SPATIAL EPIDEMIOLOGY OF SEASONAL INFLUENZA IN THE UNITED STATES

by

CODY DAILEY

(Under the Direction of JUSTIN BAHL)

ABSTRACT

The epidemiology of seasonal influenza is shaped by mechanisms across ecological scales, from molecular interactions to global climate patterns. Misaligned data may greatly impact analytical inference, but spatial constructs characterizing larger scales, e.g., regions, lack concrete, standard definitions and, consequently, are often overlooked in influenza research.

In this dissertation, I analyze patterns in human mobility, disease incidence, and viral genetic evolution to holistically characterize spatial structuring within the United States related to seasonal influenza. In Chapter 2, I model commuting flows and influenza-like illness (ILI). Using an estimated critical distance of ~150km or ~93mi, I show that simple summary metrics of local mobility from county-level commutes informs some variation in state-level ILI epidemic intensity. In Chapter 3, I evaluate numerous regional delineations of the US for their ability to capture important patterns of worker commutes, ILI incidence, and viral population structure. From this network science community analysis, I find evidence suggesting that the US may be best

represented with ~8 subnational regions which are not precisely captured by existing administrative regional delineations. In Chapter 4, I systematically describe local outbreaks of four seasonal influenza viruses across a decade of flu seasons in the US. I show that the average isolate diversities of local outbreaks exhibit weak spatial autocorrelation, and marginally, local outbreaks in more populous states tended to have less diverse viral isolates which may suggest either impactful differences in transmission patterns or isolate sampling.

Taken together, these analyses suggest that there is inherent structuring of local and regional scales within the US. Given these findings, I speculate that much of the observed variation in seasonal influenza epidemiology at the regional level could be explained by the underlying spatial organization of local populations. Additionally, this work shows that even with simple methodologies and crude conceptualizations of scale, we can abstract information from data at higher resolutions which is salient to patterns at larger scales and coarser resolutions. With continued effort, we may be able to identify systematic sources of variation in outbreak dynamics and viral evolution which would be invaluable when modeling an otherwise largely chaotic infectious disease system.

INDEX WORDS: SEASONAL INFLUENZA, SPATIAL EPIDEMIOLOGY, EPIDEMIC INTELLIGENCE, COMMUNITY ECOLOGY, GENOMIC EPIDEMIOLOGY, PHYLODYNAMICS, HUMAN INFLUENZA

SPATIAL EPIDEMIOLOGY OF SEASONAL INFLUENZA IN THE UNITED STATES

by

CODY DAILEY

BS, University of Georgia, 2015

MPH, University of Georgia, 2017

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2024

SPATIAL EPIDEMIOLOGY OF SEASONAL INFLUENZA IN THE UNITED STATES

by

CODY DAILEY

Major Professor:	Justin Bahl
Committee:	Andreas Handel
	Christopher Whalen
	Liliana Salvador

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
December 2024

DEDICATION

To my nephews and nieces whose unadulterated joie de vivre can cut through the worst of attitudes

ACKNOWLEDGEMENTS

Thank you to my family and friends for being there and vaulting me to new heights.

Thank you to my cohort homies for growing with me; you have each moved on to exciting things, but remember, age before beauty and all that.

Thank you to my lab-mates for not only putting up with my incessant and tedious lines of questioning but also for taking the time to teach me so many things, including how to blow off steam.

Thank you to the people of EpiBios for helping whenever I came running.

Thank you to my funding sources for paying my way and training me in the various domains of public health.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

OVERVIEW

The epidemiology of infectious diseases, such as seasonal influenza, is shaped by mechanisms across ecological scales, from the molecule, cell, and organism to the population and metapopulation (1). However, spatial constructs characterizing larger scales lack concrete definitions and, consequently, are often overlooked in influenza research (2). To explain briefly, a metapopulation is defined as a collection of local populations separated by space but connected by migration (3); while distinctions of the metapopulation patches housing local populations may be clear at some scales, e.g., host cells or organisms as patches for parasites, what constitutes a "local" population is less well-defined at larger scales, e.g., host populations as patches. This dearth of knowledge in spatial structuring may both bias inference (4) and inhibit accurate modeling (5), cf., the ecological fallacy and cross-level bias (6) or gerrymandering (7).

In this dissertation, I attempt to address this problem in part by characterizing sub-national spatial structuring in the United States (US) and its impact on epidemiological patterns of seasonal influenza. To achieve this goal, I set forth three specific aims: characterize regional patterns in human mobility and their associations with the epidemic intensity of influenza-like illness; identify specific influenza transmission zones (ITZs) within the US; and describe the spatial variation in the phylogenetic signal of local

influenza outbreaks. Altogether, this work constitutes a holistic characterization of the spatial epidemiology of seasonal influenza on a subnational, metapopulation scales in the US.

This introduction gives background on seasonal influenza and observed epidemiological patterns, particularly focusing on transmission across scales and the importance of researching this topic.

INFLUENZA VIRUS & DISEASE CHARACTERIZATION

Influenza Illness

Influenza is a respiratory illness caused by influenza virus infection (8). Influenza infections and the illness they cause are acute in nature, i.e., quickly onset and relatively short-lived. The natural history of an illness describes the progression of disease from exposure to resolution in the absence of medical intervention; for influenza, this typically spans 4-7 days for the average, uncomplicated case (8). Following transmission of influenza virus particles, an infected person can begin feeling sick after an average of 2 days and can shed virus and become contagious, up to 1 day before or without ever feeling ill (9). Once symptoms begin to show, an infected person will typically feel worst over the next 1-2 days and be sick for 3-7 days total (9).

Influenza illness presents in a range of disease severity. Some may experience an asymptomatic, subclinical infection, while, for others, influenza infection may be a primary cause of death. Clinical presentation in uncomplicated cases may include fever, cough, sore throat, chills, or malaise (8). More severe cases may have more numerous or debilitating symptoms and additional complications may develop not limited to the upper

respiratory system, such as encephalitis (brain inflammation), pneumonia, or even sepsis (10); these complications may be the result of the influenza virus infection alone or related to influenza-associated secondary bacterial co-infections (10,11).

Virology

Influenza viruses are orthomyxoviruses, a family of negative-sense RNA viruses with segmented genomes and host-derived envelopes (17). There are two genera of influenza viruses that cause significant disease in humans, *influenzavirus* A and *influenzavirus* B, more colloquially referred to as type A and type B. Influenza B viruses (IBV) are solely human pathogens, while influenza A viruses (IAV) have a broad range of hosts and, relatedly, similarly broad genomic diversity (18). Further narrowing it down, there are two IAV subtypes, H3 and H1, and two IBV lineages, Victoria and Yamagata, that circulate widely causing seasonal outbreaks in the US; these four viruses are referred to collectively as seasonal influenza viruses.

IAV and IBV share similarities in their viral structure, genomic composition, and viral protein functions. Both types of influenza viruses have genomes composed of single-stranded RNA and consist of eight homologous gene segments. The genomes encode similar proteins essential for the virus's structure and replication, including two surface antigens, hemagglutinin (HA) and neuraminidase (NA). HA is responsible for binding the virus to host cells and is the primary target of human adaptive immunity (19). Opposing the binding function of HA, NA cleaves the binding between HA and sialic acid residues which helps the virus escape from infected cells as well as avoid mucociliary clearance (20,21).

Box 1. Clinical Case Management of Influenza

The acute nature of influenza infection, illness, and the risk of complications makes an accurate and timely diagnosis paramount in clinical case management. Chief complaints and clinical presentation are not uniquely characterized for / specific to influenza. For example, influenza-like illness syndrome (ILI), often used in surveillance based on triaged signs, is defined as a fever (temperature 100F/37.8C or higher) and a cough and/or a sore throat (12). These criteria are non-specific, and, consequently, patients with ILI are not necessarily influenza cases. Many other seasonal, respiratory infectious agents cause similarly characterized respiratory illness which may be captured in syndromic surveillance programs. Even so, aggregate analysis of signs and symptoms can help to identify influenza cases via clinical decision/prediction rules, though algorithm performance and accuracy can vary in practical settings, e.g., using patient reported signs and symptoms in telemedicine contexts (13), and, ultimately, these approaches still need to be validated (14). Influenza diagnosis can be confirmed using one of many available diagnostic tests. These tests identify influenza by detection of viral antigens (e.g., rapid diagnostic tests), viral genetic material (e.g., polymerase chain reaction assays), or viable virus (cultures) in samples collected from suspected cases (15). Testing for influenza is recommended for clinicians if the results may influence clinical management of disease (16). Diagnostic testing does not preclude influenza treatment using antivirals, though, but it can still be otherwise useful for understanding prognostic risk profiles and recommending non-pharmaceutical interventions for outbreak control (16).

DESCRIPTIVE EPIDEMIOLOGY

Person

Each year in the United States, approximately 8% of people experience influenza illness; this amounts to estimates of 9.3 million – 41 million incident cases annually since 2010 (22,23). Most infections are self-limiting, resolving without medical intervention, and some may even be subclinical with no or very mild symptoms. Still, many influenza infections cause significant disease; surveillance figures indicate as much, with annual estimates of 100,000 – 710,000 hospitalizations and 4,900 – 51,000 deaths (23).

Everyone is susceptible to influenza virus infection, but the risks of infection, symptomatic illness, and more severe illness with complications are influenced by characteristics specific to the individual person as well as characteristics of the infecting viral strain. Typical host-related risk factors include age (being young or elderly), having a compromised immune system, and having numerous other comorbid conditions which may all influence an individual's probability of developing more severe illness and/or complications (10). These risk factors are all closely interrelated and may ultimately underlie factors of the immune system and its ability to react effectively (quickly and efficiently) to the influenza viruses (24). For example, younger children and the elderly may have less effective immune responses to influenza challenges, though for differing reasons, i.e., naïveté versus senescence, respectively. Conversely, a well-adapted immune system may be considered a protective factor, potentially even to the extent of sterilizing immunity, i.e., infection-preventing immunity. As it is thoroughly ingrained into human ecology, most people have repeated influenza infections over the course of their lives with adaptive immunity developing early in childhood (24,25).

Place & Time

The incidence of influenza illness is not consistent throughout the year. Rather, the burden felt during influenza outbreaks is concentrated over several months. In the US, influenza outbreaks align with the winter season, usually beginning around October, peaking sometime between December and February, and dissipating around May (26). This routine occurrence, or seasonality, in influenza incidence is not limited to the US, but common among other regions with temperate climates. Although coinciding with behavioral shifts during the winter holidays, the pattern of seasonal outbreaks is largely attributed to changing climatic conditions which directly impact the transmissibility of influenza viruses (27,28).

Influenza outbreaks happen at different times in different places. These heterogeneities in the regularity and relative timing of influenza outbreaks around the world may combine to facilitate the spread and persistence of influenza virus lineages (29). For example, on the global scale, the northern and southern hemispheres experience winter at different points in time in a calendar year. In fact, these seasons are complementary over the course of a year; winter in the northern hemisphere corresponds to summer in the southern hemisphere, as does northern summers and southern winters. As more ideal climatic conditions oscillate back and forth between northern and southern hemispheres over the course of a year, influenza virus lineages can follow these suitable environmental conditions by hopping between (or causing) asynchronous outbreaks (29). This conveniently timed range expansion offers opportunities for influenza virus lineages to avoid seasonal transmission bottlenecks (25,29).

Variable Disease & Outbreak Dynamics

Although clearly important to influenza epidemiology, these observed patterns and aspects of person, place, and time are moderated by features more directly related to the virus itself. Generally, IAV, particularly H3N2, are more notorious, contributing to both higher rates of infection and severe disease; IBV also cause significant disease, but overall incidence is less than IAV (3 IAV to every 1 IBV confirmed case) and demographic profiles of incidence and mortality skew relatively younger for IBV (30,31).

IAV and IBV co-circulate in human populations causing seasonal outbreaks in the US. Viruses from either influenza type may be the predominant cause of influenza illness within an outbreak or aggregate season, but identifying which strain will dominate in an upcoming season is not so trivial. Further compounding uncertainty, the dynamics and epidemiological profiles of outbreaks caused can vary considerably among populations and seasons, even when caused by closely related strains. For example, consider that there is a four-fold discrepancy in the range of point estimates for single season burdens of symptomatic illness, i.e., 9.3 million versus 41 million symptomatic illnesses estimated for the 2011-2012 and 2017-2018 influenza seasons, respectively (23). These aspects directly impact public health preparedness, particularly in strain selection for seasonal influenza vaccine design as mismatches between strains included in vaccines and those circulating can contribute to reduced vaccine effectiveness (32).

Box 2. Reproduction Numbers

The basic reproductive number, R_0 (R-naught), is an estimated figure of the average number of secondary cases generated by a single infectious case (of any communicable disease) in a completely susceptible population. Influenza is thoroughly ingrained in human ecology making it difficult to calculate R_0. The effective reproductive number, R_e, is a similar metric without the stipulation of a population being completely susceptible, a more immediate and practical measure of transmission. For seasonal influenza, R_e is estimated to be around 1.19-1.37 (33,34); this means that seasonal influenza cases generally transmit to 1-2 others. This measure of transmission has shown to vary among influenza strains; R_e estimates are around 1.47-2.27 for the 1918 influenza pandemic and 1.3-1.7 for the 2009 influenza pandemic (33). More recently, Parino et al (2024) estimates of the maximum value of R_e within a seasonal outbreak to be ~2.25 for IAV and ~1.5 for IBV (35). Additionally, while reproduction numbers vary among influenza strains, they also can vary among the specific populations which harbor influenza outbreaks (36).

TRANSMISSION ACROSS SCALES

Molecular Basis

Antigenic novelty significantly impacts the fitness of influenza viruses through a complex interplay of immune escape and functional constraints. Antigenic drift allows viruses to evade host immunity, driving the continuous replacement of circulating strains (37,38). Viral HA is robust to mutation (39), and seemingly minor differences in protein sequences can correspond to distinct antigen profiles (37,40). As the immunodominant

antigen, specific mutations encoding changes in HA antigenicity can confer an advantage to novel variants as the ability to evade adaptive immunity may correspond to an increased viral fitness. However robust and efficient, evolutionary trajectories are limited as viral function must be maintained. Aside from mutations impeding specific protein viability or function, evolution is further constrained by the need to balance complex relationships and interactions among viral components, e.g., opposing mechanisms of binding and release between HA and NA (41). Additionally, the fitness landscape is complex and dynamic, with some mutations only becoming deleterious in later strains, and molecular evolution does not always follow locally optimal pathways (41,42).

IAV and IBV differ in their potential to generate diversity. Similar influenza viruses can trade whole genome segments, but IAV and IBV have diverged extensively and reassortment across types is no longer possible (43,44). This mechanism of genomic change, referred to as antigenic shift, contributes to pandemic IAV, but more subtle genomic change based on the accumulation of mutations, called antigenic drift, contributes to the ability of seasonal influenza viruses to repeatedly invade human populations (40). IAV and IBV differ in their mutation rates, which is suggested to be rooted in RNA polymerase differences (25); H3N2 strains tend to drift more than H1N1 and IBV, which have more stable antigenic profiles (25,40). Although seemingly subtle, these molecular scale differences are the basis for substantial differences observed at larger scales in influenza epidemiology.

Spatial Heterogeneity

Global patterns of circulation differ among strains of influenza viruses. H3N2 circulates globally, while H1N1 and both IBV have more limited geographic ranges and may even persist locally during the off season (40). Southeast Asia, China, and India have been shown to act as important sources of influenza viral diversity (45). Global circulation patterns do not strictly adhere to a source-sink model of viral gene flow. Rather, influenza viruses will sometimes exhibit dynamic metapopulation structuring with lineages traversing the globe along geographic pathways outside of those expected from source-sink dynamics (46). Further supporting this notion and reinforcing the concept of dynamic metapopulation, evidence suggests that no influenza virus strains persist in the local contexts of outbreaks (45); instead, influenza viruses may persist by jumping from outbreak to outbreak within or across regions (29,45), reinforcing the concept of dynamic metapopulation and the importance of spatial epidemiology . However, much of what contributes to a region's ability to incubate variant lineages and act as a global transmission corridor remains only speculated.

Human mobility is a well-characterized driver of spatial spread across scales. Modes of transport and other factors underlying mobility patterns, e.g., motivating reason and distance, seem to have differential importance depending on the focal scale. Passenger air travel significantly influences the global spread of influenza, especially that related to H1N1 and H3N2 (47,48), and domestic airline travel volume, particularly around Thanksgiving, can help to predict the rate of influenza spread in the US (49). However, at increasingly local levels, more geographically limited mechanisms of mobility, e.g., work commutes and non-routine travel, are stronger predictors of influenza

spread than air traffic (50–52). Simulation modeling efforts recapitulate the impact of scale on the dynamics showing that population interconnectivity via passenger air travel alone is insufficient to reproduce observed patterns of spatial spread within regions (53,54). Continuing down the scale, the patterns of spread between cities in the US exhibit even more local patterns with stronger relationships with geographic distance than already locally biased work commutes (52).

At the regional level and below, other factors in addition to mobility have been shown to impact spatial patterns in seasonal influenza epidemiology. Beyond mobility, spatial hierarchies in epidemic spread, first described in the context of measles epidemics (55), may be related to differences in populations' abilities to host outbreaks, e.g., population size and density or gradients of seasonal forcing (27,28,50,56). Along the lines of host-density-dependent transmission (57), it is intuitive that influenza outbreaks in more populous locations would be larger and more extensive. However, Dalziel et al (2018) showed that seasonal influenza epidemics tend to be more diffuse, or spread out over time, in cities with larger populations and more crowding, suggested to result from increased off season transmission (58).

In addition to aspects more directly related to host population organization and mixing, ecological interactions can shape disease and outbreak dynamics. Seasonal influenza viruses exhibit complex ecological interactions, including competition and cooperation among different types and subtypes. Studies have found evidence suggesting interference between influenza strains (59–61) and between influenza strains and other respiratory pathogens, such as respiratory syncytial virus (RSV) (62,63). Interactions as these impact population dynamics potentially through immune-mediated interference

(63,64). Though potentially less impactful to transmission dynamics, coinfections with bacterial pathogens, such as *Streptococcus pneumoniae*, are an inherent risk of influenza and can impact an individual's disease severity and outcomes (65).

