symbol strings for a realisation of $I$ and $J$ are aligned one above the other (see the symbol strings in Fig 5.3). The frequency with which this pattern appears is

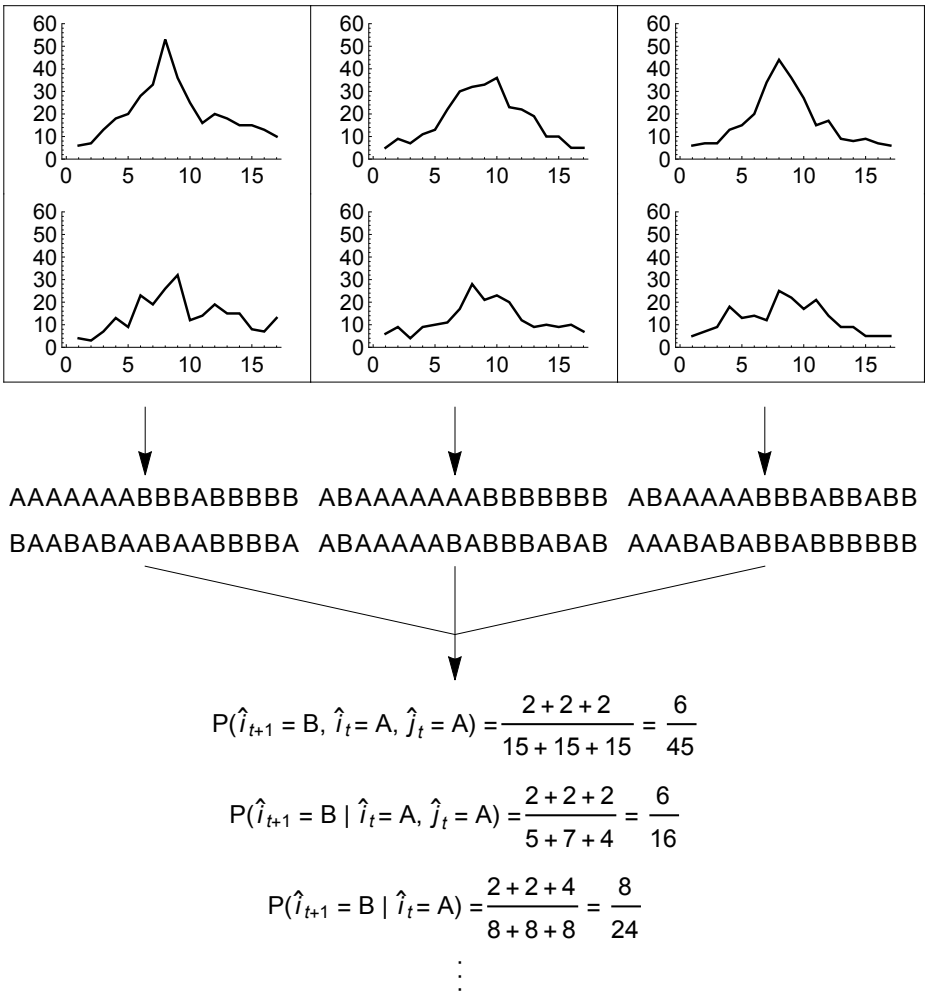



Fig. 5.3 Summary of how the STE is calculated from multiple realisations of an epidemic process with two age groups. The three columns in the top pane portray $k = 3$ realisations from an epidemic process, which in this case is the Poisson-type process described in §5.2.4. The columns consist of two time series plots each, which depict the simulated case counts for the two age groups (call them $I$ and $J$) across 17 time steps. The time series are symbolised with a symbol length of $m = 2$, yielding the character strings beneath the first set of arrows. An $A$ represents an increase and a $B$ represents a decrease. The probabilities that make up the STE sum, Eq 5.9, are estimated using the relative frequencies of the symbols in the character strings. A few of these probabilities are calculated in the lower section of this figure. See the main text (§5.1.3) for more details on how these probability estimates are computed.

calculated for each of the $k$ pairs of symbol strings, yielding a set of counts $c_1, \ldots, c_k$. Since there are $n_S - 1$ possible positions for the pattern to appear within each pair of symbol strings, the overall joint probability is estimated as

$$\frac{c_1 + c_2 + \cdots + c_k}{k(n_S - 1)}. \tag{5.10}$$

Conditional probabilities are estimated in a similar way. For example, $P(\hat{i}_{t+1} = B | \hat{i}_t = A, \hat{j}_t = A)$ is estimated as the number of times a $B$ is observed in the $I$ process, given that the previous observations in the $I$ and $J$ processes were both $A$. The counts $c_1, \ldots, c_k$ are the same as before, but rather than dividing by the total length of the symbolised time series, we instead divide by the number of times the pattern

$$A$$
$$A$$

is observed. Denote these counts $\hat{c}_1, \ldots, \hat{c}_k$. Then, the conditional probability is

$$\frac{c_1 + c_2 + \cdots + c_k}{\hat{c}_1 + \hat{c}_2 + \cdots + \hat{c}_k}. \tag{5.11}$$

The calculation is easier for the probabilities conditioned on just one term. The quantity $P(\hat{i}_{t+1} = B | \hat{i}_t = A)$, for example, is simply the number of times an $AB$ appears in the symbolised time series for $I$, divided by the number of times an $A$ appears in the symbolised time series for $I$.

The above steps are demonstrated with a specific example in Fig 5.3. The joint and conditional probabilities in Eq 5.9 may be estimated using the steps outlined above, adjusting the specific symbols as necessary.

The STE has been used to study epileptogenic neural signals and the dissemination of information through social networks [24, 223]. These studies rely on the STE's robustness to amplitude-adjusting effects, including point-wise random noise and process-wide vertical shifts. This robustness also makes STE well-suited for studying age-stratified ILI data, which features point-wise noise from non-influenza respiratory illness, and which is subject to broad-scale amplitude shifts due to differences in physician-seeking behaviour between age groups.

Estimating STE probabilities requires large volumes of data, as a trade-off for the method's minimal assumptions about the underlying process dynamics. When sufficient data are available, STE offers a way of identifying interactions between processes governed by

complicated dynamics from noisy data. The following sections consider how STE behaves when confronted with data from both simulated and true epidemics.

## 5.2 Symbolic transfer entropy and epidemiological processes

To my knowledge, there has been no systematic evaluation of whether STE can detect true differences in transmission strengths from age-structured epidemiological time series. This section presents a range of tests to check whether STE can reliably infer which age groups drive transmission in simulated outbreaks. First, the 'contextual STE' is introduced, which is the STE measured forward from a point in an epidemiological time series with known amplitude and known underlying dynamics. This is presented to ensure that symbolising the time series does not ignore too much of the information encoded in the exact amplitude values, and to verify that the most transmissive age groups can be consistently identified for a range of epidemiologically feasible parameters.

Then, the full STE is calculated for a variety of age-structured epidemic simulations. For the first set of simulations, the STE is calculated using an individual-based two-age-class SIR model, where dependence between the age classes varies continuously from none, to equivalent within- and between-group transmission, to strong asymmetric transmission from one age group. Then, due to the computational difficulties of including more age groups in such a detailed epidemic model, the same calculations are performed using a Poisson-type epidemic model, giving virtually equivalent results. Next, the STE is calculated on two sets of four-age-class Poisson-type epidemic models. The first set incorporates uniformly-varying reporting rates across all age groups to measure the ability of STE to detect true asymmetry in transmission rates despite incomplete reporting. The second set examines how STE responds to age-variable reporting rates when within- and between-group transmission strengths are equal, to see whether spurious differences in STE from unequal reporting rates are possible. Finally, STE estimates are made on two twelve-age-class Poisson-type models, one with strong asymmetric transmission from children and uniform reporting rates, and one with symmetric transmission but age-varying reporting rates. These last simulations match the resolution of the available IMS-ILI data, to be considered in §5.3. To summarise, STE estimates are made under the following simulation scenarios:

- 2-age-class individual-based SIR simulations with varying coupling between age groups (none to symmetric to asymmetric)

- 2-age-class Poisson-type simulations with varying coupling between age groups (none to symmetric to asymmetric), to ensure agreement with the SIR-based results
- 4-age-class Poisson-type simulations with asymmetric coupling between age groups and uniform reporting across age groups
- 4-age-class Poisson-type simulations with symmetric coupling between age groups and non-uniform reporting rates between age groups
- 12-age-class Poisson-type simulations with asymmetric coupling between age groups and uniform reporting across age groups
- 12-age-class Poisson-type simulations with symmetric coupling between age groups and non-uniform reporting rates between age groups.

### 5.2.1 The contextual STE

The STE assumes that that there are Markovian transitions between the symbols of the symbolised time series – that is, the probability of observing a particular symbol, given the symbol that precedes it (and possibly given the preceding value in some other time series), is some constant value, regardless of the exact amplitude of the underlying process. However, for disease outbreaks, the amplitude of the underlying process (i.e. the case count) clearly does matter. For an outbreak with case counts binned into regular time windows, the probability that an increase in cases occurs in a given age group from one time step to the next is related to the number of cases in each age group in the previous time step, the overall reproduction number, and the relative rates of infection between age groups. When the reproduction number is greater than 1, an increase in cases is more likely if there are, for example, 1000 cases in the current week than if there are 10. In the vocabulary of STE, when using symbols of length $m = 2$ (thus encoding simple increases/decreases in the time series), the probability of transitioning from the symbol $i_t < i_{t+1}$ to the (same) symbol $i_{t+1} < i_{t+2}$ is higher when $i_t$ is 1000 than when $i_t$ is 10, if the reproduction number is greater than 1. This violates the Markovian assumption embedded into the STE calculation, which assumes that the joint and conditional probabilities in Eq 5.9 are consistent throughout the entire process.

To test whether the STE can provide reliable information about age-structured transmission dynamics despite this deviation from the Markov property, we here calculate a version of the STE for a range of epidemiologically reasonable reproduction numbers, between-age-group infection rates, and true numbers of infected individuals. To lay the groundwork,

consider an outbreak where the number of cases per unit time in two age groups are described by the discrete random processes $I = \{i_1, \ldots, i_n\}$ and $J = \{j_1, \ldots, j_n\}$. Define the contextual STE as

$$T^S_{J\to I,t} = \sum p(\hat{i}_{t+1}, \hat{i}_t, \hat{j}_t | i_{t-m+1}, j_{t-m+1}) \log\Big(\frac{p(\hat{i}_{t+1}|\hat{i}_t, \hat{j}_t, i_{t-m+1}, j_{t-m+1})}{p(\hat{i}_{t+1}|\hat{i}_t, i_{t-m+1}, j_{t-m+1})}\Big). \tag{5.12}$$

where the sum is over all possible values (symbols) of $\hat{i}_{t+1}$, $\hat{i}_t$, and $\hat{j}_t$, and $m$ is the symbol length. Values with hats denote symbols, and values without hats denote actual case counts. The contextual STE is equivalent to the full STE (Eq 5.9), conditional on the true amplitudes of processes $I$ and $J$ at time $t-m-1$, or the 'context' of the underlying process. If the contextual STE gives accurate insight into the underlying epidemiological process regardless of the context, then there is reason to believe that the full STE will give similarly accurate insight.

Next, we define an epidemiological model that describes the distribution of cases in age groups $I$ and $J$ at (discrete) time $t$, given the number of cases in each age group at time $t-1$:

$$P(i_t | i_{t-1}, j_{t-1}) = f^I(i_t; i_{t-1}, j_{t-1}, \boldsymbol{\lambda}) \tag{5.13}$$

$$P(j_t | i_{t-1}, j_{t-1}) = f^J(j_t; i_{t-1}, j_{t-1}, \boldsymbol{\lambda}) \tag{5.14}$$

$$\tag{5.15}$$

where $f^I$ and $f^J$ are probability mass functions, and $\boldsymbol{\lambda}$ is a set of parameters specifying the rates of infection between age groups $I$ and $J$. The case counts $i_t$ and $j_t$ are assumed to be nonnegative integers. We first simplify notation by setting

$$f^I(i_t; i_{t-1}, j_{t-1}, \boldsymbol{\lambda}) = f^I_{t-1}(i_t) \quad \text{and} \tag{5.16}$$

$$f^J(j_t; i_{t-1}, j_{t-1}, \boldsymbol{\lambda}) = f^J_{t-1}(j_t) \tag{5.17}$$

and similarly defining the corresponding CDFs

$$F^I_t(x) = \sum_{k=0}^{x} f^I_t(k) \tag{5.18}$$

$$F^J_t(x) = \sum_{k=0}^{x} f^J_t(k) \tag{5.19}$$

where $x$ is a nonnegative integer.

A common choice is to model the disease case counts as Poisson random variables, each with a rate that is a linear combination of the case counts in the previous time step in each age group:

$$P(i_t|i_{t-1}, j_{t-1}) \sim Poisson(\lambda_{11} i_{t-1} + \lambda_{12} j_{t-1}) \tag{5.20}$$

$$P(j_t|i_{t-1}, j_{t-1}) \sim Poisson(\lambda_{21} i_{t-1} + \lambda_{22} j_{t-1}) \tag{5.21}$$

giving

$$f^I_{t-1}(i_t) = \text{Exp}[-(\lambda_{11} i_{t-1} + \lambda_{12} j_{t-1})] \frac{(\lambda_{11} i_{t-1} + \lambda_{12} j_{t-1})^{i_t}}{i_t!} \qquad \text{and} \tag{5.22}$$

$$f^J_{t-1}(j_t) = \text{Exp}[-(\lambda_{21} i_{t-1} + \lambda_{22} j_{t-1})] \frac{(\lambda_{21} i_{t-1} + \lambda_{22} j_{t-1})^{j_t}}{j_t!}. \tag{5.23}$$

If the length of the time steps matches the disease's generation interval, the next-generation matrix is simply

$$NGM = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix}, \tag{5.24}$$

and the reproduction number is the dominant eigenvalue of this matrix.

We now seek to express the contextual STE, Eq 5.12, in terms of the epidemiological process, defined by $f^I$ and $f^J$. For now, consider a symbol length of $m = 2$, and hence $2! = 2$ symbols. The contextual STE sum has $2 \times 2 \times 2 = 8$ terms, one for each set of possible symbols $\{\hat{i}_{t+1}, \hat{i}_t, \hat{j}_t\}$. The derivation will be done for a single term of that sum. The derivations for the other terms follow nearly identical steps.

Consider the term of Eq 5.12 corresponding to two consecutive decreases in process $I$ with a concurrent decrease in process $J$:

$$P(i_{t+1} < i_t, i_t < i_{t-1}, j_t < j_{t-1}|i_{t-1}, j_{t-1}) \log\Big(\frac{P(i_{t+1} < i_t|i_t < i_{t-1}, j_t < j_{t-1}, i_{t-1}, j_{t-1})}{P(i_{t+1} < i_t|i_t < i_{t-1}, i_{t-1}, j_{t-1})}\Big) \tag{5.25}$$

where the inequalities now take the place of the symbols (the terms with hats) from Eq 5.12. We seek to express these probabilities in terms of $f^I$ and $f^J$. To avoid cluttering notation, assume that $i_{t-1}$ and $j_{t-1}$ are given throughout. First, consider the joint probability (the first

term, before the logarithm in Eq 5.25):

$$P(i_{t+1} < i_t, i_t < i_{t-1}, j_t < j_{t-1}) = \tag{5.26}$$

$$P(i_t < i_{t-1}, j_t < j_{t-1}) P(i_{t+1} < i_t | i_t < i_{t-1}, j_t < j_{t-1}) = \tag{5.27}$$

$$P(i_t < i_{t-1}) P(j_t < j_{t-1}) P(i_{t+1} < i_t | i_t < i_{t-1}, j_t < j_{t-1}) = \tag{5.28}$$

$$P(i_t < i_{t-1}) P(j_t < j_{t-1}) \sum_{j_t=0}^{j_{t-1}} \sum_{i_t=0}^{i_{t-1}} P(i_{t+1} < i_t | j_t, i_t) P(i_t | i_t < i_{t-1}) P(j_t | j_t < j_{t-1}) = \tag{5.29}$$

$$= F^I_{t-1}(i_{t-1}) F^J_{t-1}(j_{t-1}) \sum_{j_t=0}^{j_{t-1}} \sum_{i_t=0}^{i_{t-1}} F^I_t(i_t) \frac{f^I_{t-1}(i_t)}{F^I_{t-1}(i_{t-1})} \frac{f^J_{t-1}(j_t)}{F^J_{t-1}(j_{t-1})} = \tag{5.30}$$

$$= \sum_{j_t=0}^{j_{t-1}} \sum_{i_t=0}^{i_{t-1}} F^I_t(i_t) f^I_{t-1}(i_t) f^J_{t-1}(j_t) \tag{5.31}$$

where (5.27) follows from a definition of conditional probability, (5.28) follows from the independence of $i_t$ and $j_t$ given $i_{t-1}$ and $j_{t-1}$, (5.29) follows from the law of total probability, (5.30) follows from substituting terms, and (5.31) follows from cancellation of the two CDFs in $I$ and $J$.

The probabilities inside the logarithm in Eq 5.25 may be similarly expressed in terms of the epidemiological process, giving

$$\sum_{j_t=0}^{j_{t-1}} \sum_{i_t=0}^{i_{t-1}} F^I_t(i_t) f^I_{t-1}(i_t) f^J_{t-1}(j_t) \log\Big( \frac{\sum_{j_t=0}^{j_{t-1}} \sum_{i_t=0}^{i_{t-1}} F^I_t(i_t) f^I_{t-1}(i_t) f^J_{t-1}(j_t)}{F^J_{t-1}(j_{t-1}) \sum_{j_t=0}^{\infty} \sum_{i_t=0}^{i_{t-1}} F^I_t(i_t) f^I_{t-1}(i_t) f^J_{t-1}(j_t)} \Big) \tag{5.32}$$

as an alternative expression for the term (5.25). The remaining seven terms in the contextual STE sum, Eq 5.12, may also be expressed in terms of the epidemiological process using similar derivations. It is now possible to explore how different epidemiological scenarios affect the contextual STE.

### 5.2.2 Contextual STE for under various epidemiological scenarios

This section considers how the contextual STE varies with reproduction number $R$ and underlying case counts $i_{t-1}$ and $j_{t-1}$, given some relative within- and between-group rates of transmission.

First, note that the contextual STE from group $J$ to group $I$ is zero when the number of cases in group $I$ does not depend on the previous number of cases in group $J$; that is, when $P(i_t | i_{t-1}, j_{t-1}) = P(i_t | i_{t-1})$. This property should be expected, since there is no transfer of

infection, and thus should be no transfer of information, from group $J$ to group $I$. To verify, consider the numerator in the logarithm in Eq 5.32. Since $F_t^I$ and $f_{t-1}^I$ do not depend on $j_t$, the sum over $j_t$ can be brought in front of the final term, yielding

$$\sum_{i_t=0}^{i_{t-1}} F_t^I(i_t) f_{t-1}^I(i_t) \sum_{j_t=0}^{j_{t-1}} f_{t-1}^J(j_t) \tag{5.33}$$

$$= \Big( \sum_{i_t=0}^{i_{t-1}} F_t^I(i_t) f_{t-1}^I(i_t) \Big) F_{t-1}^J(j_{t-1}) \tag{5.34}$$

Similarly, in the denominator, the sum over $j_t$ may be brought in front of the final term, giving

$$F_{t-1}^J(j_{t-1}) \sum_{i_t=0}^{i_{t-1}} F_t^I(i_t) f_{t-1}^I(i_t) \sum_{j_t=0}^{\infty} f_{t-1}^J(j_t) \tag{5.35}$$

$$= F_{t-1}^J(j_{t-1}) \sum_{i_t=0}^{i_{t-1}} F_t^I(i_t) f_{t-1}^I(i_t) \tag{5.36}$$

sine the infinite sum is equal to 1. The numerator (5.34) and denominator (5.36) are equal, giving the logarithm an argument of 1 and making the overall term's value zero. This happens for all eight terms in the contextual STE expression (Eq 5.12), yielding a contextual STE that is exactly equal to zero. So, it is guaranteed that when one age group does not contribute infection to another, the contextual STE in that direction will be zero, as required.

To explore further characteristics of the contextual STE, consider the Poisson-type epidemiological model given in Eqs 5.20-5.21. For three different sets of within- and between-group rates of transmission, we check how the contextual STE varies with reproduction number $R$ and contextual case counts $i_{t-1}$ and $j_{t-1}$.

Before beginning, we seek an expression for the next-generation matrix $\boldsymbol{\lambda}$ in terms of the reproduction number $R$ and some relative rates of within- and between-group transmission. Define the relative rate matrix

$$\boldsymbol{r} = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix} \tag{5.37}$$

where element $r_{ij}$ is the relative rate at which an infected individual in class $j$ infects individuals in class $i$. So, for example, if $r_{12} = 2r_{21}$, then the infection rate from group 2 to group 1 is double the infection rate from group 1 to group 2. Let $\rho$ be the dominant

eigenvalue of $\boldsymbol{r}$. The next-generation matrix is

$$\boldsymbol{\lambda} = \frac{R}{\rho}\boldsymbol{r}. \tag{5.38}$$

This ensures that the dominant eigenvalue of $\boldsymbol{\lambda}$ is $R$, and that the relative magnitudes of the elements of $\boldsymbol{\lambda}$ match the relative magnitudes of the elements of $\boldsymbol{r}$. Note that using a constant multiple of the relative rate matrix $\boldsymbol{r}^* = c\boldsymbol{r}$ will still yield the same next-generation matrix $\boldsymbol{\lambda}$, since the dominant eigenvalue $\rho^*$ of $\boldsymbol{r}^*$ will simply divide out the constant $c$ again.

First, consider equal within- and between-group rates of transmission; that is,

$$\boldsymbol{r} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}. \tag{5.39}$$

This describes mean-field dynamics between the age groups. The group that dominates transmission should therefore be the group with the most cases, at least during the phase roughly prior to the epidemic peak (when $R_{eff} > 1$), since there are no intrinsic differences in transmission rates. That is, if there are (somehow) 100 infected children and just one infected adult, children will be responsible for the bulk of new cases, even though each child individually has exactly the same transmission potential as the infected adult. We calculate the contextual STE for reproduction number $R$ between 0.6 and 1.5, which covers a range of possible reproduction numbers for influenza (see, for example, Smieszek *et al.* (2011) [218]). The number of cases in class $I$ at time $t-1$, $i_{t-1}$, is held fixed at 25, and the number of cases in class $J$ at time $t-1$, $j_{t-1}$, is allowed to vary from 5 to 50. Though these are small numbers of cases, they coincide with the weekly numbers of recorded ILI cases in some of the smaller age groups in the IMS-ILI dataset, especially in infants and the elderly (see Fig 5.23).

Fig 5.4 depicts $T^S_{I\to J,t-1} - T^S_{J\to I,t-1}$ for $R \in [0.6, 1.5]$ and $j_{t-1} \in \{5,\ldots,50\}$. Fig 5.4 is generated by considering pairs of $j_{t-1}$ and $R$, with $j_{t-1}$ between 5 and 50 in steps of 5, and $R$ between 0.6 and 1.5 in steps of 0.05. For each pair of $j_{t-1}$ and $R$, the contextual STE (Eq 5.12) is calculated in both directions ($I \to J$ and $J \to I$), using the summation terms expressed as in Eq 5.32. Contours depict the difference between these contextual STEs. Note that $i_{t-1}$, $j_{t-1}$, $\boldsymbol{r}$, and $R$ are sufficient to specify $F^I$, $F^J$, $f^I$, and $f^J$, using Eq 5.22–5.23.

When $R > 1$, the age group with more cases transfers the most information. That is, when $j_{t-1} < i_{t-1} = 25$ (Fig 5.4, upper left quadrant), $T^S_{I\to J,t-1} > T^S_{J\to I,t-1}$, and when $j_{t-1} > i_{t-1} = 25$ (upper right quadrant), $T^S_{I\to J,t-1} < T^S_{J\to I,t-1}$. This matches with what one might expect from the epidemiological dynamics: when $R > 1$, the age group with more infected individuals causes the majority of new infections, and so has a higher contextual

STE. When $R < 1$, interestingly, the reverse is true; the contextual STE is higher from the group with fewer cases. This is a spurious result arising from the STE not taking into account the underlying epidemic process. Clearly, epidemic decay is not 'driven' in the same way transmission is; decay is governed by individual recovery rates, and not by interactions between individuals. STE, however, searches simply for patterns in one process that predict patterns in another. During an epidemic's decay, one age group's dropping case counts may well anticipate drops in another group's case counts simply because the depletion of susceptibles in the first age group occurs earlier than in the second. The contextual STE identifies this relationship, and identifies the process with fewer cases as the one that is driving the decay. This suggests that caution is warranted when interpreting differences in STE; it must always be done with reference to the underlying epidemiological dynamics. In future work, it may be worthwhile to consider incorporating epidemiological intuition explicitly into the STE formulation; the simplest way to do this may be to restrict attention to only symbols that represent rises in amplitude. To summarise, when intrinsic within- and between-group rates of transmission are equal ($\boldsymbol{r}$ given by Eq 5.39), the relative number of cases at time $t-1$ dictates which group transfers the most information to the other. When $R > 1$, the group with more cases at time $t-1$ transfers the most information; when $R < 1$, the group with fewer cases at time $t-1$ transfers the most information.

Next, consider a case in which the within-group transmission rate for group $I$ is twice that of group $J$, but all other infection rates are equal. That is,

$$\boldsymbol{r} = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}. \tag{5.40}$$

In this scenario, sub-population $I$ does not directly contribute to infection in sub-population $J$ any more than $J$ contributes to itself or to $I$. However, the growth rate in the number of infections in group $I$ is higher than for group $J$, and so group $I$ will eventually account for the bulk of transmission. As before, the contextual STE is calculated using the relative rate matrix, Eq 5.40, for $R \in [0.6, 1.5]$ and $j_{t-1} \in \{5, \ldots, 50\}$, with $i_{t-1}$ fixed at 25. The contour plot in Fig 5.5 depicts $T^S_{I\to J,t-1} - T^S_{J\to I,t-1}$ under this scenario. The epidemiological intuition is supported; for nearly the full range of parameters, $T^S_{I\to J,t-1} > T^S_{J\to I,t-1}$. For large $j_{t-1}$ and $R > 1$ (upper right), and also for small $j_{t-1}$ and $R < 1$ (lower left), the dominant transfer of information is reversed ($T^S_{I\to J,t-1} < T^S_{J\to I,t-1}$), but only by a small amount. The $I \to J$ contextual STE dominates most when $j_{t-1}$ is small compared to $i_{t-1}$ and when $R > 1$ – that is, when the transmission-dominant age group ($I$) has more cases and the outbreak is on the upswing.

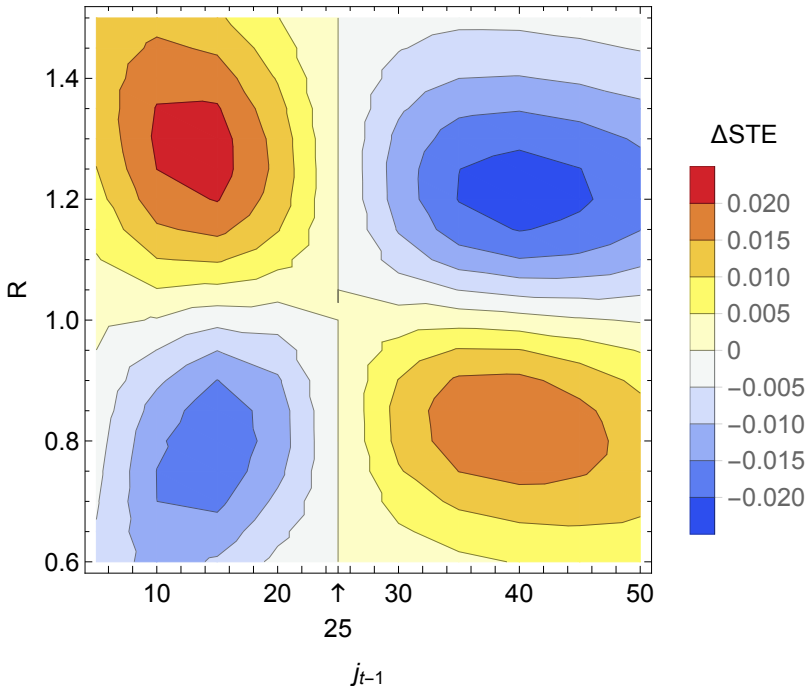



Fig. 5.4 Difference in contextual STE, $\Delta\text{STE} = T^S_{I\to J,t-1} - T^S_{J\to I,t-1}$, for $j_{t-1}$ between 5 and 50 and $R$ between 0.6 and 1.5, with $i_{t-1}$ fixed at 25. Units for the vertical scale are in bits. When $T^S_{I\to J,t-1} > T^S_{J\to I,t-1}$ (redder colours), there is evidence that process $I$ drives process $J$ more strongly than process $J$ drives process $I$, and vice-versa. The relative rates of within- and between-group transmission are equal, as specified by the rate matrix $\boldsymbol{r}$ (Eq 5.39). The $I \to J$ contextual STE is higher when $i_{t-1} > j_{t-1}$ and $R > 1$ (upper left quadrant), and when $i_{t-1} < j_{t-1}$ and $R < 1$ (lower right quadrant). The $I \to J$ and $J \to I$ contextual STEs are approximately equal when $i_{t-1} = j_{t-1}$ and when $R = 1$.

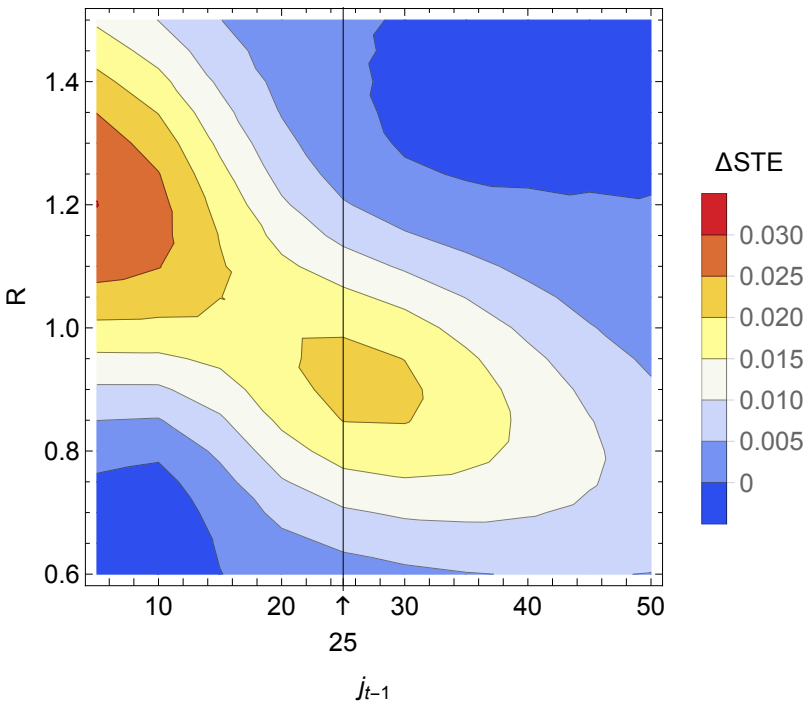



Fig. 5.5 Difference in contextual STE, $\Delta\text{STE} = T^S_{I\to J,t-1} - T^S_{J\to I,t-1}$, for $j_{t-1}$ between 5 and 50 and $R$ between 0.6 and 1.5, with $i_{t-1}$ fixed at 25. Units for the vertical scale are in bits. The within-group rate of infection for group $I$ is double the other infection rates (see Eq 5.40). The $I \to J$ contextual STE dominates in most of the parameter space, especially when $j_{t-1} < i_{t-1} = 25$ and $R > 1$. The $J \to I$ contextual STE is slightly higher than the $I \to J$ contextual STE for large $j_{t-1}$ and high $R$ (upper right), and also for small $j_{t-1}$ and low $R$ (lower left), with ΔSTE reaching a minimum of $-0.0032$ bits at $j_{t-1} = 10$ and $R = 0.65$.

Finally, consider the case in which the within-group transmission rate for group $I$ is quadruple the transmission rate from group $J$ to group $I$, and the transmission rate from group $I$ to group $J$ is double the transmission rate from group $J$ to group $I$. The within-group transmission rate for group $J$ is equal to the transmission rate from group $J$ to group $I$. That is,

$$\boldsymbol{r} = \begin{bmatrix} 4 & 1 \\ 2 & 1 \end{bmatrix}. \tag{5.41}$$

This corresponds to strong driving of transmission from group $I$. Again, the contextual STE is calculated using the relative rate matrix Eq 5.41 for $R \in [0.6, 1.5]$ and $j_{t-1} \in \{5, \ldots, 50\}$, with $i_{t-1}$ fixed at 25. The contour plot in Fig 5.6 depicts $T^S_{I\to J,t-1} - T^S_{J\to I,t-1}$ under this scenario. The $I \to J$ contextual STE is dominant throughout the parameter space, and is most dominant when $i_{t-1} > j_{t-1}$ and $R > 1$. Here, the contextual STE correctly identifies that group $I$ dominates transmission for the full range of epidemiologically-feasible parameter values.

### 5.2.3 Simulations on a two-age-class SIR model

We now shift attention from the contextual to the full STE. Since analytically evaluating the full STE with respect to an underlying epidemiological model becomes exceedingly difficult due to the complicated relationship between the symbolic transition probabilities and the underlying epidemic time series, simulation studies are used instead. This section verifies that the STE can identify meaningful epidemiological relationships between age groups using a two-age-class individual-based stochastic SIR model.

Consider two age groups $I$ and $J$. For each age group, let $X_I(t)$ and $X_J(t)$ be the number of susceptible individuals, $Y_I(t)$ and $Y_J(t)$ be the number of infected individuals, and $Z_I(t)$ and $Z_J(t)$ be the number of recovered individuals at (continuous) time $t$. Let $N$ be the total population size, equal to the total number of individuals in all age classes and disease states. $N$ remains constant throughout the simulation; there are no births and no deaths. The population sizes of each age group, $N_I$ and $N_J$, also remain constant throughout the simulation; individuals may not transition between the two age groups on the scale of this epidemic. Individuals transition from susceptible to infected to recovered according to the rates given in Table 5.1. The parameter $\beta_{mn}$ specifies the transmission rate from age group $n$ to age group $m$. The recovery rate $\gamma$ is assumed constant across age groups.