DEFINING SCALE WITHIN THE US FOR SEASONAL INFLUENZA

Necessary Scale Considerations

Much of the variation in patterns of seasonal influenza epidemiology has a molecular basis, but there are also numerous larger-scale, ecological factors that influence influenza epidemiology and viral population biology; that is, influenza disease dynamics are shaped by mechanisms across scales from either direction. Seasonal influenza epidemiology is complex and dynamic, coupled with viral evolution and dependent on scale. This nature of the infectious disease system (1) has continually frustrated public health efforts towards prevention and control, e.g., vaccine effectiveness, and, relatedly, (2) necessitates cross-scaling perspectives and study. However, cross-scaling studies are inherently challenging, owing partly to difficulties in compiling fragmented, disparate data and the need for multi-disciplinary approaches (66,67). An additional part of the challenge in cross-scaling approaches is that the concept of scale is somewhat abstract and lacks a concrete definition, an aspect that is needed for practical applications, e.g., units of observation or analysis. Individual hosts are well-defined units [of infection], but distinctions become less clear at larger scales, e.g., individual populations or regions. There is an additional challenge when theoretical constructs do not align well with the practical constructs or units of aggregation found in data. For example, cities have been referred to as "the natural unit of an outbreak" (50),

but their administrative/geopolitical boundaries may poorly represent the local population organization and intermixing (53), cf., core-based statistical areas (68).

Regionalizing the United States

The US is the third largest country in the world with respect to both land area and population, with substantial spatial heterogeneity in geography and demography (69). Regional delineations of the US are plentiful. This topic has received much attention in the field of economics research in the descriptions of labor markets (70,71) as well as from the Office of Management and Budget in the characterizations of core-based statistical areas (CBSAs) (72). Rosensteel et al (2021) found similar issues and used a complex network approach to identify an epidemiological geography of the US, and suggested 3-5 epidemiologically distinct regions per flu season (2). Largely, these efforts towards regional delineations work to identify agglomerations of county or county-equivalent areas, irrespective of state borders. However, we recognize the tendency for data on seasonal influenza to be less spatially resolved.

Publicly available surveillance data on seasonal influenza, e.g., reported case counts and location metadata of sampled viral isolates, are reported at the state-level, rather than for smaller geographic units. Regional delineations respective of state borders are still plentiful. However, none are directly related to nor derived with respect to influenza. For example, the US Census Regional Divisions first arose to describe geographical groupings of the colonies (73) and the Department of Human and Health Services Regions function to facilitate governance and communication between federal and local administrations (74). These two examples of regional schema are found in

influenza surveillance and research, e.g., Centers for Disease Control and Prevention FluView (75), but their definitions of regions conflict with each other, i.e., group states differently. So, while many regional delineations are conveniently available, the choice of which to use for influenza research is not trivial. Without due consideration of spatial units of analysis, researchers may open their inferences to unforeseen bias (70); for example, misrepresented spatial heterogeneity could obscure effect size estimates in epidemiological association studies, and improper population partitioning could influence gene flow estimates from phylogeographic studies.

APPROACH & DISSERTATION ORGANIZATION

In this dissertation, I work to address this problem of defining scale within the US. To characterize scale within the US related to seasonal influenza, I set forth three specific aims.

In Aim 1, I characterize regional patterns of human mobility within the US and quantify their associations between mobility and influenza-like illness (ILI) epidemic intensity. Human mobility has been the topic of modeling studies from a variety of disciplines. Recently, Alessandretti et al (2020) described scales of human mobility relating them to hierarchical containers; this represents a paradigm shift away from the scale-free properties of human mobility, a characteristic suggested by the authors to be the result of data aggregation (76). Mobility models have been often used in influenza research, though they tend to be comparatively simple, e.g., gravity-based formulations. Researchers use mobility models to generate synthetic networks which both capture essential but minimal characteristics of human mobility and are able to capture

relationships important to outbreak patterns. Independent studies have found a disjointed relationship relating mobility to distance, particularly in the US, and that models are improved by explicitly modeling short- and long-distance commutes (50,77). As Alessandretti et al (2020) suggest that data aggregation may obscure scale, I believe that this disjointed nature of distance distributions in the US reflects an inherent scale. Furthermore, I hypothesize that mobility patterns summarized according to scale may relate to patterns in influenza outbreaks. To test this, I first identify a critical distance threshold using gravity models fit to county-level commuting flows. Next, I use the identified distance threshold to summarize the commuting data to the state level, the resolution of ILI data. Finally, I explore the association between nested mobility patterns and ILI epidemic intensity using linear regression models.

In Aim 2, I generate data-driven regional delineations of the US and, along with other existing delineations, evaluate their suitability and validity in characterizing influenza transmission zones (ITZs). Geopolitical borders and boundaries, or other administrative geographical units, may not be suitable for describing epidemiologically relevant partitions between populations (2,53). Conversely, treating the US as a single entity or data point effectively ignores substantial spatial heterogeneity, cf., the World Health Organization's North American Influenza Transmission Zone (78). To address this issue, I take a similar approach to Rosensteel et al (2021). However, I expand upon this methodology in several key ways, including a holistic assessment of alternative schemes. I begin by conducting a specific spatial clustering analysis incorporating both ILI incidence data and human mobility to enumerate specific groupings of states that exhibit similar incidence patterns, i.e., weekly rates of change in ILI cases. Next, I

compile clustering results into a pairwise adjacency matrix creating a network representation of incidence patterns. Along with commuting networks, I analyze these networks for community structuring using several iterations of community detection algorithms. The resulting regional delineations are then compared to one another and to other existing delineations by quantifying their ability to capture elements of (1) the commuting networks, (2) ILI clustering, and (3) phylogenetic grouping of H3N2. By incorporating aspects of mobility, disease incidence, and pathogen ancestry in this way, I assess and evaluate the validity of regional delineations to represent influenza transmission zones.

In Aim 3, I investigate spatial variation in the phylogenetic signal of local outbreaks caused by co-circulating seasonal influenza viruses. Seasonal influenza exhibits considerable spatial variation in outbreak dynamics, e.g., epidemic intensity (58). However, it is unclear whether this corresponds to variation in the underlying transmission. Molecular surveillance and genomic epidemiology have become important tools in public health practice and infectious disease research, and, consequently, molecular sequence data have accumulated to a considerable degree. As artifacts of transmission are imprinted into the genomic sequences of pathogens (79,80), this means that systematic characterizations of transmission may now be possible. Furthermore, ecological interactions among and between influenza viruses, and other pathogens exist and shape outbreak dynamics (60,62–64), but the spatial extent of these interactions, or scale, has not been described. To explore these gaps, I take a phylogenetic approach to identify local transmission clusters. The phylogenetic trees of these local transmission clusters are then summarized to quantify the mean pairwise patristic distance, a measure

of phylogenetic diversity. I then correlate the realized diversity among local transmission clusters at various spatial, temporal, and spatiotemporal extents, as well as across influenza subtypes/lineages. In doing so, I characterize implicit relationships and statistical dependence in the evolutionary patterns, or more simply, the transmission chains of local outbreaks.

This dissertation is organized accordingly with the next three chapters corresponding to these three aims. Finally, I conclude with a discussion of the overall thesis in a final chapter, drawing conclusions from the aggregate work and suggesting promising future directions of this work.

REFERENCES

1. Johnson PTJ, de Roode JC, Fenton A. Why infectious disease research needs community ecology. *Science*. 2015;349(6252):1259504.
2. Rosensteel GE, Lee EC, Colizza V, et al. Characterizing an epidemiological geography of the United States: influenza as a case study. 2021;2021.02.24.21252361. (https://www.medrxiv.org/content/10.1101/2021.02.24.21252361v1). (Accessed July 11, 2024)
3. Hanski I. Metapopulation dynamics. *Nature*. 1998;396(6706):41–49.
4. Fowler CS, Jensen L. Bridging the gap between geographic concept and the data we have: The case of labor markets in the USA. *Environ Plan A*. 2020;52(7):1395–1414.
5. Turtle J, Riley P, Ben-Nun M, et al. Accurate influenza forecasts using type-specific incidence data for small geographic units. *PLOS Computational Biology*. 2021;17(7):e1009230.
6. Morgenstern H. Ecologic Studies in Epidemiology: Concepts, Principles, and Methods. *Annual Review of Public Health*. 1995;16(Volume 16, 1995):61–81.
7. Gerrymandering | Definition, Litigation, & Facts | Britannica. 2024;(https://www.britannica.com/topic/gerrymandering). (Accessed October 7, 2024)
8. CDC. Key Facts About Influenza (Flu). *Centers for Disease Control and Prevention*. 2024;(https://www.cdc.gov/flu/about/keyfacts.htm). (Accessed August 13, 2024)

9. Morris SE, Nguyen HQ, Grijalva CG, et al. Influenza virus shedding and symptoms: Dynamics and implications from a multi-season household transmission study. 2024;2024.03.04.24303692. (https://www.medrxiv.org/content/10.1101/2024.03.04.24303692v2). (Accessed July 11, 2024)
10. CDC. Flu Symptoms & Complications. *Centers for Disease Control and Prevention*. 2022;(https://www.cdc.gov/flu/symptoms/symptoms.htm). (Accessed August 13, 2024)
11. Smith AM, McCullers JA. Secondary Bacterial Infections in Influenza Virus Infection Pathogenesis. *Influenza Pathogenesis and Control - Volume I*. 2014;385:327–356.
12. U.S. Influenza Surveillance: Purpose and Methods | CDC. 2023;(https://www.cdc.gov/flu/weekly/overview.htm). (Accessed August 13, 2024)
13. Billings WZ, Cleven A, Dworaczyk J, et al. Use of Patient-Reported Symptom Data in Clinical Decision Rules for Predicting Influenza in a Telemedicine Setting. *J Am Board Fam Med*. 2023;36(5):766–776.
14. Ebell MH, Rahmatullah I, Cai X, et al. A Systematic Review of Clinical Prediction Rules for the Diagnosis of Influenza. *J Am Board Fam Med*. 2021;34(6):1123–1140.
15. Diagnosing Flu | CDC. 2022;(https://www.cdc.gov/flu/symptoms/testing.htm). (Accessed August 13, 2024)
16. Information for Clinicians on Influenza Virus Testing | CDC. 2023;(https://www.cdc.gov/flu/professionals/diagnosis/index.htm). (Accessed August 13, 2024)

17. Orthomyxoviruses: Structure of Antigens. 2016;(https://www.sciencedirect.com/science/article/pii/B9780128012383957210). (Accessed August 21, 2024)
18. Krammer F, Smith GJD, Fouchier RAM, et al. Influenza. *Nat Rev Dis Primers*. 2018;4(1):1–21.
19. Chen X, Liu S, Goraya MU, et al. Host Immune Response to Influenza A Virus Infection. *Front. Immunol.* [electronic article]. 2018;9. (https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2018.00320/full). (Accessed September 2, 2024)
20. Prachanronarong KL, Canale AS, Liu P, et al. Mutations in Influenza A Virus Neuraminidase and Hemagglutinin Confer Resistance against a Broadly Neutralizing Hemagglutinin Stem Antibody. *Journal of Virology*. 2019;93(2):10.1128/jvi.01639-18.
21. Zost SJ, Wu NC, Hensley SE, et al. Immunodominance and Antigenic Variation of Influenza Virus Hemagglutinin: Implications for Design of Universal Vaccine Immunogens. *The Journal of Infectious Diseases*. 2019;219(Supplement_1):S38–S45.
22. Tokars JI, Olsen SJ, Reed C. Seasonal Incidence of Symptomatic Influenza in the United States. *Clinical Infectious Diseases*. 2018;66(10):1511–1518.
23. CDC. Burden of Influenza. *Centers for Disease Control and Prevention*. 2024;(https://www.cdc.gov/flu/about/burden/index.html). (Accessed July 11, 2024)
24. Topham DJ, DeDiego ML, Nogales A, et al. Immunity to Influenza Infection in Humans. *Cold Spring Harb Perspect Med*. 2021;11(3):a038729.

25. Petrova VN, Russell CA. The evolution of seasonal influenza viruses. *Nat Rev Microbiol*. 2018;16(1):47–60.

26. CDC. Learn more about the flu season. *Centers for Disease Control and Prevention*. 2022;(https://t.cdc.gov/C03). (Accessed July 11, 2024)

27. Neumann G, Kawaoka Y. Seasonality of influenza and other respiratory viruses. *EMBO Molecular Medicine*. 2022;14(4):e15352.

28. Moriyama M, Hugentobler WJ, Iwasaki A. Seasonality of Respiratory Viral Infections. *Annual Review of Virology*. 2020;7(Volume 7, 2020):83–101.

29. Nelson MI, Simonsen L, Viboud C, et al. Phylogenetic Analysis Reveals the Global Migration of Seasonal Influenza A Viruses. *PLOS Pathogens*. 2007;3(9):e131.

30. Heikkinen T, Ikonen N, Ziegler T. Impact of Influenza B Lineage-Level Mismatch Between Trivalent Seasonal Influenza Vaccines and Circulating Viruses, 1999–2012. *Clinical Infectious Diseases*. 2014;59(11):1519–1524.

31. Paul Glezen W. Editorial Commentary: Changing Epidemiology of Influenza B Virus. *Clinical Infectious Diseases*. 2014;59(11):1525–1526.

32. Vaccines. (https://www.who.int/teams/global-influenza-programme/vaccines). (Accessed October 7, 2024)

33. Biggerstaff M, Cauchemez S, Reed C, et al. Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature. *BMC Infect Dis*. 2014;14(1):480.

34. Uyeki TM, Hui DS, Zambon M, et al. Influenza. *The Lancet*. 2022;400(10353):693–706.

35. Parino F, Gustani-Buss E, Bedford T, et al. Integrating dynamical modeling and phylogeographic inference to characterize global influenza circulation. 2024;2024.03.14.24303719. (https://www.medrxiv.org/content/10.1101/2024.03.14.24303719v1). (Accessed July 11, 2024)
36. Delamater PL, Street EJ, Leslie TF, et al. Complexity of the Basic Reproduction Number (R0). *Emerg Infect Dis*. 2019;25(1):1–4.
37. Smith DJ, Lapedes AS, de Jong JC, et al. Mapping the Antigenic and Genetic Evolution of Influenza Virus. *Science*. 2004;305(5682):371–376.
38. Ferguson NM, Galvani AP, Bush RM. Ecological and immunological determinants of influenza evolution. *Nature*. 2003;422(6930):428–433.
39. Thyagarajan B, Bloom JD. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *eLife*. 2014;3:e03300.
40. Bedford T, Suchard MA, Lemey P, et al. Integrating influenza antigenic dynamics with molecular evolution. *eLife*. 2014;3:e01914.
41. Gong LI, Suchard MA, Bloom JD. Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife*. 2013;2:e00631.
42. Wu NC, Otwinowski J, Thompson AJ, et al. Major antigenic site B of human influenza H3N2 viruses has an evolving local fitness landscape. *Nat Commun*. 2020;11(1):1233.
43. Chastagner A, Hervé S, Bonin E, et al. Spatiotemporal Distribution and Evolution of the A/H1N1 2009 Pandemic Influenza Virus in Pigs in France from 2009 to 2017:

Identification of a Potential Swine-Specific Lineage. *Journal of Virology*. 2018;92(24):10.1128/jvi.00988-18.

44. Sreenivasan CC, Sheng Z, Wang D, et al. Host Range, Biology, and Species Specificity of Seven-Segmented Influenza Viruses—A Comparative Review on Influenza C and D. *Pathogens*. 2021;10(12):1583.
45. Bedford T, Cobey S, Beerli P, et al. Global Migration Dynamics Underlie Evolution and Persistence of Human Influenza A (H3N2). *PLOS Pathogens*. 2010;6(5):e1000918.
46. Bahl J, Nelson MI, Chan KH, et al. Temporally structured metapopulation dynamics and persistence of influenza A H3N2 virus in humans. *Proceedings of the National Academy of Sciences*. 2011;108(48):19359–19364.
47. Lemey P, Rambaut A, Bedford T, et al. Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2. *PLOS Pathogens*. 2014;10(2):e1003932.
48. Cheng C, Li J, Liu W, et al. Modeling analysis revealed the distinct global transmission patterns of influenza A viruses and their influencing factors. *Integrative Zoology*. 2021;16(6):788–797.
49. Brownstein JS, Wolfe CJ, Mandl KD. Empirical Evidence for the Effect of Airline Travel on Inter-Regional Influenza Spread in the United States. *PLOS Medicine*. 2006;3(10):e401.
50. Viboud C, Bjørnstad ON, Smith DL, et al. Synchrony, Waves, and Spatial Hierarchies in the Spread of Influenza. *Science*. 2006;312(5772):447–451.

51. Stark JH, Cummings DAT, Ermentrout B, et al. Local Variations in Spatial Synchrony of Influenza Epidemics. *PLOS ONE*. 2012;7(8):e43528.

52. Charu V, Zeger S, Gog J, et al. Human mobility and the spatial transmission of influenza in the United States. *PLOS Computational Biology*. 2017;13(2):e1005382.

53. Balcan D, Colizza V, Gonçalves B, et al. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*. 2009;106(51):21484–21489.

54. Balcan D, Gonçalves B, Hu H, et al. Modeling the spatial spread of infectious diseases: The GLobal Epidemic and Mobility computational model. *Journal of computational science*. 2010;1(3):132–145.

55. Grenfell BT, Bjørnstad ON, Kappey J. Travelling waves and spatial hierarchies in measles epidemics. *Nature*. 2001;414(6865):716–723.

56. Morris SE, Blasio BF de, Viboud C, et al. Analysis of multi-level spatial data reveals strong synchrony in seasonal influenza epidemics across Norway, Sweden, and Denmark. *PLOS ONE*. 2018;13(5):e0197519.

57. Disease Ecology | Learn Science at Scitable. (https://www.nature.com/scitable/knowledge/library/disease-ecology-15947677/). (Accessed October 7, 2024)

58. Dalziel BD, Kissler S, Gog JR, et al. Urbanization and humidity shape the intensity of influenza epidemics in U.S. cities. *Science*. 2018;362(6410):75–79.

59. Suzuki A, Mizumoto K, Akhmetzhanov AR, et al. Interaction Among Influenza Viruses A/H1N1, A/H3N2, and B in Japan. *International Journal of Environmental Research and Public Health*. 2019;16(21):4179.

60. Perofsky AC, Huddleston J, Hansen C, et al. Antigenic drift and subtype interference shape A(H3N2) epidemic dynamics in the United States. *eLife* [electronic article]. 2024;13. (https://elifesciences.org/reviewed-preprints/91849). (Accessed July 11, 2024)

61. Chen Y, Tang F, Cao Z, et al. Global pattern and determinant for interaction of seasonal influenza viruses. *Journal of Infection and Public Health*. 2024;17(6):1086–1094.