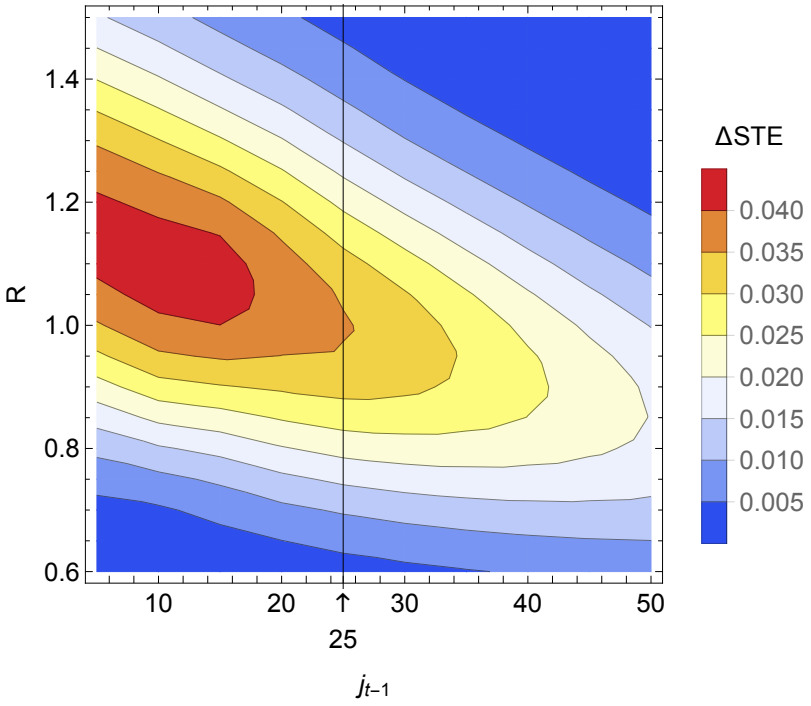



Fig. 5.6 Difference in contextual STE, $\Delta\text{STE} = T^S_{I\to J,t-1} - T^S_{J\to I,t-1}$, for $j_{t-1}$ between 5 and 50 and $R$ between 0.6 and 1.5, with $i_{t-1}$ fixed at 25. Units for the vertical scale are in bits. The within-group rate of infection for group $I$ is quadruple the within-group rate of infection for group $J$, and the infection rate from group $I$ to group $J$ is double that from group $J$ to group $I$ (see Eq 5.41). The $I \to J$ contextual STE dominates throughout of the parameter space ($\Delta\text{STE} > 0$ everywhere), especially when $j_{t-1} < i_{t-1} = 25$ and $R > 1$.

Table 5.1 Infection rates for the two-age-class SIR model

| Transition | Rate |
|---|---|
| $X_I \to Y_I$ | $\beta_{11}X_IY_I/N + \beta_{12}X_IY_J/N$ |
| $X_J \to Y_J$ | $\beta_{21}X_JY_I/N + \beta_{22}X_JY_J/N$ |
| $Y_I \to Z_I$ | $\gamma Y_I$ |
| $Y_J \to Z_J$ | $\gamma Y_J$ |

The next-generation matrix for this model is

$$NGM = \frac{1}{\gamma}\begin{bmatrix}\beta_{11}p_I & \beta_{12}p_I \\ \beta_{21}p_J & \beta_{22}p_J\end{bmatrix} \tag{5.42}$$

where $p_I = N_I/N$ and $p_J = N_J/N$. The within- and between-group transmission dynamics therefore depend on the infection rates $\beta_{mn}$ and the relative population sizes $N_I$ and $N_J$.

Let $\boldsymbol{r}$ be a matrix in which the $i, j^{th}$ entry gives the relative rate of infection from age class $j$ to age class $i$. For a given basic reproduction number $R_0$ and recovery rate $\gamma$, the transmission rates can be calculated:

$$\boldsymbol{\beta} = \begin{bmatrix}\beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22}\end{bmatrix} = \frac{R_0\gamma}{\rho}\text{diag}(1/p_I, 1/p_J)\cdot\boldsymbol{r} \tag{5.43}$$

where $\rho$ is the dominant eigenvalue of $\boldsymbol{r}$. This yields a next-generation matrix with dominant eigenvalue equal to $R_0$, in which the proportional differences between the terms match those of the rate matrix $\boldsymbol{r}$.

Table 5.2 lists the parameter values used for the individual-based SIR simulations. The basic reproduction number $R_0$ is fixed at 1.5, which is consistent with estimated values of $R_0$ for 2009 pandemic influenza [122, 257]. The average time to recovery $1/\gamma$ is 3.5 days, which is in line with estimates of the infectious period for 2009 pandemic influenza [257]. The population sizes $N_I$ and $N_J$ are small enough to ensure clearly stochastic dynamics (the dynamics become nearly deterministic as the population sizes increase), and are in line with the population sizes of the smaller age groups in the IMS-ILI dataset.

Table 5.2 Parameter values for the two-age-class SIR model

| Parameter | Value | Description | Units |
|---|---|---|---|
| $R_0$ | 1.5 | Basic reproduction number | people |
| $\gamma$ | 1/3.5 | Recovery rate | 1/day |
| $N_I$ | 400 | Number of individuals in age group $I$ | people |
| $N_J$ | 400 | Number of individuals in age group $J$ | people |

Simulations are implemented using the Gillespie algorithm, starting with one infected individual in age group $I$. Since $R_0$ is relatively low, there is a high chance of early epidemic die-out, so only outbreaks that last for at least 12.5 weeks are recorded. Once an outbreak is simulated, infections are binned into half-week intervals. Poisson noise is added to each bin

at a rate of 0.5% of the population size to simulate background non-influenza ILI. This rate is just below the 0.6% ILI ratio cutoff used in Gog *et al.* (2014) [91] to define out-of-season, low ILI activity weeks. Fig 5.7 depicts five simulated epidemic time series from this model.

To study how STE detects the transition from completely decoupled age groups to mean-field dynamics with equal within- and between-group transmission rates, consider a relative rate matrix of form

$$\boldsymbol{r} = \begin{bmatrix} 1 & z \\ z & 1 \end{bmatrix} \tag{5.44}$$

with $z \in [0,1]$. For each value of $z$ between 0 and 1 in steps of 0.1, one hundred ensembles of 800 outbreaks each are simulated. The epidemic time series are then symbolised using symbols of length $m = 3$. We do not consider longer symbol lengths because estimating the necessary joint and conditional probabilities becomes impractical. With $m = 4$, there are 24 possible symbols, which begins to defeat the purpose of symbolising to reduce the size of the state space. To my knowledge, $m = 4$ is the largest symbol length that has been considered in practice, and that was using a dataset of millions of Twitter tweets [24].

For each of the 100 ensembles, $T^S_{I\to J}$ and $T^S_{J\to I}$ are estimated using the relative symbol frequencies in the 800 symbolised time series. This gives 100 STE estimates for each value of $z$. The left-hand plot in Fig 5.8 depicts the mean STE and 95% confidence interval in both directions ($I \to J$ and $J \to I$) as a function of $z$. The $I \to J$ (blue) and $J \to I$ (black) STEs overlap for all values of $z$, correctly identifying the balanced influence between the age groups. The STE is near zero for $z = 0$ and increases steadily to approximately 0.06 bits as $z$ approaches 1, correctly identifying the increasing degree of coupling between the two age groups as $z$ increases.

Next, consider a rate matrix of form

$$\boldsymbol{r} = \begin{bmatrix} 1+3k & 1 \\ 1+k & 1 \end{bmatrix} \tag{5.45}$$

with $k \in [0,1]$. When $k = 0$, this is equivalent to the previous rate matrix, Eq 5.44, with $z = 1$. When $k = 1$, the within-group transmission rate for age class $I$ is four times the baseline transmission rate, and the transmission rate from group $I$ to group $J$ is twice the baseline transmission rate. This captures a continuum between the mean-field scenario and a scenario with strong forcing from age class $I$. As before, 100 ensembles of 800 epidemics each are simulated for values of $k$ between 0 and 1 in steps of size 0.1. For each ensemble, the STE from $I$ to $J$ and from $J$ to $I$ is calculated. The right-hand plot in Fig 5.8 depicts the mean STE and 95% confidence interval in both directions for $k$ between 0 and 1. For $k = 0$, the mean

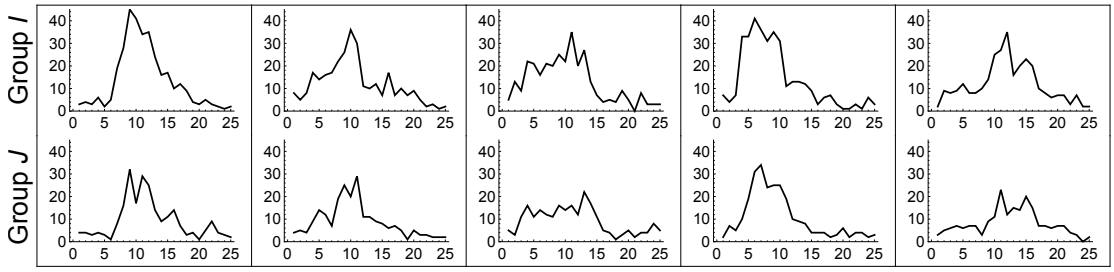



Fig. 5.7 Five simulations from the two-age-class individual-based SIR model, implemented using the Gillespie algorithm. Each column depicts the output from a single epidemic simulation, separated into case counts for group $I$ (top) and group $J$ (bottom). Vertical axes have units of case counts, and horizontal axes have units of half weeks, which is also assumed to be the length of the disease's generation interval. Transition rates are given in Table 5.1, with parameter values in Table 5.2. The transmission rates are specified by the relative rate matrix 5.45 with $k = 1$, for which age group $I$ (upper row) has quadruple the within-group transmission rate as group $J$, and for which the $I \to J$ transmission rate is twice the $J \to I$ transmission rate.

STE is approximately 0.06 bits, the same as it is for $z = 1$. As $k$ increases, $T^S_{I\to J}$ increases steadily and $T^S_{J\to I}$ decreases steadily, correctly identifying the increasing forcing from group $I$ to group $J$.

### 5.2.4 Simulations on a two-age-class Poisson model

The simulations in §5.2.3 provide evidence that the STE can capture meaningful epidemiological interactions between age groups. However, the epidemic sizes are still relatively small, and the transmission model only accounts for two age groups. Modelling larger population sizes and more age groups using an individual-based model demands large amounts of computational time. So, we now consider a more efficient Poisson-type model like the one in Eq 5.20-5.21 that specifies the number of cases in consecutive time steps as a function of the number of cases in each age class in the previous time step.

In particular, let the number of infected individuals in age classes $I$ and $J$ at time $t$ follow

$$i_t \sim Poisson(\lambda_{11,t} i_{t-1} + \lambda_{12,t} j_{t-1}) \tag{5.46}$$

$$j_t \sim Poisson(\lambda_{21,t} i_{t-1} + \lambda_{22,t} j_{t-1}) \tag{5.47}$$

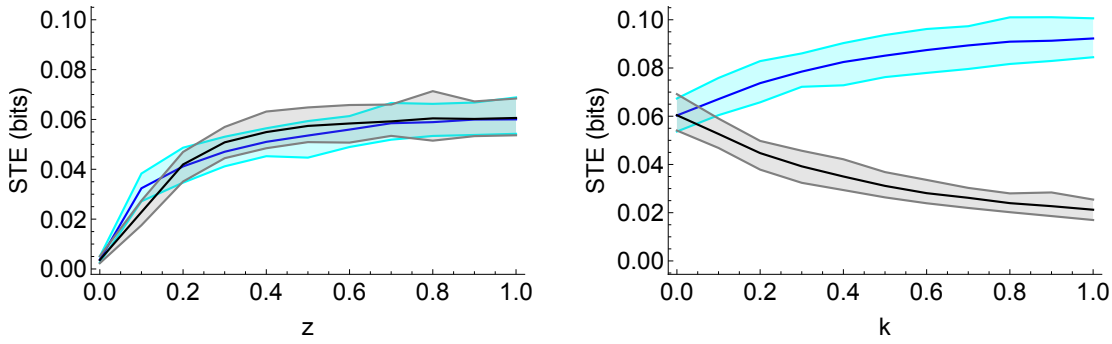



Fig. 5.8 Mean $T^S_{I\to J}$ (solid blue line) and $T^S_{J\to I}$ (solid black line) values for a range of within- and between-group transmission rates, as estimated from 100 ensembles of 800 epidemics for each value of $z$ and $k$, simulated using an individual-based SIR model. The shaded bands provide the approximate 95% confidence intervals of the STE estimates. The left-hand plot depicts the STE for the relative rate matrix given in Eq 5.44, which transitions smoothly from completely decoupled age groups to mean-field dynamics where within- and between-group transmission rates are equal. The right-hand plot depicts the STE for the relative rate matrix given in Eq 5.45, which transitions smoothly from mean-field dynamics to strong forcing from group $I$. In the left-hand plot, the STE increases steadily with $z$, and the $I \to J$ and $J \to I$ STE estimates overlap, correctly capturing the symmetric coupling between age groups that increases with $z$. In the right-hand plot, the STE correctly identifies statistically significant forcing from group $I$ to group $J$ as $k$ increases.

where $i_t$ and $j_t$ are case counts, and $\lambda_{mn,t}$ is the infection rate from age class $n$ to class $m$ at (discrete) time $t$. Unlike in the SIR model in 5.2.3, the infection rates $\lambda_{mn,t}$ are allowed to vary in time.

For the simulations, the length of the time steps is assumed to coincide with the generation interval for the disease. This way, the next-generation matrix is simply

$$NGM_t = \boldsymbol{\lambda}_t = \begin{bmatrix} \lambda_{11,t} & \lambda_{12,t} \\ \lambda_{21,t} & \lambda_{22,t} \end{bmatrix}. \tag{5.48}$$

For a given matrix $\boldsymbol{r}$ where element $r_{ij}$ denotes the relative rate of infection from group $j$ to group $i$, and a given reproduction number $R$, the next-generation matrix may be expressed as

$$\boldsymbol{\lambda}_t = \frac{R}{\rho}\boldsymbol{r} \tag{5.49}$$

where $\rho$ is the dominant eigenvalue of the matrix $\boldsymbol{r}$. As before, this ensures that the dominant eigenvalue of the next-generation matrix is equal to the reproduction number, and that the relative magnitudes of the elements of the next-generation matrix match the relative magnitudes of the elements of $\boldsymbol{r}$.

Epidemics are simulated by placing a single initial infected individual in either group $I$ or group $J$ with probability 0.5, and then simulating the numbers of infected individuals in each group for subsequent weeks according to draws from the Poisson distributions given in Eq 5.46-5.47. For the first eight time steps, R is fixed at 1.5. After the eighth time step, $R$ is decreased to 0.8 for the rest of the epidemic. This yields outbreaks of similar size as the ones simulated using the individual-based SIR model presented in §5.2.3 (compare Figs 5.7 and 5.9). Only outbreaks in which at least 400 people become infected are recorded. Five simulations from this two-age-class Poisson-type model are depicted in Fig 5.9.

To check consistency with the individual-based SIR model, 100 ensembles of 800 epidemics each were simulated for the rate matrices given in Eq 5.44 and Eq 5.45, with $z$ and $k$ ranging from 0 to 1 in steps of 0.1. The mean STE estimates from these simulations are depicted in Fig 5.10 (compare with Fig 5.8). The same trends hold: the $I \to J$ and $J \to I$ STE estimates overlap and increase as $z$ increases from 0 to 1, and the $I \to J$ STE quickly dominates over the $J \to I$ STE as $k$ increases from 0 to 1.

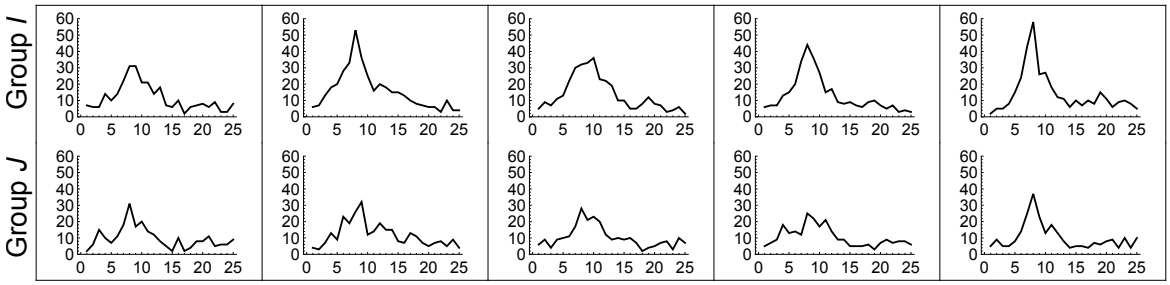



Fig. 5.9 Five simulations from the two-age-class Poisson-type simulation algorithm, Eqs 5.46-5.47. Each column depicts the output for a single epidemic simulation, separated into case counts from group $I$ (top) and group $J$ (bottom). The reproduction number $R$ is 1.5 for the first eight weeks of the outbreak, and then drops to 0.8 for the rest of the epidemic. The transmission rates are specified by the relative rate matrix 5.45 with $k = 1$, for which age group $I$ (upper row) has quadruple the within-group transmission rate as group $J$, and for which the $I \to J$ transmission rate is twice the $J \to I$ transmission rate.

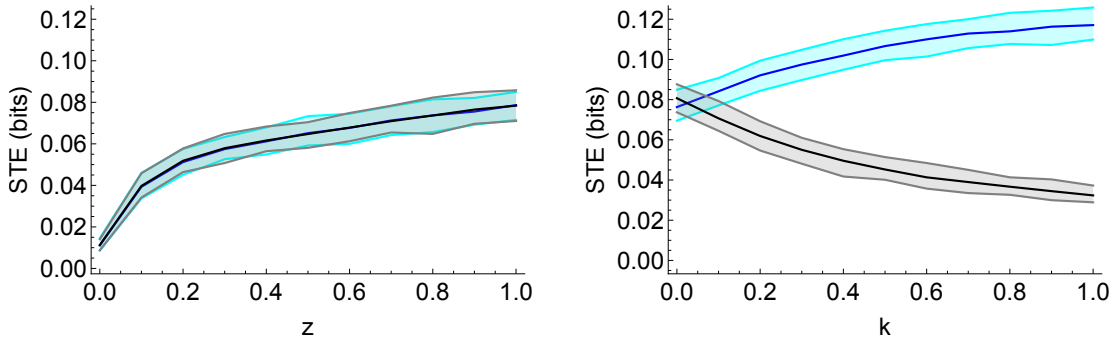



Fig. 5.10 Mean $T^S_{I\to J}$ (solid blue line) and $T^S_{J\to I}$ (solid black line) values for a range of within- and between-group infection rates, as estimated from 100 ensembles of 800 epidemics for each value of $z$ and $k$ between 0 and 1 in steps of 0.1, simulated using the two-age-class Poisson-type model (Eq 5.46-5.47). The shaded bands provide the approximate 95% confidence intervals for the STE estimates. The left-hand plot depicts the STE for the relative rate matrix given in Eq 5.44, which transitions smoothly from completely decoupled age groups to mean-field dynamics where within- and between-group transmission rates are equal. The right-hand plot depicts the STE for the relative rate matrix given in Eq 5.45, which transitions smoothly from mean-field dynamics to strong forcing from group $I$. The trends in these plots match those in Fig 5.8, showing agreement between STE estimates when using either the individual-based SIR model or the more computationally efficient Poisson-type model.

### 5.2.5 Simulations on a four-age-class Poisson model with variable reporting rates

Having demonstrated agreement between STE estimates based on an individual-based SIR model and a more efficient Poisson-type model, we now consider larger epidemics with more age classes. We also consider the relationship between reporting rates and STE.

First, define the $n$-age-class Poisson-type epidemic model as

$$y_{i,t} \sim Poisson(\lambda_{i1,t} y_{1,t-1} + \lambda_{i2,t} y_{2,t-1} + \cdots + \lambda_{in,t} y_{n,t-1}) \tag{5.50}$$

where $y_{i,t}$ is the number of infected cases in age group $i$ at (discrete) time $t$, and $\lambda_{ij,t}$ specifies the infection rate from group $j$ to group $i$ at time $t$. As before, the time steps are assumed to be equal to the generation interval of the disease, so that the next-generation matrix is

$$NGM_t = \boldsymbol{\lambda}_t = \begin{bmatrix} \lambda_{11,t} & \lambda_{12,t} & \dots & \lambda_{1n,t} \\ \lambda_{21,t} & \lambda_{22,t} & \dots & \lambda_{2n,t} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{n1,t} & \lambda_{n2,t} & \dots & \lambda_{nn,t} \end{bmatrix}. \tag{5.51}$$

For a given matrix $\boldsymbol{r}$ whose $i, j^{th}$ entry specifies the relative rate of infection from age group $j$ to group $i$, and a given reproduction number $R$, the relationship in Eq 5.49 still holds.

For the scenario with four age classes ($n = 4$), epidemics are initiated by placing a single infected individual into any of the four age groups with equal probability. The case counts in subsequent weeks are drawn from the Poisson distributions specified by Eq 5.50. To allow for larger epidemics than considered in the two-age-class scenario, the reproduction number $R$ is fixed at 1.5 for fourteen, rather than eight, time steps. After the fourteenth time step, $R$ reduces to 0.8 as before. Fig 5.11 (grey lines) depicts five simulations from this model.

In a real outbreak, only a fraction of cases are reported and recorded. To simulate this, a vector of reporting rates $\boldsymbol{c} = (c_1, \dots, c_n)$ is introduced, where $c_i$ is the reporting rate for age class $i$. The number of reported case counts for age group $i$ in time step $t$ follows

$$y^{obs}_{i,t} \sim Binomial(y_{i,t}, c_i). \tag{5.52}$$

Fig 5.11 depicts the original (grey) and observed (black) time series for five outbreaks on four age classes using this model, with $c_i = 0.5$ for all age classes.

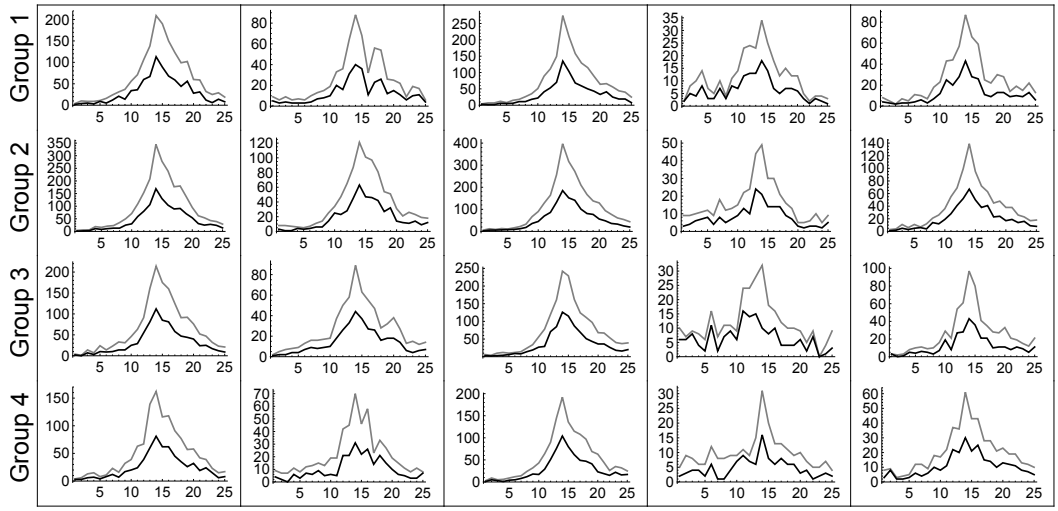



Fig. 5.11 Five simulations from the 4-age-class Poisson-type model (Eq 5.46-5.47) with relative rate matrix $r$ given in Eq 5.53. Columns correspond to a single epidemic simulation, separated into case counts for the four age groups. Grey lines depict the full case counts simulated from the Poisson model, and black lines depict the case counts obtained by using a Binomial observation model (Eq 5.52) with a reporting rate of $c_i = 0.5$ for all age groups. Note the different vertical scales.

To explore how reporting rates influence the STE, two scenarios are considered. First, we consider how varying the reporting rate equally across all age groups affects the identifiability of true differences in transmission rate. Second, we consider how different reporting rates between age groups affects the STE, when the true transmission rates are all equal.

For the first scenario, consider a relative rate matrix of form

$$\boldsymbol{r} = \begin{bmatrix} 1 & 2 & 1 & 1 \\ 1 & 4 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \tag{5.53}$$

so that the within-group transmission rate for group 2 is four times the baseline transmission rate, and the group 2 to group 1 and group 2 to group 3 transmission rates are double the baseline transmission rate. If the four age groups correspond to infants, children, adults, and elderly, then this would match a scenario with high transmission among children and intermediate transmission between children and infants and between children and adults.

To study how varying reporting rates uniformly across all age classes affects the ability of the STE to identify children as the dominant transmitters of disease, 100 ensembles of

800 epidemics each were simulated for reporting rates $c_i$ between 0.1 and 1 in steps of 0.1, with equal reporting rates across all age groups. Fig 5.12 depicts the STE from each age group to every other age group as a function of $c_i$. Even for reporting rates as low as 0.1, the STE values from group 2 are higher than those from any other group. As the reporting rates increase, the differences become more pronounced. The left-hand plot in Fig 5.13 depicts the mean pairwise STE values with $c_i = 0.5$ in all age groups, which is essentially a vertical slice from the plots in Fig 5.12 at $c_i = 0.5$. Matrix 5.56 provides the mean STE values used to make the left-hand plot in Fig 5.13, with their 95% confidence intervals. There is significantly elevated transmission from group 2 to all other groups.

For the second scenario, with mean-field transmission dynamics and reporting rates that differ by age, the relative rate matrix is

$$\boldsymbol{r} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \tag{5.54}$$

so that all within- and between-group rates of transmission are equal. Roughly following the age-structured reporting rates reported by Biggerstaff *et al.* (2012) [22], the reporting rate vector is fixed at

$$\boldsymbol{c} = (0.4, 0.6, 0.4, 0.4) \tag{5.55}$$

corresponding to elevated reporting rates in group 2 (children). To be more consistent with the reporting rates in Biggerstaff *et al.* (2012) [22], the reporting rate for group 1 (infants) should also be approximately 0.6, but for illustrative purposes we for now consider elevated reporting in just one age group. A more realistic scenario, with elevated reporting rates for both infants and children, is considered later in the 12-age-class simulations.

To estimate how age-variable reporting rates affect the STE, 100 ensembles of 800 epidemics each were simulated using the rate matrix Eq 5.54 and the reporting rates Eq 5.55. The pairwise STE between each age group was calculated for each ensemble. The right-hand plot in Fig 5.13 depicts the pairwise mean STE values, and matrix 5.57 provides the mean values and 95% confidence intervals. A 60% reporting rate in group 2 and a 40% reporting rate in all other age groups yields significantly higher STE estimates from group 2 than from any other age group. However, the first scenario, with elevated transmission from group 2 but constant 50% reporting across all age groups, yields even higher STE estimates from

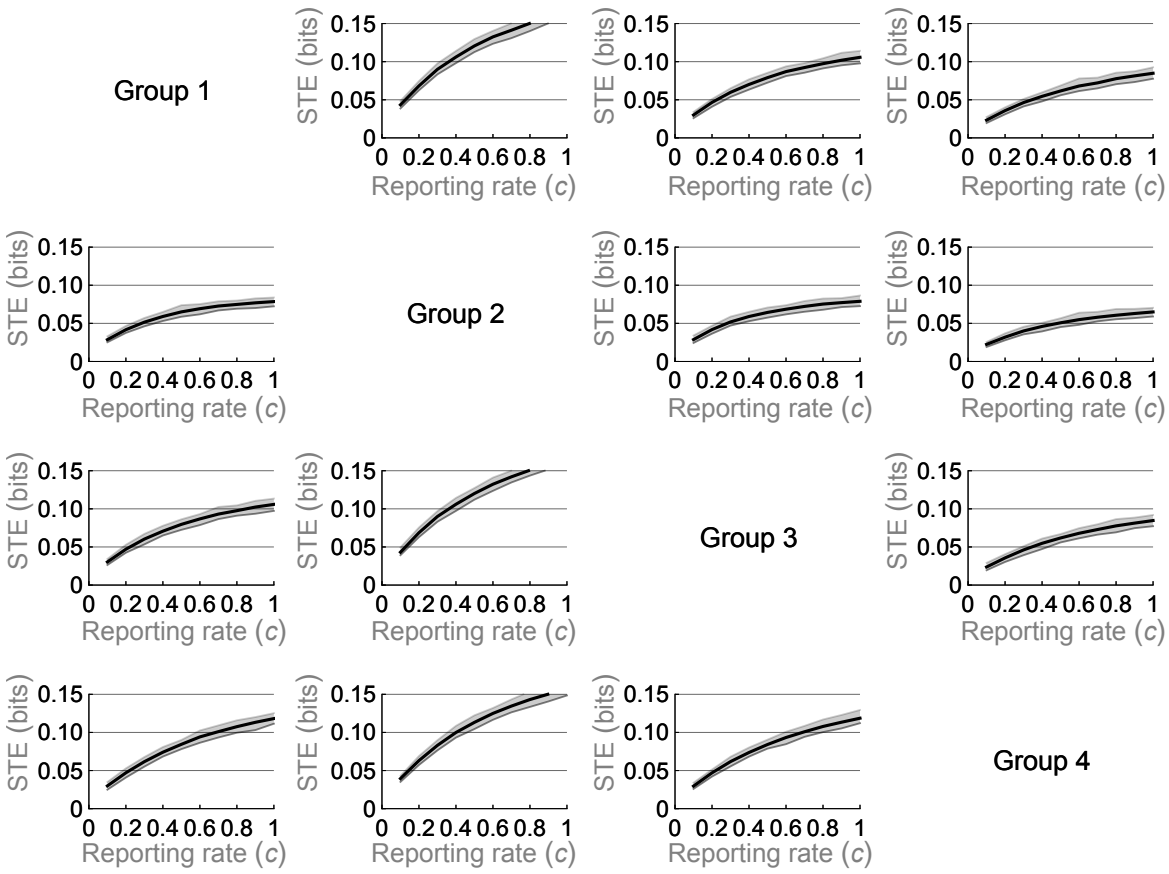



Fig. 5.12 Mean pairwise STE values (solid lines) with 95% confidence intervals (shaded bands) for 100 ensembles of 800 simulated epidemics, for reporting rates $\boldsymbol{c}$ (horizontal axis) between 0 and 1 in steps of 0.1. The vertical axis corresponds to the estimated STE, in units of bits. Reporting rates for the simulations are uniform across all age groups. The relative rate matrix that specifies within- and between-group transmission rates is given by Eq 5.53. The plot in row $i$ and column $j$ depicts the STE from group $j$ to group $i$. Even with reporting rates near 0.1, group 2 is correctly identified as the primary driver of transmission. As reporting rates increase, the dominant transmission from group 2 becomes clearer. The estimated STE increases with reporting rate for all age groups, but more quickly for group 2 than for the other age groups. According to Biggerstaff *et al.* (2012), true reporting rates for ILI in the US during the 2009 pandemic were between 0.4 and 0.6.

group 2. For comparison, the mean pairwise STE values for the two scenarios are depicted side-by-side in Fig 5.13.

$$\begin{bmatrix} 0 & 12.0(11.1,12.9) & 7.9(7.1,8.5) & 6.2(5.5,6.8) \\ 6.5(5.9,7.2) & 0 & 6.5(5.9,7.2) & 5.1(4.5,5.7) \\ 8.0(7.2,8.7) & 12.0(11.2,12.9) & 0 & 6.2(5.5,6.8) \\ 8.5(7.7,9.2) & 11.4(10.5,12.1) & 8.5(7.8,9.1) & 0 \end{bmatrix} \times 10^{-2} \tag{5.56}$$

Mean STE values (95% CI) for simulated epidemics with strong transmissive forcing from group 2 (relative rate matrix given by Eq 5.53) and with 50% reporting for all age groups. The $i, j^{th}$ entry gives the STE from group $j$ to group $i$. The STE values from group 2 to all other age groups are significantly higher than the STE values from any other age group. The mean values are also depicted in the left-hand plot of Fig 5.13.

$$\begin{bmatrix} 0 & 8.1(7.2,8.8) & 6.2(5.7,6.8) & 6.2(5.7,6.7) \\ 5.9(5.5,6.5) & 0 & 5.9(5.3,6.5) & 6.0(5.3,6.6) \\ 6.2(5.6,6.9) & 8.1(7.4,8.9) & 0 & 6.2(5.6,6.7) \\ 6.2(5.5,6.7) & 8.1(7.4,8.7) & 6.2(5.7,6.7) & 0 \end{bmatrix} \times 10^{-2} \tag{5.57}$$

Mean STE values (95% CI) for simulated epidemics with constant within- and between-group transmission rates (relative rate matrix given by Eq 5.54), and with 60% reporting rate from group 2 and 40% reporting rate for all other age groups. The $i, j^{th}$ entry gives the STE from group $j$ to group $i$. The STE values from group 2 to all other age groups are significantly higher than the STE values from any other age group, though lower than the values in matrix 5.56 obtained from the explicitly forced simulations. The mean values are also depicted in the right-hand plot of Fig 5.13.

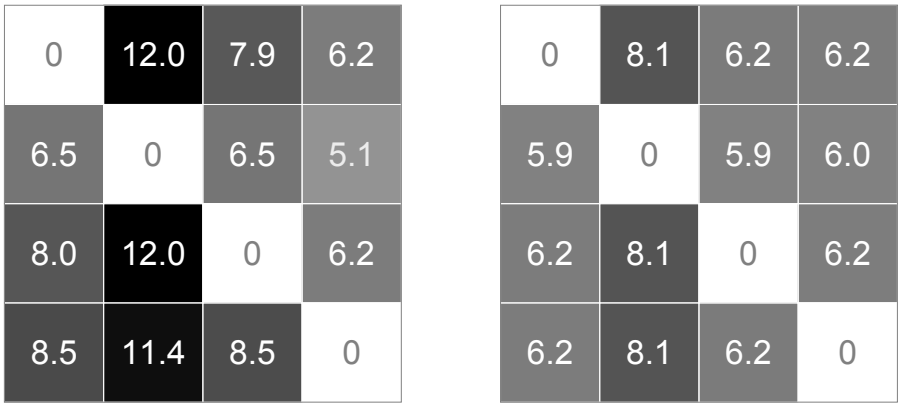



Fig. 5.13 Mean pairwise STE values in bits, multiplied by a factor of $10^2$, for 100 ensembles of 800 epidemics using relative rate matrix Eq 5.53 and 50% reporting rate for all age groups (left), and using relative rate matrix Eq 5.54 with 60% reporting rate in group 2 and 40% reporting rate in all other age groups (right). A box in row $i$ and column $j$ corresponds to the mean STE from group $j$ to group $i$, with darker shades corresponding to higher STE. In both scenarios, the STE from group 2 to all other age groups is elevated. The elevation is higher for the scenario with explicitly higher transmission rates from group 2 (left) than for the scenario with equal transmission rates and elevated reporting in group 2 (right). Mean values and confidence intervals are given in matrices 5.56 and 5.57.