62. Chen J, Gokhale DV, Liu L, et al. Characterizing Potential Interaction Between Respiratory Syncytial Virus and Seasonal Influenza in the U.S. 2023;2023.10.04.23296424. (https://www.medrxiv.org/content/10.1101/2023.10.04.23296424v1). (Accessed July 11, 2024)

63. Nickbakhsh S, Mair C, Matthews L, et al. Virus–virus interactions impact the population dynamics of influenza and the common cold. *Proceedings of the National Academy of Sciences*. 2019;116(52):27142–27150.

64. Yang W, Lau EHY, Cowling BJ. Dynamic interactions of influenza viruses in Hong Kong during 1998-2018. 2019;19008987. (https://www.medrxiv.org/content/10.1101/19008987v1). (Accessed September 2, 2024)

65. Shrestha S, Foxman B, Weinberger DM, et al. Identifying the Interaction Between Influenza and Pneumococcal Pneumonia Using Incidence Data. *Science Translational Medicine*. 2013;5(191):191ra84-191ra84.

66. Morgan OW, Abdelmalik P, Perez-Gutierrez E, et al. How better pandemic and epidemic intelligence will prepare the world for future threats. *Nat Med*. 2022;28(8):1526–1528.

67. Grenfell BT, Pybus OG, Gog JR, et al. Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. *Science*. 2004;303(5656):327–332.

68. Bureau UC. Statistical Areas. *Census.gov*. (https://www.census.gov/programs-surveys/metro-micro/about.html). (Accessed September 2, 2024)

69. United States. *The World Factbook*. 2024;(https://www.cia.gov/the-world-factbook/countries/united-states/). (Accessed July 11, 2024)

70. Fowler CS, Jensen L. Bridging the gap between geographic concept and the data we have: The case of labor markets in the USA. *Environ Plan A*. 2020;52(7):1395–1414.

71. Nelson GD, Rae A. An Economic Geography of the United States: From Commutes to Megaregions. *PLOS ONE*. 2016;11(11):e0166083.

72. Bureau UC. Metropolitan and Micropolitan Statistical Areas Map (March 2020). *Census.gov*. (https://www.census.gov/geographies/reference-maps/2020/geo/cbsa.html). (Accessed May 17, 2024)

73. US Census Bureau CHS. Regions and Divisions - History - U.S. Census Bureau. (https://www.census.gov/history/www/programs/geography/regions_and_divisions.html). (Accessed April 27, 2022)

74. Affairs (IEA) O of I and E. HHS Regional Offices. 2006;(https://www.hhs.gov/about/agencies/iea/regional-offices/index.html). (Accessed October 7, 2024)

75. FluView Interactive | CDC. 2023;(https://www.cdc.gov/flu/weekly/fluviewinteractive.htm). (Accessed September 2, 2024)
76. Alessandretti L, Aslak U, Lehmann S. The scales of human mobility. *Nature*. 2020;587(7834):402–407.
77. Truscott J, Ferguson NM. Evaluating the Adequacy of Gravity Models as a Description of Human Mobility for Epidemic Modelling. *PLOS Computational Biology*. 2012;8(10):e1002699.
78. Influenza Transmission Zones. (https://www.who.int/publications/m/item/influenza_transmission_zones). (Accessed July 11, 2024)
79. Colijn C, Gardy J. Phylogenetic tree shapes resolve disease transmission patterns. *Evolution, Medicine, and Public Health*. 2014;2014(1):96–108.
80. Hayati M. Tree shape statistics and their applications. 2019;(https://summit.sfu.ca/item/19964). (Accessed July 11, 2024)

CHAPTER 2

REGIONAL COMMUTING PATTERNS AND INFLUENZA-LIKE ILLNESS IN THE UNITED STATES[1]

[1] Cody Dailey. To be submitted to a peer-reviewed journal.

ABSTRACT

Human mobility shapes the spread of communicable diseases such as seasonal influenza. Although mobility has often been described as scale-free, several studies modeling commuting patterns have found it necessary to include scale-dependent components, e.g., piece-wise distance functions, to more accurately simulate influenza spread. Whether mobility is scale-free in nature or not, spatial scales are an inherent part of observation and data, e.g., spatial resolution, and spatial misalignment challenges analyses integrating disparate data. However, it remains unclear how to best address spatial misalignment of data, and analytical solutions are often unique for every given application. This challenge is exemplified in commuting and influenza incidence data which are available at the US-county- and US-state-levels, respectively. I hypothesized that by more explicitly considering spatial scale, it may be possible to translate information across the differing spatial resolutions. To investigate this, I first estimate a critical distance threshold distinguishing local and long-distance commutes. Then, I explore the associations between regional summaries of these local commuting patterns and regional influenza-like illness epidemic dynamics. I identified a fairly consistent distance threshold, ~150km, across US Census Regions from separate gravity model fits. Distance-based commuting summaries, e.g., proportions of total commutes that were local, showed a curvilinear relationship with ILI epidemic intensity, with intermediate values of commuting summaries often corresponding to more intense seasonal epidemics. These results suggest that there is an inherent local scale in commuting flows which may

be useful for characterizations of local population mixing and its impact on regional outbreak dynamics.

INTRODUCTION

Patterns in human mobility shape the spatial spread of communicable infectious disease, including seasonal influenza. Traditionally, human mobility has been characterized by scale-free properties, suggesting that movement patterns do not adhere to a specific scale and may be well described with power-law distributions (1–3). More recently, inherent scales of human mobility have been described, and observed scale-free properties are an artifact of data aggregation (4,5).

As a result, mobility models have been extensively used in influenza research. Studies have incorporated mobility in various ways and across scales (3,6–8). Many mechanisms of mobility contribute to influenza epidemics, but the importance of any one is determined by the spatial scale. Influenza viruses spread quite effectively around the globe along complex and dynamic networks of geographic pathways (9–11) well-characterized by passenger air travel (7,11,12). At smaller, subnational spatial scales, the population interconnectivity, and therefore influenza spread, is governed by smaller scale yet more frequent mobility mechanisms such as those related to work and school commutes (6,7,13). Mobility models fit to worker commutes in the US have revealed a somewhat disjointed relationship across space. For example, both Viboud et al (2006) and Truscott et al (2012) fit gravity models to US data using separate terms for short and long distances represented in the data; the gravity model of Viboud et al (2006) fit

commutes at distances less than and greater than 119 km separately (6), and Truscott et al (2012) estimated critical distances of ~150 km and ~300 km for separate model formulations (8).

Here, we posit that this disjointed relationship may be suggestive of an important regional scale within the US and, furthermore, hypothesize that mobility patterns at this scale may impact influenza epidemic dynamics. To test this hypothesis, we take a two-stage approach. First, we estimate a distance threshold distinguishing between short- and long-distance commuting flows. Then, we use this distance threshold to summarize state-level commuting flows into several metrics which are then assessed for their associations with state-level influenza-like illness (ILI) epidemic intensity.

METHODS

Data

All data included in this analysis are publicly available. Data on worker commutes (14), county and state population sizes (15), county and state spatial coordinates (16) and boundaries (17), and regional classifications (18) come from the US Census Bureau. Data on influenza-like illness incidence are from Centers for Disease Control and Prevention (CDC) FluView and the Florida Department of Public Health (19). All data management and analyses were conducted using R (version 4.3.0) and RStudio (20). Scripts are compiled in a reproducible format on GitHub (daileyco/Mobility-Models & daileyco/Influenza-like-Illness).

Mobility Model Fitting

Data Management

Commuting data from two time periods, 2011-2015 and 2016-2020, were combined with population and spatial data all at the county-level. Population size estimates for midpoint years were used to align with the commuting data; population size estimates for 2013 and 2018 were joined with commuting data for 2011-2015 and 2016-2020, respectively. Additional aspects of data alignment are described in further detail in the supplemental methods. Population center coordinates were used to calculate pairwise distances between locations, using Haversine or Great Circle distances.

Altogether, this dataset contains observations of commuting flows (i.e., the number of workers estimated to commute) between pairs of counties or county-equivalent areas, population estimates and coordinates of population centers for both origin/resident and destination/work locations, and the distances for each commuting flow.

Gravity Models

Commuting flows between two locations were modeled using a series of gravity models. The basic formulation of the gravity model characterizes the number of workers, T_{ij}, commuting from origin/resident location i to destination/work location j as

$$T_{ij} = C\frac{P_i^{\beta_1}P_j^{\beta_2}}{d_{ij}^{\beta_3}},$$

where Tij is the commuter flux (number of people) between locations i and j, the origin and destination, respectively, C is a constant / intercept, Pi the origin population size, Pj the destination population size, dij the distance between origin and destination, and $\beta_{1,2,3}$ are power parameters.

We extend this basic gravity model four-fold by including three-way interaction terms using indicators of long-distance commutes and commutes between two large populations, similar to Truscott and Ferguson (2012) (8). The indicator term distinguishing short- and long-distances is set by an additional distance threshold hyperparameter which we estimate as the focal point of this analysis. County population size tertiles were calculated, and commuting flows between counties whose population sizes were both in the upper tertile range were categorized as "commutes between two large populations." Altogether, the base gravity model is estimated separately for four subgroups: (1) short-distance commutes between two large populations, (2) long-distance commutes between two large populations, (3) short-distance commutes between all other population pairs, and (4) long-distance commutes between all other population pairs.

The intercept and power parameters of the gravity model were estimated in tandem with the distance threshold hyperparameter. The gravity model was fit using log-linear regression models. The distance threshold parameter was optimized against the root mean square error (RMSE) of the gravity model predictions; the RMSE is calculated on the log scale comparing model predictions with observed commutes. We estimate these parameters for the aggregate US and for subsets of the data corresponding to the US census region of the origin locations to explore potential regional variation.

Epidemic Intensity Regression

Data Management

In the next stage of our analysis, we investigate patterns in the epidemic intensity of ILI at the state-level. Epidemic intensity was calculated similar to Dalziel et al (2018) (21). Briefly, we calculate the relative distribution of ILI cases over the course of an influenza season and summarize this distribution using Shannon's entropy. The reciprocal of entropy is scaled to a unit interval using the observed minimum and maximum; epidemic intensities closer to zero correspond to more diffuse outbreaks with cases more evenly distributed among weeks, and epidemic intensities closer to one correspond to intense outbreaks with cases more concentrated / distributed among fewer weeks.

We combine data on epidemic intensities with data on population sizes and several spatial area descriptors, e.g., average county size and total state size. We calculate several metrics summarizing the commuting patterns observed at the county-level within each state. Using the distance thresholds characterized in our previous analyses, we categorize county-level commutes based on their distances, referred to as extents. Commuting extents are either internal / intracounty, short-distance, or long-distance, and we summarize the county-level commutes accordingly in counts, proportions, and ratios.

Univariate and bivariate distributions of data were inspected using histograms and scatter plots. Some variables were transformed to mitigate the effects of skewing and extreme values on model fit. Chiefly, the transformations include a square root transformation of epidemic intensity, quarter root transformations of ratios of commuting

extents, and natural logarithm transformations for most others. Additionally, prior to model fitting, each of the covariates were centered and scaled.

Regression

To analyze the variation in epidemic intensity, linear mixed effects regression models were fit to the data. The seasonal ILI epidemic intensity for each state served as the outcome of interest, or response variable, for this analysis. Independent variables included population size and various metrics summarizing spatial organization and mobility. We include two independent random effects for states (space) and influenza season (time).

We first fit a base model including only the random effects for location and season and a single fixed effect for population size. From this base model, we investigate the variation explained by a single additional predictor. That is, each covariate is assessed independently from the others but controlling for location, season, and population size. Point and 95% confidence interval estimates were calculated for each covariate term. As covariate terms were standardized, save those for peak week and county counts, the magnitudes of parameter coefficients are directly comparable and can be interpreted as changes to epidemic intensity values estimated for one standard deviation increases in covariate value.

RESULTS

Three-hundred thirty million twenty-three thousand two-hundred forty-eight people are estimated to have lived in the US in 2018, with over 68 million, 59 million, 124 million, and 77 million people in the Midwest (MW), Northeast (NE), South (S), and West (W), respectively (Supplementary Table A.1). The population is distributed among 3222 (3220 in 2013) county or county-equivalent areas, with 1055 (32.8%), 296 (9.2%), 1422 (44.1%), and 449 (13.9%) counties in the MW, NE, S, and W census regions, respectively. The slight imbalance of both population and spatial units suggests that models fit to the aggregate data would be biased towards the S region.

The commuting data are extensive with over 258 840 observations total accounting for nearly 300 million workers' trips over the ten-year period (Supplementary Table 1). Most workers in the data reported working in their resident county (72.5%); we refer to these as internal / intracounty / zero-distance commutes. The relative frequencies of internal commutes varied slightly across regions with a larger share of workers in the W (83%) and a slightly lower share of workers in the NE (65%) working in their resident county, compared with 71% in both the MW and S. Commuting distances ranged from zero to over 9400 kilometers; the median distances of commutes were 155 km for the MW, 190 km for the NE, 172 km for the S, and 519 km for the W. These marginal distributions suggest some differences with respect to the extent of commutes across regions.

Gravity model fit improved when successively stratified by a distance indicator (i.e., short- vs long-distance), an assortative population size indicator (i.e., between two

large populations or not), the census region of the origin/residence location, and the time period of data (all $p<0.001$, Supplemental Table A.2).

Trips longer than 120 km, 138 km, 136 km, and 189 km were identified as long-distance commutes in the MW, NE, S, and W regions, respectively (Figure 2.1). The estimated distance thresholds seem to be relatively consistent across the two time periods of data for all regions, save the NE region whose estimated distance threshold is <100 km for commutes from 2016-2020. This finding is sensitive to changes in the objective function used in the optimization procedure, particularly in whether the error between observed and predicted values is calculated for the data on a log scale or as counts (Supplementary Figure A.1). This suggests that there may be some regional heterogeneity as to a threshold distinguishing short- and long-distances in commuting flows. Comparing estimates of the gravity model power parameters, we find some slight differences among modeled subgroups, including regions. Power parameter estimates seem most heterogeneous across short-distance and long-distance commutes (Supplemental Figure A.2).

Using these distance thresholds, we summarize the commuting data by categorizing commuting flows as internal, short-distance, or long-distance. Across all locations, short-distance commutes accounted for an average of 25% (SD=11.3%) of all commutes and long-distance commutes accounted for an average of 1.33% (SD=0.54%) (Supplemental Table A.4). Comparing commuting extent within each state, the average ratio of short-distance to internal commutes is 0.37 (SD=0.23), the average ratio of long-distance to short-distance commutes is 0.13 (SD=0.34), and the average ratio of long-distance to

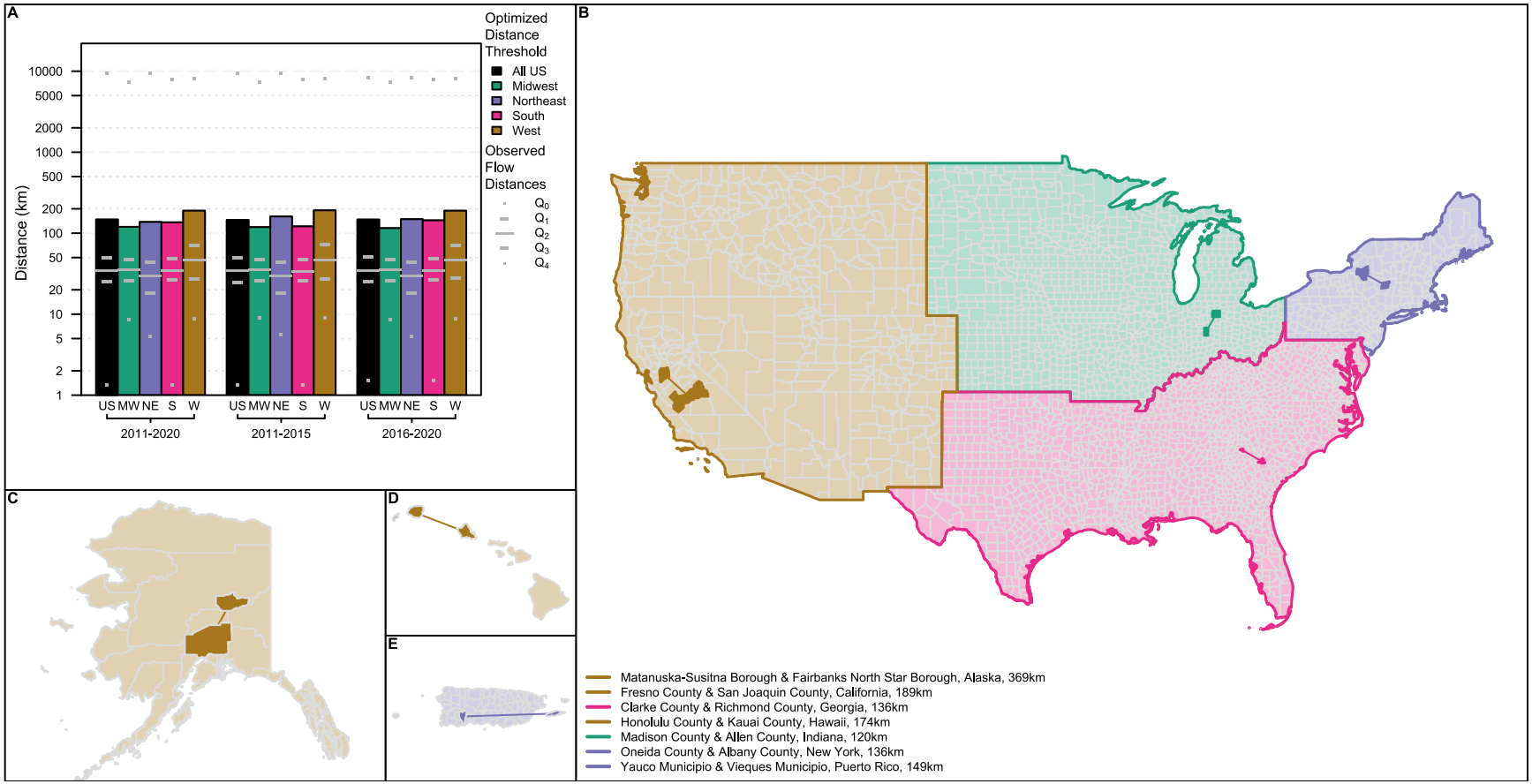


Figure 2.1. Estimated Distance Thresholds used in Gravity Models (A) and Select Examples of Commuting Flows near the Distance Threshold in the mainland US (B), Alaska (C), Hawaii (D), and Puerto Rico (E). Distance thresholds represent a hyperparameter distinguishing the piece-wise components of the distance decay terms in a gravity model. The selected examples of commuting flows are at distances similar to the estimated distance thresholds, highlighting the relative differences in scale between local mobility, US counties, and US Census Regions. Note that the size of county-equivalent areas in Alaska and the selected commuting flow distance are much larger than found elsewhere.

internal commutes is 0.02 (SD=0.007). This suggests that the commuting patterns within each state could be heterogeneous.