### 5.2.6 Simulations on a twelve-age-class Poisson model

For the final set of simulations, we consider ensembles of 834 outbreaks on twelve age classes (<2, 2-4, 5-9, 10-14, 15-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80+). This matches the resolution of the IMS-ILI data. First, 10 ensembles of 834 outbreaks each are

simulated using the Poisson-type epidemic model, Eq 5.50, with relative rate matrix

$$\boldsymbol{r} = \begin{bmatrix} 1 & 1 & 2 & 2 & 2 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 2 & 2 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 4 & 4 & 4 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 4 & 4 & 4 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 4 & 4 & 4 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 2 & 2 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 2 & 2 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 2 & 2 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 2 & 2 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \tag{5.58}$$

which corresponds to strong transmission between children (5-19 years), moderate transmission from children to infants (0-4 years) and from children to adults (20-59 years), and baseline transmission within and between all other age groups. Reporting rates are held constant at 50% for all age groups, and reported cases are simulated using the Binomial reporting rate model (Eq 5.52). The mean pairwise STEs between all age groups are depicted in the left-hand plot in Fig 5.14.

Also, 10 ensembles of 834 epidemics each were simulated using a $12 \times 12$ mean-field relative rate matrix $\boldsymbol{r}$ with all entries equal to 1, and with 60% reporting rate for age groups from 0 to 19 years old, and 40% reporting rates for age groups 20 years old and older. This aligns with the ILI reporting rates reported by Biggerstaff *et al.* (2012) [22]. The mean pairwise STEs between all age groups under this scenario are depicted in the right-hand plot in Fig 5.14. As before, differences in reporting rates are enough to yield detectable differences in STE, but the explicitly forced model with uniform reporting rates yields more dramatic differences in STE than the mean-field model with age-varying reporting rates. In Fig 5.14, there is little row-wise variation in STE. This is due in part to the effects discussed in §5.2.2, in which it was found that age groups with more infected members tend to transfer more information. So, even though the child-to-elderly transmission rate used to produce the left-hand plot is no different than the elderly-to-elderly transmission rate, the elevated numbers of children who become infected due to the high within-child transmission rate yields a high STE from children to the elderly. In the right-hand plot, elevated reporting rates

for infants and children makes the case counts in those age groups appear artificially high, which manifests in higher apparent STE from those age groups to all other age groups.

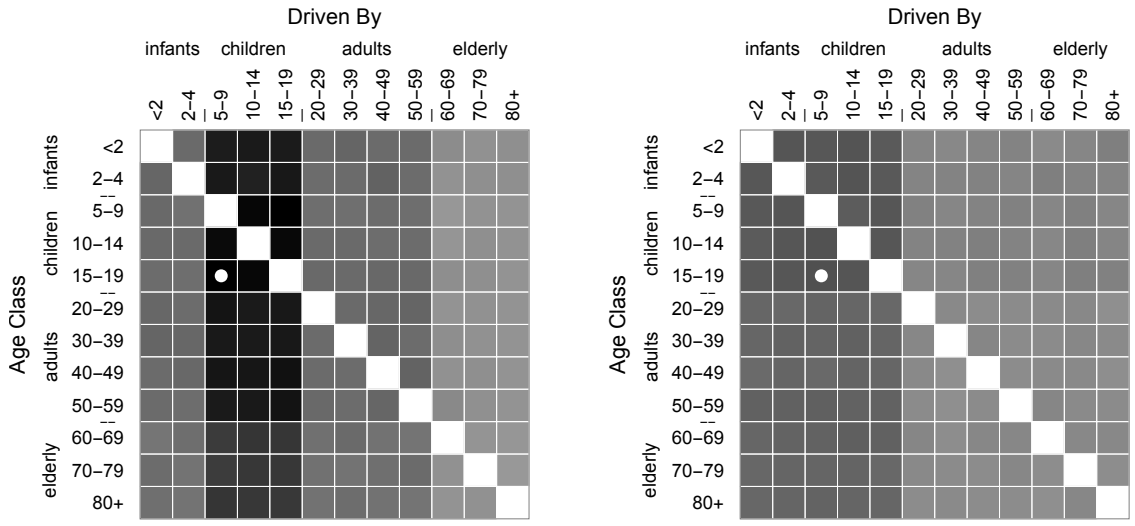



Fig. 5.14 Pairwise mean STE between 12 age groups from 10 ensembles of 834 simulated epidemics using a Poisson-type outbreak model (Eq 5.50). A box in row $i$ and column $j$ corresponds to the STE from group $j$ to group $i$, where darker shades corresponds to higher STE. For the left-hand plot, epidemics are simulated with high transmission among children and moderate transmission between children and infants and between children and adults (see Eq 5.58). Reporting rates are fixed at 50% across all age groups. For the right-hand plot, epidemics are simulated with uniform transmission rates within and between all age groups, but with 60% reporting rates for infants and children and 40% reporting rates for adults and elderly. Higher STE is associated with each of elevated transmission and elevated reporting rates. In the left-hand plot, the maximum pairwise STE is 0.052 bits, from 5-9 year-olds to 15-19 year-olds. In the right-hand plot, the maximum pairwise STE is 0.036 bits, also from 5-9 year-olds to 15-19 year-olds. The cells corresponding to these maximum values are marked with a dot.

## 5.3 STE to identify dominant age groups in transmission of the 2009 A/H1N1pdm influenza pandemic in the United States

Having established a relationship between relative transmission rates, reporting rates, and STE on simulated outbreaks, we now calculate the STE between age groups during the 2009 A/H1N1pdm influenza pandemic in the United States using the IMS-ILI data. First, the STE is computed between age groups within the same ZIP. Then, the STE is estimated between age groups in ZIPs between which infection likely spread, as identified by the geographic transmission model in Chapter 3. In both cases, there is elevated STE from 5-19 year-olds to most other age groups, suggesting that school-aged children contributed most to both within-city transmission and between-city transmission of the outbreak.

### 5.3.1 Age-group differences in within-city transmission

Weekly ILI data are available for 834 3-digit ZIP codes across the United States (see Chapter 2). For each of these ZIP codes, the ILI data are further separated into 12 age groups. The ILI ratios (ILI counts divided by number of physician visits in each age group) are symbolised for the 12 age groups in the 25 weeks between 12 July 2009 and 27 December 2009, using a symbol length of $m = 3$. The age-aggregated ILI ratios for each ZIP are also symbolised in the same way. It is assumed that each outbreak represents an instance of an underlying epidemic process that is consistent for all ZIPs; that is, we assume that age-related transmission and reporting rates remain constant across ZIPs. To the best of my knowledge, this should be a reasonable assumption, since the overall demographics of the US are fairly consistent throughout the country. This way, the STE transition probabilities may be estimated as simple relative frequencies of symbol combinations in a particular age group across all locations, as they were in the outbreak simulations.

Fig 5.15 depicts the within-ZIP STE from each age group to the ZIP's age-aggregated time series. There is a clear peak in STE for 5-19 year-olds, with the highest STE from 10-14 year-olds. There is also a smaller peak in STE from 70-79 year-olds. It is unclear whether this represents a true signal or is just a spurious effect, since the time series for the elderly age groups are noisy (see Fig 5.23).

Fig 5.16 depicts the pairwise within-ZIP STE between all age groups. The elevated STE from 5-19 year-olds to all ages from infants through adults, depicted as the dark shades in columns 3-5, provides evidence that there was systematically elevated transmission from

school-aged children to these other age groups within cities. The adult-adult STE is also moderately elevated, suggesting that adults may have played some role in transmitting the outbreak amongst themselves, though this could also be explained by elevated transmission from children alone. Compare, for example, to the left-hand plot in Fig 5.14. In that simulation, only transmission from children is elevated, but it causes a moderate elevation in the STE from adults and infants.

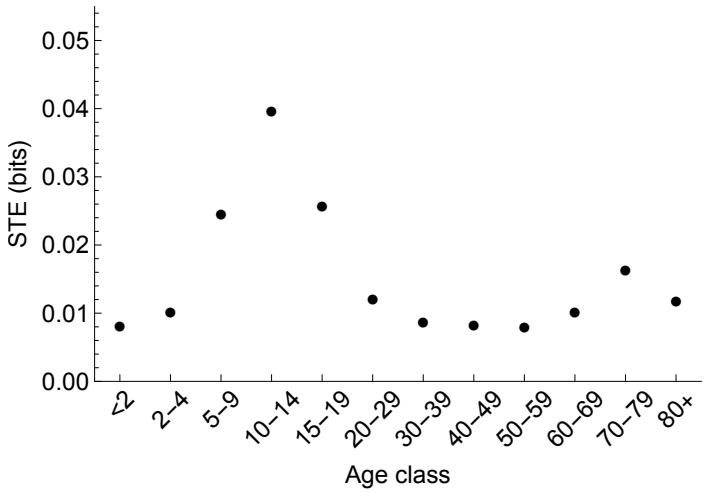



Fig. 5.15 Within-ZIP STE from each age group to the age-aggregated symbolised ILI time series. School-aged children (ages 5-19 years) have the highest transfer of information to the age-total time series.

### 5.3.2 Age-group differences in geographic transmission

In addition to examining within-city relationships between the age groups, it is possible to measure the extent to which the age groups in one city may have contributed to transmission in a different city. Here, the STE is calculated between age groups in pairs of cities between which infection likely spread. The $i, j^{th}$ entry of matrix $\boldsymbol{\tau}$, where $\boldsymbol{\tau}$ is defined in §4.2.2, gives the probability that the outbreak in ZIP $i$ was triggered by transmission from ZIP $j$, according to the transmission model developed in Chapter 3. For each row $i$ of matrix $\boldsymbol{\tau}$, the index of the largest entry coincides with the ZIP that most likely infected ZIP $i$. The ZIPs for which $\max(\tau_{i1}, \tau_{i2}, \ldots, \tau_{in}) < \sigma_i$ (that is, for which external seeding is the most likely source of infection) are excluded.

First, the STE is calculated from each age band in ZIP $i$'s most likely infector to the age-aggregated symbolised ILI time series for ZIP $i$. These STE values are depicted in Fig

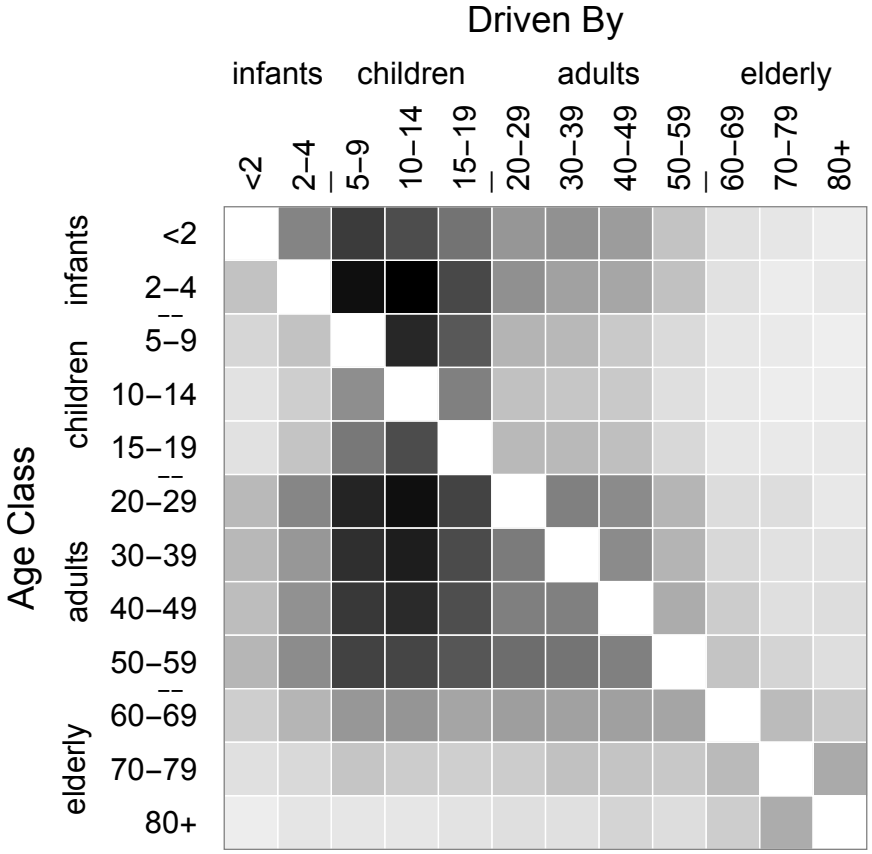



Fig. 5.16 Within-ZIP STE from the age classes along the top to the age classes along the left, estimated from the IMS-ILI data. Darker boxes indicate higher STE. The highest STE value is from 10-14 year-olds to 2-4 year-olds, at 0.084 bits.

5.17. Again, the STE from school-aged children is elevated, with 5-9 year-olds providing the most information about the age-aggregated ILI in the nearby infected ZIP. The STE plateaus for 20-49 year-olds, before decreasing again for 50+ year-olds. This may reflect moderate between-city transmission from working adults who commute between ZIPs.

Next, the pairwise STE is computed from each age group in each ZIP $i$'s maximum-likelihood infector to each age group in ZIP $i$. These pairwise STEs are depicted in Fig 5.18. The overall picture is similar to the one in Fig 5.16, with elevated STE from children to infants through adults, and moderate STE from adults to other adults.

Note that the transmission model developed in Chapter 3 and the between-ZIP pairwise STE considered here each characterise a different type of transmission. The transmission model from Chapter 3 describes epidemic 'sparks' that travel between ZIPs, while the pairwise STE is better interpreted as a measurement of which age groups sustain transmission over the course of the outbreak. The maximum-likelihood infector ZIP, as identified by the mechanistic transmission model, may not be the ZIP most responsible for sustaining transmission in the recipient ZIP, and conversely, the age groups that drive transmission according to STE are not necessarily the ones that most likely sparked transmission in a nearby ZIP. However, the maximum-likelihood infector pairs considered in this section provide a proxy for locations that are likely epidemiologically coupled in some way. A discussion of the differing conclusions that may be drawn from the mechanistic transmission model versus from the STE may be found in this chapter's Discussion.

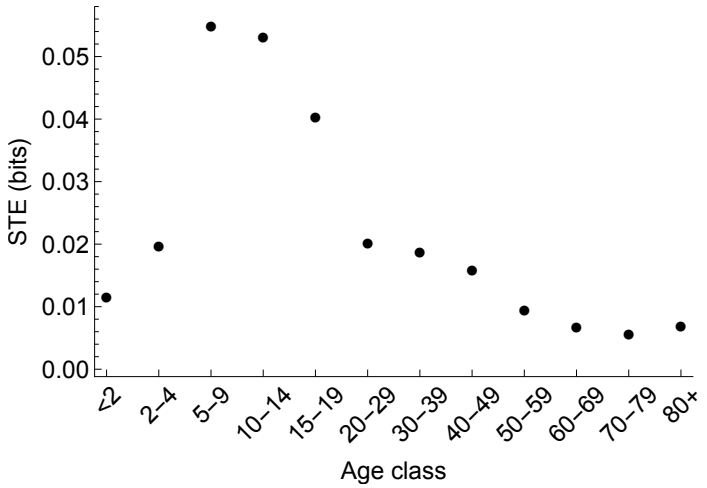



Fig. 5.17 STE from each age group to the age-aggregated time series in maximum-likelihood donor/recipient-of-infection ZIP pairs. As for within-ZIP transmission, school-aged children (ages 5-19) have the highest STE to the age-total time series.

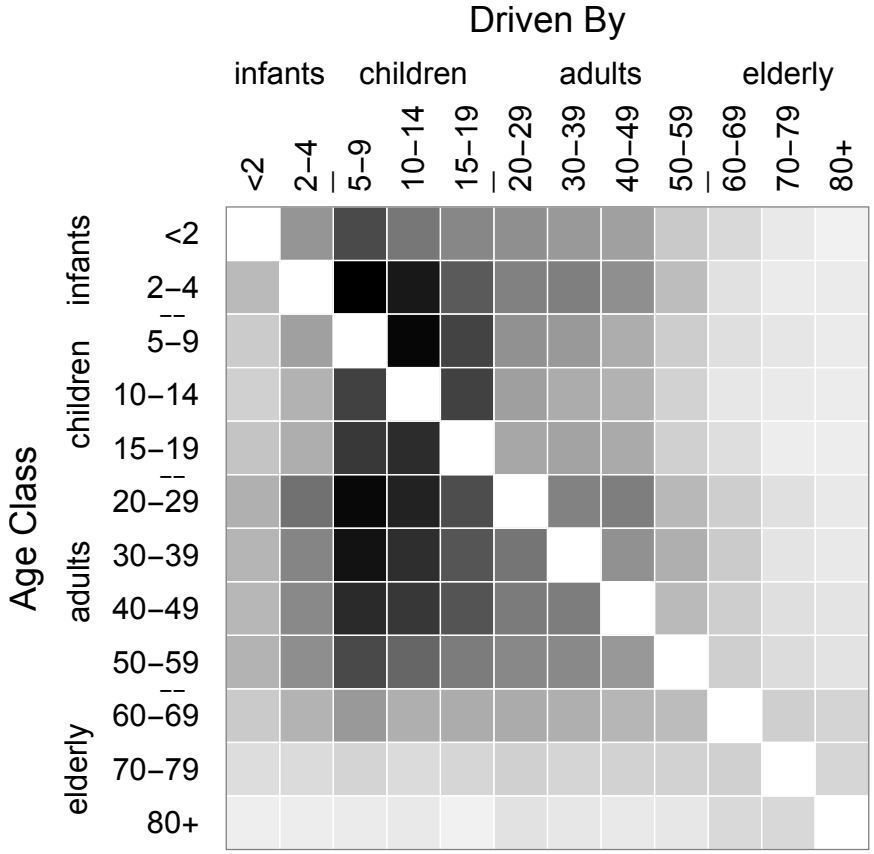



Fig. 5.18 STE from the age classes listed along the top to the age classes listed along the left, for maximum-likelihood donor/recipient-of-infection ZIP pairs. Darker boxes indicate higher transfer of information. The STE is especially elevated in the column beneath school-aged children, which may indicate a high degree of transmission between school-aged children and infants, other school-aged children, and adults in nearby cities. The highest transfer of information is from 5-9 year-olds to 2-4 year-olds, at 0.060 bits.

A quick way to validate these methods is to calculate the STE between ZIPs that are very far apart. These STE values should be lower, since there should be very little epidemiological coupling, and thus very little transfer of information, between the ZIPs. To check this, for each ZIP, a partnering ZIP is randomly drawn from the set of ZIPs at least 1000 km away. The transfer entropies between the age-specific time series in the original ZIPs and the age-aggregated time series in the distant ZIPs are calculated. These are depicted in Fig 5.19. There is still some evidence of elevated STE from school-aged children, possibly due to synchrony between outbreaks in distant locations [239] or elevated reporting rates in these age groups. The STE values are lower than in Fig 5.17, however, as expected.

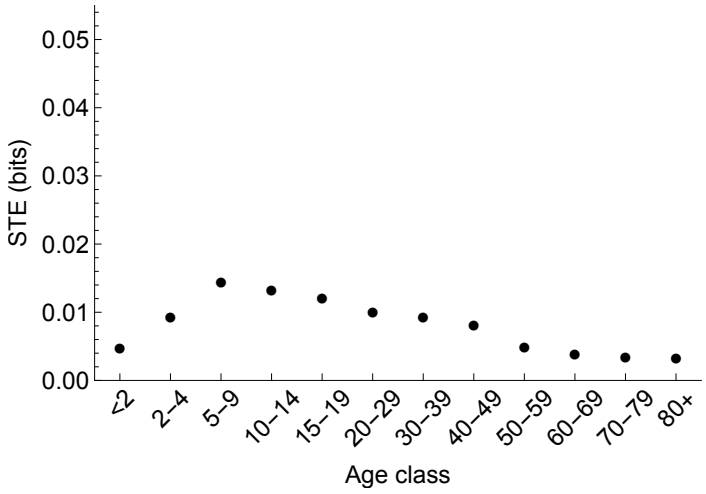



Fig. 5.19 STE from each age group to the age-aggregated time series in a randomly-selected ZIP at least 1000 km away. School-aged children (ages 5-19) still transfer the most information to the age-total time series, but the magnitude of the STE is much smaller than within ZIPs (Fig 5.15) or between maximum-likelihood donor/recipient-of-infection pairs (Fig 5.17). This provides further evidence that the STE captures epidemiologically relevant coupling, which is expected to be low at long distances.

Taken together, these findings suggest that school-aged children may have been the primary drivers of within-city transmission, and may also have contributed disproportionately to between-city transmission, of the 2009 A/H1N1pdm influenza pandemic in the United States.

### 5.3.3 Robustness to variation in reporting rates

It remains to be seen whether the observed differences in STE in the ILI data could be fully explained by differences in reporting rates between age groups. To check this, pre-reporting

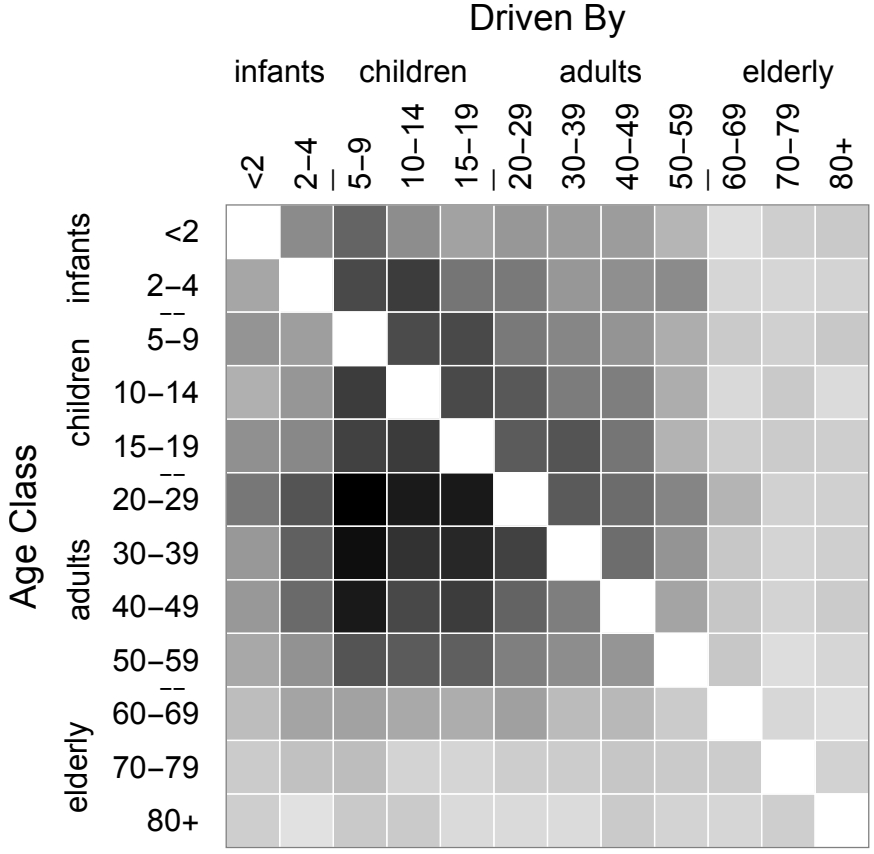



Fig. 5.20 STE from the age classes listed along the top to the age classes listed along the left, between randomly-selected pairs of ZIPs at least 1000 km apart. Darker boxes indicate higher transfer of information. The STE from children to other age groups is elevated, but not as markedly as it is in the within-ZIP and the maximum-likelihood infector pair scenarios (Figs 5.16 and 5.18). The maximum pairwise STE in this case is also lower than for either of the two previous scenarios, at 0.015 bits, from 5-9 year-olds to 20-29 year-olds.

counts of influenza-like illness can be roughly inferred from the observed ILI counts, and the STE may be re-calculated. If the STE trends persist, then they are likely robust to reporting uncertainty. To accomplish this, it is necessary to first demonstrate that the STE differences observed in the ILI data still hold when estimated using ILI counts rather than the ILI ratios. Figs 5.21 and 5.22 depict the STE from each age group to the age-aggregated time series, and the pairwise STE between each age group, calculated using the within-ZIP ILI counts rather than the ILI ratios. The same patterns hold, with school-aged children emerging as the primary contributors of information to most other age groups' time series.

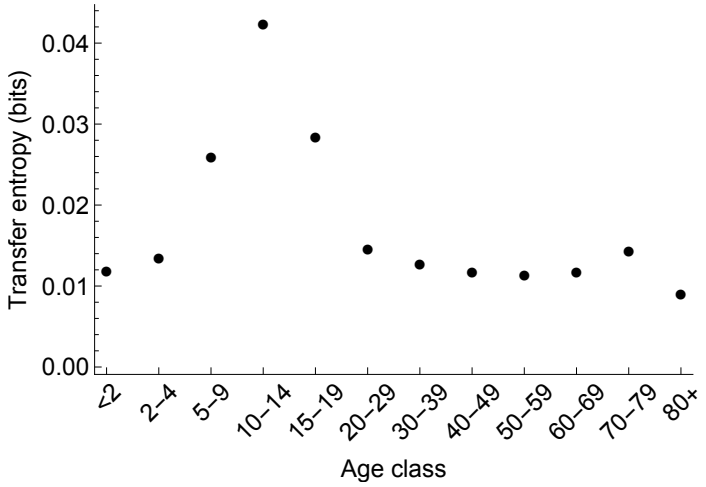



Fig. 5.21 Within-ZIP STE from each age band to the age-aggregated time series, using ILI counts rather than ILI ratios. The values are almost identical to the ones obtained using the ILI ratios (see Fig 5.15).

Next, it is possible to reconstruct a distribution of possible true case counts from the observed case counts, given a particular reporting rate. First, for a given location, as in §5.2.5, we assume that the observed ILI count $y_{i,t}^{obs}$ in age band $i$ in week $t$ represents a binomial sample from $y_{i,t}$ total ILI cases in that week. That is,

$$y_{i,t} \sim Binomial(y_{i,t}, c_i) \tag{5.59}$$

where $c_i$ is the reporting rate for age band $i$. Given the number of observed cases $y_{i,t}^{obs}$ and reporting rate $c_i$, the normalised likelihood for the true number of cases in week $t$ is

$$L(y_{i,t}; y_{i,t}^{obs}, c_i) = \binom{y_{i,t}}{y_{i,t}^{obs}} c_i^{y_{i,t}^{obs}+1} (1-c_i)^{y_{i,t}-y_{i,t}^{obs}} \tag{5.60}$$

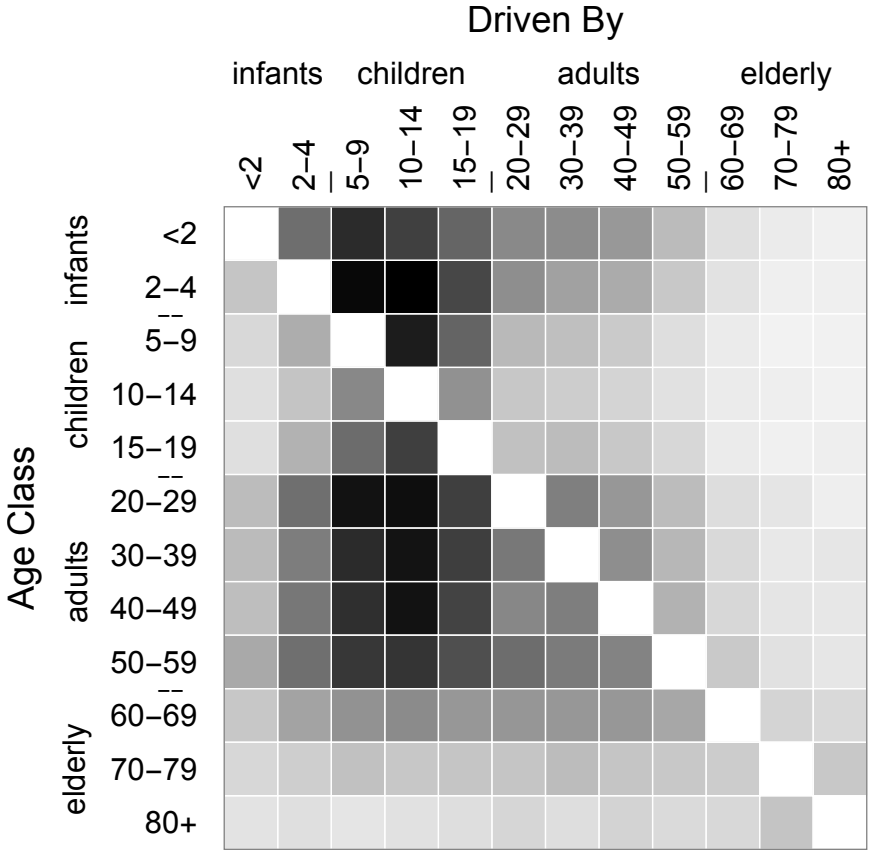



Fig. 5.22 Within-ZIP pairwise STEs between age bands, using ILI counts rather than ILI ratios. Darker boxes correspond to higher values. The values are almost identical to the ones obtained using ILI ratios (see Fig 5.16).

which satisfies

$$\sum_{y_{i,t}=0}^{\infty} L(y_{i,t}; y_{i,t}^{obs}, c_i) = 1. \tag{5.61}$$

The normalised likelihood for $y_{i,t}$, Eq 5.60, may be interpreted as a probability distribution, from which possible true numbers of cases in week $t$ can be drawn.

To test the robustness of the STE results to reporting uncertainty, possible true case counts are drawn using Eq 5.60 for each week and each age group in each location, and the pairwise STE is re-calculated. This is repeated 100 times. The reporting rate $c$ is assumed to be 60% for age groups between 0 and 19, and 40% for age groups 20 and above, following the ILI reporting rate estimates from the 2009 A/H1N1pdm pandemic in the US reported by Biggerstaff *et al.* (2012) [22]. Fig 5.23 depicts four reconstructed full case counts (grey) from four age-group time series (rows) in four different locations (columns). Fig 5.24 depicts the mean pairwise within-ZIP STE between age groups using the reconstructed case counts. The pattern of elevated STE from children to infants through adults persists.

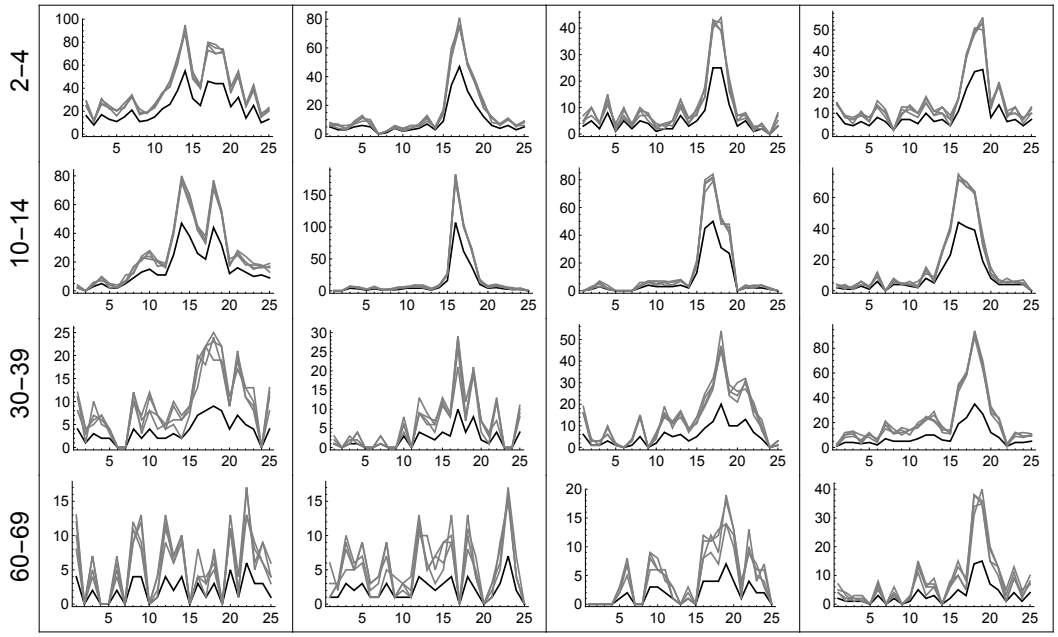



Fig. 5.23 True counts (black) with four binomial reconstructions (grey) of the true numbers of counts. It is assumed that children (under 20 years) have 60% reporting rates and adults (20+) have 40% reporting rates, following [22].

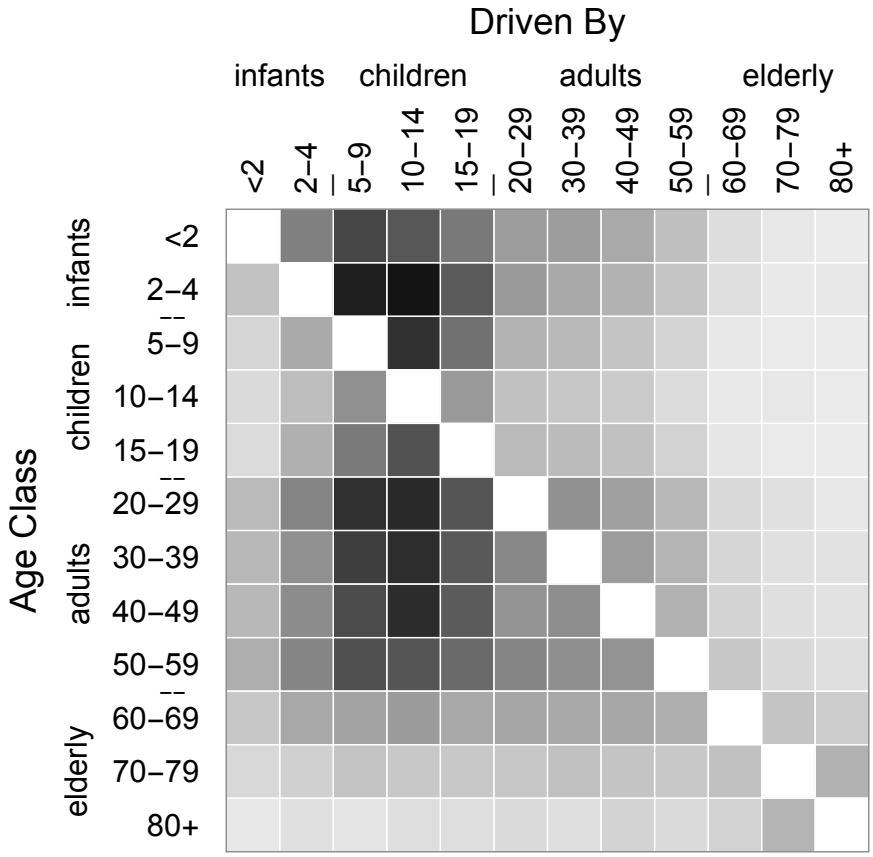



Fig. 5.24 Mean pairwise within-ZIP STE values estimated from 100 reconstructed ILI case-count time series, assuming a 60% reporting rate in infants and children and a 40% reporting rate in adults and elderly. Darker boxes correspond to higher values. School-aged children still transfer the most information to the other age groups, consistent with all of the previous results.

### 5.3.4 Maximum-information symbols

As a brief aside, it is possible to identify which symbols carry the most information in a given stochastic process. For example, if one time series has the symbol $A$ at time $t$, it is possible to identify whether this provides more information about which symbol another time series will display at time $t+1$ than if the first time series had a different symbol at time $t$. This is done by separating the STE sum in Eq 5.9 into parts. The amount of information that an $A$ in process $J$ at time $t-1$ provides about whether the symbol in process $I$ is a $B$ at time $t$ is

$$T_{J\to I}^{S,AB} = \sum_{\hat{i}_t} p(\hat{i}_{t+1}=B,\hat{i}_t,\hat{j}_t=A)\log\Big(\frac{p(\hat{i}_{t+1}=B|\hat{i}_t,\hat{j}_t=A)}{p(\hat{i}_{t+1}=B|\hat{i}_t)}\Big). \tag{5.62}$$

The amount of information that each symbol contributes to every other symbol can thus be calculated. Using the IMS-ILI data, the symbol-specific STE is calculated from the three school-aged-children age groups (5-19, 10-14, and 15-19 years) to the age-aggregated time series in their same ZIP. The values are depicted in Fig 5.25. A symbol $A$ (steadily increasing) in the child time series provides a lot of information that there will be an $A$ in the overall time series at the next time step, and similarly for the symbol $F$ (steadily decreasing). Symbols that end on an increase/decrease tend to predict symbols that also end on an increase/decrease, though there are exceptions. Overall, this suggests that the most informative symbols are strong increases and strong decreases (symbols $A$ and $F$), which matches with the intuition that repeated increases and decreases in ILI tend to correspond to takeoffs of infection or declines in infection – epidemiologically relevant events that should affect the dynamics in other time series and thus transmit information to them. Other symbols are more likely to arise through stochastic noise, and so there is lower transfer of information between them.

## 5.4 Fitting a mechanistic geographic transmission model with age class data

As an alternative to the methods developed above, the mechanistic geographic transmission model developed in Chapter 3 can be adjusted to include information from the age-specific time series. Re-fitting the model using these more refined ILI data can shed light on which age groups best predict the geographic transmission of influenza.

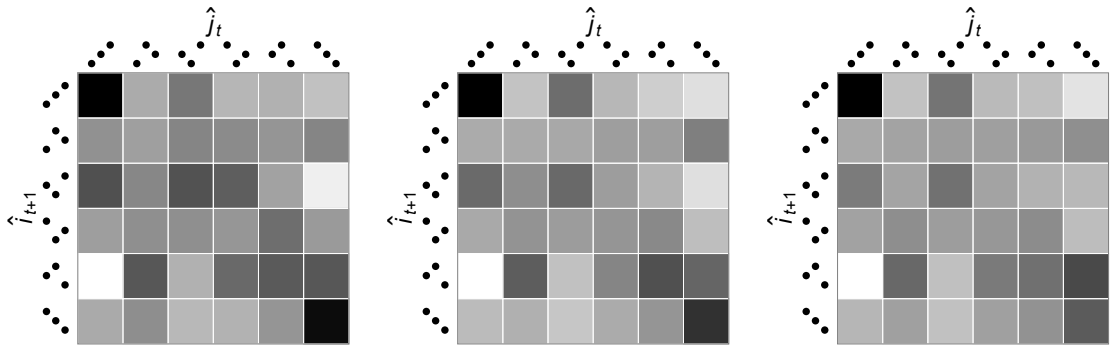



Fig. 5.25 Information that each symbol $\hat{j}_t$ in the child time series carries about the next symbol $\hat{i}_{t+1}$ the age-aggregated time series, for each group of school-aged children (ages 5-9, 10-14, and 15-19, from left to right). The more information the symbol in the child time series carries about the symbol in the overall time series, the darker the corresponding square. The steadily-increasing and steadily-decreasing symbols ($A$ and $F$) carry the most information, and tend to predict the same symbol in the age-total time series.

### 5.4.1 Adjusting the data and the model

The geographic transmission model for the force of infection on a ZIP $i$, developed in §3.2.2, takes as one of its inputs the ILI ratio from neighbouring infected ZIPs. It is possible to infer which age groups predict the geographic transmission of the outbreak by replacing this overall ILI ratio with the ILI ratio from a particular age group, and testing for an improvement in model fit. The ILI ratios from multiple age groups can also be combined to identify which combination of age groups best predict onward geographic spread.

Before the age-specific ILI time series can be aggregated, however, some missing entries in the data must be filled in. The number of physician visits is missing in each age group for all weeks when there were no recorded ILI cases in that age group. This affects the denominator of the ILI ratio when data from multiple age groups are aggregated; in some weeks, there will appear to be an unrealistically small number of physician visits, artificially amplifying the ILI ratio. To address this, any missing records of physician visits are replaced with the median number of weekly physician visits in that age band between July and December 2009. The weeks containing Labour Day and Thanksgiving, however, must be treated differently, since there are generally fewer physician visits in those weeks across all locations. To fill in these values, the median percentage drop in cases for each holiday is calculated for each age group. This is done by dividing the number of physician visits on the holiday by the median number of physician visits in July-December of 2009, for all locations/age groups where there were recorded physician visits on the holiday, and then

taking the median of these fractions for each age group. Then, for all locations in each age group for which the number of physician visits on the holiday is missing, the missing value is replaced by the median number of visits in that location and age group multiplied by the median fractional decrease in visits for that age group, rounded to the nearest whole number. Table 5.3 provides the fraction of missing entries in each age band from July-December 2009, as well as the median fractional decrease in physician visits for Labour Day and Thanksgiving.

Table 5.3 Fraction of missing physician visits entries, with the median fractional decrease in the number of visits in the weeks containing Labour Day and Thanksgiving, for each age group

| Age Group | Fraction of entries missing | Labour Day frac. decrease in cases | Thanksgiving frac. decrease in cases |
|---|---|---|---|
| $<2$ | 0.08 | 0.85 | 0.83 |
| 2-4 | 0.09 | 0.83 | 0.73 |
| 5-9 | 0.11 | 0.84 | 0.70 |
| 10-14 | 0.15 | 0.86 | 0.64 |
| 15-19 | 0.16 | 0.85 | 0.74 |
| 20-29 | 0.15 | 0.85 | 0.73 |
| 30-39 | 0.17 | 0.85 | 0.70 |
| 40-49 | 0.17 | 0.84 | 0.69 |
| 50-59 | 0.18 | 0.83 | 0.70 |
| 60-69 | 0.26 | 0.84 | 0.69 |
| 70-79 | 0.39 | 0.86 | 0.68 |
| 80+ | 0.49 | 0.88 | 0.72 |

After filling in the missing entries, the data are ready for input into the geographic transmission model. First, for a given set of age groups, the numbers of ILI cases are summed across all the age groups in each week for each location. The same is done for the number of physician visits. The aggregated ILI cases are divided by the aggregated physician visits in each week, giving an age-aggregated weekly ILI ratio time series for each location. Following the methods presented in §3.2.2, the aggregated ILI ratio for each location is normalised by the mean ILI ratio in that location between July and December 2009. These aggregated, normalised time series are then used as the $n_{j,t}$ in the transmission model, Eq 3.37. Note that the outbreak onset times to which the model is fit are not changed; these are still the outbreak onset times from the fully age-aggregated time series considered in Chapters 2 and 3. Through this model formulation, we are essentially asking which age groups provide the

best prediction of the overall (age-aggregated) outbreak onset time in neighbouring ZIPs. Model parameters are fit by maximum likelihood, as described in §3.2.2. This is done for all possible combinations of age groups, excluding those over 60 years old, since the time series for those age groups become extremely noisy and hinder the model fits, and because so many data entries in those age groups are missing (see Table 5.3). This yields a total of 511 new model fits.

### 5.4.2 Geographic transmission model fits using age-specific ILI data

The maximum-likelihood model uses normalised ILI ratios $n_{j,t}$ obtained by aggregating the ILI data from <2 year-olds, 5-9 year-olds, and 20-29 year-olds. The AIC for this model is 4259.8, which is a 15.8-point improvement over the model that uses the fully aggregated time series (see Table 3.3). The parameter values for this best-fit model are given in Table 5.4.

Table 5.4 Estimated parameter values for the maximum likelihood transmission model, Eq 3.37, using aggregated ILI data from <2 year-olds, 5-9 year-olds, and 20-29 year-olds to generate $n_{j,t}$, with parameter values estimated using the age-aggregated time series (see Chapter 3, Table 3.4)

| Parameter | Estimated value (95% CI) | Previous value (95% CI) | Units |
|---|---|---|---|
| $\beta_0$ | 0.00041 (0.00014, 0.00085) | 0.00043 (0.00015, 0.00087) | $(\Delta t)^{-1}$ |
| $\beta_d$ | 0.60 (0.50, 0.72) | 0.61 (0.53, 0.70) | $(\Delta t)^{-1}(km)^{1-\varepsilon}$ |
| $\mu$ | 0.31 (0.23, 0.40) | 0.32 (0.24, 0.40) | none |
| $\rho$ | 58 (41, 84) | 66 (48, 96) | $km$ |
| $\gamma$ | 7.0 (4.8, 18) | 8.9 (5.5, 74) | none |
| $\varepsilon$ | 0.97 (0.89, 1.0) | 1.0 (fixed) | none |
| $\theta$ | 0.70 (0.48, 0.93) | 0.56 (0.35, 0.77) | none |

There are five other models within two log-likelihood units of the maximum-likelihood model. Table 5.5 lists their log likelihoods and specifies which age groups are included in each. All six models include the <2 year-old and 5-9 year-old ILI. The 2-4 year-old ILI is sometimes included, as are the 20-29 year-old and 40-49 year-old ILI, but only one set of adult ILI is ever included at a time. There are no model fits that fall between 1 and 2 log-likelihood units below the optimal model; the next-best model, after the ones included in Table 5.5, is 2.2 log-likelihood units below the maximum-likelihood model.

Table 5.5 Age groups included in $n_{i,t}$ for the geographic transmission model fits within two log likelihood units of the best model

| $\Delta$ log likelihood | <2 | 2-4 | 5-9 | 20-29 | 40-49 |
|---|---|---|---|---|---|
| 0 | • | | • | • | |
| −0.10 | • | | • | | |
| −0.13 | • | • | • | • | |
| −0.58 | • | | • | | • |
| −0.73 | • | • | • | | |
| −0.91 | • | • | • | | • |

## 5.5 Discussion

This chapter adopts STE as a means of identifying which age groups contribute most to the transmission of infectious disease outbreaks. STE is chosen for its robustness to point-wise noise and broad-scale amplitude shifts in time series, which particularly affect the ILI data stream due to non-influenza respiratory illness and incomplete reporting. Simulation studies indicate that STE can correctly identify asymmetric transmission patterns between age groups. However, STE is also positively associated with reporting rates, which can partially confound estimates of asymmetric transmission. STE estimates on ILI time series data from July-December 2009 in the United States indicate that 5-19 year-olds were primarily responsible for driving the autumn wave of the A/H1N1pdm pandemic outbreak. These estimates were made possible by the fine geographic and age detail in the IMS-ILI dataset, since STE calculations are "data hungry", requiring many replicates of the observed process (in this case, ZIP-level age-stratified ILI outbreaks) to obtain reliable estimates of the probabilities in Eq 5.9. It is unlikely that reporting rates alone can account for the elevated STE from these ages. Fitting a geographic transmission model using age-stratified ILI data reveals that ILI incidence in young infants (<2 years), young children (5-9 years), and young adults (20-29 years) best predicts the geographic transmission of the autumn wave of the 2009 A/H1N1pdm influenza pandemic in the United States.

The identification of school-aged children as the primary drivers of transmission of the 2009 influenza pandemic in the United States is in line with most other studies on age-specific transmission of both seasonal and pandemic influenza [91, 170, 218, 245]. Elevated transmission from school-aged children can likely be explained by the elevated number of contacts in these age groups. Mossong *et al.* (2008) [170] estimate that 10-19 year-olds have more contacts per day than any other age group, and the contact patterns of 5-9 year-olds

are tightly coupled with these older children. The highest number of daily contacts is made by 10-14 year-olds, who also display the highest within-ZIP STE (see Fig 5.15). Smieszek *et al.*'s (2011) [218] estimates of infection rate by age during the 2003 influenza outbreak in Switzerland are qualitatively similar to the within-ZIP STE estimates presented in Fig 5.15, with a peak infection rate in 10-14 year-olds and elevated infection rates in 5-9 and 15-19 year-olds. Unfortunately, little data exists on the movement patterns of children, so it is difficult to say why the maximal STE shifts to 5-9 year-olds when comparing the ILI time series from ZIPs between which infection likely spread, according to the geographic transmission model developed in Chapter 3.

It is unlikely that differences in reporting rate alone can explain the observed differences in STE between the age groups. Besides the checks presented in this chapter, Biggerstaff *et al.* (2012) [22] report that 0-4 year-olds had the highest reporting rates for ILI in 2009, yet the STE from 0-4 year-olds is consistently low. If reporting rates alone could explain the observed differences in STE, the STE from infants should be at least as high as the STE from school-aged children.

The geographic transmission model fits reveal a slightly different picture than the STE. The inclusion of ILI data from $<2$ year-olds and, in slightly less optimal models, from 2-4 year-olds is in agreement with Brownstein *et al.* (2005) [27], who find that respiratory illness in children under 3 is the best predictor of overall influenza-related mortality, and that children under 5 generally act as sentinels of respiratory illness, even if school-aged children predominately drive infection once an outbreak has taken hold. The inclusion of ILI from 5-9 year-olds suggests that transmission from school-aged children is nevertheless a key determinant of geographic spread. The 5-9 year-old age group is also associated with the highest short-distance between-city STE (see Fig 5.17). The other two school-aged-children age groups (10-14 and 15-19 year-olds) may not be included in the best geographic transmission models because they provide redundant information and dampen the 5-9 year-old signal. The 5-9 year-old ILI ratio signal is consistently the strongest of the three school-aged children age groups; the peak ILI ratio for 5-9 year-olds is the highest of the three in 680 of the 834 ZIP in the span between July and December of 2009. Furthermore, the youngest three age groups ($<2$, 2-4, and 5-9 years) have the fewest number of missing physician visit data entries (see Table 5.3), so their time series are the most accurate overall, further justifying their inclusion in the geographic transmission model. Finally, including adult ILI may help account for between-city transmission from individuals who commute for work. The ILI signal for 20-29 year-olds is also consistently the strongest of the adult age groups, with a peak ILI ratio that is higher than the peak in 30-39, 40-49, and 50-59 year-olds

in 739 of the 834 ZIPs. These findings are all overshadowed by the fact that many missing counts of physician visits needed to be filled in. More complete data would be extremely helpful for verifying these results.

The geographic transmission model fits are not necessarily in contradiction with the STE estimates. The STE ranks age groups according to their contributions to transmission throughout an outbreak, while the geographic transmission model reveals which age groups best predict epidemic onset times in neighbouring ZIPs. So, elevated ILI in young infants, young children, and young adults may best herald the onset of an outbreak in a city, while school-aged children sustain transmission once it has begun.

Despite the apparent well-suitedness of STE for making inferences on ILI data, its epidemiological relevance currently remains limited. Indeed, the essential lack of epidemiological insight in the formulation of STE makes it surprising that STE is as capable of identifying true epidemiological interactions between age groups as the simulation studies presented in this chapter suggest. The next-generation matrix is the key object for characterising age-structured, or more generally population-structured, transmission dynamics, and yet there is no obvious direct link between STE estimates and the NGM. It is possible that further simulation studies could help identify such a link; even though the explicit STE values seem to bear little meaning apart from the relative ordering that they yield, it is possible that regressing the inferred STE values on the underlying transmission matrix could connect the pairwise STE matrix with the NGM under certain conditions. However, it appears unlikely that a simple link exists, especially since STE can say nothing about transmission within a single age group, which is necessary for filling in the diagonal entries of the next-generation matrix. STE and related methods such as convergent cross-mapping that do not explicitly incorporate mechanistic descriptions of the underlying physical system are unlikely to be able to reveal more than an approximate hierarchy of driving processes. Nevertheless, such a hierarchy can contain valuable information, especially if developing and fitting a mechanistic model is too demanding to be practicable. Extensions to STE could also enhance its relevance for epidemiological inference. Local transfer entropy [153] and state-dependent transfer entropy [249], like the contextual STE, are intended to make the traditional transfer entropy more flexible and general, by considering how information transfer may change under varying conditions. These may yield better insight into epidemic processes, which are inherently nonlinear and context-dependent, than the more established measurements of transfer entropy can provide.

Perhaps the most important challenge confronting both transfer entropy- and convergent cross mapping-based approaches is deciding how to measure power and significance. STE

calculations rely on a middle level of stochasticity; for a deterministic system, the STE will always be exactly zero, while for a stochastic system with too much within-sequence noise, the small-scale variation in amplitudes will likely mask important patterns from which the transfer of information might be inferred. This acceptable range of stochasticity has not been clearly defined. In other words, it is unclear how to measure how much statistical power is present in a given experiment for distinguishing true differences in STE. Similarly, it is unclear how best to measure when a difference in STE should be called statistically significant. Though this is recognised as an open and difficult problem [12, 13], it may be possible to make some progress by assuming that the underlying process follows certain epidemiological, or otherwise well-specified, dynamics.

A different way forward would be to extend existing methods for next-generation matrix inference to accommodate data with lower specificity and more age classes. There are two strategies that appear especially promising. The first is to follow the groundwork laid by Wallinga and Teunis (2004) [244] and Glass *et al.* (2011) [90]. Their approaches use branching process theory to infer the likely number of cases triggered by each observed case in an epidemic. The methods rely on age-specific estimates of the generation interval of the disease, which are currently under-studied for influenza (see [149] for one of the few examples of age-stratified generation intervals for influenza). It also assumes that accurate case counts are available; to my knowledge, the methods cannot currently accommodate ILI data. Furthermore, the method proposed by Glass *et al.* (2011) [90] makes strict assumptions on the form of the next-generation matrix, which become increasingly unrealistic as the number of age classes increases. Leveraging data from multiple nearby outbreaks and assuming some geographic dependence between them may be one way to relax these assumptions and obtain more refined estimates of the next-generation matrix.

The second strategy follows in the line of Ionides *et al.* (2006) [121], Shaman and Karspeck (2012) [210], and Yang *et al.* (2015) [258]. This approach uses sequential Monte Carlo methods to fit explicit mechanistic transmission models to time series data. Yang *et al.* (2015) [258] use this strategy to infer key transmission parameters for influenza outbreaks over the course of nine years. That study incorporates age-specific transmission rates into the underlying model, but estimates these rates *a priori* from the POLYLMOD contact matrices, rather than inferring them from epidemiological data (it appears that age-stratified ILI data were not available). Though implementing such a model for age-structured inference would be extremely computationally intensive, it may hold the most promise for linking age-structured ILI data directly with a next-generation matrix.

An interesting and important way to extend all of the aforementioned methods would be to incorporate other sources of heterogeneity in transmission rate, such as viral strain. An age- and strain-structured model would likely have to accommodate non-trivial interactions between host age and viral strain, since different influenza strains preferentially infect different age groups, often due to differences in previous exposure [18]. Though this is likely unimportant for the 2009 pandemic, when a single viral strain caused the vast majority of infections, such an approach could provide valuable insight into the transmission of seasonal outbreaks when multiple strains may co-circulate.

## 5.6 Summary

In this chapter, symbolic transfer entropy (STE) is used to infer which age groups may have contributed most to the transmission of the autumn wave of the 2009 A/H1N1pdm influenza pandemic in the United States. The STE overcomes limitations associated with existing strategies to infer from time series data how disease transmission varies by age group. The contextual STE is introduced to verify that symbolisation does not omit too much relevant information from epidemiological time series. It is shown that the contextual STE gives valid insight into the underlying disease transmission process for a range of epidemiologically feasible parameters. Using a two-age-class individual-based stochastic SIR model, it is shown that the STE can detect both increasing symmetric coupling and asymmetric coupling between age groups. Equivalent results are obtained using a more computationally efficient Poisson-type epidemic model. Simulations on a set of four-age-class Poisson-type epidemic models demonstrate that STE is positively associated with both transmission strength and reporting rate, and also that STE can identify true differences in transmission rates even when reporting rates are low. Simulations on 12-age-class Poisson-type epidemic models reinforce the results from the four-age-class simulations, and provide a point of comparison with available ILI data from July-December 2009 in the United States. STE calculations on this ILI data provide evidence that 5-19 year-olds were primarily responsible for within-city and short-distance between-city transmission of the autumn wave of the 2009 A/H1N1pdm influenza pandemic in the US, in agreement with previous studies. A sensitivity analysis based on reconstructing true case counts from observed data indicates that the results are robust to reporting uncertainty. As an alternative check of the role of different age groups in the transmission of the pandemic outbreak, age-specific time series are incorporated into a mechanistic geographic transmission model. The optimal model predicts city-level epidemic onset times as a function of ILI intensity in young infants ($<$2 years), young children (5-

9 years), and young adults (20-29 years) from nearby cities, suggesting that school-aged children alone may not fully account for the geographic transmission of the outbreak, despite likely sustaining the majority of local transmission.

# Chapter 6

# Seasonal variation in the geographic transmission of influenza in the United States

In this chapter, a geographic disease transmission model is fit to epidemic onset times inferred from city-level ILI data from the 2003-04 and 2007-08 seasonal influenza outbreaks in the United States. These were the two outbreaks with the highest overall peak ILI incidence in the decade preceding the 2009 pandemic, according to the IMS-ILI dataset. Following the methods developed in Chapters 3 and 4, transmissibility surfaces and transmission hubs are identified for both outbreaks. Differences in transmission strength by age group are inferred using the methods developed in Chapter 5. For the 2007-08 outbreak, in which three distinct viral strains circulated, geo-tagged antigenic data is used to infer which strains may have seeded the outbreaks in each of the hubs.

## 6.1 Background

Though pandemic influenza has arguably received more attention than seasonal influenza in the epidemiological modelling literature, there are a number of studies that consider the transmission characteristics of seasonal influenza at the international, continent, and country scales. Internationally, genetic analyses by Russell *et al.* (2008) [204] and Bedford *et al.* (2015) [18] reveal that seasonal outbreaks of A/H3N2 influenza tend to be seeded from southeast Asia, while outbreaks of other strains may follow more complicated resurgence dynamics. The continent-scale transmission of seasonal influenza in Europe is considered for

example by Paget *et al.* (2007) [184], who find evidence of frequent north-easterly geographic waves of transmission across Europe. Within the US, Viboud *et al.* (2006) [239] use 30 years' worth of mortality data in the United States to demonstrate a high degree of synchrony in the timing of influenza outbreaks between US states. Yang *et al.* (2015) [258] use ILI data to calculate key epidemiological parameters for influenza outbreaks in the US between 2003 and 2013. Charu *et al.* (2017) [48] fit gravity-type mechanistic transmission models, like the one presented in Chapter 3, to outbreak onset times from eight influenza seasons in the US using a spatially-aggregated version of the IMS-ILI data considered in this thesis. Though this chapter focuses on just two seasonal outbreaks, it extends the work by Charu *et al.* (2017) [48] in that it considers the IMS-ILI data in its full spatial detail, presents a detailed comparison of the transmission kernels for each outbreak, identifies the outbreaks' transmission hubs using the methods developed in Chapter 4, and uses symbolic transfer entropy to infer which age groups contributed most to transmission. Attention is restricted to the 2003-04 and 2007-08 outbreaks because these seasons, along with the autumn 2009 pandemic outbreak, featured the sharpest rises in ILI out of all influenza outbreaks between 2001 and 2010, permitting outbreak onset times to be successfully estimated for most ZIPs. The analyses that follow in this chapter are in theory applicable to all seasons included in the IMS-ILI dataset, but care must be taken to address the reduction in data clarity for seasons beyond the three considered here. Consideration of those additional seasons is left for future work.

### 6.1.1 The 2003-04 influenza outbreak in the United States

In 2003, a novel strain of influenza subtype A/H3N2 emerged, possibly from southeast Asia [88, 204]. The strain quickly became the most prevalent of the 2003-04 seasonal influenza outbreak in the United States and across much of the world [34, 88]. The influenza vaccine for that season was a poor match to the novel strain, and provided little reduction in the overall burden of ILI [35]. In the United States, the 2003-04 seasonal influenza outbreak began in October and peaked in late November, earlier than most flu seasons [34, 36]. Texas was the first state to report widespread influenza activity, and was the state from which the first viral strains of the season were isolated [37]. Yang *et al.* (2015) estimate that this outbreak had the highest basic reproduction number $R_0$ (2.04) of any influenza outbreak in the United States between 2003 and 2013, including the 2009 A/H1N1pdm pandemic [257].

### 6.1.2 The 2007-08 influenza outbreak in the United States

In 2007, elevated regional influenza activity in the United States was first reported in December from Texas [38]. The outbreak peaked in mid-February [39]. Three distinct influenza strains co-circulated, with a shift in dominant strain from type A/H1N1 to type A/H3N2 to type B over the course of the season [38]. Overall, type A/H3N2 was most prevalent during the season [38]. The season's vaccine was a good match for the A/H1N1 strain, but not for the A/H3N2 or B strains [38]. Yang *et al.* (2015) estimate that this outbreak had the second-highest basic reproduction number $R_0$ (2.03) of any influenza outbreak in the United States between 2003 and 2013, including the 2009 A/H1N1pdm pandemic [257].

## 6.2 Outbreak onset times for the 2003-04 and 2007-08 seasonal influenza outbreaks

Using the methods developed in Chapter 2, ZIP-level outbreak onset times may be estimated for the 2003-04 and 2007-08 seasonal outbreaks in the United States using the breakpoint method applied to the IMS-ILI dataset. Fig 6.1 depicts the ZIP-level outbreak onset times for the 2003-04 seasonal influenza outbreak in the United States. To generate the onset times for 2003-04, the maximum ILI incidence is sought in each ZIP between 28 September 2003 and 25 January 2004, and the breakpoint method is used to calculate the ZIP's outbreak onset time using the ILI ratios from $n = 17$ weeks prior to and including this epidemic peak. Like the 2009 pandemic outbreak, there is evidence of radial spread from an epicentre in the southern US. The spread is faster than in 2009, with 14.5 weeks between the earliest and latest onsets, over a month shorter than the autumn 2009 outbreak. Over 98% (723) of the 734 outbreaks with detectable onset times have onset in the 10 weeks following 5 October 2003.

Fig 6.2 depicts the ZIP-level outbreak onset times for the 2007-08 seasonal influenza outbreak in the United States. To generate these onset times, the maximum ILI incidence is sought in each ZIP between 9 December 2007 and 20 April 2008, and the breakpoint method is used to calculate the ZIP's outbreak onset time using the ILI ratios from $n = 17$ weeks prior to and including the epidemic peak. A radial spread pattern is less apparent for this outbreak than for 2003-04 or 2009. Rather than spreading from a clear epicentre, it appears that infection was introduced in various locations and spread outward locally. The onset times span a total of 15 weeks, similar to the length of the 2003-04 outbreak, but shorter than

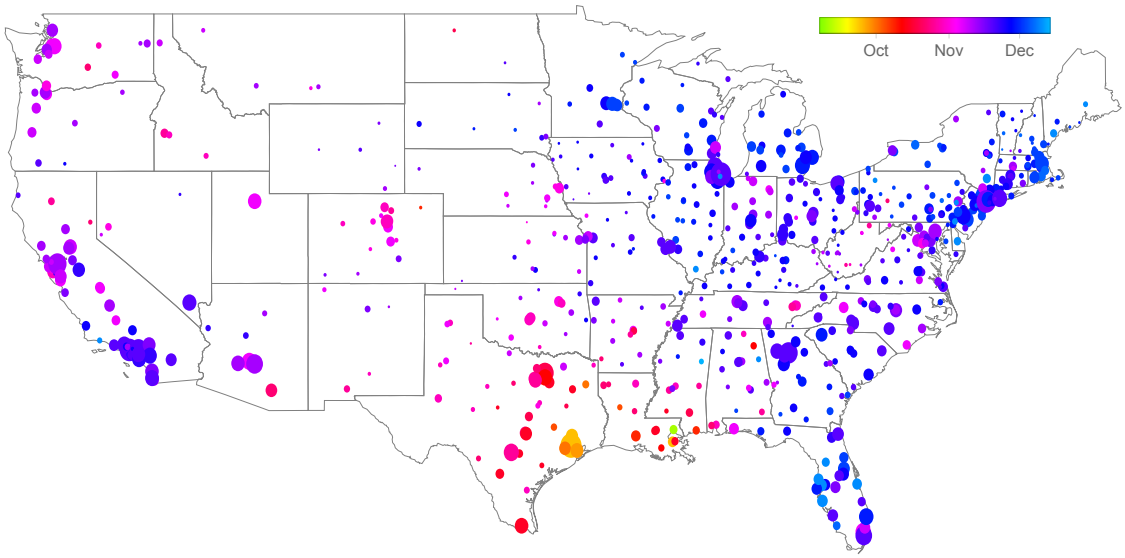



Fig. 6.1 ZIP-level outbreak onset times for 2003-04. Discs represent ZIPs, and disc area is proportional to the ZIP's population size. Each disc is coloured according to the corresponding ZIP's outbreak onset time, with green/yellow representing outbreaks early in the epidemic and purple/blue representing outbreaks late in the epidemic. The epidemic appears to spread radially from southern Louisiana and Texas.

the 2009 pandemic. Over 98% (708) of the 716 outbreaks with detectable onset times have onset in the 12 weeks following 25 November 2007.

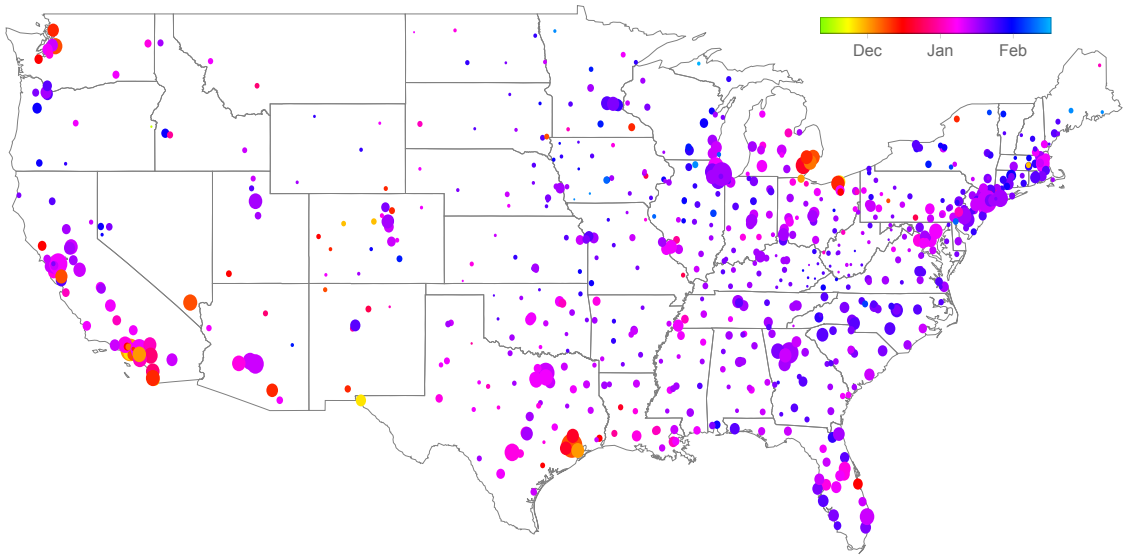



Fig. 6.2 ZIP-level outbreak onset times for 2007-08. Discs represent ZIPs, and disc area is proportional to the ZIP's population size. Each disc is coloured according to the corresponding ZIP's outbreak onset time, with green/yellow representing outbreaks early in the epidemic and purple/blue representing outbreaks late in the epidemic. There appear to be multiple clusters of ZIPs with early onsets near Houston TX, Detroit MI, and Los Angeles CA. There is still some evidence of radial spread from these epicentres, though the signature is less apparent than for 2003-04 and 2009. The epidemic also took place much later in the year than the 2003-04 and 2009 outbreaks.

## 6.3 The seasonal influenza transmission model

To model the geographic transmission of influenza in the United States in 2003-04 and 2007-08, the transmission model developed in Chapter 3 for pandemic influenza (Eq 3.27) is fit to outbreak onset times from these two seasonal outbreaks. The most parsimonious model for both seasons matches the form of the best model for the 2009 pandemic, indicating that the key drivers of all three outbreaks were similar. However, some of the model parameter values differ significantly between the seasons, reflecting possible differences in strain severity and human behaviour.

### 6.3.1 Model selection and parameter estimation

The transmission model developed in Chapter 3, Eq 3.27, is fit to outbreak onset times from the 2003-04 and 2007-08 influenza seasons, using the methods presented in §3.2.2. Both outbreaks began after schools were in session across the country, so $\beta_d$ is fixed at 0 and $I_a$ to 1 for both seasons.

For both seasons, the best model in terms of AIC omits the donor population size parameter $\nu$, and leaves all other parameters free. That is, the force of infection $\lambda$ on location $i$ in half-week $t$ is given by

$$\lambda_i(t) = \beta_0 + \beta_{ds} N_i^{\mu} \frac{\sum_{j \in \Lambda_t} n_{j,t}^{\theta} \kappa(d_{i,j})}{[\sum_{j \neq i} \kappa(d_{i,j})]^{\varepsilon}} \tag{6.1}$$

where $\Lambda_t$ is the set of locations with outbreak onset prior to half-week $t$; $n_{j,t}$ is the ILI ratio in location $j$ in half-week $t$, normalised by the mean ILI ratio in location $j$ throughout the epidemic season, and fixed at 0 for all $t$ prior to location $j$'s outbreak onset time; and $N_i$ is the mean-normalised population size of the recipient location $i$. As before, the parameter $\beta_0$ is the background force of infection due to long-distance seeding; $\beta_{ds}$ is the transmissive strength of the disease; $\mu$ is the recipient-population gravity model exponent; $\theta$ modulates the importance of the ILI time series in neighbouring ZIPs; $\varepsilon$ modulates population density dependence; and $\kappa(d_{i,j}) = \left(1 + \frac{d_{i,j}}{\rho\gamma}\right)^{-\gamma}$ captures the decay in epidemiological connectivity between locations as a function of distance $d_{i,j}$ with length scale $\rho$ and power law parameter $\gamma$. The model selection procedures for 2003-04 and 2007-08 yield optimal models of the same overall form as the best model for the autumn wave of the 2009 pandemic; compare Eq 6.1 and Eq 3.37.