ILI data were mostly complete, save for Puerto Rico in the 2011-2012 and 2012-2013 influenza seasons (Supplemental Figure A.3). The average epidemic intensity for all locations across all seasons was 0.227 (Supplemental Table A.3); season averages ranged from 0.152 for the 2011-2012 season to 0.288 for the 2017-2018 season. On average, Delaware experienced the most intense epidemics, 0.59 averaged across all seasons, and the District of Columbia experienced the most diffuse epidemics, 0.053 averaged across all seasons. Season specific trends in epidemic intensity are seen somewhat universally across all locations; similarly, but to a lesser extent, many states seem to exhibit consistent patterns of epidemic intensity across all seasons (Supplementary Figure A.4). Taken together, this suggests that autocorrelation is substantial within seasons and within locations.

These commuting extent summary metrics, along with various others, were included as fixed effects in mixed-effects linear regression models of epidemic intensity (main variables depicted in Figure 2.2). Upon noticing potentially non-monotonic, curvilinear relationships during data exploration, we decided to include polynomial terms, through third-order/cubic terms, for the assessed independent variables. When compared to a baseline model including random effects for season and location (Supplementary Figures A.5 & A.6) and a fixed effect for population size, we observe significant ($p<0.05$) model fit improvements from analyses of variance when including terms for the proportion of internal commutes ($p=0.035$), the ratio of short-distance to internal commutes ($p=0.047$), and the ratio of long-distance to internal commutes

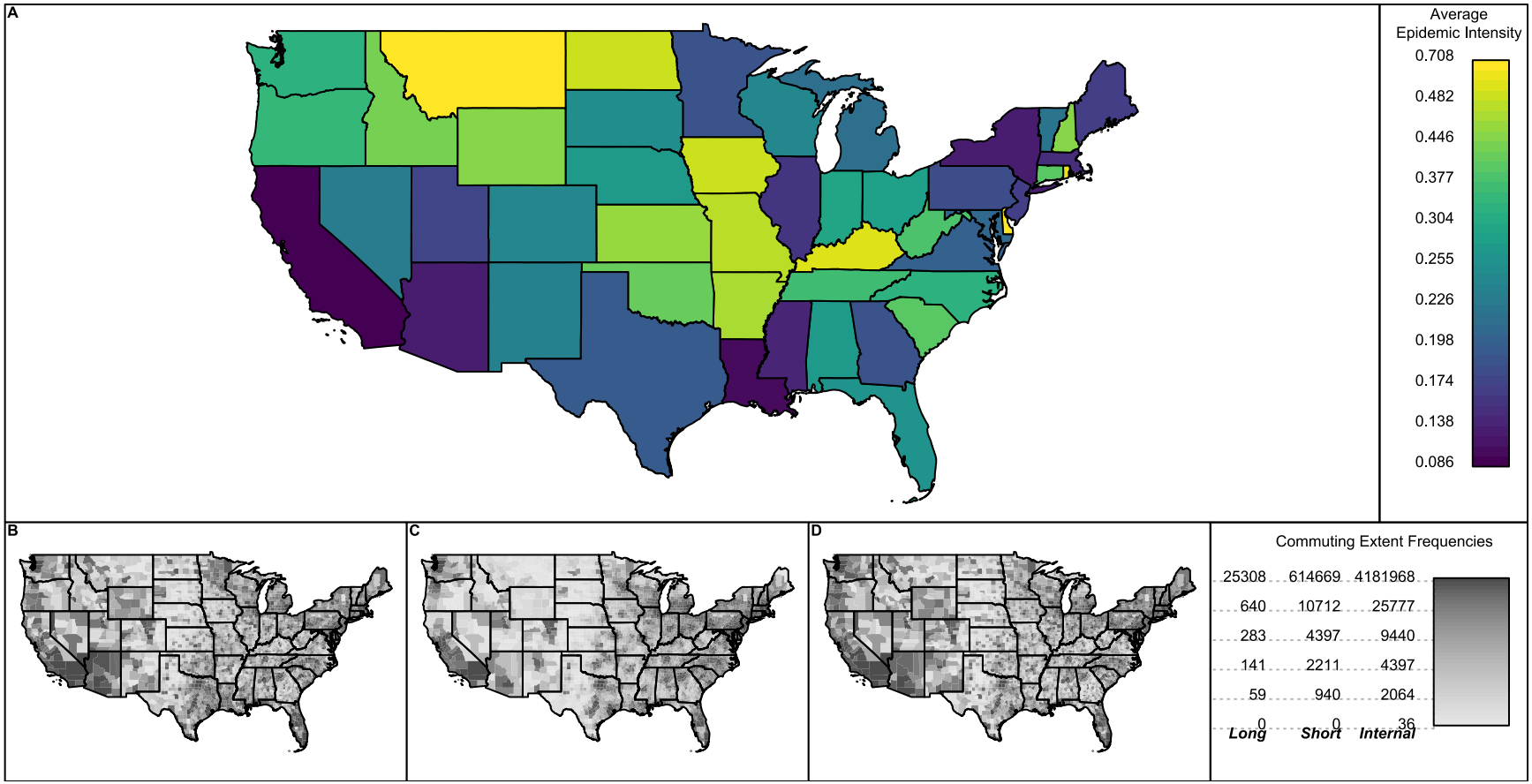


Figure 2.2. State-level Influenza-like Illness Seasonal Epidemic Intensity (A) and County-level Commuting Summaries of Long-distance (B), Short-distance (C), and Zero-distance / Internal Commutes (D). Epidemic intensity is Shannon's entropy rescaled to the unit interval where values closer to one correspond to more intense outbreaks with incidence concentrated over relatively few weeks in an influenza season. The graph depicts the epidemic intensity for each state averaged over nine influenza seasons. Commuting extent frequencies show the county-level categorizations of commuting flows which were then summarized to the state-level for use as predictors in regression models of ILI epidemic intensity.

(p=0.002) (Supplementary Table A.5). Additionally, we observe marginally significant ($0.05<p<0.1$) model fit improvement when including terms for the proportion of short-distance commutes (p=0.051) and the proportion of long-distance commutes (p=0.097). For each of these models, the coefficient estimates for the population size term stayed relatively consistent. Though, for the models including the proportion of long-distance commutes and the ratio of long-distance to internal commutes, the coefficient estimates for the population size term (β=-0.041 and β=-0.045, respectively) are slightly lower than estimates for other models (e.g., β=-0.051 when including peak week). This suggests that these commuting extent summary metrics are associated with epidemic intensity and that the effects of long-distance commutes may be somewhat related to those of population size.

By plotting the model curves, we can more clearly see the non-monotonic relationships between the commuting extent summary metrics and epidemic intensity (Figure 3). Even though some cubic terms are statistically significant (Supplementary Table A.5), the prevailing trend in the data seems to be an inverted U-shaped curve. That is, epidemic intensity values tend to be lower valued at either extreme, while they are higher at middling, intermediate values of the commuting extent summary metrics. For example, in states where either relatively few or many (small or large proportions, respectively) workers commute within their residence county, there tend to be more diffuse ILI epidemics. Though, there seems to be substantial unexplained variability (Figure 3 B-D, F-H) and, at most, a moderately sized effect. This suggests that summaries of commuting extent have a slight impact on epidemic intensity, states with

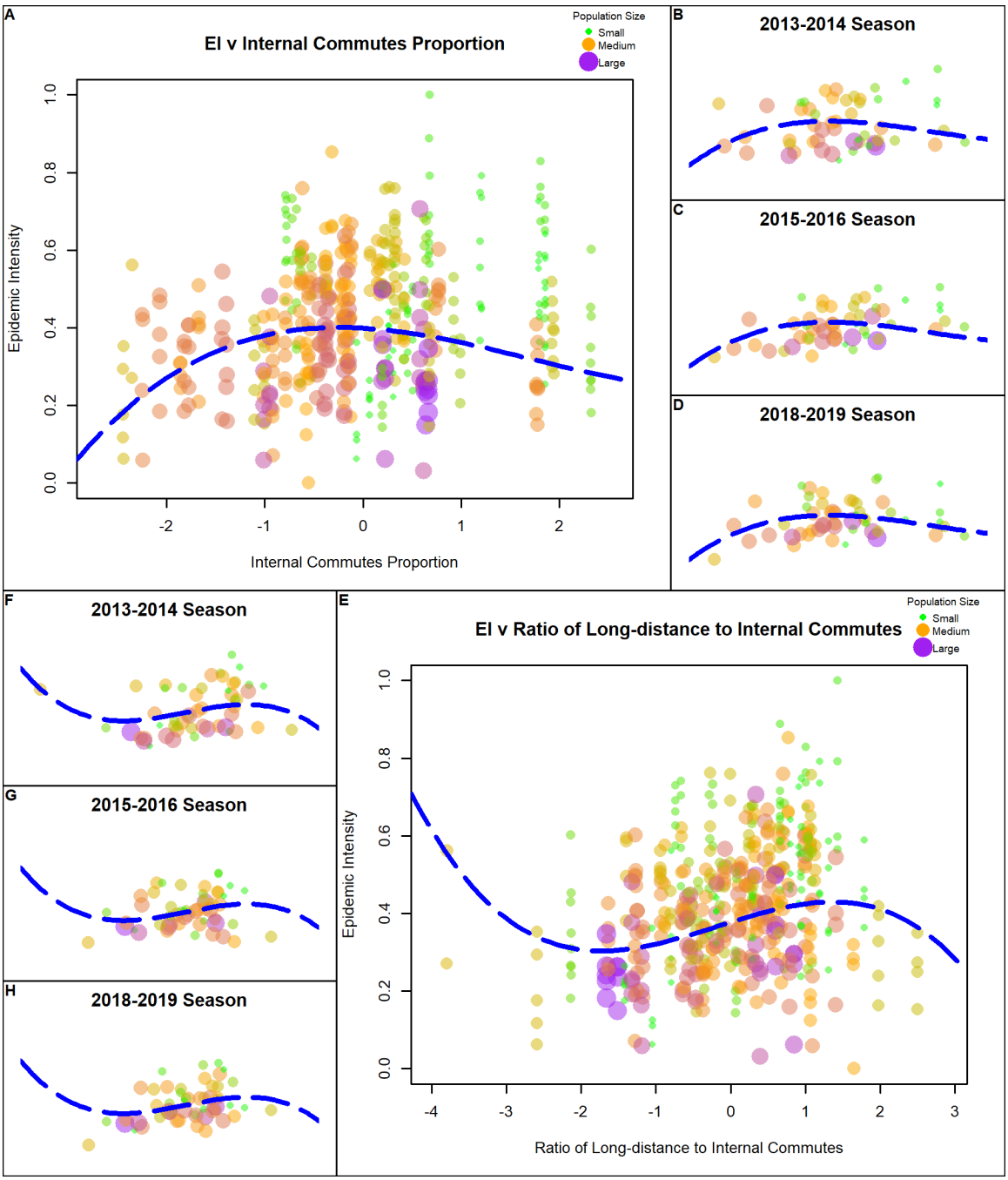


Figure 2.3. Estimated Model Curves for the Relationships between Influenza-like Illness Epidemic Intensity and (A-D) the Proportion of Internal Commutes and (E-H) the Ratio of Long-distance to Internal Commutes. Data points represent a single influenza season for a single US state. As the commuting data are less temporally resolved, observations for each state tend to stack along the x-axis. The curvilinear relationship is weak, and high-leverage values seem to be quite influential in the shape of the model curves. Generally, the relationship between population size and epidemic intensity can be seen along a top-bottom gradient.

intermediately valued commuting patterns experience more intense epidemics, and that the relationship is likely well-described with quadratic terms.

DISCUSSION

Influenza epidemic dynamics are complex and influenced by a variety of mechanisms across scales. Human mobility is one such factor that has been well-characterized as a driver of influenza spread and outbreak dynamics. In this study, we attempted to characterize patterns of human mobility at a sub-national, regional spatial scale within the US and investigate their associations with ILI epidemic intensity. We identified a potentially region-defining distance of ~150 km or ~93 mi by fitting gravity models to worker commutes. This distance may vary slightly across broad regions of the US and may be subject to shifting over time. Moreover, commuting extent summary metrics seemed to both vary substantially among US states and weakly correlate with ILI epidemic intensity.

The estimated distance thresholds seem to agree with estimates from others who employed gravity models fit to commuting data (6,8) as well as to a much more complex mobility model fit to extensive mobile-phone data (4). As such, it seems likely that ~150 km or ~93 mi is a region-defining distance. Human mobility is complex and can be influenced by a variety of factors. The built environment greatly impacts mobility patterns, e.g., via characteristics such as walkability or the availability of public transport (22,23). These aspects may underlie identified scales of human mobility (4). This also means that patterns in human mobility, such as what constitutes a regional scale, may be

spatially heterogeneous, depending on transportation infrastructure. Our results may offer evidence of heterogeneous scale definition as we observed slightly different distance thresholds among US Census regions. However, there may be an alternative explanation for the observed variation. Rather than fundamental differences in a scale-defining distance, there may be several alternative explanations, including data imbalance, heterogeneity in spatial organization of populations, and observation censoring. Particularly, we suspect that interval censoring of spatial distances may have a particularly strong impact. For example, we estimate a larger distance threshold for the W region. However, qualitatively, the size of counties increases along an East-to-West gradient, and many of the counties or county-equivalent areas in the West are large enough to completely contain short-distance displacements, e.g., Matanuska-Susitna Borough, Alaska (Figure 2.1C).

Infectious disease data are often only available at coarser resolutions, e.g., state versus county, whether for reasons of limitations in observation or concerns of privacy. This discrepancy presents a challenge to consolidate disjointed / misaligned data in cross-scaling analyses as it is not trivial to align theoretical structuring with practical constructs/units of aggregation found in data. Generally, this can represent a challenge in integrating disparate data and effectively quantifying nested distributions across scales. Aggregation to a common spatial unit is a simple fix. This is often done in studies that focus on the coupling patterns *between* larger geographic regions found in mobility data (6); in doing so, much of the heterogeneity in mobility nested *within* larger geographic areas is effectively lost (c.f., mixing assumptions in infectious disease models). Some researchers have found ways to translate information across scales. For example, Dalziel

et al (2018) characterize a city's baseline transmission potential in terms of population crowding within city districts (21). Here, we suggest that nested mobility patterns at a regional scale may also be impactful to influenza outbreak dynamics. A simple explanation for this may relate to population mixing. We observed more diffuse ILI epidemics in places where either a small or a large proportion of workers commuting short distances. These places may represent either a higher- or lower-degree of population mixing, respectively. More extensive population mixing may correspond to increased baseline transmission of influenza which relates to more diffuse influenza epidemics, as in the findings of Dalziel et al (2018) (21). Conversely, less extensive population mixing may be interpreted as relative isolation, or a modular/fragmented population through which influenza struggles to spread consistently. Either way, the intermediate values in commuting extent summary metrics may represent a sort of "Goldilocks" scenario with more intense influenza epidemics.

The greatest limitation of our study lies with limitations in the data, specifically in the granularity. The influenza data was well-resolved temporally, with weekly incidence estimates, yet relatively spatially coarse, aggregated to the state level. On the other hand, the commuting data was temporally coarse, 5-year aggregates, yet more spatially granular, county level. Additionally, the commuting data is limited in the spatial relationships it quantifies; less than 2% of all potential / unique county pairings $\binom{3222}{2}$ are represented, and other mechanisms of mobility, e.g., leisure, or mobility of different segments of the population, e.g., children, are not captured. Additionally, as these data are made available using the geopolitical units of county and state, it was not possible for us to assess these as "containers" which impact human mobility scales (4), though other

researchers have noted that state borders may not represent epidemiologically relevant population partitions.

Despite these limitations in the data, we still were able to characterize a sub-national regional scale within the US and relate patterns of mobility about that scale to patterns in ILI epidemic intensity.

REFERENCES

1. Brockmann D, Hufnagel L, Geisel T. The scaling laws of human travel. *Nature*. 2006;439(7075):462–465.

2. Rhee I, Shin M, Hong S, et al. On the Levy-Walk Nature of Human Mobility. *IEEE/ACM Transactions on Networking*. 2011;19(3):630–643.

3. Barbosa H, Barthelemy M, Ghoshal G, et al. Human mobility: Models and applications. *Physics Reports*. 2018;734:1–74.

4. Alessandretti L, Aslak U, Lehmann S. The scales of human mobility. *Nature*. 2020;587(7834):402–407.

5. Boucherie L, Maier BF, Lehmann S. Decomposing geographical and universal aspects of human mobility. 2024;(http://arxiv.org/abs/2405.08746). (Accessed September 2, 2024)

6. Viboud C, Bjørnstad ON, Smith DL, et al. Synchrony, Waves, and Spatial Hierarchies in the Spread of Influenza. *Science*. 2006;312(5772):447–451.

7. Balcan D, Colizza V, Gonçalves B, et al. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*. 2009;106(51):21484–21489.

8. Truscott J, Ferguson NM. Evaluating the Adequacy of Gravity Models as a Description of Human Mobility for Epidemic Modelling. *PLOS Computational Biology*. 2012;8(10):e1002699.

9. Bedford T, Cobey S, Beerli P, et al. Global Migration Dynamics Underlie Evolution and Persistence of Human Influenza A (H3N2). *PLOS Pathogens*. 2010;6(5):e1000918.

10. Bahl J, Nelson MI, Chan KH, et al. Temporally structured metapopulation dynamics and persistence of influenza A H3N2 virus in humans. *Proceedings of the National Academy of Sciences*. 2011;108(48):19359–19364.

11. Parino F, Gustani-Buss E, Bedford T, et al. Integrating dynamical modeling and phylogeographic inference to characterize global influenza circulation. 2024;2024.03.14.24303719. (https://www.medrxiv.org/content/10.1101/2024.03.14.24303719v1). (Accessed July 11, 2024)

12. Lemey P, Rambaut A, Bedford T, et al. Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2. *PLOS Pathogens*. 2014;10(2):e1003932.

13. Charu V, Zeger S, Gog J, et al. Human mobility and the spatial transmission of influenza in the United States. *PLOS Computational Biology*. 2017;13(2):e1005382.