While the model forms match for all three seasons, the parameter values differ. Table 6.1 provides the maximum likelihood parameter values for the outbreaks in 2003-04, 2007-08, and 2009. The baseline force from external seeding, $\beta_0$, is similarly small across all three seasons. Comparing the exact values of $\beta_0$ across seasons is not especially illuminating, since the estimated value depends strongly on the date that is arbitrarily chosen as the start of the flu season. The other parameters are not as strongly affected by this choice, and so can be compared across seasons. The 2003-04 and 2009 outbreaks, both of which were caused by antigenically novel strains of influenza, have similar transmissibility ($\beta_{ds}$). The transmissibility term for the 2007-08 outbreak is significantly smaller. The boost in susceptibility from the recipient location's population size ($\mu$) is fairly consistent across all seasons, though it is slightly higher for the 2009 pandemic than for the 2003-04 and 2007-08

outbreaks. The characteristic distance of transmission, $\rho$, is smallest in 2003-04 and largest in 2009, but the uncertainties on the parameter are large, and the parameter's confidence intervals for all three seasons overlap substantially. The power kernel parameter $\gamma$ is similar for 2003-04 and 2007-08, but is significantly larger in 2009. The distance kernels for the two seasonal outbreaks therefore have thicker tails than the kernel for the pandemic outbreak, or in other words, long-range jumps were relatively more likely during the seasonal outbreaks than during the 2009 pandemic. The population density dependence parameter $\varepsilon$ is high (close to 1) in 2003-04, as it is in 2009. It is somewhat lower in 2007-08. The ILI intensity factor $\theta$ is consistent across all three seasons, with a value close to 0.5.

Table 6.1 Estimated parameter values for the most parsimonious transmission model, Eq 6.1, fit to outbreak onset times from the 2003-04, 2007-08, and autumn 2009 influenza outbreaks.

| Parameter | 2003-04 value (95% CI) | 2007-08 value (95% CI) | 2009 value (95% CI) | Interpretation | Units |
|---|---|---|---|---|---|
| $\beta_0$ | 8.6E-5 (5.0E-6, 3.7E-4) | 6.0E-4 (1.6E-4, 1.4E-3) | 4.3E-4 (1.5E-4, 8.7E-4) | External seeding risk | $(\Delta t)^{-1}$ |
| $\beta_{ds}$ | 0.72 (0.57, 0.88) | 0.31 (0.17, 0.49) | 0.61 (0.53, 0.70) | Transmissibility factor | $\frac{(km)^{1-\varepsilon}}{(\Delta t)}$ |
| $\mu$ | 0.24 (0.15, 0.34) | 0.23 (0.14, 0.33) | 0.32 (0.24, 0.40) | Gravity model exponent | none |
| $\rho$ | 32 (20, 49) | 54 (23, 111) | 66 (48, 96) | Characteristic distance | $km$ |
| $\gamma$ | 3.1 (2.6, 4.1) | 2.5 (1.8, 4.3) | 8.9 (5.5, 74) | Power law decay factor | none |
| $\varepsilon$ | 0.89 (0.78, 1.0) | 0.66 (0.51, 0.81) | 1.0 (fixed) | Density correction | none |
| $\theta$ | 0.46 (0.21, 0.72) | 0.59 (0.30, 0.87) | 0.56 (0.35, 0.77) | ILI intensity exponent | none |

Rather than comparing the parameters side-by-side, however, it is helpful to consider how they jointly affect range of geographic transmission for each outbreak. It is possible to define a distance kernel for a hypothetical city with median population size and median surrounding population density. This kernel has form

$$K(x) = \beta_{ds} \bar{N}^{\mu} \frac{\kappa(x)}{\bar{\Delta}^{\varepsilon}} \tag{6.2}$$

where $\bar{N}$ is the median normalised ZIP population size, and $\bar{\Delta}$ is the median population density divisor; that is,

$$\bar{\Delta} = \text{Median}_i(\sum_{j \neq i} \kappa(d_{i,j})). \tag{6.3}$$

This distance kernel is depicted in Fig 6.3 for the 2003-04, 2007-08, and 2009 outbreaks, obtained by substituting the maximum likelihood parameter estimates from Table 6.1 into Eq 6.2. The shaded areas in Fig 6.3 depict parameter uncertainty. The boundaries of these regions are calculated by choosing a subset of distances $x$ and, for each of these distances, maximising/minimising Eq 6.2 while constraining the log-likelihood of the parameters to be within 1.92 units of the maximum log-likelihood. This provides an effective 95% confidence interval for the transmission kernel at each distance. If the distances are sampled densely enough, they produce the curved boundaries of the regions.

From Fig 6.3, it may be seen that the estimated gravity-driven transmission strength of the 2003 outbreak is higher than for the autumn 2009 pandemic at short distances, under about 75km. After 75km, the confidence bands for the two kernels intersect, and by a distance of about 200km the maximum-likelihood kernel values are nearly the same. This agrees with the observed behaviour of the two outbreaks: the higher transmission strength of the 2003-04 outbreak at short distances is indicative of its faster spread, while the similar thinness of both kernels' tails agrees with the observation that both outbreaks spread primarily in coherent geographic waves, rather than through a series of long-distance jumps. On the other hand, the estimated transmission strength of the 2007-08 outbreak is higher than for both the 2003-04 outbreak and the 2009 pandemic at all distances. In particular, the 2007-08 outbreak's kernel retains a thick tail at long distances, in agreement with the outbreak's rapid and geographically patchy spread.

### 6.3.2 Transmissibility surfaces

Following the methods developed in §3.2.3, the spatial and temporal variation in transmissibility are estimated for the 2003-04 and 2007-08 seasonal influenza outbreaks. Given the model

$$\lambda_i(t) = \beta_0 + \beta_{ds} \, \text{Exp}[\xi_t^T + \xi_i^S] N_i^\mu \frac{\sum_{j \in \Lambda_t} n_{j,t}^\theta \kappa(d_{i,j})}{\sum_{j \neq i} \kappa(d_{i,j})}, \tag{6.4}$$

posterior values of $\boldsymbol{\xi}^T$ and $\boldsymbol{\xi}^S$ are estimated by drawing candidate values from Gaussian process priors with squared-exponential covariance function, and accepting with probability proportional to the ratio of the proposed to the previous model's likelihood (see Eq 3.36). For consistency, the length scales of the SE covariance functions are set at the same values as

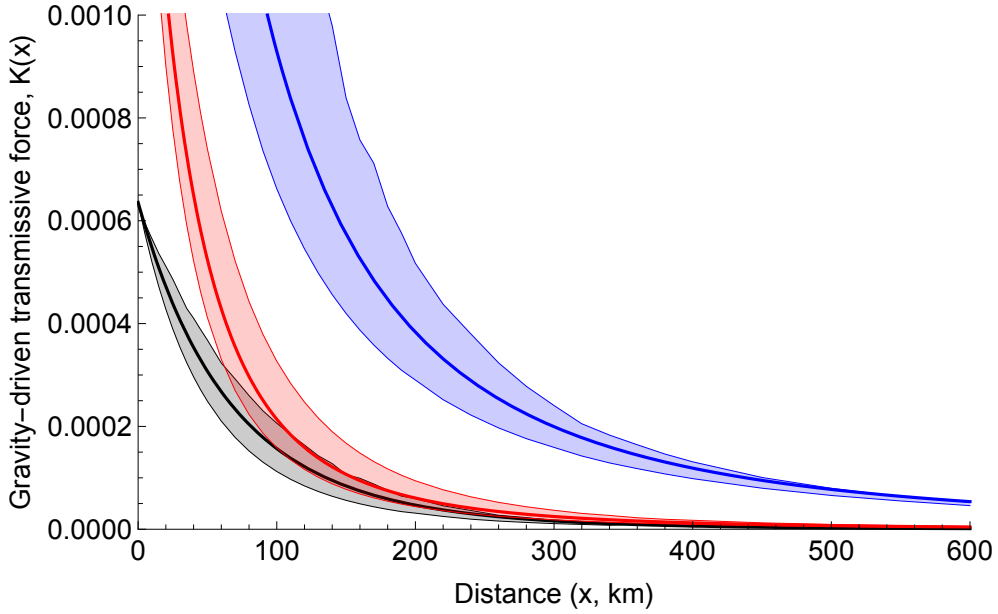



Fig. 6.3 Maximum-likelihood gravity transmission kernels for the 2003-04 (red), 2007-08 (blue), and autumn 2009 (black) influenza outbreaks in the United States. The kernels quantify the amount of infective force that an infected ZIP would contribute to a hypothetical ZIP with median population size and median surrounding population density when situated some distance $x$ away. The shaded regions depict uncertainty in the kernel, obtained by maximising/minimising the kernel function (Eq 6.2) at a set of distance values $x$, while constraining the log-likelihood of the parameters to be within 1.92 units of the maximum log-likelihood. The maximum-likelihood kernels for both 2003-04 and 2007-08 sit above the maximum-likelihood kernel for 2009, which agrees with the observation that both of these outbreaks spread more quickly than the 2009 pandemic. The 2007-08 kernel's thick tail is indicative of that outbreak's patchy geographic spread with multiple long-distance jumps of infection. The 2003-04 and 2007-08 kernels are cut off at $K(x) = 0.001$ to better depict the relationships between the three curves. The 2003-04 kernel intersects the vertical axis at a value of $K(0) = 0.0018$, and the 2007-08 kernel intersects the vertical axis at a value of $K(0) = 0.016$.

they were for the analysis of the 2009 pandemic, $l = 8$ half-weeks for $\boldsymbol{\xi}^T$ and $l = 200$km for $\boldsymbol{\xi}^S$. The MCMC algorithm is run four times for each season, for 10,000 iterations each. Following Gelman *et al.* (2013) [86], the first 5,000 iterations of each run are discarded to avoid effects from the burn-in period. For 2003-04, the Gelman-Rubin statistic for all chains is below 1.2, and for 2007-08 the Gelman-Rubin statistic for all chains is below 1.6, which suggest that the chains have converged acceptably. The Gelman-Rubin statistic is likely higher for 2007-08 than for 2003-04 because the higher number of introduction sites and patchier spread of the 2007-08 outbreak makes it more difficult for the Gausssian process to settle into a good fit. The final 5,000 iterations for each of the four runs are combined, yielding 20,000 draws of $\boldsymbol{\xi}^T$ and $\boldsymbol{\xi}^S$ for each season.

Fig 6.4 depicts the estimated temporal variation in transmissibility, $\boldsymbol{\xi}^T$, for the 2003-04 and 2007-08 seasonal influenza outbreaks. In 2003-04, there is little evidence of temporal variation in transmissibility. In 2007-08, on the other hand, the transmissibility reaches a minimum in late December and rises again by mid-January. The shape of this curve is similar to the one identified for the 2009 A/H1N1pdm pandemic (see Fig 3.10), though smaller in absolute magnitude. The dip in transmissibility could reflect changes in contact rates associated with the winter holiday season.

Fig 6.5 depicts the mean values of $\boldsymbol{\xi}^S$ by geographic location for 2003-04 and 2007-08. In 2003-04, the geographic transmissibility surface has a banded structure, with a region of higher-than-average transmissibility in Indiana, Ohio, West Virginia, and Virginia, and lower-than-average transmissibility in the southeast and the northeast. The model also identifies higher-than-average transmissibility in California, especially in Los Angeles. In 2007-08, the transmissibility surface is somewhat patchier, with a small area of lower-than-average transmissibility near the borders of Kentucky, West Virginia, and Virginia, as well as in the northeast. Transmissibility in the southeast is higher than average, while transmissibility in Florida and much of California is near average. The overall geographic variation in transmissibility is smaller for 2003-04 than it is for 2007-08, and is smaller for both of those seasons than it is for the 2009 pandemic (compare with Fig 3.11).

The variability in the spatial and temporal transmissibility adjustments for both seasons is greater than the variation observed from simulated outbreaks with constant $\beta_d$ across all locations (see §3.16), but only just. This suggests that, while there may have been some spatiotemporal variation in the transmissibility of the 2003-04 and 2007-08 influenza outbreaks, it was not especially pronounced, and likely did not have as large of an influence on the geographic transmission of those outbreaks as it did on the spread of the 2009 A/H1N1pdm pandemic.

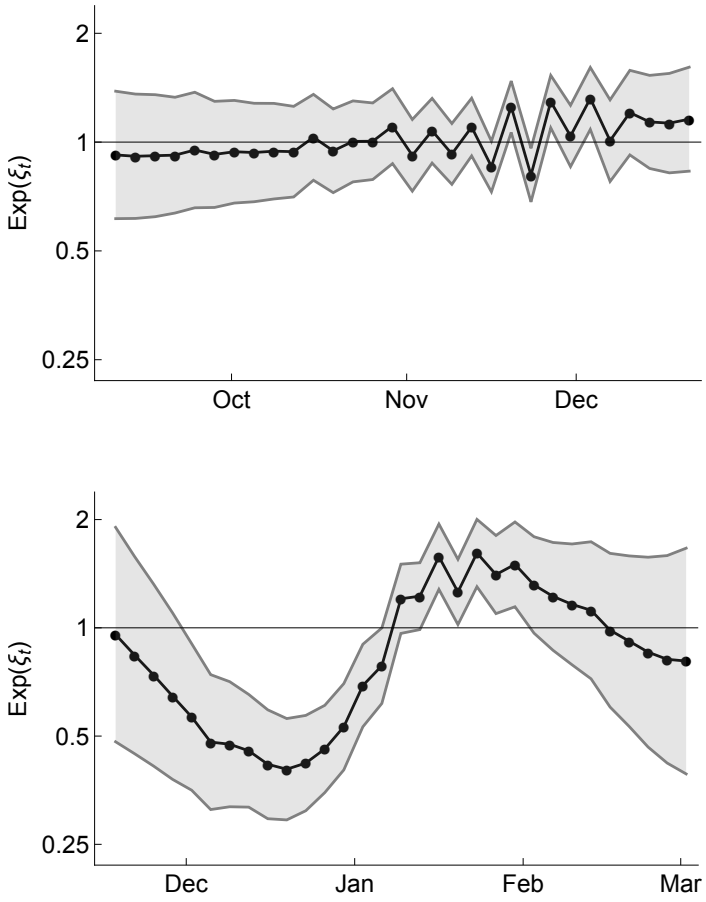



Fig. 6.4 Temporal variation in transmissibility, Exp[$\boldsymbol{\xi}^T$], for the 2003-04 (top) and 2007-08 (bottom) seasonal influenza outbreaks in the United States. The exponentiated transmissibility adjustment depicted here is a multiplicative factor for the transmissibility term, $\beta_{ds}$, in Eq 6.1; values greater than 1 indicate higher-than-average transmissibility and values less than 1 indicate lower-than-average transmissibility. In 2003-04, there is no evidence of temporal variation in transmissibility. In 2007-08, the transmissibility dips to a minimum in late December, and then rises again in mid-January, possibly reflecting variation in contact rates due to the winter holiday season.

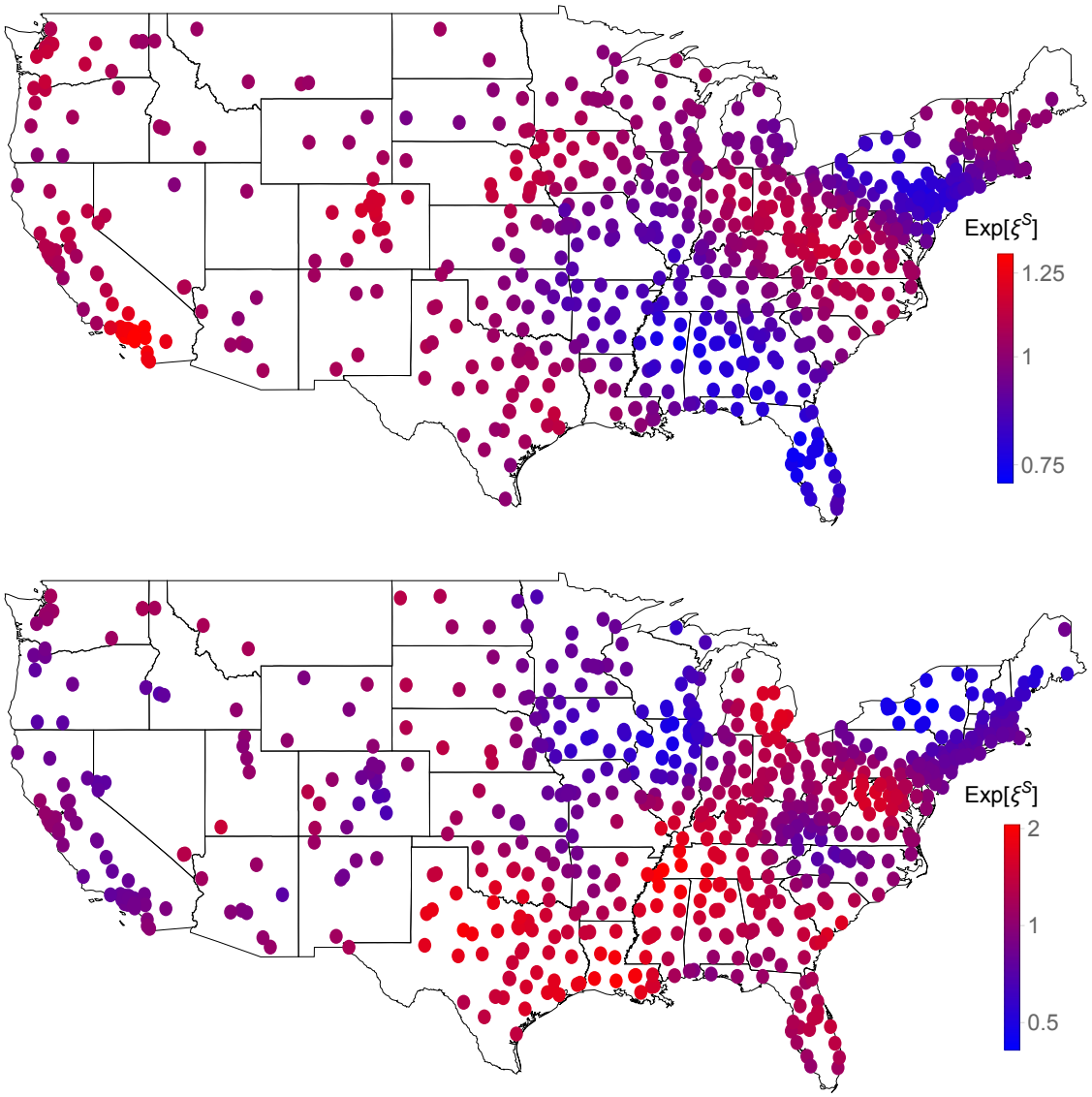



Fig. 6.5 Geographic variation in transmissibility, Exp[$\boldsymbol{\xi}^S$], for the 2003-04 (top) and 2007-08 (bottom) seasonal influenza outbreaks in the United States. The exponentiated transmissibility adjustment depicted here is a multiplicative factor for the transmissibility term, $\beta_d$, in Eq 6.1; values greater than 1 indicate higher-than-average transmissibility and values less than 1 indicate lower-than-average transmissibility. In 2003-04, there is a band of increased transmissibility between the southeast and the northeast, and in Los Angeles. In 2007-08, there are patches of high transmissibility near Detroit and Washington, DC, as well as in the southern states of Texas, Louisiana, Alabama, and Kentucky.

## 6.4 Transmission hubs of the 2003-04 and 2007-08 influenza outbreaks

Following the methods developed in §4.2, transmission hubs are identified for the 2003-04 and 2007-08 seasonal influenza outbreaks in the United States. Following §4.2, the proportion of the force of infection due to external seeding on each ZIP at its time of onset ($\sigma$) is calculated. These values are depicted in Figs 6.6-6.7 for both seasons. Then, transmission is probabilistically traced back to a set of most likely points of introduction, or hubs. The extent to which external seeding in each ZIP contributed to all other outbreaks through gravity-driven onward transmission ($C$) is depicted in Figs 6.8-6.9 for both seasons. Transmission hubs are taken to be the locations with $\sigma$ of at least 0.5 and $C$ at least 5.

In 2003-04, nearly all transmission ($C = 679.1$ of 734 locations) is traced back to seeding in a single transmission hub, in Mandeville, LA (pop. 423,850). From the map of outbreak onset times in 2003-04 (Fig 6.1), it appears that the overall epidemic was seeded in Mandeville, jumped quickly to Houston, and then spread outward from that part of the southeastern US. Fig 6.10 depicts the estimated probability that each ZIP's outbreak can be attributed to gravity-driven transmission from Mandeville. As one might expect, the outbreaks in ZIPs close to Mandeville can be traced back to seeding in Mandeville with higher probability than those that are far away, but even the outbreaks in ZIPs in the far northeast and northwest of the country can be traced back to Mandeville with a probability of over 80%.

In 2007-08, five transmission hubs are identified. These are listed in Table 6.2. Unlike both the 2003-04 and 2009 outbreaks, all of the transmission hubs in 2007-08 lie in or near major cities. The Inglewood Vicinity/Torrance CA hub is identified by the probabilistic back-tracing procedure as two separate hubs, but the two ZIPs have onset in the same week and lie just 11 km apart, suggesting that they represent a single seeding event near Los Angeles. In Table 6.2, the overall contributions from seeding in the two ZIPs ($C$) are combined, but the probabilities of seeding ($\sigma$) are reported separately. Taken independently, the transmission model estimates that seeding in Inglewood Vicinity CA triggered $C = 61.0$ outbreaks through gravity-driven onward transmission, and that seeding in Torrance CA triggered $C = 63.8$ outbreaks. Recall that $C$ measures the effective number of outbreaks triggered by a hub, so in reality, a single seeding event near Inglewood Vicinity or Torrance likely triggered approximately 125 ($\approx 61 + 63.8$) downstream outbreaks.

Fig 6.11 depicts the basins of infection for each of the transmission hubs in 2007-08. The transmission hubs in Cleveland OH, Inglewood Vicinity/Torrance CA, and Boise (ID) West

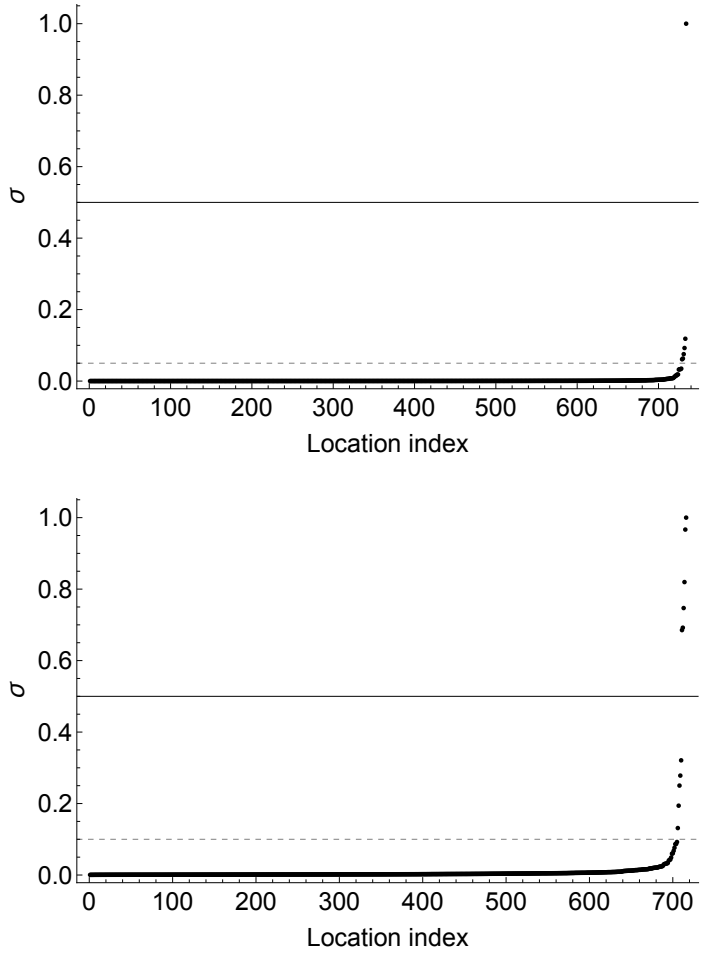



Fig. 6.6 Probability $\sigma$ that each ZIP's outbreak was caused by external seeding, in ascending order, for 2003-04 (top) and 2007-08 (bottom). In both seasons, there is a large gap between ZIPs with $\sigma$-values above and below 0.5 (solid horizontal bar). In 2003-04, only Mandeville LA surpasses this cutoff. Five ZIPs – Johnstown North PA, Reno (NV) West CA, Yakima WA, Brighton East CO, and New Rochelle NY – surpass a lower natural cutoff, at $\sigma = 0.05$ (dashed). These ZIPs are depicted geographically in the upper map of Fig 6.7, and labelled with names parentheses. In 2007-08, six ZIPs surpass the cutoff at $\sigma = 0.5$. These happen to be the transmission hubs identified by the back-tracing procedure, and so are listed in Table 6.2. Their geographic locations are depicted with solid black discs in Fig 6.7. Five ZIPs – Denver West CO, North Houston Southeast TX, Glenwood Springs CO, Sioux Falls Main SD, and Orlando East FL – surpass a lower natural cutoff, at $\sigma = 0.1$ (dashed). These are also depicted geographically in the lower map of Fig 6.7, and labelled with names parentheses.


Fig. 6.7 Probability $\sigma$ that external seeding triggered each ZIP's outbreak, by geographic location, for 2003-04 (upper map) and 2007-08 (lower map). Disc area is proportional to $\sigma$. In 2003-04, only Mandeville LA has a seeding probability $\sigma$ of over 0.5. Five other ZIPs, labelled with names in parentheses, have $\sigma$ over 0.05. The rest have $\sigma < 0.05$. In 2007-08, six ZIPs have $\sigma$ greater than 0.5. These are represented by the larger black discs, and labelled without parentheses. Five additional ZIPs have $\sigma$ greater than 0.1. These are labelled with names in parentheses.

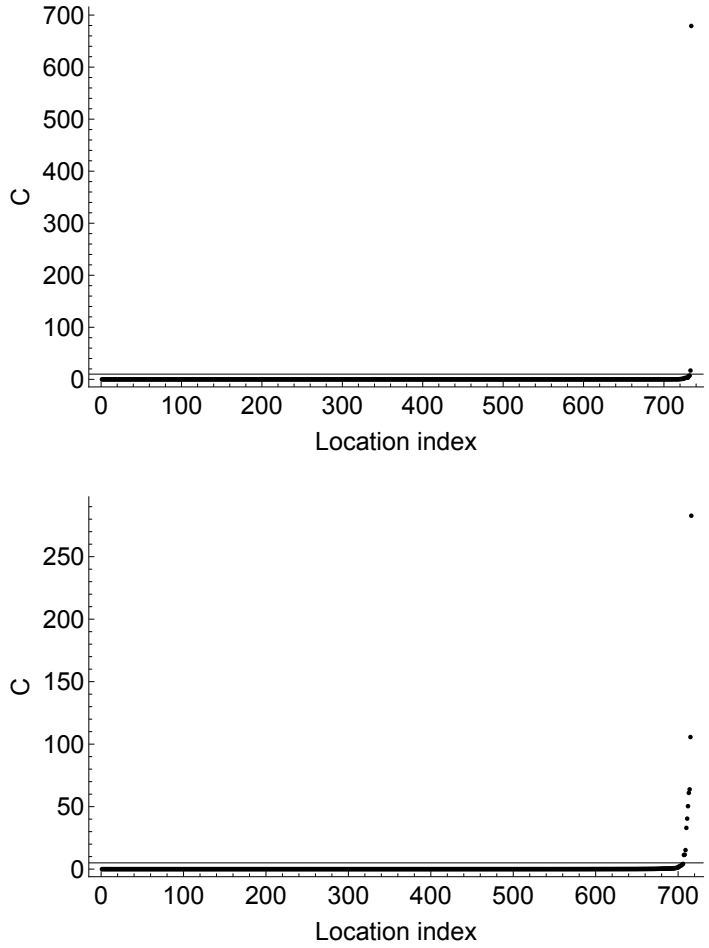



Fig. 6.8 Total contribution $C$ of external seeding in each ZIP to all other ZIP-level outbreaks via downstream gravity-driven transmission, in ascending order, for 2003-04 (top) and 2007-08 (bottom). The value $C$ may be interpreted as the effective number of outbreaks triggered by seeding in a ZIP. For reference, there were 734 ZIPs in the analysis of the 2003-04 outbreak, and 716 ZIPs in the analysis of the 2007-08 outbreak. In 2003-04, Mandeville LA accounts for the vast majority of geographic transmission; it is represented by the single dot at the upper right corner of the upper plot. Only four ZIPs – Mandeville LA, Johnstown North PA, North Houston North TX, and Houston Main (1) TX – surpass the cutoff at $C = 5$. The geographic locations of these ZIPs are depicted in Fig 6.9. In 2007-08, ten ZIPs surpass the cutoff at $C = 5$. These are labelled with their geographic locations in Fig 6.9.


Fig. 6.9 Contributions $C$ of seeding in each ZIP to all other outbreaks via downstream gravity-driven transmission, by geographic location, for 2003-04 (top) and 2007-08 (bottom). The value $C$ may be interpreted as the effective number of outbreaks triggered by seeding in a ZIP. Disc area is proportional to $C$. In 2003-04, nearly all geographic transmission can be traced back to Mandeville LA. Seeding in three other ZIPs (also labelled) contributed to just over five ZIP-level outbreaks each through downstream gravity-driven geographic transmission. In 2007-08, seeding in ten ZIPs (labelled) contributed to at least five downstream outbreaks.

OR, are the most important, together effectively accounting for 513.3 of the 716 observed outbreaks through gravity-driven onward spread.

Table 6.2 Transmission hubs of the 2007-08 influenza outbreak in the United States

| Name | ZIP | Pop. size | $C$ | $\sigma$ | Onset date |
|---|---|---|---|---|---|
| Cleveland Vicinity OH | 440 | 854036 | 282.8 | 0.97 | 25 Nov |
| Inglewood Vicinity/Torrance CA | 902/905 | 1453468 | 124.8 | 0.75/0.82 | 25 Nov |
| Boise (ID) West OR | 979 | 32039 | 105.7 | 1.0 | 18 Nov |
| Worcester Main MA | 016 | 180202 | 40.4 | 0.69 | 29 Nov |
| El Paso Main TX | 799 | 706832 | 50.4 | 0.69 | 25 Nov |

## 6.5 Age-structured transmission of the 2003-04 and 2007-08 seasonal influenza outbreaks

Using the methods developed in Chapter 5, we calculate the symbolic transfer entropy (STE) between age groups during the 2003-04 and 2007-08 seasonal outbreaks. To do so, a span of 25 weeks that contains the outbreak of interest in each season is isolated. In 2003-04, this span is from 17 Aug 2003 to 1 Feb 2004, and in 2007-08 the span is from 21 Oct 2007 to 6 Apr 2008. The twelve age-stratified time series and the age-aggregated time series for each ZIP within this timespan are symbolised with a symbol length of $m = 3$. Using these symbolised time series, the within-ZIP STE is calculated between each age group's time series and the age-aggregated time series. These are depicted for 2003-04 and 2007-08 in Fig 6.12. For both epidemics, the pairwise STE is less pronounced than for 2009 (compare with Fig 5.15). For 2003-04, the 15-19 year-old age group appears to contribute most to overall transmission, while for 2007-08, the 10-14 year-old age group contributes most to overall transmission, as was the case for the 2009 pandemic. The within-ZIP pairwise STE is also calculated between each age group. These values are depicted for 2003-04 and 2007-08 in Fig 6.13. Again, the differences are less pronounced than they are for the 2009 pandemic (compare with Fig 5.16), though there is still some evidence that children contributed disproportionately to infection in most other age groups during both seasonal outbreaks.

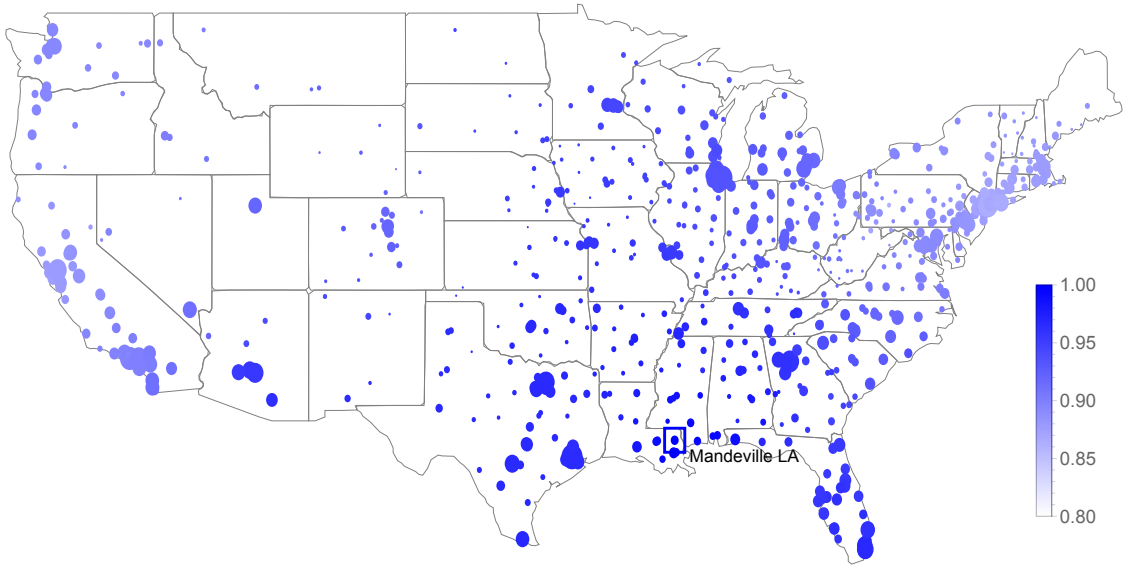



Fig. 6.10 Probability with which each ZIP's outbreak can be traced back to seeding in Mandeville LA in 2003-04. According to the geographic transmission model, Eq 6.1, downstream gravity-driven transmission from Mandeville can account for every other ZIP's outbreak with at least 80% probability. Dark blue indicates a high (close to 1) probability that downstream transmission from Mandeville sparked the ZIP's outbreak, and light blue indicates a lower (close to 0.8) probability that downstream transmission from Mandeville sparked the ZIP's outbreak.