14. Bureau UC. Commuting Flows. *Census.gov*. (https://www.census.gov/topics/employment/commuting/guidance/flows.html). (Accessed July 11, 2024)

15. Bureau UC. County Population Totals and Components of Change: 2020-2023. *Census.gov*. (https://www.census.gov/data/datasets/time-series/demo/popest/2020s-counties-total.html). (Accessed September 2, 2024)

16. Bureau UC. Centers of Population. *Census.gov*. (https://www.census.gov/geographies/reference-files/time-series/geo/centers-population.html). (Accessed July 11, 2024)

17. Bureau UC. Cartographic Boundary Files. *Census.gov*. (https://www.census.gov/geographies/mapping-files/time-series/geo/cartographic-boundary.html). (Accessed July 11, 2024)

18. US Census Bureau CHS. Regions and Divisions - History - U.S. Census Bureau. (https://www.census.gov/history/www/programs/geography/regions_and_divisions.html). (Accessed July 11, 2024)

19. FluView Interactive | CDC. 2023;(https://www.cdc.gov/flu/weekly/fluviewinteractive.htm). (Accessed September 2, 2024)

20. Posit team. RStudio: Integrated development environment for R. Boston, MA: Posit Software, PBC; 2023.(http://www.posit.co/)

21. Dalziel BD, Kissler S, Gog JR, et al. Urbanization and humidity shape the intensity of influenza epidemics in U.S. cities. *Science*. 2018;362(6410):75–79.

22. Barbosa H, Hazarie S, Dickinson B, et al. Uncovering the socioeconomic facets of human mobility. *Sci Rep*. 2021;11(1):8616.

23. Southworth M. Designing the Walkable City. *Journal of Urban Planning and Development*. 2005;131(4):246–257.

CHAPTER 3

INFLUENZA TRANSMISSION ZONES WITHIN THE UNITED STATES[1]

[1]Cody Dailey, Lambodhar Damodaran, & Guppy Stott. To be submitted to a peer-reviewed journal.

ABSTRACT

Seasonal influenza viruses spread around the world. The World Health Organization (WHO) categorizes countries into Influenza Transmission Zones (ITZs). However, these classifications may not appropriately capture the spatial heterogeneity found in large countries. For the United States, I hypothesize there may be important subnational ITZs each capable of harboring influenza outbreaks and shaping influenza epidemiology. This study aims to identify and validate US regional delineations relevant to influenza epidemiology. Utilizing network science community detection, I generate various regional delineations and evaluate their alignment with human mobility, disease incidence, and viral evolution to propose subnational influenza transmission zones in the US. Out of the 173 regional delineations evaluated, many with 8-13 regions showed an increased signal in modularity for commuting networks. A similar signal is seen when comparing fits to influenza-like illness clustering networks, but schema with fewer regions had greater modularity scores, perhaps indicative of a resolution limit given a more sparsely connected network. Tip-trait association indices between regional delineations and H3 phylogenetic trees may suggest better alignment for schema with six or fewer regions, but the delineations with 5-14 regions all fit similarly well. Overall, these results suggest that the US may be comprised of ~8 subnational ITZs. Furthermore, data driven regional delineations produced herein indicate slightly different spatial structuring than existing administrative regional delineations such as the Department of Health and Human Services Regions which is commonly used in influenza research.

INTRODUCTION

Seasonal influenza viruses routinely spread among many regions of the world. Recognizing the global and diffuse nature of influenza transmission, the World Health Organization (WHO) classified nations and territories among 18 influenza transmission zones (ITZs) which represent "geographical groups of countries, areas or territories with similar influenza transmission patterns" (1). While a regional classification as proposed by the WHO is able to aid in international coordination in the prevention and control of influenza on the global scale, its validity in describing *transmission zones* remains to be shown.

The timing and extent of seasonal influenza outbreaks has been used to investigate alternative classifications of ITZs. Caini et al (2017) studied the WHO European Region and report a simpler scheme wherein two ITZs were suggested over the five ITZs outlined by WHO (2). This work has, in turn, garnered its own scrutiny. In an opinion piece, Shin and Manuel (2017) voiced concerns of bias in the representation of large countries as single data points; specifically, they discuss Russia, the largest country in the world with considerable geographic and demographic variation (3). This critique resonates with the delineation of the North American ITZ which is comprised of Bermuda, Canada, Greenland, Saint Pierre and Miquelon, and the United States of America (US) (1). Such a coarse classification misrepresents the heterogeneity of the region and its ability to harbor multiple influenza outbreaks. The US alone is the third largest country in the world in terms of both total area and population (4). As such, the US likely constitutes several separate, subnational ITZs.

Regional delineations of the US are plentiful. This topic has received much attention in the field of economics research in the descriptions of labor markets (5,6) as well as from the Office of Management and Budget in the characterizations of core-based statistical areas (CBSAs) (7). Largely, these efforts towards regional delineations work to identify agglomerations of county or county-equivalent areas, irrespective of state borders. Similarly, Rosensteel et al (2021) developed a county-level "epidemiological geography" of influenza through the use of proprietary data on influenza-like illness (8). However, we recognize the tendency for publicly available data on seasonal influenza to be less spatially resolved. Publicly available surveillance data on seasonal influenza, e.g., reported case counts and location metadata of sampled viral isolates, are often reported at the state-level, rather than for smaller geographic units. Regional delineations respective of state borders are still plentiful. However, none are directly related to nor derived with respect to influenza. For example, the US Census Regional Divisions first arose to describe geographical groupings of the colonies (9). So, while many regional delineations are conveniently available, the choice of which to use for influenza research is not trivial. Without due consideration of spatial units of analysis, researchers may open their inferences to unforeseen bias (5). In this study, we aim to address this gap by both identifying regional delineations of the US that are relevant to the epidemiology of seasonal influenza and validating them as putative influenza transmission zones. We do this in two stages. First, we generate many variations of US regional delineations using a network science community detection approach. Then, we evaluate these regional delineations in their alignment with spatial constructs found in human mobility, disease

incidence, and viral evolution to identify subnational influenza transmission zones in the US.

METHODS

Data

Data for these analyses are publicly available and consist of ILI incidence, commuting flows, geographic centers of population, regional classifications, and phylogenetic trees of seasonal influenza subtype H3N2.

Data on ILI incidence are from Centers for Disease Control and Prevention (CDC) FluView (10) and the Florida Department of Public Health. These data contain weekly counts of reported ILI cases and the population of healthcare recipients from which the ILI cases originated. These counts cover the 50 US states, the District of Columbia, and Puerto Rico (henceforth, collectively referred to as "states") and span October 2011 to September 2020.

State-level cartographic boundary files were downloaded from the US Census Bureau (11). Questionnaire responses concerning the origin and destinations of commuting flows from the American Community Survey (ACS) are summarized as tables and made available by the US Census Bureau (12); tables for 2011-2015 and 2016-2020 were downloaded and included in analysis. The (ACS) commuting data is essential for understanding patterns of daily movement among populations, which can be indicative of economic activity, urban planning needs, and regional connectivity. Data on the spatial distributions of population are also from the US Census (13). These data consist of single point locations, geographic coordinates, representing the geographic

center of population for each state at the time of the decennial census, 2010 and 2020. The 2010 estimates of population center coordinates were aligned with the 2011-2015 estimates of ILI and commuting, and the 2020 estimates were aligned with the 2016-2020 ILI and commuting estimates. State classifications into Census Regions, Census Divisions, and Department of Health and Human Services (HHS) Regions were abstracted and used in comparisons with our data-driven regional delineations. Of note, the US Census Regions nor US Census Divisions include Puerto Rico in their classifications, so, when necessary, we include Puerto Rico as its own separate region within the US Census Regions and US Census Divisions classifications.

Damodaran et al (2023) investigated the phylogeography of seasonal influenza virus H3N2 in the US(14). In their work, a set of empirical trees was sampled from a posterior distribution derived from Markov chain Monte Carlo (MCMC) computation via a Bayesian Evolutionary Analysis Sampling Trees (BEAST) analysis (25). In our study, we use the resulting set of 500 empirical H3N2 trees from Damodaran et al (2023) to compare regional delineations. In doing so, we aim to identify and validate the regional structures that best describe the spread of the virus, providing insights into the dynamics of influenza transmission.

All data management and analysis were conducted using R programming language (version 4.3.0) in the RStudio/Posit interactive developer environment(15), unless otherwise specified. Processing and analytical scripts are made available in GitHub repositories (link daileyco/Influenza-like-Illness-Clustering and daileyco/Spatial-Structuring) to facilitate reproducibility.

Community Detection

We work to characterize US ITZs by generating regional delineations from human mobility and ILI incidence data. To generate regional delineations of the US, we investigate community structuring in a network science framework. Networks were generated from the data to represent coupling between locations. Broadly, we analyze two types of networks: commuting networks and clustering networks. Commuting networks are generated from the commuting flows data, and clustering networks are generated from a focused clustering analysis.

To generate commuting networks, we aggregate the commuting flow data to the state level separately for each time period of data. While our clustering analyses only included *undirected*, state-level commuting networks, in this community analysis, we also analyze the commuting flow data as *directed*, state-level commuting networks. Additionally, we include modified versions of each network which have rescaled edge weights. The edges in the networks were rescaled by standardizing the edge weights according to the total number of commuters originating from each location/node in the directed networks; that is, edge weights in the directed networks were divided by the total sum of edge weights for each origin node to transform edge weights into proportions. As with the original directed commuting network, these scaled networks were aggregated for each unique pair of locations to generate undirected networks.

In addition to networks based purely on commuting data, we generate networks based on the incidence of ILI using a spatial clustering analysis. Briefly, clusters were identified using scan statistics via the SaTScan software (16) by comparing bi-weekly change in the counts of ILI between locations. Our approach to this clustering analysis is

discussed in more detail in the Supplementary Information. Our clustering results consist of two types of clusters, i.e., spatial clusters and commuting network clusters. Each set of clusters was transformed into an adjacency matrix wherein two locations were considered adjacent if a single cluster included both locations. The weights in these adjacency matrices correspond to the frequencies of clustering, i.e., how many times two states clustered together. These adjacency matrices were used to generate two networks: one spatial clusters network and one commuting clusters network.

Altogether, we analyze ten separate networks: two directed commuting networks (one for each time period of data, 2011-2015 and 2016-2020), two undirected commuting networks, two scaled/directed commuting networks, two scaled/undirected commuting networks, and two ILI clustering networks. Network management and analyses are carried out using the igraph package in R (17).

To generate regional delineations of the US, i.e., potential ITZs, we investigate the community structuring in each included network. We do this by passing each network to an array of community detection algorithms. A community in network science refers to a group of nodes more strongly or densely connected to each other than to nodes belonging to other groups; in our case, as nodes correspond to spatial locations, the generated communities reflect groups of states or regions. We use three community detection algorithms: edge-betweenness, Louvain, and Spinglass which all vary in how communities are generated but may each be useful in delineating transmission regions (reviewed in (18)). The Louvain and Spinglass algorithms both contain a hyper-parameter, called resolution, r, and gamma, γ, respectively, which effectively controls the number of communities detected within a network. We implemented a grid search

approach in which these hyperparameters were set to a range of values (r, γ Î [0,0.5,1,1.5,2]) to generate an array of regional delineations which vary in the total number of regions/communities. With the 10 included networks and the 11 variations of community detection algorithms, we attempted 110 independent runs to generate regional delineations of the US. In addition to these data-driven regional delineations, we include three administrative classifications in our comparisons: US Census Regions, US Census Divisions (9), and HHS Regions (19).

Comparison of Regional Delineations

As our goal is to identify US ITZs, we compare our data-driven regional delineations in terms of their fit to the underlying networks from which they were generated, their alignment with grouping patterns in seasonal influenza phylogenies, and their relative balance in community memberships.

To quantify the fit of the regional delineations to commuting and clustering networks, we use network modularity. Clauset, Newman, and Moore (2004) (20), following Newman and Girvan (2003) (21), define the modularity of a weighted network as

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j),$$

where A_{ij} represents the weight of the edge between nodes i and j, $k_i = \sum_j A_{ij}$ is the sum of the weights of the edges attached to, or degree of, node i, $m = \sum_{ij} A_{ij}$ is the total sum

of edge weights in the network, c_i is the community to which node i is assigned, and the δ function $\delta(c_i, c_j)$ is 1 if $c_i = c_j$ and 0 otherwise. Modularity values are strictly less than one; positive values indicate that community ties are stronger than would be expected by random chance, zero values indicate no deviation from randomness, and negative values indicate community ties are weaker than expected by random chance. Simply, higher values of modularity correspond to better alignment of community structure with the underlying relationships in the network. We calculated the modularity for each combination of regional delineation and network in our analysis. Modularity scores were averaged for commuting and clustering networks, separately, to have two composite scores for each regional delineation. These scores were used to rank the regional delineations in their alignments with patterns of either human mobility or disease incidence.

To quantify the alignment between regional delineations and phylogenetic groupings of seasonal influenza viruses, we use the tip-trait association index and parsimony score. Following Wang et al (2001) (22), Parker, Rambaut, and Pybus (2008) (23) define the association index statistic (AI) as

$$AI = \sum_{i=1}^{k} \frac{1 - f_i}{2^{m_i - 1}},$$

where k is the number of internal nodes in a phylogeny, f_i is the frequency of the majority trait among all descendant tips of internal node i, and m_i is the total number of descendant tips of internal node i. Simply, larger values of AI correspond to worse

alignment between trait classifications, here ITZ, and the grouping structure in the phylogeny. The significance of observed phylogeny-trait associations can be tested against null distributions of association index values for a given phylogeny. These null distributions are generated by permuting trait labels among phylogenetic tree tips and recalculating the association index. Furthermore, by averaging values across a set of trees generated from posterior distributions in Bayesian analyses, it is possible to incorporate phylogenetic uncertainty. We use methods similar to those employed in the Bayesian Tip-association Significance testing (BaTS) software (23,24). Briefly, we calculate AI of the observed data with each of the 500 phylogenies in the empirical tree set found in Damodaran et al (2023) (14). The observed trait data are permuted 1000 times and subsequently used to generate a null distribution of AI. We then record the proportion of trees in the null distribution that have an AI value less than or equal to the median AI of the observed trait data and the H3N2 phylogenies. Additionally, we follow the same procedure with the data in the calculations of parsimony scores (PS) to relate the regional delineations to phylogenetic grouping. Calculations of AI and PS were done using Leke Lyu's R package, TTAT (https://github.com/lyu-leke/TTAT). Together, we use the metrics of AI and PS to rank regional delineations in the alignments with patterns of ancestry for an important seasonal influenza virus.

To quantify the relative balance in community membership among regional delineations, we calculate the Shannon entropy for two metrics. First, we use entropy to assess the balance in membership frequencies of states among regions, i.e., does each region have a similar number of constituent states. Second, we use entropy to assess the balance of maximum clade sizes (MCS) (23,24) of each region in the H3N2 empirical

tree set. Damodaran et al (2024) included a subsample of 1000 sequences based on the overall phylogenetic diversity in all available sequences for their phylogenetic reconstructions; as such, we feel that by assessing the balance in MCS, we also indirectly quantify the extent of shared phylogenetic diversity within each region. MCS and respective calculations of entropy are done for each phylogeny in the empirical tree set. Following, we take the median MCS entropy across all trees for a single metric per regional delineation. We use both membership and MCS entropy values to rank regional delineations in their balancing of region size.

Altogether, we ranked each of the regional delineations with respect to their ability to capture spatial patterns in human mobility networks, ILI incidence, and H3N2 phylogenetic grouping while also balancing the sizes of each region with respect to the number of constituent states and, by proxy, the shared phylogenetic diversity. We take each of the four ranking metrics equally to create a composite ranking of the regional delineations. We use this composite ranking in two ways. First, we select the overall best regional delineation. Second, we take the top 50 ranking regional delineations and overlay their boundaries for a composite view of the extent/magnitude of partitioning between states.

RESULTS

ILI data were mostly complete over the 469-week study period, from 2 October 2011 to 20 September 2020. Puerto Rico was the only location with missing data with ~22% of dates missing observations; these missing values mostly correspond to the 2011-

2012 and 2012-2013 influenza seasons, with complete data 2014 onwards. Observations with missing data were simply excluded from clustering analyses.

Using this ILI data, we identified 730 and 671 potential spatial and commuting network clusters, respectively (Supplementary Table 1). Of these, 67 (9.2%) of spatial clusters and 46 (6.9%) of commuting clusters were deemed significant based on permutation tests.

Significant spatial and commuting clusters were used to generate two cluster networks. The spatial clusters network was not fully connected, i.e., not all nodes were connected to another node by an edge and had a total of 892 edges (Supplemental Table 2). The commuting clusters network was fully connected by its 1 060 edges. On average, the clustering networks seem to have relatively larger edge weights than the other commuting networks. The undirected commuting networks have more edges than clustering networks, as do the directed commuting networks, expectedly.

Eighty-five of our 110 runs (77%) of the community detection algorithms identified a total of 144 regional delineations (schematized in Figure 1). All runs of the edge-betweenness community detection algorithm yielded results. The Louvain algorithm runs yielded results for every network except the two directed commuting networks as the algorithm works for undirected networks only. The Spinglass algorithm does not work for graphs that aren't fully connected and, thus, did not yield results for the spatial clustering network. The hierarchy of communities detected from the edge-betweenness algorithm was extracted and all levels were included in our comparisons; this is compared to a single regional delineation for each successful run of the Louvain and Spinglass algorithms which include hyperparameters to similarly yield a set of

community schemes varying in resolution. In addition to our 144 data-driven regional delineations we include 4 comparisons: US Census Regions, US Census Divisions, US HHS Regions, and a regional delineation where each state is its own region (e.g., using the two-letter state code as the classification).

Our regional delineations had a variable number of regions, ranging from 1 up to 51 separate regions with most integers covered in between. Eleven of our regional delineations contained only a single region and 2 others contained only two regions, one of which was solely comprised of Puerto Rico.

When examining the commuting network modularity scores, we find a signal of increased modularity for regional delineations which have a total number of 4-14 separate regions which peaks with delineations having approximately 6-8 regions (Figure 3.1). The administrative regional delineations (US Census Regions, Census Divisions, and HHS Regions) fall within this range and have similarly large modularity values; though, some of our other data-driven regional delineations seem to outperform the administrative delineations with respect to modularity of the commuting networks. This pattern seems consistent across all commuting networks, i.e., for both time periods and regardless of the network modifications we explored. Averaging modularity across all commuting networks, we rank the regional delineations in terms of their alignment with human mobility. The US Census Regions, US Census Divisions, and HHS Regions rank 46^{th}, 54^{th}, and 52^{nd}, respectively, in commuting network modularity.