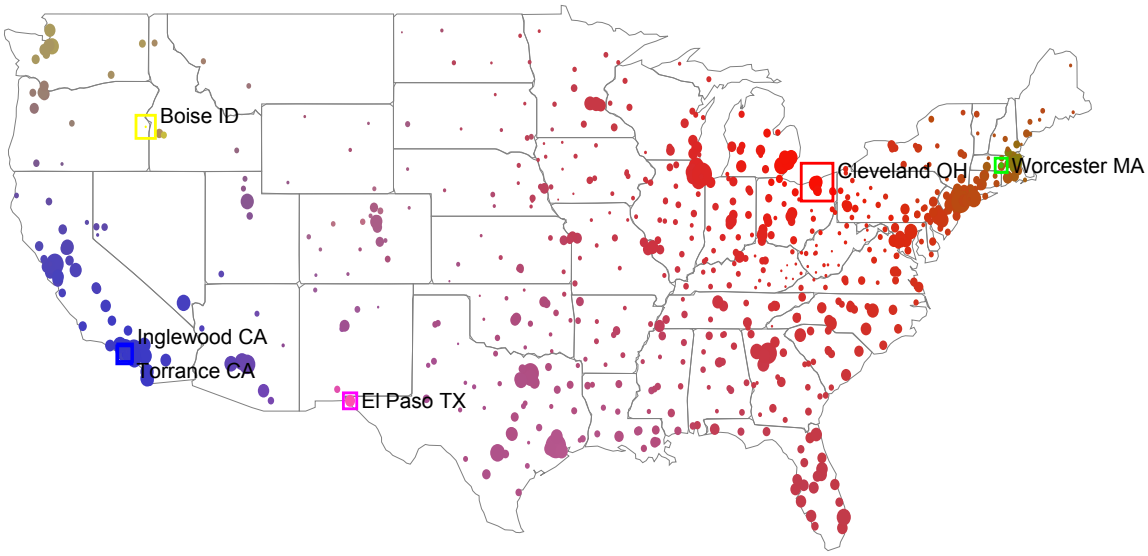



Fig. 6.11 Transmission hubs (boxed) and basins of infection for the 2007-08 seasonal influenza outbreak. Seeding events in each hub are assigned colours, indicated by the colour of the surrounding box. Discs are then coloured according to the probability with which their outbreak can be traced back to the seeding event in the hub of the corresponding colour. Colours are allowed to mix to depict mixed influence from multiple hubs. The prevailing red in the northeast indicates a predominant influence from Cleveland OH. The blue in California indicates a predominant influence from Inglewood Vicinity/Torrance CA.The green in the northwest indicates mixed influence from Boise ID and Inglewood Vicinity/Torrance CA, and the green in east Texas indicates mixing from Inglewood Vicinity/Torrance CA, El Paso TX, and Cleveland OH.

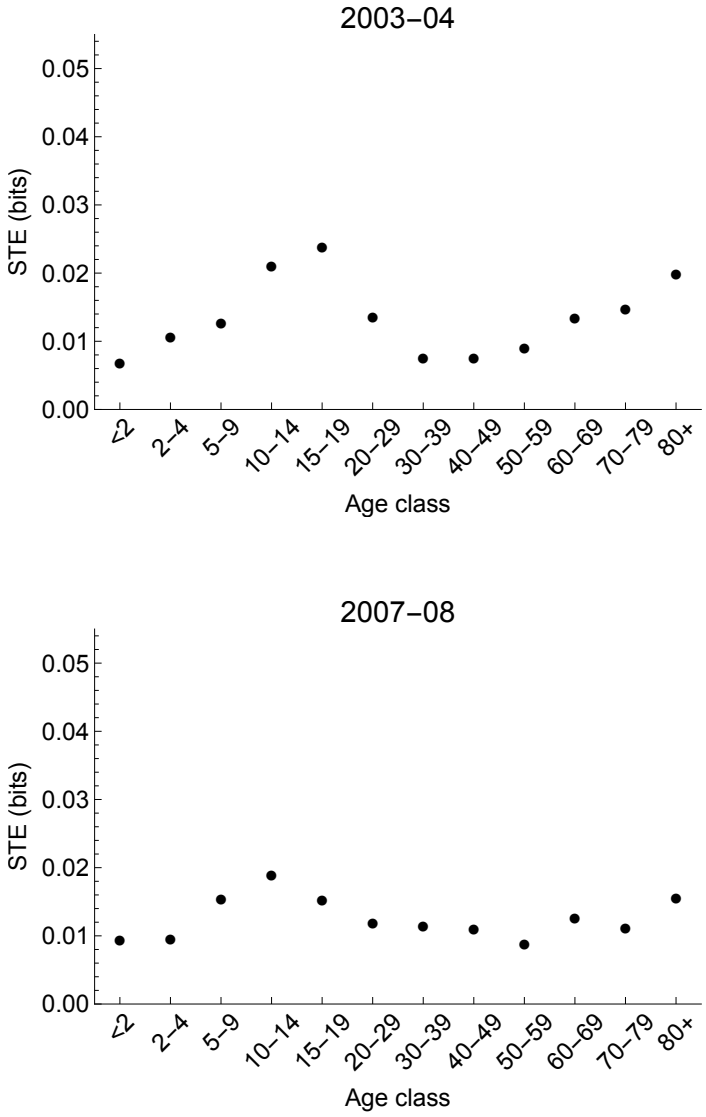



Fig. 6.12 Within-ZIP STE from each age group to the age-aggregated time series, for the 2003-04 (top) and 2007-08 (bottom) seasonal outbreaks. For both seasons, the peak STE is smaller than for the 2009 pandemic (compare with Fig 5.15). In 2003-04, the STE is highest from the 15-19 year-old age group, while for 2007-08, the STE is highest from the 10-14 year-old age group.

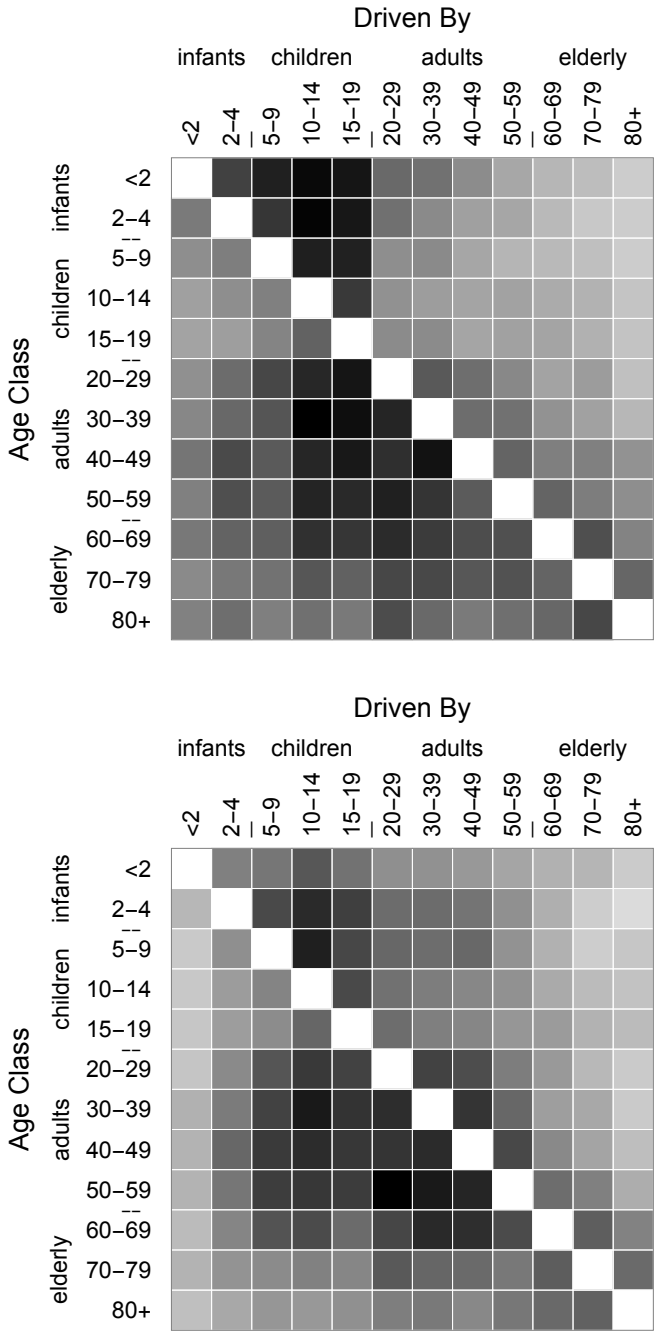



Fig. 6.13 Within-ZIP pairwise STE between age groups, for the 2003-04 (top) and 2007-08 (bottom) seasonal outbreaks. Darker colours correspond to higher STE. For both seasons, the differences in STE between the age groups are less pronounced than for the 2009 pandemic (compare with Fig 5.16). However, both still show some evidence of elevated STE from children to the other age groups. The maximum pairwise STE for 2003-04 is 0.060 bits, and the maximum pairwise STE for 2007-08 is 0.052 bits.

## 6.6 Correlations between antigenic prevalence and seeding from a hub

The Centers for Disease Control and Prevention (CDC) make available data on the weekly number of laboratory-confirmed influenza cases, together with the sampled viruses' antigenic types, for 10 regions in the United States. In 2007-08, three distinct antigenic subtypes co-circulated, making it possible to identify geographic differences in the prevalences of each strain. In this section, estimates of regional antigenic prevalences from the CDC data are correlated with the transmissive influence from each of the five hubs in 2007-08, providing a first hypothesis of which strains may have infected which hubs. Since the 2003-04 outbreak was caused predominately by a single strain and had a single transmission hub, the corresponding analysis for that outbreak is trivial and will not be considered further here.

### 6.6.1 Curating the antigenic data

The antigenic data available from the CDC (see §2.7) contain weekly counts of laboratory-confirmed influenza cases with their antigenic subtype (A/H1, A/H3, A/Unsubtyped, or B) collected from the 10 US Department of Health and Human Services (HHS) regions in the United States. Fig 6.14 depicts these samples for 2007-08 as a time series for each region. The incidence of each antigenic subtype varies by geographic region and over time. Region 4 (southeast), for example, is dominated throughout by antigenic subtype A/H3, while the outbreaks in Regions 9 and 10 (west) are dominated early by subtype A/H1 and later by subtype A/H3 and type B. Also depicted in Fig 6.14 is the date before which the earliest 10% of each region's ZIP-level outbreak onset times lie, to give a rough estimate of the epidemic onset time in the region as a whole. Visually, these onset times match well with the initial rises in incidence in the virologic data.

Since the geographic transmission model only captures the dynamics of the initial invasion wave of infection, we restrict our attention to the relative antigenic prevalences in each region prior to the epidemic peak. This helps to avoid effects from potential later waves of infection that may have featured different subtypes. In particular, we identify the cumulative incidence of each strain in each region prior to the week in which the maximum number of laboratory-confirmed cases of all subtypes was observed. For a region $h$, denote the number of observed pre-peak laboratory-confirmed cases of each antigenic subtype as

$$\boldsymbol{n}_h = (n_{h\mathrm{AH1}}, n_{h\mathrm{AH3}}, n_{h\mathrm{AU}}, n_{h\mathrm{B}}). \tag{6.5}$$

These observation vectors will be used to determine the antigenic subtypes that may have been responsible for triggering the outbreaks in the transmission hubs in 2007-08. The final results are essentially unaffected if the statistical analysis is performed instead using the cumulative incidences of each subtype during the first two months after $10^{\text{th}}$-percentile ZIP-level outbreak onset in each region (vertical bars in Fig 6.14) or across the entire season; the directions all trends remain the same, but some significances diminish.

### 6.6.2 Regression analysis

If each hub's outbreak is assumed to have been triggered by a single influenza virus subtype, one might expect that subtype to be highly represented in downstream outbreaks triggered by that hub. If this is true, one might then also expect that the relative number of laboratory-confirmed cases of that subtype in each HHS region would correlate with the expected number of outbreaks the hub triggered in each region. Here, the relative cumulative incidence of each antigenic subtype in each HHS region is regressed against the relative expected transmissive contribution from each hub, as estimated by the mechanistic transmission model, Eq 6.1. For a given hub and antigenic subtype, a positive correlation between expected transmissive influence from the hub and prevalence of the subtype gives evidence that that subtype may have been responsible for infecting the hub.

The first step requires calculating the expected fraction of ILI cases in region $h$ caused by gravity-driven transmission from seeding in hub $j$. This is

$$f_{h,j} = \frac{\sum_{i \in H} \mathbf{P}_{i,j} N_i}{\sum_{i \in H} N_i}, \tag{6.6}$$

where $H$ is the set of all ZIPs in region $h$, matrix element $\mathbf{P}_{i,j}$ gives the proportion of ZIP $i$'s outbreak attributable to seeding in ZIP $j$ (see derivation in §4.2), and $N_i$ is the population size of ZIP $i$.

Next, the fraction of pre-peak laboratory-confirmed cases caused by each strain are identified. For this preliminary analysis, it is assumed that the ratio of A/H1 cases to A/H3 cases in each region's A/Unsubtyped group is the same as the ratio of A/H1 cases to A/H3 cases in the subtyped cases. That is, the effective number of pre-peak laboratory-confirmed A/H1 cases in region $h$ is

$$\tilde{n}_{h\text{AH1}} = n_{h\text{AH1}} + \left( \frac{n_{h\text{AH1}}}{n_{h\text{AH1}} + n_{h\text{AH3}}} \right) n_{h\text{AU}}, \tag{6.7}$$

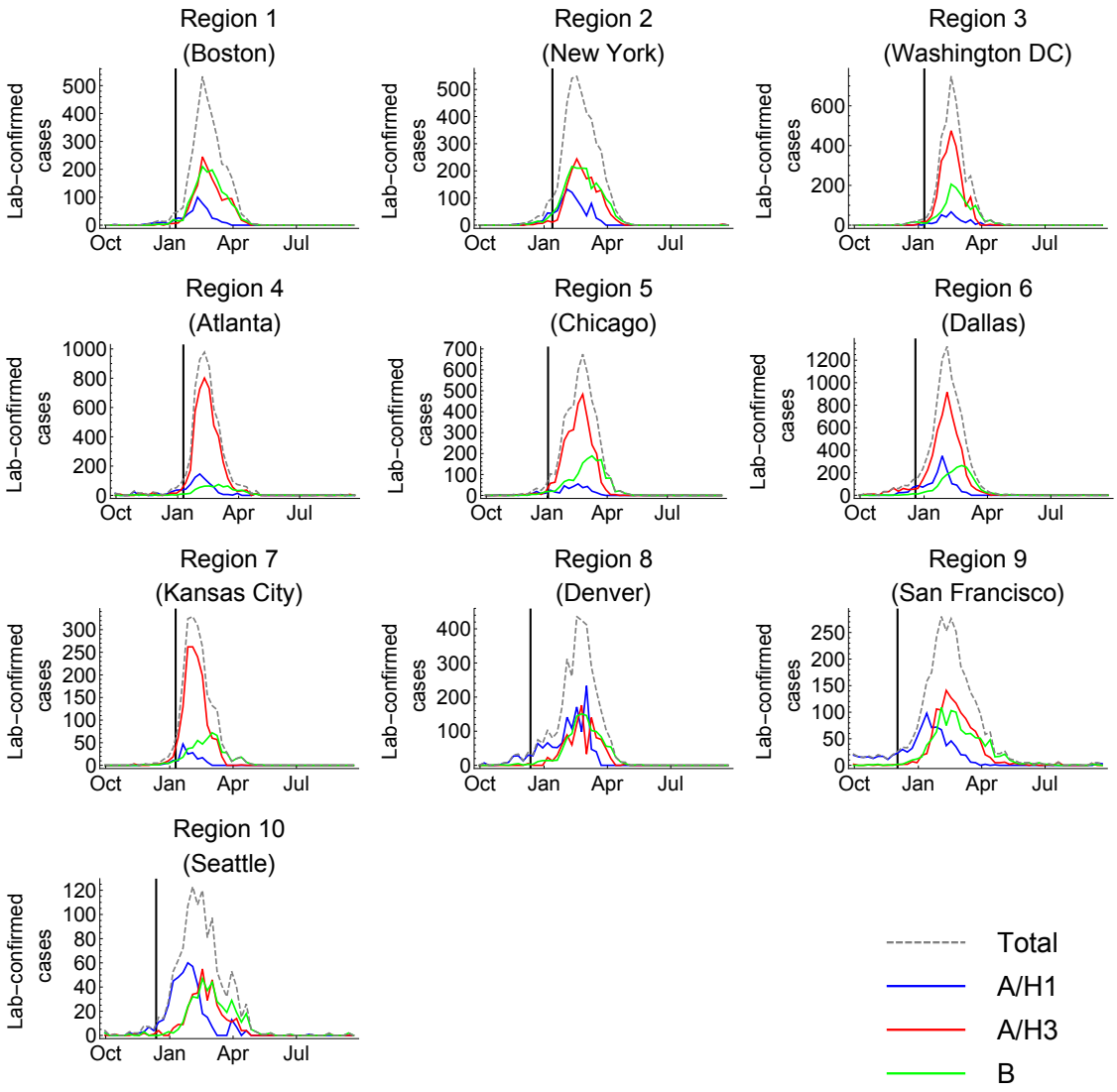



Fig. 6.14 Weekly counts of laboratory-confirmed cases of influenza subtypes A/H1 (blue), A/H3 (red), and B (green) in 2007-08 in the 10 HHS regions, as reported by the CDC [45]. Some type A strains processed by clinical laboratories do not undergo further subtyping (see §2.7); these are depicted in yellow. Dashed grey lines depict the cumulative strain counts. Subtype A/H3 was dominant in Regions 4-7, while Regions 9 and 10 suffered early outbreaks of subtype A/H1. The vertical black bars depict the $10^{\text{th}}$ percentile ZIP-level breakpoint outbreak onset times for each region, giving a rough estimate of the epidemic onset time in the region as a whole.

and the effective number of pre-peak A/H3 cases is

$$\tilde{n}_{h\mathrm{AH3}} = n_{h\mathrm{AH3}} + \left(\frac{n_{h\mathrm{AH3}}}{n_{h\mathrm{AH1}} + n_{h\mathrm{AH3}}}\right) n_{h\mathrm{AU}}. \tag{6.8}$$

After assigning the A/Unsubtyped cases in this way, the relative prevalences of each strain type in each region are calculated as

$$\frac{\tilde{n}_{h\mathrm{AH1}}}{\tilde{n}_{h\mathrm{AH1}} + \tilde{n}_{h\mathrm{AH3}} + n_{h\mathrm{B}}}, \quad \frac{\tilde{n}_{h\mathrm{AH3}}}{\tilde{n}_{h\mathrm{AH1}} + \tilde{n}_{h\mathrm{AH3}} + n_{h\mathrm{B}}}, \quad \text{and} \quad \frac{n_{h\mathrm{B}}}{\tilde{n}_{h\mathrm{AH1}} + \tilde{n}_{h\mathrm{AH3}} + n_{h\mathrm{B}}} \tag{6.9}$$

for types A/H1, A/H3, and B, respectively.

Finally, these relative prevalences are plotted and regressed against the expected fraction of cases $f_{h,j}$ in region $h$ triggered by hub $j$, for each of the five hubs in 2007-08. These scatters are depicted in Fig 6.15. The plotted lines indicate associations with a $p$-value below 0.1.

There is a strong positive correlation between transmission from Worcester MA and a high prevalence of antigenic type B, between transmission from Cleveland OH and a high prevalence of antigenic type A/H3, and between transmission from Inglewood Vicinity/Torrance CA and Boise ID and a high prevalence of antigenic type A/H1. There is a strong negative correlation between transmission from Cleveland OH and a high prevalence of antigenic type A/H1, and a strong negative correlation between transmission from El Paso TX and a high prevalence of antigenic type B. There are moderate negative correlations between transmission from Inglewood Vicinity/Torrance CA and Boise ID and a high prevalence of antigenic type A/H3, and a weak negative correlation between transmission from Boise ID and a high prevalence of antigenic type B. The $p$-values for all of these regressions are given in Table 6.3. This suggests that subtype A/H1 may have been responsible for seeding the transmission hubs in Boise ID and Inglewood Vicinity/Torrance CA, that subtype A/H3 may have been responsible for seeding the transmission hub in Cleveland OH, and that subtype B may have been responsible for seeding the transmission hub in Worcester MA. Subtype A/H1 was probably not responsible for seeding the outbreak in Cleveland OH, subtype A/H3 was probably not responsible for seeding the outbreaks in Inglewood Vicinity/Torrance CA or Boise ID, and subtype B was probably not responsible for seeding the outbreaks in El Paso TX or Boise ID.

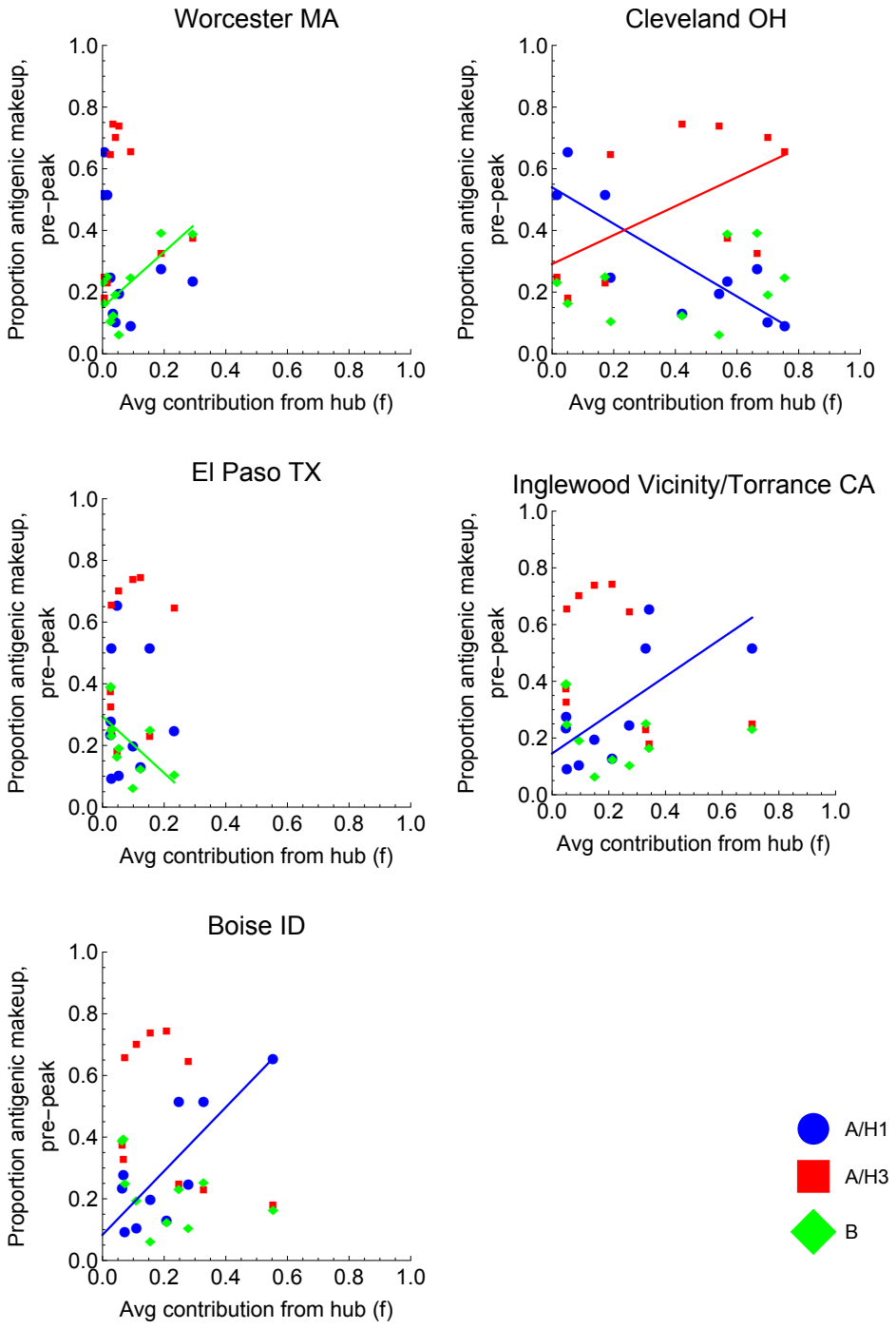



Fig. 6.15 Scatter plots of the relative prevalence of antigenic types A/H1 (blue circles), A/H3 (red squares), and B (green diamonds) in each HHS region vs. the relative transmissive contribution from a given transmissive hub to that region, for all five transmission hubs of the 2007-08 influenza outbreak. The plotted lines depict the linear least squares fits for all regressions with $p$-value below 0.1. Line colour corresponds to the antigenic type, with blue for A/H1, red for A/H3, and green for B. There is a positive relationship between prevalence of antigenic type B and transmissive influence from Worcester MA, between prevalence of antigenic type A/H3 and transmissive influence from Cleveland OH, and between prevalence of antigenic type A/H1 and transmissive influence from Inglewood Vicinity/Torrance CA and Boise ID. There is a negative relationship between prevalence of antigenic type A/H1 and transmissive influence from Cleveland OH.

Table 6.3 Regression $p$-values and direction of relationship (positive or negative) between antigenic prevalence and transmissive influence from each hub, across the 10 HHS regions in the United States. Significant ($p < 0.1$) values are in bold.

| Transmission hub | A/H1 | A/H3 | B |
|---|---|---|---|
| Worcester MA | 0.35 (–) | 0.82 (–) | **0.010 (+)** |
| Cleveland OH | **0.0026 (–)** | **0.089 (+)** | 0.40 (+) |
| El Paso TX | 0.94 (–) | 0.40 (+) | **0.081 (–)** |
| Inglewood Vicinity/Torrance CA | **0.024 (+)** | 0.17 (–) | 0.48 (–) |
| Boise ID | **0.0049 (+)** | 0.16 (–) | 0.22 (–) |

## 6.7 Discussion

In this chapter, the geographic transmission model developed in Chapter 3 is fit to epidemic onset times from the 2003-04 and 2007-08 influenza outbreaks in the United States. In spite of the many differences between these two seasonal outbreaks and the 2009 pandemic, including epidemic timing, duration, and responsible strain(s), the best transmission model for all three epidemics takes the same general form. In all three seasons, transmission is described by gravity-driven transmission modulated by recipient, but not donor, ZIP population size, as well as by surrounding population density. However, the values for the model parameters differ between seasons, capturing differences in transmission strength and the characteristic distance of gravity-driven geographic spread. The transmissibility ($\beta_d$) is highest in 2003-04 and lowest in 2007-08. The high transmissibility in 2003-04 may be explained by the fact that the outbreak was caused by a novel type of an especially virulent influenza strain, A/H3, to which underlying population immunity was low. Furthermore, the outbreak, though strong, was not classified as a pandemic, so the general public's vigilance may not have been as high as it was in 2009, leading to higher contract rates and increased transmission. On the other hand, in 2007-08, all three circulating strains were related to strains that had circulated in previous years, so underlying immunity was likely higher than in 2003-04 or in 2009, reducing transmissibility. The distance kernels in 2003-04 and in 2007-08 feature shorter characteristic distances ($\rho$) than the 2009 pandemic and thicker tails (smaller $\gamma$). This could be explained by differences in social behaviour during the pandemic vs. the seasonal outbreaks. Individuals may have been more likely to make long-distance trips during the seasonal outbreaks, leading to a lower $\gamma$. Also, the average age of infection is generally higher for outbreaks dominated by strain type A/H3 (the dominant strain in 2003-04 and 2007-08) than by types A/H1 (the dominant strain in 2009) and B [18]. This may have

further increased the impact of long-distance travel, which is likely more frequent for older individuals, during the two seasonal outbreaks. The population density normalisation term ($\varepsilon$) is smaller in 2007-08 than it is in 2003-04 or in 2009. Smaller $\varepsilon$ is associated with a higher force of infection for ZIPs in densely-populated areas relative to equivalent ZIPs in sparsely-populated areas. This difference in $\varepsilon$ may be attributed to underlying immunity, which was present in 2007-08 but largely absent in 2003-04 and 2009. For a strain to spark an outbreak in a population that already has some underlying immunity to it, an especially high frequency of interpersonal contacts (which are likely more frequent in densely-populated areas) may be required, compared with a strain to which essentially everyone is vulnerable. The estimated values for the gravity population size factor $\mu$ and the ILI time series normalisation factor $\theta$ are fairly consistent across all three seasons; transmission is roughly related to the fourth root of the recipient population size and the square root of the normalised ILI intensity.

It appears that the 2003-04 and autumn 2009 outbreaks have much in common, including the abnormally early time of year when they began circulating in the US, their transmission hubs in small- to mid-sized cities, and their marked wave-like geographic transmission patterns emanating from the southeastern US. The 2007-08 outbreak differed from these other two outbreaks on all three counts. It remains unclear why this might be the case. One possibility is that the 2003-04 and 2009 strains were able to invade earlier in the year, due to low levels of underlying immunity. This means that the 2003-04 and 2009 outbreaks may have spread during a time when influenza transmission was 'sub-optimal'; that is, indoor crowding, humidity, and temperature, among other factors, may not have been as conducive to influenza transmission as they tend to be in the later winter months. Studying the geographic transmission patterns of other influenza seasons may help unpick which of these factors may matter most to the geographic spread of influenza, and whether there is indeed a split regime between 2003-04- and 2009-like outbreaks and other 'regular' seasonal outbreaks.

It would be interesting to see whether the same model form persists across all seasons for which the IMS-ILI data are available, and how the model parameters compare across the seasons. In seasons for which the initial rise in ILI intensity is less severe, the breakpoint onset time estimates are less certain and, in general, fewer ZIPs can be retained in the analysis. However, one way to avoid omitting data entirely would be to account for neighbouring ILI intensity in the model's $n_{j,t}$ term, even if ZIP $j$'s onset was undetectable. Also, fitting the model parameters using an MCMC scheme, rather than the maximum likelihood strategy employed here, may be especially useful, since it would more explicitly account for uncertainty in outbreak onset time. In the weaker seasons, this onset uncertainty is likely to have a greater impact on the model inferences.

The transmissibility surfaces fit to 2003-04 and 2007-08 reveal further temporal and geographic variation in transmission in both of those seasons. In 2003-04, there is a band of increased transmissibility between the southeastern US and the northeast. The map of outbreak onset times (Fig 6.1) for that season reveals relatively early jumps to that band of the US, suggesting that the transmissibility surface is capturing a true effect. In 2007-08, there is a small patch of decreased susceptibility near the borders of Kentucky, West Virginia, and Virginia, as well as near Chicago and in the northeast, around New York City. These might correspond to underlying immunity in these areas due to previous outbreaks.

There is essentially no significant temporal variation in the transmissibility of the 2003-04 outbreak (see Fig 6.4). For the 2007-08 outbreak, however, there is a clear dip in transmissibility in late December, followed by a slight increase in early January. This might follow the school term; as schools closed for the winter vacation, transmissibility dipped, and then increased again when schools opened for the new term at the beginning of the calendar year. The 2003-04 outbreak, which struck significantly earlier, would not have been impacted by this effect.

Remarkably, only one transmission hub is identified for the 2003-04 outbreak. Since the outbreak spread quickly, especially compared with the 2009 pandemic, there may have simply been less time for multiple introductions of infection. Also, the parameter fits for the transmission model indicate that the transmission kernel has a thicker tail than the kernel for the 2009 pandemic, so any mid-distance jumps that occurred are more likely to be attributed to gravity-driven spread, rather than seeding. Following the pattern observed in the 2009 pandemic, the transmission hub for 2003-04 is a mid-sized town in the southeast. However, Houston TX was infected quickly thereafter, potentially facilitating onward spread. This agrees with data from the CDC that indicates that Texas was the first state to report influenza activity in 2003 [37].

Five transmission hubs are identified for the 2007-08 outbreak. In contrast to the 2003-04 outbreak and the 2009 pandemic, each of these cities is a major regional population centre. Worcester MA is close to Boston, and after Boston is the second most populous city in New England. Inglewood and Torrance CA both lie within greater Los Angeles. Cleveland and Boise are the largest and second-largest cities in their respective states. El Paso TX is a major southwestern city, and lies adjacent to Ciudad Juárez, the largest city in the Mexican state of Chihuahua. This matches more closely with the conventional notion that influenza spreads hierarchically from major cities into surrounding areas. It is unclear why a traditional seasonal outbreak might follow this pattern, while outbreaks of novel strains, as in 2003-04 and 2009, do not. This is an important area for further research.

Measuring the symbolic transfer entropy between the age-stratified time series for the 2003-04 and 2007-08 outbreaks suggests that the role of children was qualitatively similar for the two seasonal outbreaks as it was for the 2009 A/H1N1pdm pandemic. In all three outbreaks, there is evidence of elevated transmission from school-aged children to most other age groups. The dominant age group for the 2003-04 outbreak may have been just older than the dominant age group for the 2007-08 and 2009 epidemics, at 15-19 years vs. 10-14 years. This may reflect an authentic difference in the average age of infection between those outbreaks; indeed, influenza subtype A/H3N2, which dominated the 2003-04 outbreak, is associated with a higher average age of infection than are subtypes A/H1N1 and B [18]. The overall ranges of the STE values for 2003-04 and 2007-08 are smaller than for the 2009 pandemic, but this does not necessarily imply that transmission during the two seasonal outbreaks was less affected by age structure. Instead, it is possible that elevated reporting rates and relatively more acute outbreaks during the pandemic lead to more pronounced observed differences in STE. In the simulation studies presented in Chapter 5, previous immunity was not taken into account, and so it is unclear how STE might respond to the underlying immunity that may have affected the transmission of the 2003-04 and 2007-08 influenza outbreaks. It is possible that this underlying immunity could explain why the between-age-group differences in STE were less pronounced in 2003-04 and 2007-08 than in 2009. This would be an interesting area for further theoretical work.

Like the 2009 pandemic, the 2003-04 outbreak was dominated by a single antigenic subtype. During the 2007-08 outbreak, however, three strains co-circulated. An exploratory statistical analysis shows that a high relative prevalence of antigenic type A/H1 is associated with regions whose outbreaks can be traced predominately to hubs in the western half of the US (Inglewood Vicinity/Torrance CA and Boise ID); that the relative prevalence of antigenic type A/H3 is associated with regions seeded predominately by Cleveland OH; and that high relative prevalence of type B is associated with regions whose outbreaks can be traced predominately back to Worcester MA.

An obvious next step is to develop a full statistical model to infer which viral subtype triggered each hub's outbreak. This would involve specifying the probability of observing some collection of laboratory-confirmed subtypes in each region, given the subtypes that infected each hub, and then working backwards to identify the strain-hub assignments that maximise the likelihood of observing the true laboratory-confirmed data. It is tempting to use the matrix $P$ to estimate the expected subtype prevalences in each region, since $P$ gives the expected onward transmissive influence from each hub. However, the matrix $P$ does not adequately account for the dependence structure between city-level outbreaks. Consider, for

example, the transmission network depicted in Fig 6.16. Circles represent cities, and arrow widths are proportional to the forces of infection between them. Consider a scenario in which locations 1 and 2 have simultaneous outbreak onsets, followed by location 3, then location 4, and so on. Calculating $P$ for this scenario would reveal that locations 1 and 2 contribute roughly equally to all of the downstream outbreaks; that is, $C$ for locations 1 and 2 would be about 3.5 each for this system of 7 cities. However, if we imagine that locations 1 and 2 are infected by different viral subtypes, and that each downstream city-level outbreak is caused by the subtype of the city that directly infected it, then the strain that manages to infect city 3 will likely be highly represented in all downstream outbreaks. So, while the *expected* relative prevalence of each strain in this system, captured by $P$, will be around 50%, in reality it is likely that six of the seven city-level outbreaks will have been caused by single strain. This must be taken into account when calculating the strain-hub likelihoods.