A similar signal in the modularity scores of clustering networks is observed with a band of observations separated from others at the baseline. However, unlike modularity in the commuting networks, the signal in the modularity of the clustering networks seems

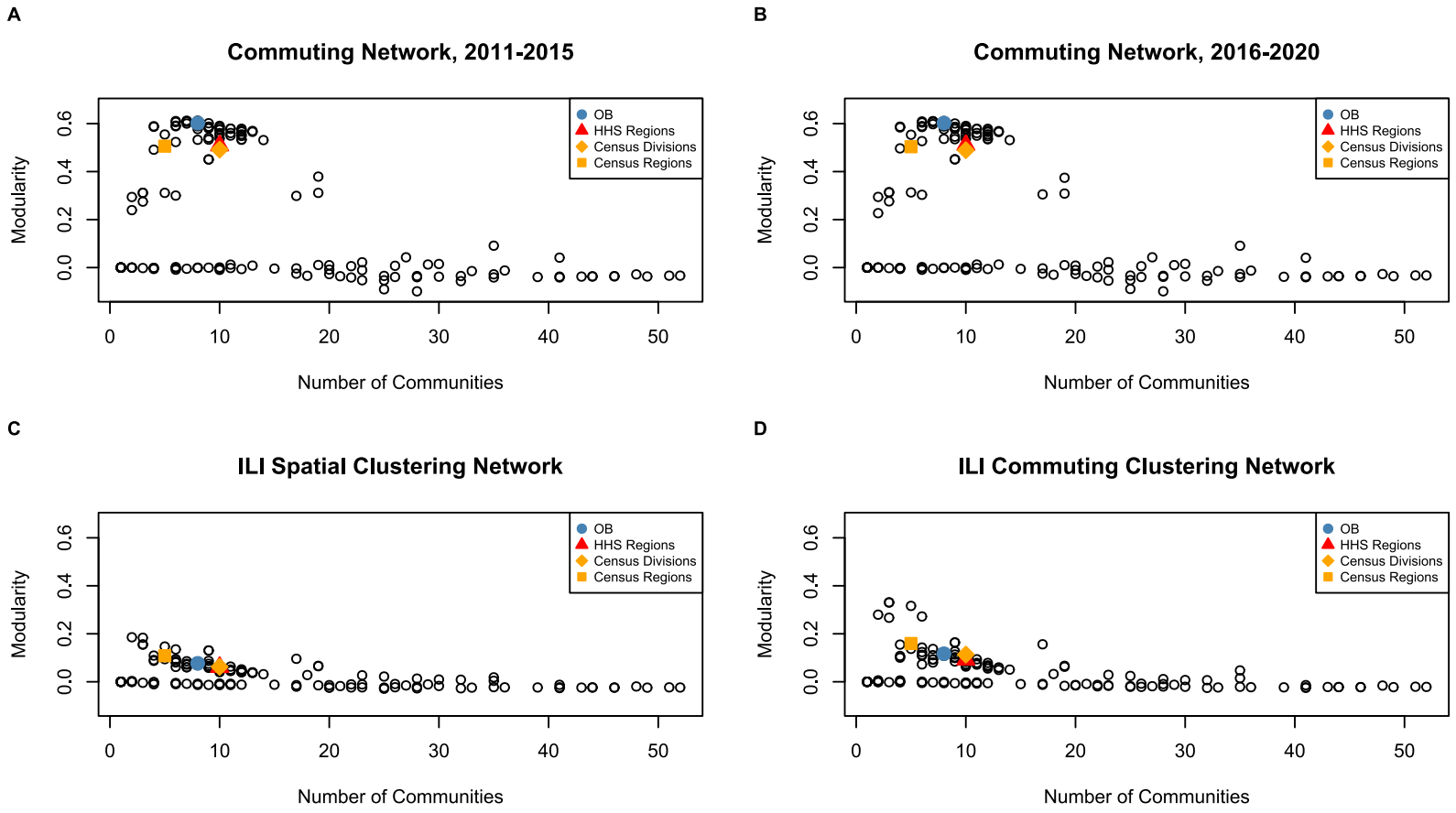


Figure 3.1. Commuting Networks (A,B) and Clustering Networks (C,D) Modularity Scores for Regional Delineations against the Number of Regions. In all networks, nodes were individual states. For the commuting networks, the edges connecting nodes represented the commuting ties between those nodes. For the clustering networks, the edges represented the number of times two nodes were in an identified disease cluster together. Higher modularity values represent a stronger community structure within the given network. The highlighted points show the scores of three administrative regional delineations and a single overall-best-performing regional delineation generated from the community detection analyses.

in favor of fewer regions, peaking for those delineations with 2-3 regions; though, this may be an artifact of a resolution limit due to the relatively sparse connections in the clustering networks compared with the commuting networks (discussed in (25)). Again, the administrative regional delineations similarly matched our regional delineations in their fit to the clustering networks. The tendency for greater clustering network modularity in delineations with fewer regions to seems to explain much of the discrepancy in modularity for the US Census Regions compared to the US Census Divisions, HHS Regions, and our delineations with a similar number of regions; that is, much of the variation is seemingly affected by the differences in the number of regions. Averaging the modularity for each delineation between the two clustering networks, we rank the regional delineations in terms of their alignment with the incidence of influenza-like illness. The US Census Regions, US Census Divisions, and HHS Regions rank 9th, 32nd, and 39th, respectively, in clustering network modularity.

Comparing the phylogenetic tip-trait association indices (AI) shows a slight favor for delineations with fewer regions (Figure 3.2). The values for the AI are less directly comparable across the delineations varying in the number regions; that is, the AI has an implicit bias for traits with fewer numbers of classification levels (Supplemental Figure B.7). The values shown for the AI reflect the proportion of values in our generated null distribution which had an AI statistic less (i.e., more extreme) than the median AI statistic across all trees for the observed data, similar to a p-value. These AI show a less pronounced signal than the modularity scores. Many of the delineations with 5-12 regions have similar values for AI. Contrary to the commuting network modularity scores, the AI for the administrative regions seem to be on the lower-valued (i.e., better) edge of these

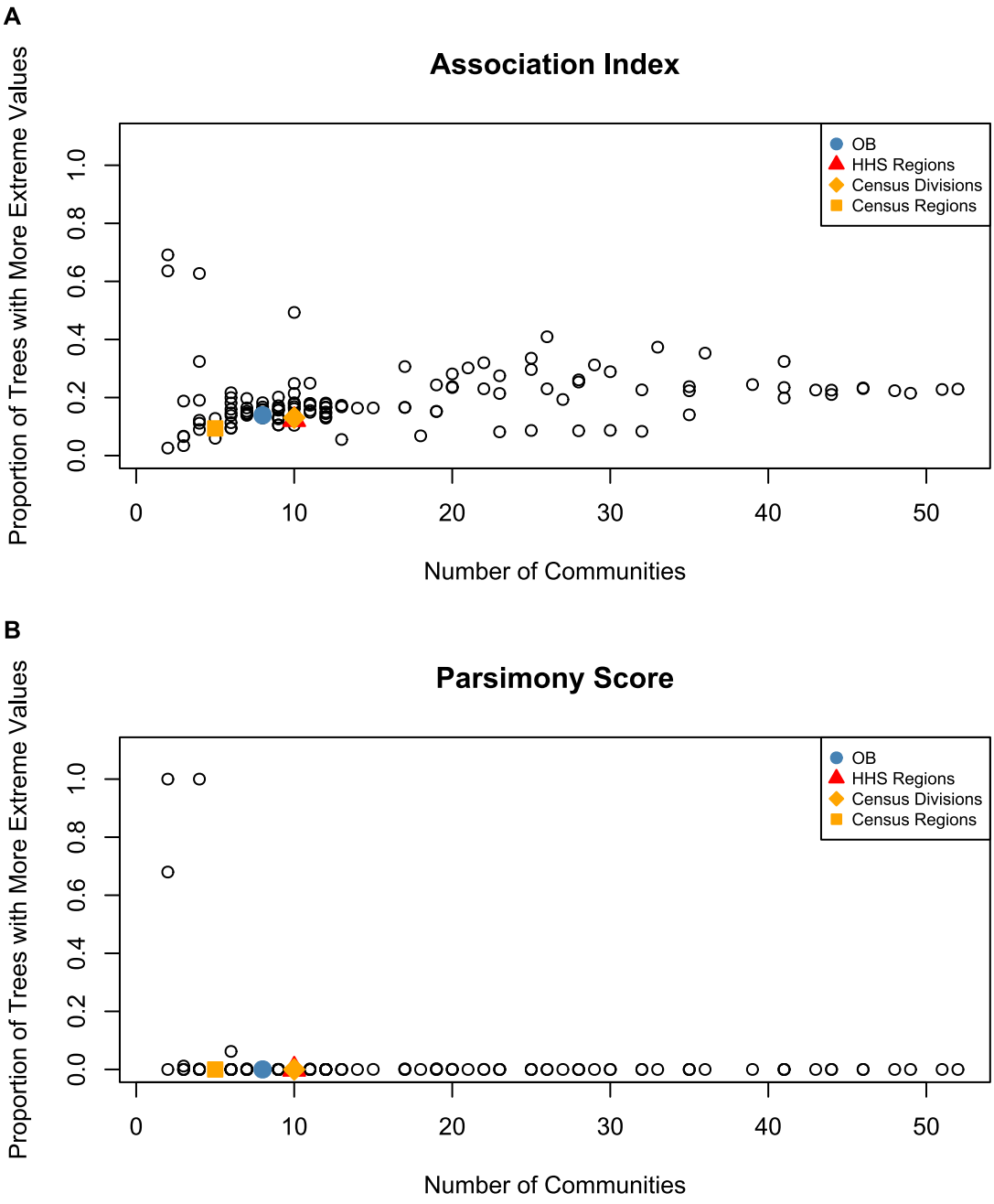


Figure 3.2. Phylogenetic Tip-trait Association of Regional Delineations with H3N2 Empirical Tree Set against the Number of Regions. (A) Association Index; (B) Parsimony Score. Association indices were calculated for each regional delineation and each of 500 phylogenetic trees sampled from a posterior distribution in a Bayesian evolutionary analysis of influenza virus subtype A/H3. Trait labels, i.e., state of isolate collection, were permuted to generate a null distribution. The median association index value was compared to this null distribution to calculated to show proportions. Lower values indicate that groupings found in the phylogeny support the spatial structuring of a given regional delineation.

points, with relatively few of our regional delineations performing better. The PS values have no signal whatsoever with all of the phylogenetic groupings being most parsimoniously described by the observed data for most regional delineations with markedly little variation. Combining these metrics of phylogenetic association, we rank the regional delineations in terms of their alignment with the phylogenetic groupings of H3N2 influenza viruses. The US Census Regions, US Census Divisions, and HHS Regions rank 14th, 27th, and 22nd, respectively, in phylogenetic tip-trait association.

The entropy scores for region memberships and maximum clade sizes (MCS) again show a pronounced signal in favor of delineations with approximately 4-14 regions (Figure 3.3). A band of observations lies separate from others at the baseline with relatively higher entropy scores. However, unlike clustering network modularity and phylogenetic tip-trait AI, entropy scores are biased in favor of delineations with greater numbers of regions. Entropy scores peak locally around 14 regions, and globally for those delineations with 40+ regions. Combining the entropy scores for region membership sizes and MCS using a simple arithmetic mean, we rank the regional delineations in terms of their balance of the frequencies of states included in each region and the aggregate apportionment of phylogenetic diversity. The US Census Regions, US Census Divisions, and HHS Regions rank 102nd, 55th, and 46th, respectively, in regional entropy/balance.

By combining each of our four comparison rankings, we create a composite score to identify candidate ITZs within the US. We use this composite score to select an overall best regional delineation and to identify specific partitions common among top-performing delineations; we include the top 50 (Figure 3.4). Our overall best regional

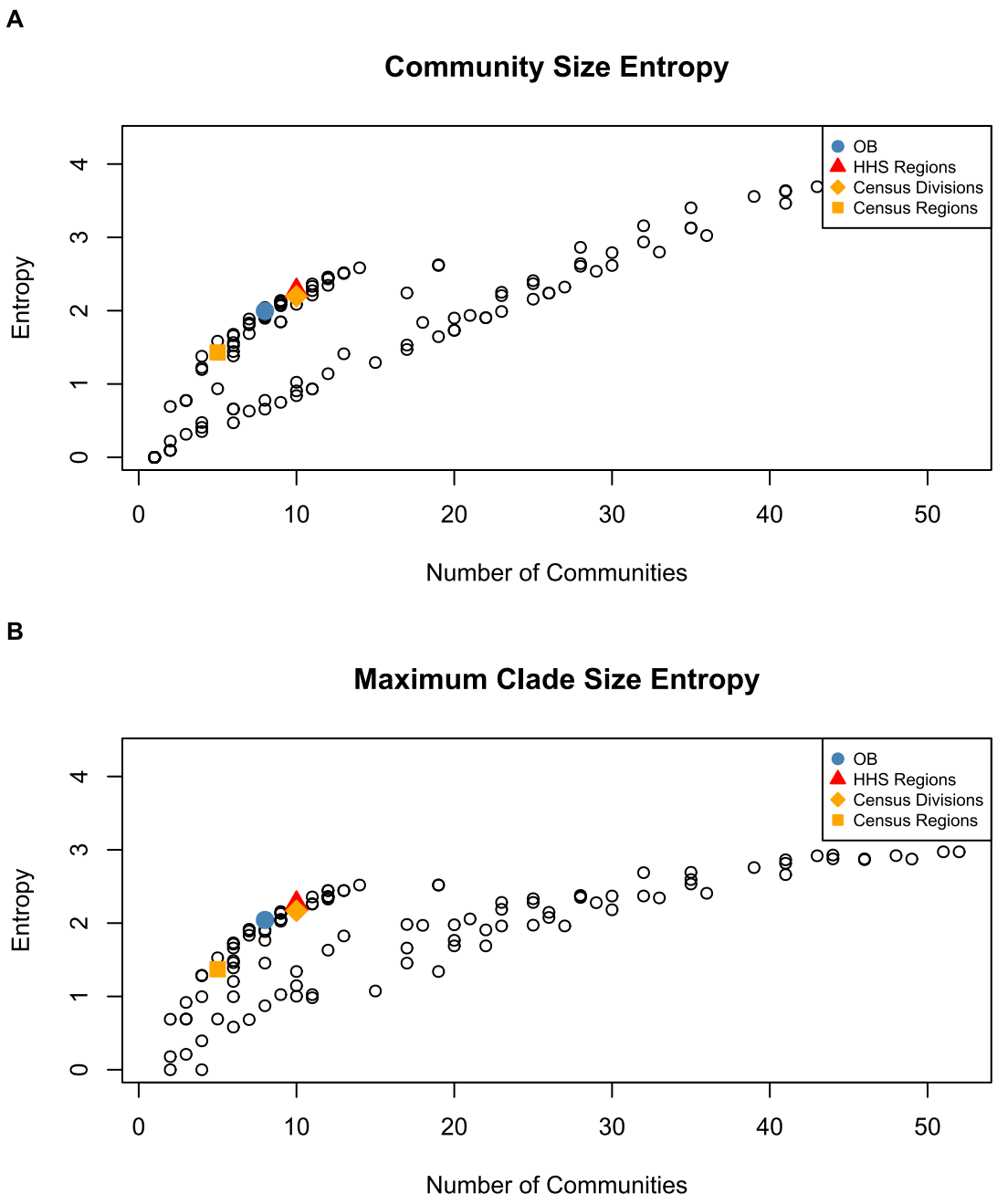


Figure 3.3. Entropy of Community Membership (A) and Maximum Clade Size (B) for Regional Delineations against the Number of Regions. I use entropy to assess the balance of regions within each delineation. If a delineation has several regions with few states and a single region with many, this would be an unbalanced regional delineation with relatively lower entropy. So, larger values of entropy indicate a more balanced regional delineation. Community size is as described with the relative frequencies of states within each region. Maximum clade size refers to the largest monophyletic grouping for a given region across a posterior set of 500 phylogenies of influenza virus subtype A/H3; here, entropy captures the relative balance of phylogenetic diversity.

delineation outlines 8 regions within the US (Figure 3.4). This overall best regional delineation ranks 10th in commuting network modularity, 21st in clustering network modularity, 35th in phylogenetic tip-trait association, and 71st in entropy. Of all the administrative regions, the regions in our overall best delineation are most similar to the HHS Regions (Rand Index (RI) = 0.793), followed by the US Census Divisions (RI = 0.788), and then the US Census Regions (RI = 0.691).