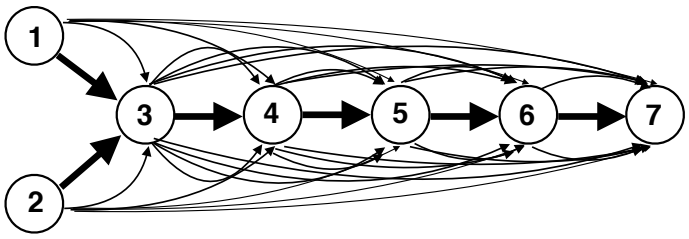



Fig. 6.16 Example transmission network to illustrate an extreme case of dependence between outbreaks. Circles represent cities, and arrows represent possible routes of transmission. Arrow widths represent the force of infection along that transmission route, as would be specified by the geographic transmission model, Eq 6.1. The epidemic depicted here begins simultaneously in cities 1 and 2, then spreads to city 3, then to city 4, and so on. In this scenario, each city's outbreak can be traced back to either city 1 or city 2 with about 50% probability each. However, this does not imply that the strains responsible for infecting cities 1 and 2 will be found in roughly equal quantities in the downstream outbreaks; instead, it is far more likely that whichever strain infects location 3 will dominate the epidemic overall. This has important implications when developing a statistical strategy to infer which strain infected which hub.

One possible strategy for incorporating this dependence structure is to simulate many possible "who infected whom" transmission trees. This may be done by randomly assigning a 'parent' to each ZIP outbreak, according to the transition probabilities in the reverse transmission network (see Fig 4.2). Then, for a given transmission tree, each ZIP's outbreak may be traced back to the hub from which its outbreak originated. Doing this for all ZIPs within an HHS region provides an estimate of the relative influence of each hub in the region, given the particular who-infected-whom transmission tree. Each hub may then be assigned

the subtype that best accounts for the observed laboratory-confirmed prevalences of each subtype across all regions. This may be repeated for many who-infected-whom networks, until a pattern emerges, with particular subtypes being repeatedly assigned to particular hubs. There are two key complications with this strategy, however. First, the number of possible epidemic trajectories is massive, on the order of $10^{2034}$, making it nearly impossible to sample the space of possible transmission trees completely enough to guarantee robust subtype-hub assignments. Second, some of the transmission trees assign all outbreaks in a region to just one or two hubs, and yet all three antigenic types (A/H1, A/H3, and B) are observed in the region. So, random effects need to be built into the model to account for outbreak "impurity", that is, that each ZIP-level outbreak does not consist of solely one antigenic type. It is difficult to say how exactly these random effects ought to be incorporated. Attempts to develop and make inferences from such a model have so far been unsuccessful, but this remains an important and interesting area for future work.

## 6.8 Summary

A geographic transmission model is fit to outbreak onset times inferred from ILI data in the United States from 2003-04 and 2007-08. The most parsimonious transmission model for both epidemics matches the overall form of the best model for the geographic transmission of the 2009 A/H1N1pdm influenza pandemic. Transmissibility of the 2003-04 outbreak was similar in magnitude to that of the 2009 outbreak, while the transmissibility of the 2007-08 outbreak was smaller, possibly reflecting an effect from underlying immunity. The gravity distance kernels for the seasonal outbreaks have thicker tails than the kernel for the pandemic outbreak, indicating a higher likelihood of long-distance jumps. Transmissibility varied geographically for both seasonal outbreaks, though the magnitude of this variation was smaller than for the 2009 pandemic. Transmissibility also varied over time in 2007-08, with a dip in late December and a peak in mid-January, possibly reflecting differences in contact rates associated with the winter holidays. Nearly all of the geographic transmission in 2003-04 can be attributed to a single transmission hub in Mandeville, LA. In 2007-08, five transmission hubs are identified, all of which are in or near large cities. Calculating the symbolic transfer entropy between age-stratified ILI time series from both the 2003-04 and 2007-08 influenza epidemics suggests that school-aged children likely contributed disproportionately to transmission during both seasons. For the 2007-08 outbreak, when three antigenically distinct strains circulated, the strains that may have infected each hub are identified using regressions. According to this analysis, cumulative incidence of strain

types A/H1 is correlated with seeding from Inglewood Vicinity/Torrance CA and Boise ID, cumulative incidence of type A/H3 is correlated with seeding from Cleveland OH, and cumulative incidence of type B is correlated with seeding from Worcester MA.

# Chapter 7

# Discussion and conclusions

By making use of a dataset that captures weekly city-level influenza-like illness incidence in the United States, I have characterised the autumn 2009 A/H1N1pdm pandemic influenza outbreak in the US in terms of its geographic transmission patterns, its establishment sites, and the age groups most responsible for sparking and sustaining its transmission. While the age-structured results align with established theory regarding the transmission of influenza – that the bulk of transmission is likely driven by children – the geographic results are somewhat more surprising. Despite the well-documented rapid international spread of the 2009 A/H1N1pdm influenza pandemic, the invasion wave of the autumn 2009 pandemic in the United States was relatively slow and cohesive, compared to the 2003-04 and 2007-08 seasonal outbreaks. Also, the geographic sites where the autumn 2009 pandemic first became established in the United States were generally mid-sized cities, not the major urban areas that conventional wisdom might expect to be epicentres of disease transmission. These findings highlight the difficulties involved with planning for pandemic influenza outbreaks, and indicate key areas for further research to prepare for the next emergence of a novel influenza virus.

## 7.1 Accounting for the establishment sites and transmission patterns of the autumn 2009 A/H1N1pdm epidemic in the US

In Chapters 3 and 4, the southeastern United States is identified as a key region for both the initial establishment and subsequent transmission of the autumn 2009 A/H1N1pdm epidemic in the US. This is not the first time the region has been identified as an important

site of influenza transmission: Charu *et al.* (2017) [48] find that seven of the eight influenza outbreaks in the US between 2003 and 2010 were likely seeded in the southeast, and Shaman *et al.* (2011) [209] note that the southeastern US suffered a third wave of pandemic influenza in 2010 from which the rest of the country was spared. Unfortunately, the IMS-ILI dataset concludes just before the onset of this third pandemic wave, so it cannot be studied in the same way as the earlier autumn 2009 wave is in this thesis. Shaman *et al.* (2011) [209] attribute the third pandemic wave to a local rise in absolute humidity in the southeastern US. Absolute humidity is not considered as a contributor to the geographic transmission of influenza in this thesis, largely because it was rejected as an important predictor of city-level influenza outbreak onset times in the geographic transmission models developed by Gog *et al.* (2014) [91] and Charu *et al.* (2017) [48]. However, the potential influence of absolute humidity on the geographic transmission of influenza is worth revisiting. It may be worthwhile to develop a 'risk of establishment' map based on potential predictors of local outbreak establishment, including absolute humidity, age-structured contact rates, and age-structured movement patterns. This could follow in the spirit of Pigott *et al.* (2014) [192], who estimate the risk of emergence of Ebola virus disease in West Africa according to a set of geographically-varying predictors. The Gaussian process transmissibility surfaces presented in Chapter 3 provide a first prediction of what this risk surface might look like. Regressing possible covariates against the posterior mean Gaussian process values may provide a way of identifying plausible predictors for the risk map. Unfortunately, little is known about age-structured contact and movement patterns in the US as a whole, let alone for particular subregions. An important area for future research, then, is the collection of geo-tagged and age-structured movement and contact data in the United States, possibly using POLYMOD-like surveys [170] and mobile phone geo-tracking. These data could help explain regional and inter-seasonal differences in the transmission dynamics of influenza, and possibly of other diseases as well.

## 7.2 Incorporating epidemiological and genetic data

In Chapter 4, a method is developed to identify an outbreak's geographic transmission hubs and their associated basins of infection. The method lays the groundwork for a new way of combining epidemiological and genetic data. The most straightforward avenue for further investigation would be to perform an independent genetic clustering analysis on 2009 A/H1N1pdm influenza virus sequences from the United States, to see if the basins of infection identified in Chapter 4 can be reproduced. Then, it may be possible to develop

a more unified statistical inferential framework, possibly using the basins of infection as a Bayesian prior estimate of the geographic variation in 2009 A/H1N1pdm strain prevalences, and using viral genome sequences to refine the estimate. This sort of analysis could help distinguish true transmission hubs from false ones, since the basins of infection associated with false transmission hubs should not be visible in the genetic data.

A second way of incorporating genetic data into the methods from Chapter 4 would be to use phylogenetic approaches to infer transmission links between the autumn 2009 transmission hubs and possible reservoirs of infection outside the US. This would follow in the vein of Lycett *et al.* (2012) [160], who infer the migration routes of 2009 A/H1N1pdm influenza between various countries and continents. Specifically, phylogenetic relationships between viral strains circulating in distinct basins of infection and in other parts of the world would be used to infer long-distance transmission routes of the autumn 2009 A/H1N1pdm outbreak. Knowing the basins of infection may enhance the inferential capability of Lycett *et al.*'s approach, since the basins are likely to be more epidemiologically relevant partitions of geographic space than geopolitical boundaries. This would flesh out the transmission history of the 2009 A/H1N1pdm influenza pandemic in the US: the genetic analysis would reveal how long-distance transmission to/between the hubs may have occurred, while the geographic transmission model from Chapter 3 would describe the onward short-distance transmission of infection from the hubs.

## 7.3 Linking individual-based and metapopulation disease dynamics

This thesis relies heavily on a metapopulation approach to disease modelling, though an individual-based simulation features briefly in the validation of the hub identification method in Chapter 4. Individual-based and metapopulation approaches are perhaps the two most common perspectives for modelling the geographic spread of infectious diseases [128, 200]. Both are likely to remain relevant; indeed, epidemiological data is nearly always collected at some scale of spatial aggregation, motivating the use of metapopulation models, yet individuals are ultimately responsible for spreading disease, so individual-based models provide a more faithful description of the underlying disease dynamics. Despite the widespread use of these two modelling paradigms, it remains unclear how epidemiological dynamics translate between them.

The transmission kernel is a concept present in both the individual-based and metapopulation perspectives, and so it offers a sensible starting place for exploring the relationship between epidemiological dynamics at the two scales. For individual-based models, the transmission kernel is usually interpreted as an individual's propensity to move over certain distances, while in the metapopulation framework, the transmission kernel describes a more abstract decay in infective force with increasing spatial separation [128]. There is no clear way of inferring the shape of a metapopulation transmission kernel from an individual-based movement kernel. The shape of a metapopulation transmission kernel is likely to depend intricately on the disease's infectiousness, the frequency and duration of individual-based movements, and the precise way in which the metapopulation patches are partitioned. Extensive simulation studies may shed light on the relationship between kernels at the two scales. In this thesis, and indeed in many epidemiological modelling analyses, the metapopulation paradigm is largely taken for granted. This thesis' findings, and spatial epidemiology as a whole, would be greatly enhanced by a better understanding of the links between individual-based and metapopulation epidemic models.

## 7.4 Inferring age-structured transmission from incidence time series

As demonstrated in Chapter 5, symbolic transfer entropy (STE) provides a way of determining which age groups drive the transmission of a disease from age-stratified incidence time series. While STE is only applied to ILI time series in this thesis, it would be interesting to see whether it can give plausible estimates of age-structured transmission for other diseases. However, STE and related methods are inherently limited; since they do not explicitly take into account underlying epidemiological dynamics, they can only provide a rough ordering of which age groups contribute most to transmission. Obtaining more precise estimates of age-structured differences in transmission strength will almost certainly require developing model-based methods to reliably estimate the next-generation matrix from incidence data that is separated into many age classes, possibly following the basic framework established by Glass *et al.* (2011) [90], or extending the inferential methods developed by Yang *et al.* (2015) [257] to accommodate multiple age groups. As classifications of population heterogeneity increase in resolution – for example, as disease incidence data become increasingly tagged by the sufferer's exact age and geographic location – it may be necessary to consider further generalisations of the next-generation matrix, which may include developing a

continuous-state next-generation operator or function. Inaba (2017) [120] provides an up-to-date discussion of work in this budding field. Ultimately, it will be necessary to unify age-structured and spatially-structured notions of the generalised reproduction number; for now, the best way of doing this remains an open question [200].

# References


[1] Abeku, T. A., Hay, S. I., Ochola, S., Langi, P., Beard, B., de Vlas, S. J., and Cox, J. (2004). Malaria epidemic early warning and detection in African highlands. *Trends in Parasitology*, 20(9):400–405.

[2] Adler, A. J., Eames, K. T., Funk, S., and Edmunds, W. J. (2014). Incidence and risk factors for influenza-like-illness in the UK: online surveillance using Flusurvey. *BMC Infectious Diseases*, 14(1):232.

[3] Aimone, F. (2010). The 1918 influenza epidemic in New York City: a review of the public health response. *Public health reports*, 125(Suppplement 3):71–9.

[4] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

[5] Andreasen, V. (1989). Disease regulation of age-structured host populations. *Theoretical Population Biology*, 36(2):214–239.

[6] Andreasen, V., Viboud, C., and Simonsen, L. (2008). Epidemiologic characterization of the 1918 influenza pandemic summer wave in Copenhagen: implications for pandemic control strategies. *The Journal of infectious diseases*, 197(2):270–278.

[7] Austerlitz, F., Dick, C. W., Dutech, C., Klein, E. K., Oddou-Muratorio, S., Smouse, P. E., and Sork, V. L. (2004). Using genetic markers to estimate the pollen dispersal curve. *Molecular Ecology*, 13(4):937–954.

[8] Baguette, M. (2004). The classical metapopulation theory and the real, natural world: A critical appraisal. *Basic and Applied Ecology*, 5(3):213–224.

[9] Bailey, N. T. (1957). *The Mathematical Theory of Epidemics*. Charles Griffin and Company Limited, London.

[10] Bajardi, P., Poletto, C., Ramasco, J. J., Tizzoni, M., Colizza, V., and Vespignani, A. (2011). Human Mobility Networks, Travel Restrictions, and the Global Spread of 2009 H1N1 Pandemic. *PLoS ONE*, 6(1):e16591.

[11] Banerjee, S. and Gelfand, A. E. (2006). Bayesian Wombling: Curvilinear Gradient Assessment Under Spatial Process Models. *Journal of the American Statistical Association*, 101(476):1487–1501.

[12] Barnett, L., Barrett, A. B., and Seth, A. K. (2009). Granger causality and transfer entropy Are equivalent for gaussian variables. *Physical Review Letters*, 103(23):2–5.

[13] Baskerville, E. B. and Cobey, S. (2017). Does influenza drive absolute humidity? *Proceedings of the National Academy of Sciences*, 114(12):E2270–E2271.

[14] Batty, M. and Sikdar, P. K. (1982a). Spatial aggregation in gravity models: 1. An information- theoretic framework. *Environment & Planning A*, 14:377–405.

[15] Batty, M. and Sikdar, P. K. (1982b). Spatial aggregation in gravity models: 2. One-dimensional population density models. *Environment and Planning A*, 14(4):525–553.

[16] Batty, M. and Sikdar, P. K. (1982c). Spatial aggregation in gravity models: 3. Two-dimensional trip distribution and location models. *Environment & Planning A*, 14(5):629–658.

[17] Batty, M. and Sikdar, P. K. (1982d). Spatial aggregation in gravity models: 4. Generalisations and large-scale applications. *Environment & Planning A*, 14(6):795–822.

[18] Bedford, T., Riley, S., Barr, I. G., Broor, S., Chadha, M., Cox, N. J., Daniels, R. S., Gunasekaran, C. P., Hurt, A. C., Kelso, A., Klimov, A., Lewis, N. S., Li, X., McCauley, J. W., Odagiri, T., Potdar, V., Rambaut, A., Shu, Y., Skepner, E., Smith, D. J., Suchard, M. a., Tashiro, M., Wang, D., Xu, X., Lemey, P., and Russell, C. a. (2015). Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature*, 523(7559):217–20.

[19] Bernoulli, D. (1766). Essai d'une nouvelle analyse de la mortalité causée par la petite verole, et des advantages de l'inoculation pour la prévenir. *Hist. et Mém. de l'Acad. Royale des Sciences de Paris*, pages 1–45.

[20] Biek, R., Pybus, O. G., Lloyd-Smith, J. O., and Didelot, X. (2015). Measurably evolving pathogens in the genomic era. *Trends in Ecology & Evolution*, 30(6):306–313.

[21] Biggerstaff, M., Cauchemez, S., Reed, C., Gambhir, M., and Finelli, L. (2014). Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature. *BMC Infectious Diseases*, 14(1):480.

[22] Biggerstaff, M., Jhung, M., Kamimoto, L., Balluz, L., and Finelli, L. (2012). Self-reported influenza-like illness and receipt of influenza antiviral drugs during the 2009 pandemic, United States, 2009-2010. *American Journal of Public Health*, 102(10):2009–2010.

[23] Boëlle, P.-Y., Ansart, S., Cori, A., and Valleron, A.-J. (2011). Transmission parameters of the A/H1N1 (2009) influenza virus pandemic: a review. *Influenza and Other Respiratory Viruses*, 5(5):306–316.

[24] Borge-Holthoefer, J., Perra, N., Goncalves, B., Gonzalez-Bailon, S., Arenas, A., Moreno, Y., and Vespignani, A. (2016). The dynamics of information-driven coordination phenomena: A transfer entropy analysis. *Science Advances*, 2(4):e1501158–e1501158.

[25] Bouvier, N. M. and Palese, P. (2008). The biology of influenza viruses. *Vaccine*, 26(SUPPL. 4):49–53.

[26] Britton, T. (2010). Stochastic epidemic models: A survey. *Mathematical Biosciences*, 225(1):24–35.

[27] Brownstein, J. S., Kleinman, K. P., and Mandl, K. D. (2005). Identifying pediatric age groups for influenza vaccination using a real-time regional surveillance system. *American Journal of Epidemiology*, 162(7):686–693.

[28] Cabinet Office of the United Kingdom (2015). National Risk Register of Civil Emergencies, 2015 Edition. Technical report, Cabinet Office of the United Kingdom, London.

[29] Caley, P., Becker, N. G., and Philip, D. J. (2007). The waiting time for inter-country spread of pandemic influenza. *PLoS ONE*, 2(1).

[30] Carrat, F., Vergu, E., Ferguson, N. M., Lemaitre, M., Cauchemez, S., Leach, S., and Valleron, A. J. (2008). Time lines of infection and disease in human influenza: A review of volunteer challenge studies. *American Journal of Epidemiology*, 167(7):775–785.

[31] Castillo-Chavez, C., Hethcote, H. W., Andreasen, V., Levin, S. A., and Liu, W. M. (1989). Epidemiological models with age structure, proportionate mixing, and cross-immunity. *Journal of Mathematical Biology*, 27(3):233–258.

[32] Cauchemez, S., Donnelly, C. A., Reed, C., Ghani, A. C., Fraser, C., Kent, C. K., Finelli, L., and Ferguson, N. M. (2009a). Household Transmission of 2009 Pandemic Influenza A (H1N1) Virus in the United States. *New England Journal of Medicine*, 361(27):2619–2627.

[33] Cauchemez, S., Ferguson, N. M., Wachtel, C., Tegnell, A., Saour, G., Duncan, B., and Nicoll, A. (2009b). Closure of schools during an influenza pandemic. *The Lancet Infectious Diseases*, 9(8):473–481.

[34] Centers for Disease Control and Prevention (2004a). 2003-04 U.S. Influenza Season Summary. Technical report, Centers for Disease Control and Prevention, Atlanta, GA.

[35] Centers for Disease Control and Prevention (2004b). Preliminary Assessment of the Effectiveness of the 2003-04 Inactivated Influenza Vaccine — Colorado, December 2003. Technical report, Centers for Disease Control and Prevention, Atlanta, GA.

[36] Centers for Disease Control and Prevention (2004c). Update: Influenza Activity – United States, 2003-04 Season. Technical report, Centers for Disease Control and Prevention, Atlanta, GA.

[37] Centers for Disease Control and Prevention (2004d). Update: Influenza Activity – United States and Worldwide, 2003-04 Season, and Composition of the 2004-05 Influenza Vaccine. Technical report, Centers for Disease Control and Prevention, Atlanta, GA.

[38] Centers for Disease Control and Prevention (2008a). Flu Season Summary (September 30 2007 - May 17 2008). Technical report, Centers for Disease Control and Prevention, Atlanta, GA.

[39] Centers for Disease Control and Prevention (2008b). Influenza Activity – United States and Worldwide, 2007-08 Season. Technical report, Centers for Disease Control and Prevention, Atlanta, GA.

[40] Centers for Disease Control and Prevention (2010). The 2009 H1N1 Pandemic: Summary Highlights, April 2009-April 2010. Technical report, Centers for Disease Control and Prevention, Atlanta, GA.

[41] Centers for Disease Control and Prevention (2011). Influenza Symptoms and the Role of Laboratory Diagnostics.

[42] Centers for Disease Control and Prevention (2014). How Flu Spreads.

[43] Centers for Disease Control and Prevention (2016a). Overview of Influenza Surveillance in the United States. Technical report, Centers for Disease Control and Prevention.

[44] Centers for Disease Control and Prevention (2016b). Types of Influenza Viruses.

[45] Centers for Disease Control and Prevention (2017a). FluView National and Regional Level Outpatient Illness and Viral Surveillance.

[46] Centers for Disease Control and Prevention (2017b). Influenza Type A Viruses.

[47] Chao, D. L., Halloran, M. E., and Longini, I. M. (2010). School opening dates predict pandemic influenza A(H1N1) outbreaks in the United States. *The Journal of Infectious Diseases*, 202(6):877–80.

[48] Charu, V., Zeger, S., Gog, J., Bjørnstad, O. N., Kissler, S., Simonsen, L., Grenfell, B. T., and Viboud, C. (2017). Human mobility and the spatial transmission of influenza in the United States. *PLOS Computational Biology*, 13(2):e1005382.

[49] Chowell, G., Echevarría-Zuno, S., Viboud, C., Simonsen, L., Tamerius, J., Miller, M. A., and Borja-AburtoVí, V. H. (2011a). Characterizing the Epidemiology of the 2009 Influenza A/H1N1 Pandemic in Mexico. *PLoS Medicine*, 8(5).

[50] Chowell, G., Viboud, C., Munayco, C. V., Gómez, J., Simonsen, L., Miller, M. A., Tamerius, J., Fiestas, V., Halsey, E. S., and Laguna-Torres, V. A. (2011b). Spatial and Temporal Characteristics of the 2009 A/H1N1 Influenza Pandemic in Peru. *PLoS ONE*, 6(6):e21287.

[51] Chowell, G., Viboud, C., Simonsen, L., Miller, M. A., and Acuna-Soto, R. (2010). Mortality patterns associated with the 1918 influenza pandemic in Mexico: evidence for a spring herald wave and lack of preexisting immunity in older populations. *The Journal of infectious diseases*, 202(4):567–575.

[52] Clark, N. and Lynch, J. (2011). Influenza: Epidemiology, Clinical Features, Therapy, and Prevention. *Seminars in Respiratory and Critical Care Medicine*, 32(04):373–392.

[53] Coburn, B. J., Wagner, B. G., and Blower, S. (2009). Modeling influenza epidemics and pandemics: insights into the future of swine flu (H1N1). *BMC Medicine*, 7(1):30.

[54] Cohen, I. B. (1984). Florence Nightingale. *Scientific American*, 250(3):128–137.

[55] Colizza, V., Barrat, A., Barthelemy, M., Valleron, A. J., and Vespignani, A. (2007a). Modeling the worldwide spread of pandemic influenza: Baseline case and containment interventions. *PLoS Medicine*, 4(1):0095–0110.

[56] Colizza, V., Pastor-Satorras, R., and Vespignani, A. (2007b). Reaction-diffusion processes and metapopulation models in heterogeneous networks. *Nature Physics*, 3(April):22.

[57] Conlan, A. J. K., Eames, K. T. D., Gage, J. A., von Kirchbach, J. C., Ross, J. V., Saenz, R. A., and Gog, J. R. (2011). Measuring social networks in British primary schools through scientific engagement. *Proceedings. Biological sciences / The Royal Society*, 278(1711):1467–75.

[58] Conlan, A. J. K. and Grenfell, B. T. (2007). Seasonality and the persistence and invasion of measles. *Proceedings. Biological sciences / The Royal Society*, 274(1614):1133–1141.

[59] Cooper, B. S., Pitman, R. J., Edmunds, W. J., and Gay, N. J. (2006). Delaying the international spread of pandemic influenza. *PLoS Medicine*, 3(6):0845–0855.

[60] Cyranoski, D. and Abbott, A. (2003). Virus detectives seek source of SARS in China's wild animals. *Nature*, 423(6939):467–467.

[61] Dawood, F. S., Iuliano, A. D., Reed, C., Meltzer, M. I., Shay, D. K., Cheng, P. Y., Bandaranayake, D., Breiman, R. F., Brooks, W. A., Buchy, P., Feikin, D. R., Fowler, K. B., Gordon, A., Hien, N. T., Horby, P., Huang, Q. S., Katz, M. A., Krishnan, A., Lal, R., Montgomery, J. M., Mølbak, K., Pebody, R., Presanis, A. M., Razuri, H., Steens, A., Tinoco, Y. O., Wallinga, J., Yu, H., Vong, S., Bresee, J., and Widdowson, M. A. (2012). Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: A modelling study. *The Lancet Infectious Diseases*, 12(9):687–695.

[62] Deyle, E. R., Maher, M. C., Hernandez, R. D., Basu, S., and Sugihara, G. (2016). Global environmental drivers of influenza. *Proceedings of the National Academy of Sciences of the United States of America*, 113(46):201607747.

[63] Didelot, X., Fraser, C., Gardy, J., Colijn, C., and Malik, H. (2017). Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Molecular Biology and Evolution*, 34(4):997–1007.

[64] Didelot, X., Gardy, J., and Colijn, C. (2014). Bayesian Inference of Infectious Disease Transmission from Whole-Genome Sequence Data. *Molecular Biology and Evolution*, 31(7):1869–1879.

[65] Diekmann, O., Heesterbeek, J. a., and Metz, J. a. (1990). On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations. *Journal of mathematical biology*, 28(4):365–382.

[66] Diekmann, O., Heesterbeek, J. A. P., and Roberts, M. G. (2010). The construction of next-generation matrices for compartmental epidemic models. *Journal of The Royal Society Interface*, 7(47):873–885.

[67] Dietz, K. (1975). Transmission and control of arboviruses. In Ludwig, D. and Cooke, K., editors, *Epidemiology*, pages 104–121. SIAM, Philadelphia.

[68] Dietz, K. and Heesterbeek, J. A. P. (2002). Daniel Bernoulli's epidemiological model revisited. *Mathematical Biosciences*, 180:1–21.

[69] Dietz, K. and Schenzle, D. (1985). Proportionate mixing models for age-dependent infection transmission. *Journal of Mathematical Biology*, 22(1):117–120.

[70] Diggle, P. (2014). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. CRC Press, Boca Raton, FL, 3rd edition.

[71] Dodd, P. J., Looker, C., Plumb, I. D., Bond, V., Schaap, A., Shanaube, K., Muyoyeta, M., Vynnycky, E., Godfrey-Faussett, P., Corbett, E. L., Beyers, N., Ayles, H., and White, R. G. (2016). Age- and Sex-Specific Social Contact Patterns and Incidence of Mycobacterium tuberculosis Infection. *American Journal of Epidemiology*, 183(2):156–166.

[72] Drummond, A. J. and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7(1):214.

[73] Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, 22(5):1185–1192.

[74] Dushoff, J., Plotkin, J. B., Levin, S. A., and Earn, D. J. D. (2004). Dynamical resonance can account for seasonality of influenza epidemics. *Proceedings of the National Academy of Sciences*, 101(48):16915–16916.

[75] Eggo, R. M., Cauchemez, S., and Ferguson, N. M. (2011). Spatial dynamics of the 1918 influenza pandemic in England, Wales and the United States. *Journal of the Royal Society, Interface*, 8(55):233–43.

[76] Ferguson, N. M., Cummings, D. A., Cauchemez, S., Fraser, C., Riley, S., Meeyai, A., Iamsirithaworn, S., and Burke, D. S. (2005). Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature*, 437(7056):209–214.

[77] Ferguson, N. M., Cummings, D. A. T., Fraser, C., Cajka, J. C., Cooley, P. C., and Burke, D. S. (2006). Strategies for mitigating an influenza pandemic. *Nature*, 442(7101):448–452.

[78] Filipe, J. A., Riley, E. M., Drakeley, C. J., Sutherland, C. J., and Ghani, A. C. (2007). Determination of the processes driving the acquisition of immunity to malaria using a mathematical transmission model. *PLoS Computational Biology*, 3(12):2569–2579.

[79] Fox, J. P. and Kilbourne, E. D. (1973). Epidemiology of Influenza : Summary of Influenza Workshop IV. *The Journal of Infectious Diseases*, 128(3):361–386.

[80] France, A. M., Jackson, M., Schrag, S., Lynch, M., Zimmerman, C., Biggerstaff, M., and Hadler, J. (2010). Household transmission of 2009 influenza A (H1N1) virus after a school-based outbreak in New York City, April-May 2009. *The Journal of infectious diseases*, 201(7):984–92.

[81] Fraser, C., Donnelly, C. A., Cauchemez, S., Hanage, W. P., Van Kerkhove, M. D., Hollingsworth, T. D., Griffin, J., Baggaley, R. F., Jenkins, H. E., Lyons, E. J., Jombart, T., Hinsley, W. R., Grassly, N. C., Balloux, F., Ghani, A. C., Ferguson, N. M., Rambaut, A., Pybus, O. G., Lopez-Gatell, H., Alpuche-Aranda, C. M., Chapela, I. B., Zavala, E. P., Guevara, D. M. E., Checchi, F., Garcia, E., Hugonnet, S., and Roth, C. (2009). Pandemic potential of a strain of influenza A (H1N1): early findings. *Science (New York, N.Y.)*, 324(5934):1557–61.

[82] Frost, S. D. W., Pybus, O. G., Gog, J. R., Viboud, C., Bonhoeffer, S., and Bedford, T. (2014). Eight challenges in phylodynamic inference. *Epidemics*, 10:88–92.

[83] Fujie, R. and Odagaki, T. (2007). Effects of superspreaders in spread of epidemic. *Physica A: Statistical Mechanics and its Applications*, 374(2):843–852.

[84] Gatrell, A. C., Bailey, T. C., Diggle, P. J., and Rowlingson, B. S. (1996). Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British Geographers*, 21(1):256–274.

[85] Gelfand, A. E., Kim, H.-J., Sirmans, C. F., and Banerjee, S. (2003). Spatial Modeling With Spatially Varying Coefficient Processes. *Journal of the American Statistical Association*, 98(462):387–396.

[86] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press, 3 edition.

[87] Germann, T. C., Kadau, K., Longini, I. M., and Macken, C. A. (2006). Mitigation strategies for pandemic influenza in the United States. *Proceedings of the National Academy of Sciences*, 103(15):5935–5940.

[88] Ghedin, E., Sengamalay, N. A., Shumway, M., Zaborsky, J., Feldblyum, T., Subbu, V., Spiro, D. J., Sitz, J., Koo, H., Bolotov, P., Dernovoy, D., Tatusova, T., Bao, Y., St George, K., Taylor, J., Lipman, D. J., Fraser, C. M., Taubenberger, J. K., and Salzberg, S. L. (2005). Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature*, 437(7062):1162–1166.

[89] Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–4.

[90] Glass, K., Mercer, G. N., Nishiura, H., McBryde, E. S., and Becker, N. G. (2011). Estimating reproduction numbers for adults and children from case data. *Journal of The Royal Society Interface*, 8(62):1248–1259.

[91] Gog, J. R., Ballesteros, S., Viboud, C., Simonsen, L., Bjørnstad, O. N., Shaman, J., Chao, D. L., Khan, F., and Grenfell, B. T. (2014). Spatial Transmission of 2009 Pandemic Influenza in the US. *PLoS Computational Biology*, 10(6):e1003635.

[92] Goldstein, E., Greene, S. K., Olson, D. R., Hanage, W. P., and Lipsitch, M. (2015a). Estimating the hospitalization burden associated with influenza and respiratory syncytial virus in New York City, 2003-2011. *Influenza and Other Respiratory Viruses*, pages n/a–n/a.

[93] Goldstein, J., Haran, M., Bjornstad, O. N., and Liebhold, A. (2015b). Quantifying Spatio-Temporal Variation of Invasion Spread. pages 1–21.

[94] Google (2017). Google Flu Trends.

[95] Grad, Y. H. and Lipsitch, M. (2014). Epidemiologic data and pathogen genome sequences: a powerful synergy for public health. *Genome biology*, 15(11):538.

[96] Grais, R. F., Ellis, J. H., and Glass, G. E. (2003). Assessing the impact of airline travel on the geographic spread of pandemic influenza. *European Journal of Epidemiology*, 18(11):1065–1072.

[97] Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37(3):424.

[98] Grenfell, B. (1997). (Meta)population dynamics of infectious diseases. *Trends in Ecology & Evolution*, 12(10):395–399.

[99] Grenfell, B. T. (2004). Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. *Science*, 303(5656):327–332.

[100] Grenfell, B. T., Bjørnstad, O. N., and Kappey, J. (2001). Travelling waves and spatial hierarchies in measles epidemics. *Nature*, 414(6865):716–23.

[101] Grunwald, G., Hyndman, R., Tedesco, L., and Tweedie, R. (1997). A unified view of linear AR(1) models.

[102] Halloran, M. E., Ferguson, N. M., Eubank, S., Longini, I. M., Cummings, D. a. T., Lewis, B., Xu, S., Fraser, C., Vullikanti, A., Germann, T. C., Wagener, D., Beckman, R., Kadau, K., Barrett, C., Macken, C. a., Burke, D. S., and Cooley, P. (2008). Modeling targeted layered containment of an influenza pandemic in the United States. *Proceedings of the National Academy of Sciences of the United States of America*, 105(12):4639–44.

[103] Hanski, I. (1994). A Practical Model of Metapopulation Dynamics. *The Journal of Animal Ecology*, 63(1):151.

[104] Hanski, I. (1998). Metapopulation Dynamics. *Nature*, 396(6706):41–49.

[105] Hanski, I. (1999). *Metapopulation Ecology*. Oxford University Press, Oxford.

[106] Hanski, I. and Gilpin, M. (1991). Metapopulation dynamics: brief history and conceptual domain. *Biological Journal of the Linnean Society*, 42(1-2):3–16.

[107] Hartley, D. M., Giannini, C. M., Wilson, S., Frieder, O., Margolis, P. A., Kotagal, U. R., White, D. L., Connelly, B. L., Wheeler, D. S., Tadesse, D. G., and Macaluso, M. (2017). Coughing , sneezing , and aching online : Twitter and the volume of influenza-like illness in a pediatric hospital. *PLoS ONE*, 12(7):1–10.

[108] Hashimoto, S., Murakami, Y., Taniguchi, K., and Nagai, M. (2000). Detection of epidemics in their early stage through infectious disease surveillance. *International Journal of Epidemiology*, 29(5):905–910.

[109] HealthMap (2017). FluNearYou.

[110] Heesterbeek, H., Anderson, R. M., Andreasen, V., Bansal, S., De Angelis, D., Dye, C., Eames, K. T. D., Edmunds, W. J., Frost, S. D. W., Funk, S., Hollingsworth, T. D., House, T., Isham, V., Klepac, P., Lessler, J., Lloyd-Smith, J. O., Metcalf, C. J. E., Mollison, D., Pellis, L., Pulliam, J. R. C., Roberts, M. G., and Viboud, C. (2015). Modeling infectious disease dynamics in the complex landscape of global health. *Science*, 347(6227):aaa4339–aaa4339.

[111] Heesterbeek, J. a. (2002). A brief history of R0 and a recipe for its calculation. *Acta Biotheor*, 50(3):189–204.

[112] Held, L., Hofmann, M., Höhle, M., and Schmid, V. (2006). A two-component model for counts of infectious diseases. *Biostatistics*, 7(3):422–437.

[113] Hens, N., Goeyvaerts, N., Aerts, M., Shkedy, Z., Van Damme, P., and Beutels, P. (2009). Mining social mixing patterns for infectious disease models based on a two-day population survey in Belgium. *BMC infectious diseases*, 9:5.

[114] Hollingsworth, T. D., Ferguson, N. M., and Anderson, R. M. (2006). Will travel restrictions control the international spread of pandemic influenza? *Nature Medicine*, 12(5):497–499.

[115] Holmes, E. C. (2004). The phylogeography of human viruses. *Mol Ecol*, 13(4):745–756.

[116] House, T., Baguelin, M., Van Hoek, A. J., White, P. J., Sadique, Z., Eames, K., Read, J. M., Hens, N., Melegaro, A., Edmunds, W. J., and Keeling, M. J. (2011). Modelling the impact of local reactive school closures on critical care provision during an influenza pandemic. *Proceedings. Biological sciences / The Royal Society*, 278(1719):2753–2760.

[117] Huang, K. E., Lipsitch, M., Shaman, J., and Goldstein, E. (2014). The US 2009 A(H1N1) Influenza Epidemic. *Epidemiology*, 25(2):203–206.

[118] Hutchinson, E. C., von Kirchbach, J. C., Gog, J. R., and Digard, P. (2010). Genome packaging in influenza A virus. *Journal of General Virology*, 91(2):313–328.

[119] IMS Health (2017). IMS Health.

[120] Inaba, H. (2017). *Age-Structured Population Dynamics in Demography and Epidemiology*. Springer Singapore, Singapore, 1 edition.

[121] Ionides, E. L., Breto, C., and King, A. A. (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 103(49):18438–18443.

[122] Jhung, M. A., Swerdlow, D., Olsen, S. J., Jernigan, D., Biggerstaff, M., Kamimoto, L., Kniss, K., Reed, C., Fry, A., Brammer, L., Gindler, J., Gregg, W. J., Bresee, J., and Finelli, L. (2011). Epidemiology of 2009 pandemic influenza a (H1N1) in the United States. *Clinical Infectious Diseases*, 52(SUPPL. 1):13–26.

[123] Jombart, T., Eggo, R. M., Dodd, P. J., and Balloux, F. (2011). Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity*, 106(2):383–390.

[124] Kaiser, A. and Schreiber, T. (2002). Information transfer in continuous processes. *Physica D: Nonlinear Phenomena*, 166(1-2):43–62.

[125] Kamiński, M., Ding, M., Truccolo, W. A., and Bressler, S. L. (2001). Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. *Biological Cybernetics*, 85(2):145–157.

[126] Karageorgopoulos, D. E., Vouloumanou, E. K., Korbila, I. P., Kapaskelis, A., and Falagas, M. E. (2011). Age Distribution of Cases of 2009 (H1N1) Pandemic Influenza in Comparison with Seasonal Influenza. *PLoS ONE*, 6(7):e21690.

[127] Keeling, M. J. (2001). Dynamics of the 2001 UK Foot and Mouth Epidemic: Stochastic Dispersal in a Heterogeneous Landscape. *Science*, 294(5543):813–817.

[128] Keeling, M. J., Danon, L., Vernon, M. C., and House, T. A. (2010). Individual identity and movement networks for disease metapopulations. *Proceedings of the National Academy of Sciences*, 107(19):8866–8870.

[129] Keeling, M. J. and Rohani, P. (2011). *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press.

[130] Keeling, M. J. and White, P. J. (2011). Targeting vaccination against novel infections: risk, age and spatial structure for pandemic influenza in Great Britain. *Journal of The Royal Society Interface*, 8(58):661–670.

[131] Kermack, W. O. and McKendrick, A. G. (1927). A Contribution to the Mathematical Theory of Epidemics. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 115(772):700–721.

[132] Khan, K., Arino, J., Hu, W., Raposo, P., Sears, J., Calderon, F., Heidebrecht, C., Macdonald, M., Liauw, J., Chan, A., and Gardam, M. (2009). Spread of a Novel Influenza A (H1N1) Virus via Global Airline Transportation. *New England Journal of Medicine*, 361(2):212–214.

[133] Kissler, S. M., Gog, J. R., Viboud, C., Charu, V., Bjørnstad, O. N., Simonsen, L., and Grenfell, B. T. (2017). Geographic Transmission Hubs of the 2009 Influenza Pandemic in the United States. *Submitted*.

[134] Klepac, P. and Caswell, H. (2011). The stage-structured epidemic: linking disease and demography with a multi-state matrix approach model. *Theoretical Ecology*, 4(3):301–319.

[135] Kucharski, A. J. and Gog, J. R. (2012a). Age profile of immunity to influenza: Effect of original antigenic sin. *Theoretical Population Biology*, 81(2):102–112.

[136] Kucharski, A. J. and Gog, J. R. (2012b). The Role of Social Contacts and Original Antigenic Sin in Shaping the Age Pattern of Immunity to Seasonal Influenza. *PLoS Computational Biology*, 8(10).

[137] Kucharski, A. J., Kwok, K. O., Wei, V. W. I., Cowling, B. J., Read, J. M., Lessler, J., Cummings, D. A., and Riley, S. (2014). The Contribution of Social Behaviour to the Transmission of Influenza A in a Human Population. *PLoS Pathogens*, 10(6).

[138] Kuiken, T., Holmes, E. C., McCauley, J., Rimmelzwaan, G. F., Williams, C. S., and Grenfell, B. T. (2006). Host Species Barriers to Influenza Virus Infections. *Science*, 312(5772):394–397.

[139] Lagacé-Wiens, P. R. S., Rubinstein, E., and Gumel, A. (2010). Influenza epidemiology—past, present, and future. *Critical Care Medicine*, 38(4):e1–e9.

[140] Lam, T. T.-Y., Wang, J., Shen, Y., Zhou, B., Duan, L., Cheung, C.-L., Ma, C., Lycett, S. J., Leung, C. Y.-H., Chen, X., Li, L., Hong, W., Chai, Y., Zhou, L., Liang, H., Ou, Z., Liu, Y., Farooqui, A., Kelvin, D. J., Poon, L. L. M., Smith, D. K., Pybus, O. G., Leung, G. M., Shu, Y., Webster, R. G., Webby, R. J., Peiris, J. S. M., Rambaut, A., Zhu, H., and Guan, Y. (2013). The genesis and source of the H7N9 influenza viruses causing human infections in China. *Nature*, 502(7470):241–4.

[141] Lampos, V., Bie, T. D., and Cristianini, N. (2010). Flu Detector - Tracking Epidemics on Twitter. In Balcazar, J., Bonchi, F., Sebag, M., and Gionis, A., editors, *Machine Learning and Knowledge Discovery in Databases*, pages 599–602. Springer-Verlag, Berlin/Heidelberg.

[142] Lau, M. S. Y., Marion, G., Streftaris, G., and Gibson, G. (2015). A Systematic Bayesian Integration of Epidemiological and Genetic Data. *PLOS COMPUTATIONAL BIOLOGY*, 11(11).

[143] Lavine, J. S., Bjørnstad, O. N., de Blasio, B. F., and Storsaeter, J. (2012). Short-lived immunity against pertussis, age-specific routes of transmission, and the utility of a teenage booster vaccine. *Vaccine*, 30(3):544–551.

[144] Lawson, A. B. (2006). *Statistical Methods in Spatial Epidemiology*. John Wiley & Sons, West Sussex, UK, 2nd edition.

[145] Lee, E., Arab, A., Goldlust, S., Viboud, C., and Bansal, S. (2017). Socio-environmental and measurement factors drive variation in influenza-like illness. *bioRxiv*.

[146] Legrand, J., Egan, J. R., Hall, I. M., Cauchemez, S., Leach, S., and Ferguson, N. M. (2009). Estimating the location and spatial extent of a covert anthrax release. *PLoS computational biology*, 5(1).

[147] Lessler, J., Reich, N. G., and Cummings, D. A. (2009). Outbreak of 2009 Pandemic Influenza A (H1N1) at a New York City School. *New England Journal of Medicine*, 361(27):2628–2636.

[148] Levins, R. (1969). Some Demographic and Genetic Consequences of Environmental Heterogeneity for Biological Control. *Bulletin of the Entomological Society of America*, 15(3):237–240.

[149] Levy, J. W., Cowling, B. J., Simmerman, J. M., Olsen, S. J., Fang, V. J., Suntarattiwong, P., Jarman, R. G., Klick, B., and Chotipitayasunondh, T. (2013). The serial intervals of seasonal and pandemic influenza viruses in households in Bangkok, Thailand. *American Journal of Epidemiology*, 177(12):1443–1451.

[150] Levy, M. Z., Small, D. S., Vilhena, D. A., Bowman, N. M., Kawai, V., Cornejo del Carpio, J. G., Cordova-Benzaquen, E., Gilman, R. H., Bern, C., and Plotkin, J. B. (2011). Retracing Micro-Epidemics of Chagas Disease Using Epicenter Regression. *PLoS Computational Biology*, 7(9).

[151] Lipsitch, M. and Viboud, C. (2009). Influenza seasonality: lifting the fog. *Proceedings of the National Academy of Sciences of the United States of America*, 106(10):3645–3646.

[152] Lizier, J. T. and Prokopenko, M. (2010). Differentiating information transfer and causal effect. *European Physical Journal B*, 73(4):605–615.

[153] Lizier, J. T., Prokopenko, M., and Zomaya, A. Y. (2008). Local information transfer as a spatiotemporal filter for complex systems. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 77(2):1–11.

[154] Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E., and Getz, W. M. (2005). Superspreading and the effect of individual variation on disease emergence. *Nature*, 438(7066):355–359.

[155] London School of Hygiene and Tropical Medicine (2017). FluSurvey.

[156] Longini, I. M., Ackerman, E., and Elveback, L. R. (1978). An optimization model for influenza A epidemics. *Mathematical Biosciences*, 38(1-2):141–157.

[157] Longini, I. M. and Halloran, M. E. (2005). Strategy for Distribution of Influenza Vaccine to High-Risk Groups and Children. *American Journal of Epidemiology*, 161(4):303–306.

[158] Lu, L., Lycett, S. J., and Leigh Brown, A. J. (2014). Determining the phylogenetic and phylogeographic origin of highly pathogenic avian influenza (H7N3) in Mexico. *PloS one*, 9(9):e107330.

[159] Lungarella, M., Ishiguro, K., Kuniyoshi, Y., and Otsu, N. (2007). Methods for Quantifying the Causal Structure of Bivariate Time Series. *International Journal of Bifurcation and Chaos*, 17(03):903–921.

[160] Lycett, S., Mcleish, N. J., Robertson, C., Carman, W., Baillie, G., McMenamin, J., Rambaut, A., Simmonds, P., Woolhouse, M., and Leigh Brown, A. J. (2012). Origin and fate of A/H1N1 influenza in scotland during 2009. *Journal of General Virology*, 93(Pt 6):1253–1260.

[161] Marsden-Haug, N., Foster, V. B., Gould, P. L., Elbert, E., Wang, H., and Pavlin, J. A. (2007). Code-based Syndromic Surveillance for Influenzalike Illness by International Classification of Diseases, Ninth Revision. *Emerging infectious diseases*, 13(2).

[162] Menezes, T. and Roth, C. (2017). Natural Scales in Geographical Patterns. *Scientific Reports*, 7(April):45823.

[163] Miller, M. A., Viboud, C., Balinska, M., and Simonsen, L. (2009). The Signature Features of Influenza Pandemics — Implications for Policy. *New England Journal of Medicine*, 360(25):2595–2598.

[164] Mills, H. L. and Riley, S. (2014). The Spatial Resolution of Epidemic Peaks. *PLoS Computational Biology*.

[165] Mollison, D., Anderson, R. M., Bartlett, M. S., and Southwood, R. (1986). Modelling Biological Invasions: Chance, Explanation, Prediction [and Discussion]. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 314(1167):675–693.

[166] Morelli, M. J., Thébaud, G., Chadœuf, J., King, D. P., Haydon, D. T., and Soubeyrand, S. (2012). A Bayesian Inference Framework to Reconstruct Transmission Trees Using Epidemiological and Genetic Data. *PLoS Computational Biology*, 8(11).

[167] Morgan, O. W., Parks, S., Shim, T., Blevins, P. A., Lucas, P. M., Sanchez, R., Walea, N., Loustalot, F., Duffy, M. R., Shim, M. J., Guerra, S., Guerra, F., Mills, G., Verani, J., Alsip, B., Lindstrom, S., Shu, B., Emery, S., Cohen, A. L., Menon, M., Fry, A. M., Dawood, F., Fonseca, V. P., and Olsen, S. J. (2010). Household transmission of pandemic (H1N1) 2009, San Antonio, Texas, USA, April-May 2009. *Emerging Infectious Diseases*, 16(4):631–637.

[168] Morgan-Capner, P., Wright, J., Miller, C. L., and Miller, E. (1988). Surveillance of antibody to measles, mumps, and rubella by age. *BMJ*, 297(6651):770–772.

[169] Moriyama, I. M., Loy, R. M., and Robb-Smith, A. H. (2011). History of the statistical classification of diseases and causes of death. Technical report, Centers for Disease Control and Prevention.

[170] Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G. S., Wallinga, J., Heijne, J., Sadkowska-Todys, M., Rosinska, M., and Edmunds, W. J. (2008a). Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases. *PLoS Medicine*, 5(3):e74.

[171] Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G. S., Wallinga, J., Heijne, J., Sadkowska-Todys, M., Rosinska, M., and Edmunds, W. J. (2008b). Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases. *PLoS Medicine*, 5(3):e74.

[172] Murray, J., Stanley, E., and Brown, D. (1986). On the spatial spread of rabies among foxes. *Proc. Royal Soc. B*, 229(1255):111–150.

[173] Mylius, S. D., Hagenaars, T. J., Lugnér, A. K., and Wallinga, J. (2008). Optimal allocation of pandemic influenza vaccine depends on age, risk and timing. *Vaccine*, 26(29-30):3742–3749.

[174] National Center for Biotechnology Information (2016). Influenza Virus Resource.

[175] National Institute of Allergy and Infectious Diseases (2016). Influenza Research Database.

[176] Nelson, G. D. and Rae, A. (2016). An economic geography of the United States: From commutes to megaregions. *PLoS ONE*, 11(11):1–23.

[177] Nelson, M. I., Tan, Y., Ghedin, E., Wentworth, D. E., St George, K., Edelman, L., Beck, E. T., Fan, J., Lam, T. T.-Y., Kumar, S., Spiro, D. J., Simonsen, L., Viboud, C., Holmes, E. C., Henrickson, K. J., and Musser, J. M. (2011). Phylogeography of the spring and fall waves of the H1N1/09 pandemic influenza virus in the United States. *Journal of virology*, 85(2):828–34.

[178] Nichol, K. L., Heilly, S. D., and Ehlinger, E. (2005). Colds and Influenza-Like Illnesses in University Students: Impact on Health, Academic and Work Performance, and Health Care Use. *Clinical Infectious Diseases*, 40(9):1263–1270.

[179] Nishiura, H., Castillo-Chavez, C., Safan, M., and Chowell, G. (2009). Transmission potential of the new influenze A(H1N1) virus and its age-specificity in Japan. *Eurosurveillance*, 14(22):1–5.

[180] Nishiura, H., Chowell, G., Safan, M., and Castillo-Chavez, C. (2010). Pros and cons of estimating the reproduction number from early epidemic growth rate of influenza A (H1N1) 2009. *Theoretical Biology and Medical Modelling*, 7(1):1.

[181] Okubo, A. and Levin, S. A. (1989). A Theoretical Framework for Data Analysis of Wind Dispersal of Seeds and Pollen. *Ecology*, 70(2):329–338.

[182] Olson, D. R., Konty, K. J., Paladini, M., Viboud, C., and Simonsen, L. (2013). Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. *PLoS Computational Biology*, 9(10):e1003256.

[183] Olson, D. R., Simonsen, L., Edelson, P. J., and Morse, S. S. (2005). Epidemiological evidence of an early wave of the 1918 influenza pandemic in New York City. *Proceedings of the National Academy of Sciences of the United States of America*, 102(31):11059–11063.

[184] Paget, J., Marquet, R., Meijer, A., and van der Velden, K. (2007). Influenza activity in Europe during eight seasons (1999-2007): an evaluation of the indicators used to measure activity and an assessment of the timing, length and course of peak activity (spread) across Europe. *BMC infectious diseases*, 7:141.

[185] Pahle, J., Green, A. K., Dixon, C. J., and Kummer, U. (2008). Information transfer in signaling pathways: a study using coupled simulated and experimental data. *BMC bioinformatics*, 9:139.

[186] Pappas, G., Kiriaze, I. J., and Falagas, M. E. (2008). Insights into infectious disease in the era of Hippocrates. *International Journal of Infectious Diseases*, 12(4):347–350.

[187] Parrott, R. H., Kim, H. W., Arrobio, J. O., Hodes, D. S., Murphy, B. R., Brandt, C. D., Camargo, E., and Chanock, R. M. (1973). Epidemiology of respiratory syncytial virus infection in Washington, D.C. *American Journal of Epidemiology*, 98(4):289–300.

[188] Paull, S. H., Song, S., McClure, K. M., Sackett, L. C., Kilpatrick, A. M., and Johnson, P. T. (2012). From superspreaders to disease hotspots: linking transmission across hosts and space. *Frontiers in Ecology and the Environment*, 10(2):75–82.

[189] Pelecanos, A. M., Ryan, P. A., and Gatton, M. L. (2010). Outbreak detection algorithms for seasonal disease data: a case study using Ross River virus disease. *BMC medical informatics and decision making*, 10(1):74.

[190] Perrotta, D., Bella, A., Rizzo, C., and Paolotti, D. (2017). Participatory online surveillance as a supplementary tool to sentinel doctors for influenza-like illness surveillance in Italy. *PLoS ONE*, 12(1):1–15.

[191] Pestre, V., Morel, B., Encrenaz, N., Brunon, A., Lucht, F., Pozzetto, B., and Berthelot, P. (2012). Transmission by super-spreading event of pandemic A/H1N1 2009 influenza during road and train travel. *Scandinavian Journal of Infectious Diseases*, 44(3):225–227.

[192] Pigott, D. M., Golding, N., Mylne, A., Huang, Z., Henry, A. J., Weiss, D. J., Brady, O. J., Kraemer, M. U., Smith, D. L., Moyes, C. L., Bhatt, S., Gething, P. W., Horby, P. W., Bogoch, I. I., Brownstein, J. S., Mekaru, S. R., Tatem, A. J., Khan, K., and Hay, S. I. (2014). Mapping the zoonotic niche of Ebola virus disease in Africa. *eLife*, 3.

[193] Poletti, P., Ajelli, M., and Merler, S. (2011). The effect of risk perception on the 2009 H1N1 pandemic influenza dynamics. *PLoS ONE*, 6(2).

[194] Polgreen, P. M., Chen, Y., Pennock, D. M., and Nelson, F. D. (2008). Using internet searches for influenza surveillance. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 47(11):1443–8.

[195] Potter, C. W. (2001). A history of influenza. *Journal of Applied Microbiology*, 91(4):572–579.

[196] Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, MA.

[197] Read, J. M., Lessler, J., Riley, S., Wang, S., Tan, L. J., Kwok, K. O., Guan, Y., Jiang, C. Q., and Cummings, D. A. T. (2014). Social mixing patterns in rural and urban areas of southern China. *Proceedings of the Royal Society B: Biological Sciences*, 281(1785):20140268–20140268.

[198] Reichert, T., Chowell, G., and Mccullers, J. A. (2012). The age distribution of mortality due to influenza : pandemic and peri-pandemic. *BMC Medicine*, 10(162):1–15.

[199] Reichert, T. A., Sugaya, N., Fedson, D. S., Glezen, W. P., Simonsen, L., and Tashiro, M. (2001). The Japanese experience with vaccinating schoolchildren against influenza. *N Engl J Med*, 344(12):889–896.

[200] Riley, S., Eames, K., Isham, V., Mollison, D., and Trapman, P. (2015). Five challenges for spatial epidemic models. *Epidemics*, 10:68–71.

[201] Riley, S. and Ferguson, N. M. (2007). Smallpox transmission and control: Spatial dynamics in Great Britain. *Pnas*, 103(33):12637–12642.

[202] Rodriguez, G. (2007). Survival Models. In *Lecture Notes on Generalized Linear Models*, chapter 7, pages 1–34. G. Rodriguez.

[203] Ross, R. (1910). *The Prevention of Malaria*. Dutton, New York.

[204] Russell, C. A., Jones, T. C., Barr, I. G., Cox, N. J., Garten, R. J., Gregory, V., Gust, I. D., Hampson, A. W., Hay, A. J., Hurt, A. C., de Jong, J. C., Kelso, A., Klimov, A. I., Kageyama, T., Komadina, N., Lapedes, A. S., Lin, Y. P., Mosterin, A., Obuchi, M., Odagiri, T., Osterhaus, A. D. M. E., Rimmelzwaan, G. F., Shaw, M. W., Skepner, E., Stohr, K., Tashiro, M., Fouchier, R. A. M., and Smith, D. J. (2008). The Global Circulation of Seasonal Influenza A (H3N2) Viruses. *Science*, 320(5874):340–346.

[205] Rvachev, L. A. (1968). Modeling experiment of a large-scale epidemic by means of a computer. *DOKLADY AKADEMII NAUK SSSR*, 180(2):294.

[206] Rvachev, L. A. and Longini, I. M. (1985). A mathematical model for the global spread of influenza. *Mathematical Biosciences*, 75(1):3–22.

[207] Sattenspiel, L. and Dietz, K. (1995). A structured epidemic model incorporating geographic mobility among regions. *Mathematical Biosciences*, 128(1-2):71–91.

[208] Schreiber, T. (2000). Measuring Information Transfer. *Physical Review Letters*, 85(2):19.

[209] Shaman, J., Goldstein, E., and Lipsitch, M. (2011). Absolute Humidity and Pandemic Versus Epidemic Influenza. *American Journal of Epidemiology*, 173(2):127–135.

[210] Shaman, J. and Karspeck, A. (2012). Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences of the United States of America*, 109(3):20425–30.

[211] Sharpe, D., Hopkins, R., Cook, R. L., and Striley, C. W. (2017). Using a Bayesian Method to Assess Google, Twitter, and Wikipedia for ILI Surveillance. In *ISDS Annual Conference Proceedings 2017*, volume 9.

[212] Shen, Z., Ning, F., Zhou, W., He, X., Lin, C., Chin, D. P., Zhu, Z., and Schuchat, A. (2004). Superspreading SARS Events, Beijing, 2003. *Emerging Infectious Diseases*, 10(2):256–260.

[213] Shmueli, G. and Burkom, H. (2010). Statistical Challenges Facing Early Outbreak Detection in Biosurveillance. *Technometrics*, 52(1):39–51.

[214] Signorini, A., Segre, A. M., and Polgreen, P. M. (2011). The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *PLoS ONE*, 6(5).

[215] Simini, F., González, M. C., Maritan, A., and Barabási, A.-L. (2012). A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100.

[216] Simon, A. K., Hollander, G. A., and McMichael, A. (2015). Evolution of the immune system in humans from infancy to old age. *Proceedings of the Royal Society B: Biological Sciences*, 282(1821):20143085.

[217] Simonsen, L., Clarke, M. J., Schonberger, L. B., Arden, N. H., Cox, N. J., and Fukuda, K. (1998). Pandemic versus epidemic influenza mortality: a pattern of changing age distribution. *The Journal of infectious diseases*, 178(1):53–60.

[218] Smieszek, T., Balmer, M., Hattendorf, J., Axhausen, K. W., Zinsstag, J., and Scholz, R. W. (2011). Reconstructing the 2003/2004 H3N2 influenza epidemic in Switzerland with a spatially explicit, individual-based model. *BMC Infectious Diseases*, 11(1):115.

[219] Smolinski, M. S., Crawley, A. W., Baltrusaitis, K., Chunara, R., Olsen, J. M., Wójcik, O., Santillana, M., Nguyen, A., and Brownstein, J. S. (2015). Flu near you: Crowdsourced symptom reporting spanning 2 influenza seasons. *American Journal of Public Health*, 105(10):2124–2130.

[220] Snow, J., Richardson, B., and Frost, W. H. (1936). *Snow on Cholera: being a reprint of two papers by John Snow, M.D.* The Commonwealth Fund, New York.

[221] Spicer, C. C. (1979). The mathematical modelling of influenza epidemics. *Br Med Bull*, 35(1):23–28.

[222] Spicer, C. C. and Lawrence, C. J. (1984). Epidemic influenza in Greater London. *Epidemiology and Infection*, 93(01):105–112.

[223] Staniek, M. and Lehnertz, K. (2008). Symbolic Transfer Entropy. *Phys. Rev. Lett.*, 100(April):158101.

[224] Steeg, G. V. and Galstyan, A. (2011). Information Transfer in Social Media. *Entropy*, 90292(1):1–8.

[225] SteelFisher, G. K., Blendon, R. J., Bekheit, M. M., and Lubell, K. (2010). The Public's Response to the 2009 H1N1 Influenza Pandemic. *New England Journal of Medicine*, 362(22):e65.

[226] Stein, R. A. (2011). Super-spreaders in infectious diseases. *International Journal of Infectious Diseases*, 15(8):e510–e513.

[227] Sugihara, G., Deyle, E. R., and Ye, H. (2017). Reply to Baskerville and Cobey: Misconceptions about causation with synchrony and seasonal drivers. *Proceedings of the National Academy of Sciences of the United States of America*, 114(12):E2272–E2274.

[228] Sugihara, G., May, R., Ye, H., Hsieh, C.-h., Deyle, E., Fogarty, M., and Munch, S. (2012). Detecting causality in complex ecosystems. *Science*, 338(6106):496–500.

[229] Takens, F. (1981). Detecting strange attractors in turbulence. In Rand, D. and Young, L.-S., editors, *Dynamical Systems and Turbulence*, pages 366–381. Springer-Verlag.

[230] Taubenberger, J. K. and Morens, D. M. (2009). Pandemic influenza–including a risk assessment of H5N1. *Rev Sci Tech*, 28(1):187–202.

[231] Truscott, J. and Ferguson, N. M. (2012). Evaluating the adequacy of gravity models as a description of human mobility for epidemic modelling. *PLoS Computational Biology*, 8(10):e1002699.

[232] Tumpey, T. M. (2005). Characterization of the Reconstructed 1918 Spanish Influenza Pandemic Virus. *Science*, 310(5745):77–80.

[233] United States Census Bureau (2010). 2010 Census Gazetteer Files.

[234] United States Census Bureau (2015). ZIP Code Tabulation Areas (ZCTAs).

[235] US Census Bureau (2012). Methodology for the Intercensal Population and Housing Unit Estimates: 2000 to 2010. Technical Report October, US Census Bureau.

[236] U.S. Centers for Medicare & Medicade Services (2017). Details for title: CMS 1500.

[237] U.S. Department of Health and Human Services (2017). Regional Offices.

[238] U.S. Postal Service Office of Inspector General (2013). The Untold Story of the ZIP Code. Technical report, United State Postal Services.

[239] Viboud, C., Bjørnstad, O. N., Smith, D. L., Simonsen, L., Miller, M. A., and Grenfell, B. T. (2006). Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science (New York, N.Y.)*, 312(5772):447–51.

[240] Viboud, C., Charu, V., Olson, D., Ballesteros, S., Gog, J., Khan, F., Grenfell, B., and Simonsen, L. (2014). Demonstrating the Use of High-Volume Electronic Medical Claims Data to Monitor Local and Regional Influenza Activity in the US. *PLoS ONE*, 9(7):e102429.

[241] Viboud, C., Nelson, M. I., Tan, Y., and Holmes, E. C. (2013). Contrasting the epidemiological and evolutionary dynamics of influenza spatial transmission. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1614):20120199–20120199.

[242] Wagner, M. M., Tsui, F. C., Espino, J. U., Dato, V. M., Sittig, D. F., Caruana, R. a., McGinnis, L. F., Deerfield, D. W., Druzdzel, M. J., and Fridsma, D. B. (2001). The emerging science of very early detection of disease outbreaks. *J Public Health Manag Pract*, 7(290):51–59.

[243] Wallinga, J. and Lipsitch, M. (2007). How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences*, 274(1609):599–604.

[244] Wallinga, J. and Teunis, P. (2004). Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures. *American Journal of Epidemiology*, 160(6):509–516.

[245] Wallinga, J., Teunis, P., and Kretzschmar, M. (2006). Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. *American Journal of Epidemiology*, 164(10):936–944.

[246] Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., Buckee, C. O., Prothero, R. M., Bejon, P., Dolgin, E., Stoddard, S. T., Lynch, C., Roper, C., Tatem, A. J., Smith, D. L., Moonen, B., Menach, A. L., Chuquiyauri, R., Tatem, A. J., Linard, C., Gilbert, M., Snow, R. W., Noor, A. M., Tatem, A. J., Tatem, A. J., Noor, A. M., von Hagen, C., Gregorio, A. D., Hay, S. I., Tatem, A. J., Noor, A. M., Hay, S. I., Noor, A. M., Gething, P. W., Mudhune, S. A., Yé, Y., Madise, N., Ndugwa, R., Ochola, S., Snow, R. W., Kasili, S., Wesolowski, A., Eagle, N., Noor, A. M., Snow, R. W., Buckee, C. O., González, M. C., Hidalgo, C. A., Barabási, A. L., Song, C., Qu, Z., Blumm, N., Barabási, A. L., Dye, C., Hasibeder, G., Sama, W., Dietz, K., Smith, T., Balcan, D., Ferguson, N. M., and Longini, I. M. (2012). Quantifying the impact of human mobility on malaria. *Science (New York, N.Y.)*, 338(6104):267–70.

[247] White, L. F., Wallinga, J., Finelli, L., Reed, C., Riley, S., Lipsitch, M., and Pagano, M. (2009). Estimation of the reproductive number and the serial interval in early phase of the 2009 influenza A/H1N1 pandemic in the USA. *Influenza and other Respiratory Viruses*, 3(6):267–276.

[248] WHO Global Influenza Programme Surveillance and Epidemiology team (2012). WHO Interim Global Epidemiological Surveillance Standards for Influenza. Technical Report July, World Health Organization.

[249] Williams, P. L. and Beer, R. D. (2011). Generalized measures of information transfer. *arXiv:1102.1507*, pages 1–6.

[250] Wilson, A. (1970). *Entropy in Urban and Regional Modelling*. Pion, London.

[251] Won, M., Marques-Pita, M., Louro, C., Gonçalves-Sá, J., Barker, W. H., Molinari, N.-A. M., Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M., Brilliant, L., Hickmann, K., Fairchild, G., Priedhorsky, R., Generous, N., Hyman, J., Deshpande, A., Santillana, M., Nguyen, A. T., Dredze, M., Paul, M. J., Nsoesie, E. O., Brownstein, J. S., Chretien, J., George, D., Shaman, J., Chitale, R., McKenzie, F., Shaman, J., Alicia, K., Moriña, D., Rhodes, C. J., Hollingsworth, T. D., Closas, P., Coma, E., Méndez, L., Cowling, B. J., Martinez-Beneito, M. A., Shaman, J., Christakis, N. A., Fowler, J. H., Pervaiz, F., Kermack, W., McKendrick, A., Trevor, J. H., Tibshirani, R. J., Friedman, Lazer, D., Kennedy, R., King, G., Vespignani, A., Olson, D. R., Konty, K. J., Paladini, M., Viboud, C., Simonsen, L., Tariq, A., Westbrook, J., Byrne, M., Robinson, M., Baysari, M. T., and Cooper, D. L. (2017). Early and Real-Time Detection of Seasonal Influenza Onset. *PLOS Computational Biology*, 13(2):e1005330.

[252] Worby, C. J., Chaves, S. S., Wallinga, J., Lipsitch, M., Finelli, L., and Goldstein, E. (2015). On the relative role of different age groups in influenza epidemics. *Epidemics*, 13:10–16.

[253] World Health Organization (2017). Influenza: into the history of influenza control...

[254] World Health Organization Regional Office for South-East Asia (2009). Pandemic H1N1 2009. Technical report, WHO Regional Office for South-East Asia, New Delhi.

[255] Wu, J. T., Cowling, B. J., Lau, E. H. Y., Ip, D. K. M., Ho, L. M., Tsang, T., Chuang, S. K., Leung, P. Y., Lo, S. V., Liu, S. H., and Riley, S. (2010). School closure and mitigation of pandemic (H1N1) 2009, Hong Kong. *Emerging Infectious Diseases*, 16(3):538–541.

[256] Xia, Y., Bjørnstad, O. N., and Grenfell, B. T. (2004). Measles metapopulation dynamics: a gravity model for epidemiological coupling and dynamics. *The American Naturalist*, 164(2):267–81.

[257] Yang, W., Lipsitch, M., and Shaman, J. (2015a). Inference of seasonal and pandemic influenza transmission dynamics. *Proceedings of the National Academy of Sciences*, 112(9):201415012.

[258] Yang, W., Zhang, W., Kargbo, D., Yang, R., Chen, Y., Chen, Z., Kamara, A., Kargbo, B., Kandula, S., Karspeck, A., Liu, C., and Shaman, J. (2015b). Transmission network of the 2014–2015 Ebola epidemic in Sierra Leone. *Journal of The Royal Society Interface*, 12(112):20150536.

[259] Yang, Y., Sugimoto, J. D., Halloran, M. E., Basta, N. E., Chao, D. L., Matrajt, L., Potter, G., Kenah, E., and Longini, I. M. (2009). The transmissibility and control of pandemic influenza A(H1N1) virus. *Science*, 326(5953):729–733.

[260] Yih, W. K., Cocoros, N. M., Crockett, M., Klompas, M., Kruskal, B. a., Kulldorff, M., Lazarus, R., Madoff, L. C., Morrison, M. J., Smole, S., and Platt, R. (2014). Automated Influenza-Like Illness Reporting-An Efficient Adjunct to Traditional Sentinel Surveillance. *Public Health Reports*, 129(February):55–63.

[261] Zipf, G. K. (1946). The P 1 P 2 D Hypothesis: On the Intercity Movement of Persons. *American Sociological Review*, 11(6):677.