Comparing our overall best and administrative regional delineations qualitatively, we note some differences which may contribute to varying performance. One of the more pronounced differences between the administrative regions and our top-performing regional delineations concerns classification of Kentucky. Our top-50-performing regional delineations indicate a strong separation between Tennessee and Kentucky while in each of the administrative regions, these two states are grouped together. Similarly, we observe strong separation of the DC-Maryland-Virginia area from either areas to the North and South; this region is grouped with Pennsylvania to the North in the HHS Regions and with North Carolina to the South in the Census Regions and Divisions, while in our overall best it stands alone. From the other perspective, our top-performing regional delineations indicate a strong grouping between Oregon, Washington, and Idaho. This grouping is found in the HHS Regions and Census Regions, though Idaho is split from the others in the Census Divisions. Similarly, we observe strong grouping between New Mexico and Texas, but the Census Regions and Divisions partition areas along their border. No single regional delineation, among our overall best and the administrative regions, seems to represent all the strong partitions or groupings found in our top 50. However, our findings do suggest that there is strong regional structuring in the US and

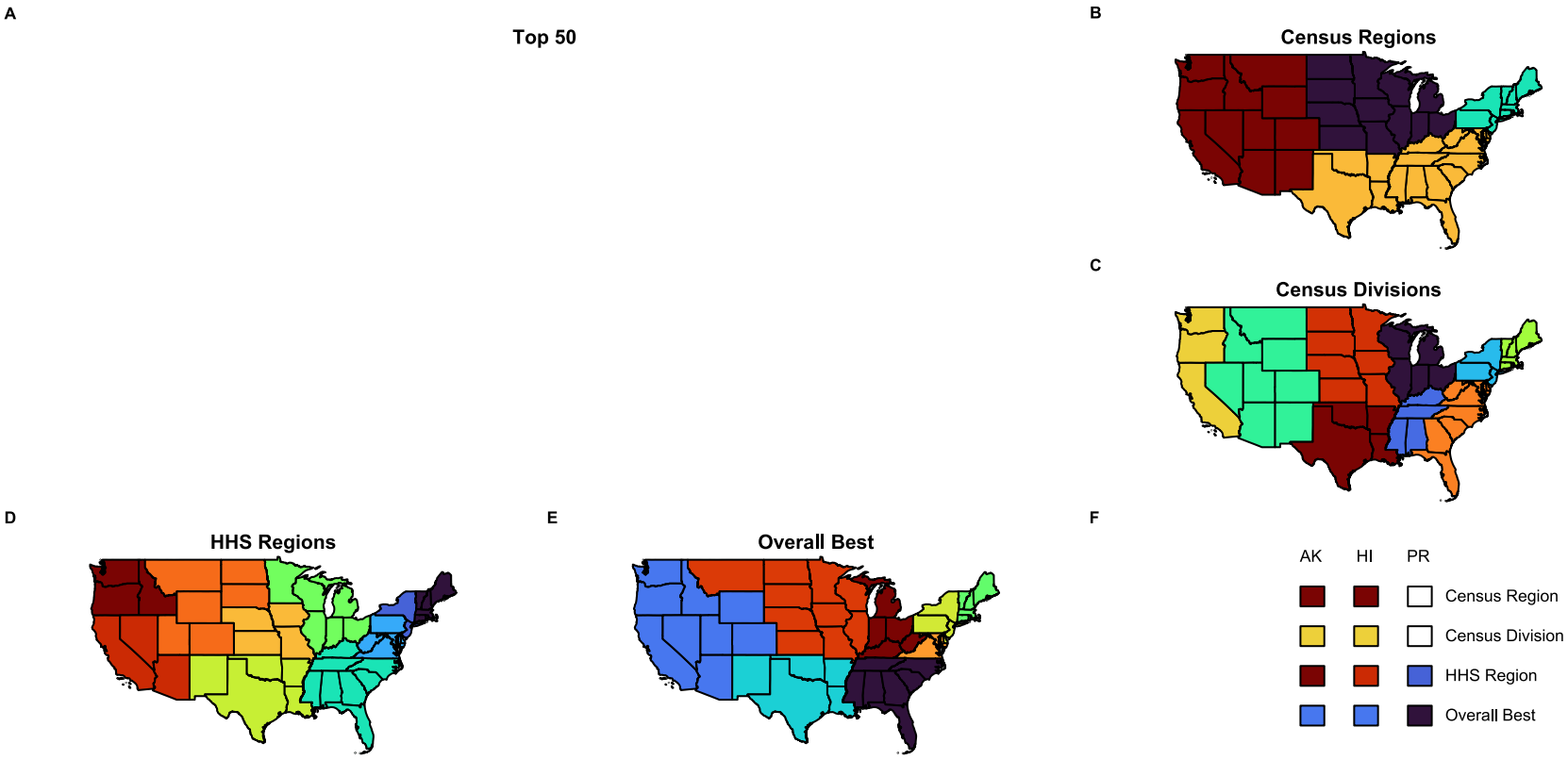


Figure 3.4. Regional Delineations of the US. (A) Overlayed borders of our Top-50-performing regional delineations, (B) Census Regions, (C) Census Divisions, (D) HHS Regions, (E) Our Overall Best regional delineation, (F) Classifications of Alaska, Hawaii, and Puerto Rico for (A,B,D,E). Combining scores across commuting network modularity, clustering network modularity, and phylogenetic tip-trait association indices(and entropy), I was able to rank the performance of all regional delineations for their ability to capture patterns across mobility, disease incidence, and viral population structure. The top 50 performing regional delineations are shown here with region borders overlayed. Therefore, darker lines correspond to divisions found more often in the top performing regional delineations; conversely, faint, lighter lines correspond to divisions less frequently found in the top performing regional delineations. Categorizations of Alaska, Hawaii, and Puerto Rico are shown as simple color-coded legend boxes.

that there may be some implicit hierarchies around core states, e.g., Georgia, Texas, and California.

DISCUSSION

Transmission of influenza viruses occurs on broad spatial scales and is not necessarily obstructed by geopolitical borders found across the world. Conversely, a lack of administrative or practical partitions in surveillance data from broad geographic areas does not necessarily correspond to a lack of epidemiologically relevant partitions within the region. As one of the largest countries in the world, the US likely constitutes several separate influenza transmission zones (ITZs). Here, we explored this notion of subnational ITZs. First, we explored patterns of clustering in the incidence of influenza-like illness (ILI). We found largely similar spatial signals in the clustering of ILI incidence in space and on commuting networks, though the temporal signals may suggest locations are linked by different mobility processes over the course of an outbreak. We then utilize our clustering results alongside representations of human mobility and seasonal influenza virus evolution to identify epidemiologically relevant regional delineations of the US, i.e., ITZs. We find a strong signal indicating that the US may be well characterized using 4-14 regions. Also, in our evaluations of many potential ITZ classifications, we present a holistic approach to characterizing the spatial epidemiology of seasonal influenza and suggest an overall best regional delineation.

Epidemiological surveillance data has been used by others in the characterization of ITZs (2,26). These studies relied on similar timing and peaks in epidemic curves to identify groups or clusters in the data. Similarly assessing aspects of epidemic timing,

Rosensteel et al (2021) developed a county-level “epidemiological geography” using a similar network community detection approach. However, we feel that the clusters identified in this way may not directly correspond to transmission clusters. For example, similarity in the timing of influenza outbreaks can occur between locations simply based on the timing of seasonal forcing / local climates. Without more concrete linkages between outbreaks in difference locations, e.g., contact tracing or viral genetic relationships, it is difficult to assess the validity of clustering as transmission linkages. We attempt to overcome this limitation in two ways: we generate regional delineations using patterns in human mobility and disease incidence; and we holistically evaluate many potential regional delineations as ITZs.

As an obligate intracellular parasite, influenza viruses rely on their hosts’ mobility for diffusion and dispersion on larger geographic scales. While global patterns of influenza spread are largely owed to passenger air travel, influenza spread within the US more strongly relate to local processes, e.g., workers' commutes and geographic distance (27,28). Commutes are largely local movements and commuter volume between locations is greatly impacted by the geographical distance that separates them, diminishing greatly as separation increases (29,30). Our results reinforce the importance of more local spatial relationships as our regional delineations were comprised of contiguous regions. By analyzing our data in a network science framework, we removed implicit spatial constraints and allowed for the possibility of ITZs to be spatially discontinuous. This was intentional as we considered the possibility that highly populated, international travel-hub states would be more tightly coupled with one another than with flanking, hinterland regions. However, we did not identify such a feature in our candidate delineations.

Rather, regions were contiguous collections of states, save the obvious exceptions for Alaska, Hawaii, and Puerto Rico; though, Alaska, Hawaii, and Puerto Rico were grouped with the closest region of the mainland in our overall best delineation.

In our clustering analyses, we use ILI rates of change as our outcome of interest. That is, instead of using aspects of onset timing or epidemic peaks to assess the coupling/relatedness of outbreaks in separate locations, we analyze the progression of an outbreak using biweekly rates of change. If two locations have similar outbreak dynamics, this may offer stronger evidence of some degree of relationship, e.g., linked in transmission, between outbreaks as opposed to relationships potentially confounded by latent variables, e.g., coinciding seasonal forcing. Additionally, the composition of each cluster is founded on relationships based on both geographic proximity and population interconnectivity, i.e., commuting ties, as opposed to relationships or similarity only found in data signals. That is, we find these clusters to be epidemiologically/biologically plausible groupings of outbreaks. Consequently, we feel this feature extends to our candidate ITZ delineations. By using patterns of human mobility and disease incidence to identify ITZs, there are elements of construct validity in each of our data-driven regional delineations.

Further adding to the validity, each of the regional delineations in our analysis was evaluated for its ability to align with spatial constructs in human mobility, disease incidence, and viral ancestry. This interdisciplinary and cross-scaling feature of our analysis is an increasingly necessary feature of infectious disease research. Of specific interest, phylogeographic analyses interrogate pathogen molecular sequences, e.g., genes, to uncover evidence of large-scale geographic transmission. One particularly common

approach in these studies is the discretization of geographic space which allows for simple and flexible inference of gene flow unconstrained by notions of distance. However, we find that the use of regional delineations in influenza research is non-standard and subject to analyst choice, e.g., CDC FluView presents "regional" data for both Census Regions and HHS Regions (31). Furthermore, conveniently available regional delineations, e.g., Census Regions, may not align well with the geographic limits of an outbreak (14,27). This misrepresentation of space can allow lurking variables to bias estimates of geographic diffusion. For example, consider the difference between the classification of Kentucky in our overall best regional delineation (i.e., with Southern states) versus the administrative regional delineations (i.e., with Midwest/Great Lake states). For a quick reminder, our top-performing regional delineations indicated a strong grouping between Kentucky and Indiana/Ohio and a strong partitioning between Kentucky and Tennessee. If we extend this to describe influenza outbreaks, we could consider there to be more shared phylogenetic diversity between Kentucky and Indiana/Ohio and less shared phylogenetic diversity between Kentucky and Tennessee. So, if we were to conduct a phylogeographic analysis using one of the administrative regional delineations which group Kentucky with other Southern states, we could artificially inflate / bias our estimates of gene flow between Southern and Midwest/Great Lake states. Of course, this type of error is common to any discretization of space (or continuous variables), and infectious disease surveillance data would first need to be available at finer geographic resolutions to even attempt to optimally mitigate misrepresentations.

We identify at least three limitations in our study. First, we use a non-specific disease indicator in influenza-like illness for our clustering analyses. There are many co-circulating respiratory viruses which are subject to common seasonal forcing each winter in the US, i.e., climate impacts transmissibility (32,33); these include SARS-CoV-2 and RSV which together with influenza comprise the "tripledemic" threatening global health (studied in (34). However, by using ILI in our analyses, we may have identified more general respiratory infectious disease transmission zones, rather than specifically ITZs. This could facilitate comparative analyses across infectious disease systems which could be a fruitful direction for future research. Second, as previously mentioned, the scope and resolution of our analysis is limited by the same characteristics of our data. We felt it necessary to conduct our analysis respective of state borders, though smaller scale areas may better align with outbreak limits. However, by using states as spatial units, we both facilitate practical spatial alignment in data and, potentially, public health intervention. Additionally, as we envision our results' use cases in phylogeographic analyses, fewer classifications (e.g., 8 regions versus 52 states versus 925 core-based statistical areas) are much preferable to avoid issues computational complexity and model identifiability. This aspect is becoming increasingly important as molecular epidemiology studies transition into an era of big data. Third, we apply a rigid definition for the concept of community / ITZ. Our community detection analyses yield non-overlapping / mutually exclusive, static regional delineations. In reality, these characteristics likely do not apply to transmission zones. Particularly, the stochastic nature of virus introduction may relate to more variable delineations from season to season as the transmission zone unfolds depending on where a local outbreak originates. However, regional delineations and ITZs

with these characteristics would have a limited utility as existing methodologies may have difficulty accounting for the added complexity.

Altogether, we conclude that the US is characterized by several ITZs. Alongside our specific results pertaining to the overall best regional delineation, we set forth a framework for aligning spatial constructs across biological scales of organization and validating given geographic constructs.

REFERENCES

1. Influenza Transmission Zones. (https://www.who.int/publications/m/item/influenza_transmission_zones). (Accessed April 18, 2024)

2. Caini S, Alonso WJ, Séblain CE-G, et al. The spatiotemporal characteristics of influenza A and B in the WHO European Region: can one define influenza transmission zones in Europe? *Euro Surveill*. 2017;22(35):30606.

3. Shin GY, Manuel R. Letter to the editor: Sampling bias should be minimised when analysing influenza transmission zones involving very large countries. *Eurosurveillance*. 2017;22(40):17.

4. United States. *The World Factbook*. 2024;(https://www.cia.gov/the-world-factbook/countries/united-states/). (Accessed May 7, 2024)

5. Fowler CS, Jensen L. Bridging the gap between geographic concept and the data we have: The case of labor markets in the USA. *Environ Plan A*. 2020;52(7):1395–1414.

6. Nelson GD, Rae A. An Economic Geography of the United States: From Commutes to Megaregions. *PLOS ONE*. 2016;11(11):e0166083.

7. Bureau UC. Metropolitan and Micropolitan Statistical Areas Map (March 2020). *Census.gov*. (https://www.census.gov/geographies/reference-maps/2020/geo/cbsa.html). (Accessed May 17, 2024)

8. Rosensteel GE, Lee EC, Colizza V, et al. Characterizing an epidemiological geography of the United States: influenza as a case study. 2021;2021.02.24.21252361. (https://www.medrxiv.org/content/10.1101/2021.02.24.21252361v1). (Accessed July 11, 2024)

9. US Census Bureau CHS. Regions and Divisions - History - U.S. Census Bureau. (https://www.census.gov/history/www/programs/geography/regions_and_divisions.html). (Accessed April 27, 2022)

10. National, Regional, and State Level Outpatient Illness and Viral Surveillance. (https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html). (Accessed May 11, 2024)

11. Bureau UC. Cartographic Boundary Files - Shapefile. *Census.gov*. (https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html). (Accessed May 11, 2024)

12. Bureau UC. Commuting Flows. *Census.gov*. (https://www.census.gov/topics/employment/commuting/guidance/flows.html). (Accessed May 11, 2024)

13. Bureau UC. Centers of Population. *Census.gov*. (https://www.census.gov/geographies/reference-files/time-series/geo/centers-population.html). (Accessed May 11, 2024)

14. Damodaran L. MOLECULAR EPIDEMIOLOGY OF CONTEMPORARY SEASONAL INFLUENZA EPIDEMICS. University of Georgia; 2023 (Accessed October 12, 2024).(https://esploro.libs.uga.edu/esploro/outputs/doctoral/MOLECULAR-EPIDEMIOLOGY-OF-CONTEMPORARY-SEASONAL-INFLUENZA/9949558923602959). (Accessed October 12, 2024)

15. Posit team. RStudio: Integrated development environment for R. Boston, MA: Posit Software, PBC; 2023.(http://www.posit.co/)

16. SaTScan - Software for the spatial, temporal, and space-time scan statistics. (https://www.satscan.org/). (Accessed May 11, 2024)

17. Csárdi G, Nepusz T, Traag V, et al. igraph: Network Analysis and Visualization. 2024;(https://cran.r-project.org/web/packages/igraph/index.html). (Accessed October 13, 2024)

18. Smith NR, Zivich PN, Frerichs L, et al. A guide for choosing community detection algorithms in social network studies: The Question-Alignment approach. *Am J Prev Med*. 2020;59(4):597–605.

19. Affairs (IEA) O of I and E. HHS Regional Offices. 2006;(https://www.hhs.gov/about/agencies/iea/regional-offices/index.html). (Accessed October 7, 2024)

20. Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. *Phys. Rev. E*. 2004;70(6):066111.

21. Newman MEJ, Girvan M. Finding and evaluating community structure in networks. *Phys. Rev. E*. 2004;69(2):026113.

22. Identification of Shared Populations of Human Immunodeficiency Virus Type 1 Infecting Microglia and Tissue Macrophages outside the Central Nervous System | Journal of Virology. (https://journals.asm.org/doi/10.1128/JVI.75.23.11686-11699.2001). (Accessed November 10, 2022)

23. Parker J, Rambaut A, Pybus OG. Correlating viral phenotypes with phylogeny: Accounting for phylogenetic uncertainty. *Infection, Genetics and Evolution*. 2008;8(3):239–246.

24. BaTS – Bayesian Tip-association Significance testing. (https://online.fliphtml5.com/yyhu/fvkl/). (Accessed May 15, 2024)

25. Fortunato S, Barthélemy M. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*. 2007;104(1):36–41.

26. The global region-specific epidemiologic characteristics of influenza: World Health Organization FluNet data from 1996 to 2021 - ClinicalKey. (https://www.clinicalkey.com/#!/content/playContent/1-s2.0-S1201971223000528?returnurl=https:%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS1201971223000528%3Fshowall%3Dtrue&referrer=https:%2F%2Fpubmed.ncbi.nlm.nih.gov%2F). (Accessed May 7, 2024)

27. Balcan D, Colizza V, Gonçalves B, et al. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*. 2009;106(51):21484–21489.

28. Charu V, Zeger S, Gog J, et al. Human mobility and the spatial transmission of influenza in the United States. *PLOS Computational Biology*. 2017;13(2):e1005382.

29. Viboud C, Bjørnstad ON, Smith DL, et al. Synchrony, Waves, and Spatial Hierarchies in the Spread of Influenza. *Science*. 2006;312(5772):447–451.

30. Charu V, Zeger S, Gog J, et al. Human mobility and the spatial transmission of influenza in the United States. *PLOS Computational Biology*. 2017;13(2):e1005382.

31. FluView Interactive | CDC. 2023;(https://www.cdc.gov/flu/weekly/fluviewinteractive.htm). (Accessed September 2, 2024)

32. Moriyama M, Hugentobler WJ, Iwasaki A. Seasonality of Respiratory Viral Infections. *Annual Review of Virology*. 2020;7(Volume 7, 2020):83–101.

33. Neumann G, Kawaoka Y. Seasonality of influenza and other respiratory viruses. *EMBO Molecular Medicine*. 2022;14(4):e15352.

34. Luo W, Liu Q, Zhou Y, et al. Spatiotemporal variations of "triple-demic" outbreaks of respiratory infections in the United States in the post-COVID-19 era. *BMC Public Health*. 2023;23(1):2452.

CHAPTER 4

SPATIAL VARIATION IN THE PHYLOGENETIC SIGNAL OF LOCAL OUTBREAKS OF SEASONAL INFLUENZA IN THE UNITED STATES[1]

[1]Cody Dailey & Lambodhar Damodaran. To be submitted to a peer-reviewed journal.

ABSTRACT

Genomic epidemiology approaches are increasingly common in influenza surveillance, offering high-resolution insights into transmission patterns. By analyzing genomic data and reconstructing pathogen ancestry through phylogenetic methods, researchers can uncover transmission dynamics that traditional case-based approaches might fail to capture. Particularly, phylogenetic trees can shed light on to the underlying transmission dynamics in an outbreak. With the accumulation of genomic data, it may now be possible to systematically characterize spatial variation in transmission dynamics of seasonal influenza in the United States. This study works towards this goal by describing the phylogenetic signals of local influenza outbreaks across the US from 2010-2020, focusing on type A (H3 and H1 subtypes) and type B (Victoria and Yamagata lineages) influenza viruses. Comparing local influenza outbreaks, I find that the mean pairwise patristic distance (MPD) among isolates of local outbreaks tends to be higher for influenza A viruses than influenza B viruses, and that there seems to be a strong seasonal fluctuation in the signals, perhaps indicative of subtype/strain dominance within a given season. The MPD of these local transmission clusters showed weak spatial dependence overall, but, comparatively, H1 and BYamagata seemed to be more consistent among neighboring outbreaks, both in terms of space, e.g., border-sharing neighbors, and time, e.g., sequential influenza seasons. Also, I find that local outbreaks in some states, e.g., California and Georgia, had marginally less diverse local outbreaks, potentially suggesting some systematic differences in transmission patterns, if not simply isolate sampling. With continued efforts towards systematic characterizations of local outbreaks

using genomic epidemiology approaches, we stand to gain new, high-resolution insights to seasonal influenza epidemiology.

INTRODUCTION

Genomic epidemiology approaches are becoming increasingly common in influenza surveillance and research efforts. The analysis of genomic data on influenza allows for the characterization of patterns that are obscured or otherwise difficult to observe using more traditional, case-based approaches alone. A particular advantage of using molecular scale approaches is the inherent high-resolution of observations. Additionally with respect to genomic sequences, relationships among observations are objectively encoded within the sequences themselves. These aspects can be leveraged such that by reconstructing the ancestral patterns of pathogens using phylogenetic methodologies, researchers can glean aspects of transmission (1,2). Phylogenetic trees contain a wealth of information, and through their detailed analysis, we have learned much about influenza. For example, phylogenetic studies of seasonal influenza viruses have empirically characterized large-scale circulation patterns (3–5), coupling in epidemiology and viral evolution (5,6), ecological interactions among co-circulating strains (7), and numerous drivers of transmission (8,9). Moreover, simulation studies have shown that simple quantitative summary metrics of phylogenies or tree shape statistics are able to discriminate host contact patterns of transmission (10), help in predictions of viral lineage persistence (11), and correlate with epidemiological quantities such as reproduction numbers (12). With the increasing quality, coverage, and availability of genomic sequence data, researchers are afforded new opportunities for

higher-resolution studies of influenza. We find one such opportunity to investigate the spatial variation of seasonal influenza outbreaks within the United States (US).

Seasonal influenza outbreaks may be caused by any one of several different seasonal influenza viruses, including type A influenza viruses (IAV), H3 and H1 subtypes, and type B influenza viruses (IBV), Victoria and Yamagata lineages (13). Spatial heterogeneities and hierarchies in seasonal influenza outbreak dynamics have been well-characterized, e.g., with respect to timing (14–18) and epidemic intensity (19). However, it is unclear how underlying patterns of transmission of local influenza outbreaks may compare across the US. Here, we attempt to address this gap by systematically characterizing the phylogenetic signal of local influenza outbreaks within the US, 2010-2020.

METHODS

Data

Data for this study are publicly available and consist of genetic sequences, spatial boundaries, and commuter flows.

Influenza genetic sequences are hosted by the Global Initiative on Sharing All Influenza Data (GISAID) platform (20). Viral hemagglutinin (HA) gene sequences were downloaded for influenza A viruses (IAV), H3 and H1 subtypes, and influenza B viruses (IBV), Victoria and Yamagata lineages. Sequences were included for influenza virus isolates sampled within the US from January 2010 through December 2020; any sequences out of the study scope or with indeterminate/missing metadata on location and

date of collection were excluded from analysis. Additionally, four isolate sequences, one for each subtype, from 2000-2001 were included to serve as outgroups.

Cartographic boundary files for state or state-equivalent areas in 2018 at the 1:5m resolution were downloaded from the US Census Bureau (21). Questionnaire responses concerning the origin and destinations of commuting flows from the American Community Survey (ACS) are summarized as tables and made available by the US Census Bureau (22); tables for 2011-2015 and 2016-2020 were downloaded and included in analysis.

All data management and analysis were conducted using R programming language (version 4.3.0) in the RStudio/Posit interactive developer environment (23), unless otherwise specified. Processing and analytical scripts are made available in a GitHub repository (link daileyco/Seasonal-Flu-Evolution) to facilitate reproducibility.

Phylogenetic Reconstructions

Sequences were aligned in a multiple sequence alignment using MAFFT (24) for each of the four influenza subtypes separately. Following, sequences were further stratified, or grouped, by location and influenza season. Locations comprise the fifty states and the District of Columbia (DC). Influenza seasons span the 2010-2011 season through the 2019-2020 season. Sequence data were partitioned using overlapping two-year intervals; for example, a stratum for the 2010-2011 influenza season would include all sequences collected from 1 January 2010 through 31 December 2011. Each data partition also included a single outgroup isolate sequence from 2000-2001. Altogether,

there are 2040 strata or combinations of the 4 subtypes, 51 locations, and 10 influenza seasons.

Strata containing at least 3 isolate sequences were used to reconstruct phylogenies. All phylogenetic trees, i.e., phylograms/evolutionary trees, were generated in a maximum likelihood framework under a general time reversible (GTR) nucleotide substitution model using the IQ-TREE2 software (25). Additionally, each phylogenetic tree was dated or rescaled using the least-squares dating (LSD2) method also available via IQ-TREE2. Outgroups were used to specify root branches then subsequently dropped in the phylogenetic dating process; that is, outgroups are not included in the resulting time trees. Additionally, phylogenetic tree reconstruction and dating were performed in duplicate with replicated analyses excluding outgroup sequences and relying on least-squares fit to identify the best root branch prior to dating.

We compare the replicate time trees to select a single representation for downstream analyses. The time trees were selected according to two criteria. First, we evaluated each time of the most recent common ancestor (tMRCA) and how close the estimated date was to the given influenza season. Second, we assessed the difference between the tMRCAs from trees generated with and without the outgroups. If a tMRCA for the time tree reconstructed without an outgroup was both closer to the given influenza season and the difference between outgroup and no-outgroup tree tMRCAs was greater than 3 years, then we included the time tree reconstructed without an outgroup in downstream analyses; otherwise, the time tree reconstructed with an outgroup was included.

Local Transmission Clusters

Following phylogenetic estimation and dating, we identified specific subtrees/clades within each phylogenetic tree that represent local transmission clusters. Phylogenies were reconstructed using genetic sequences from isolates collected over an entire two-year period. As such, trees potentially include relationships (i.e., nodes and edges) between isolates beyond the scope of a single influenza season. So, we "pruned" or partitioned each phylogenetic tree into subtrees/components based on the alignment in the timing of virus ancestry with that of the focal influenza season. We define influenza seasons as starting in calendar (or epidemiological/epi-) week 30 (~end of July) in year 1 and ending in calendar week 18 (~beginning of May) in year 2; for example, the 2010-2011 influenza season was defined as starting 25 July 2010 and ending 7 May 2011. In each of the resulting pruned subtrees, included taxa correspond to isolates which were both collected within the focal season and descendant from a single common ancestor estimated to have existed within the focal season; that is, phylogenetic trees were pruned/cut or partitioned by identifying separate, co-circulating lineages that diverged sometime before the beginning of the given influenza season. In this way, each resulting subtree conveys patterns of ancestry, or diversification, specific to each location and each influenza season, potentially in replicate for multiple clades or co-circulating lineages within each subtype. These subtrees are referred to as local transmission clusters.

Spatiotemporal Lags

Local transmission clusters were quantitatively summarized using tree shape statistics, metrics of phylogenetic signal. Chiefly, we focus on the mean pairwise patristic

distance or mean pairwise tree distance (MPD) to quantify the relatedness, or diversity, of isolates represented in a phylogeny. These MPD values were treated as time-series and subsequently analyzed to quantify spatial, temporal, and cross-subtype dependence in phylogenetic signals. In other words, we correlate the phylogenetic signals of local transmission clusters with those of neighboring locations, sequential seasons, and other co-circulating influenza subtypes.

State borders and commuting ties were used to define spatial and network neighbors, respectively, for all locations. Locations were considered spatial neighbors, or spatially adjacent, if they shared a border determined by the spdep R package's poly2nb() function using binary encoding (i.e., neighbor or not)(26). Spatial neighbors were determined up to three degrees of separation, or three spatial lags, e.g., neighbors of neighbors correspond to two spatial lags. Similarly, commuting flow data were transformed into weighted adjacency matrices to represent commuting network neighbors where weights correspond to the number of people estimated to have participated in the given, undirected commuting flow between two locations; note, network neighbor weights were only calculated for the single degree of separation, or one network lag, as the state level commuting networks are nearly fully connected. Two network adjacency matrices were created corresponding to the two time periods of commuting data.

Using the adjacency matrices, MPD values for local transmission clusters were spatially lagged. Spatial and network lagged values of MPD were calculated by averaging the MPD values observed in neighboring locations. That is, non-missing values were averaged across all a location's neighbors to yield a single MPD value for the given spatial lag, subtype, and season; for the network neighbors, network lagged MPD values

were weighted averages of neighbor values. Following, all MPD values were temporally lagged up to three degrees of separation, i.e., three seasons.

Correlating Phylogenetic Signals

Pearson correlation coefficients were computed using these MPD values to characterize temporal, spatial, and spatiotemporal dependence within and among the phylogenetic signals of each circulating influenza virus lineage. As there were multiple local transmission clusters / co-circulating strains / subtrees for some combinations of subtype, season, and location, a single local transmission cluster was randomly sampled within each stratum before generating the spatiotemporally lagged values and computing the correlations. The data were repeatedly resampled for 1000 replications. The resulting distributions of correlation coefficients were summarized using quantiles and are presented as medians and 95% confidence intervals; a correlation was deemed significant if the confidence interval did not include zero.

RESULTS

Over 50 000 influenza hemagglutinin (HA) gene sequences were downloaded from GISAID. After exclusions and quality control, N = 42 113 isolate sequences remained with an additional 4 virus isolate sequences for the outgroups (Supplementary Figure C.1). Sequences for IAV subtype H3 (n = 18 829) were most numerous, followed closely by subtype H1 (n = 12 243) and distantly by both IBV lineages (n = 6 371 for B Victoria, and n = 4 670 for B Yamagata).

When fully stratified by subtype, season, and location, 1 806 of 2 040 (~88.5%) strata contained enough HA sequences to attempt phylogenetic reconstruction. Suitable data coverage varied with influenza subtypes, as with the frequency of isolate sequences (Supplementary Figure C.2). The data coverage was more complete for IAV compared to IBV. Also, data were more complete for more recent seasons for both types of influenza viruses with a notable increase in coverage/frequency from ~2015 onwards. For all locations and influenza seasons, there were enough H3 isolate sequences to estimate phylogenies, 510 of 510 strata (100%). Comparatively, there were 490 (96.1%), 409 (80.2%), and 397 (77.8%) strata with enough data for H1, B Victoria, and B Yamagata, respectively. Phylogenies were reconstructed successfully for all but one set of B Victoria sequences.

Each stratum had a variable number of sequences, and, consequently, the number of sequences used to reconstruct each phylogeny varied. Phylogenies included a median of 41.5 [Q1=20, Q3=71], 25 [11, 48], 14 [6, 32], and 14 [7, 29] sequences for H3, H1, B Victoria, and B Yamagata, respectively (Supplemental Table C.1).

The estimated time trees tMRCAs indicated a potentially poor fit to a molecular clock model for many strata. When comparing the tMRCAs among the full phylogenetic trees, H3 lineages coalesced a median of 5 years before the circulating season, while H1, BVic, and BYam trees coalesced approximately 2.1, 3.1, and 3.3 years, respectively, before the given influenza season. There were 577 phylogenies which had an estimated tMRCA that was over 5 years before the start of the season in which the isolates were collected (Supplementary Figure C.7); excluding those, tMRCA medians are

approximately 3.3, 1.9, 2.0, and 2.5 years before the circulating season for H3, H1, BVic, and BYam, respectively.

The average mutation rates estimated were 2.7x10^-3, 3.8x10^-3, 1.8x10^-3, and 2.2x10^-3 substitutions per site per year for H3, H1, BVic, and BYam, respectively; excluding the phylogenies with extreme tMRCAs, the average mutation rates for H3, H1, BVic, and BYam are 3.2x10^-3, 4.1x10^-3, 2.3x10^-3, and 2.9x10^-3 mutations/site/year. The distributions of the estimated mutation rates across seasons were somewhat stable, though some individual seasons for each subtype showed more broad distributions with many potential outlying estimates (Supplemental Figures C.3 & C.6). With the time-scaled phylogenies, we were able to identify local transmission clusters (Figure 4.1).

The relative frequencies of local transmission clusters (extracted subtrees) were similar to the frequencies of available sequence data across influenza subtypes. Frequencies were greatest for H3 (n=1 897), followed by H1 (n=1 419), BYam (n=852), and BVic (n=831; Supplemental Table C.2). Some phylogenies did not have an internal node estimated to have existed within the given seasons, and, as such, the frequencies of missing data slightly increased following local transmission cluster identification. Local transmission clusters seemed well aligned with each season following partitioning as, on average, the MRCAs were estimated to have existed just before the new year (i.e., 1 January) within each influenza season (Supplemental Table C.2, Supplemental Figure C.5). The subtrees are more balanced in the number of tips than the full phylogenetic trees with a median of 3 tips across subtrees from all subtypes. As such, the comparisons of tree shape statistics may be more meaningful.

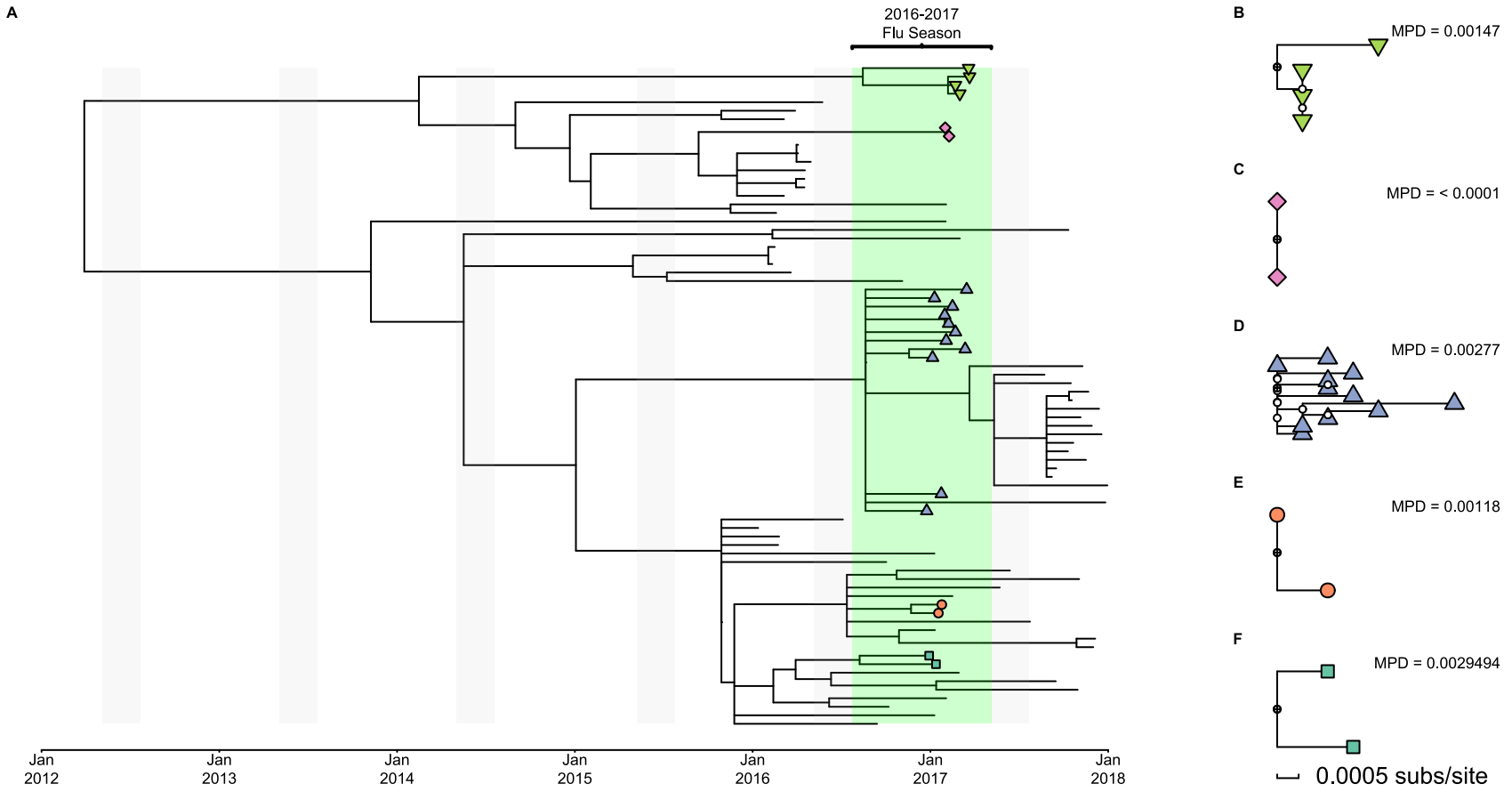


Figure 4.1. Time-scaled Phylogenetic Tree of Influenza A Subtype H3 using Isolate Sequences Collected from 2016 to 2018 in Georgia (A) and Identified Local Transmission Clusters (B-F). Hemagglutinin gene sequences from viral isolates sampled across a few years were used to reconstruct phylogenies. Using a least-squares dating procedure, nodes and branches were scaled in time. By isolating internal nodes and descendant tips that exist within a focal flu season window, I identify isolates comprising local transmission clusters, or local outbreaks. Each of these local transmission clusters can be simply summarized by calculating the average branch length between pairs of tips, a measure of phylogenetic diversity.

The MPD of H3 subtrees was an average of 1.2x10^-3 substitutions per site compared to 1.4 x10^-3, 0.7 x10^-3, and 0.9 x10^-3 substitutions per site for H1, BVic, and BYam, respectively. MPD values show a decreasing trend over the course of an individual influenza season with the earliest transmission clusters in each season having larger values than transmission clusters identified later in the season (Figure 4.2). There does seem to be some season-to-season variability in MPD across all subtypes and locations (Figure 4.3).

Additionally, I compared MPD values across states using a standardized mean differences comparing local transmission clusters within each state to the seasonal average for each subtype; to get a marginal summary, each of these mean differences were averaged across all subtypes and seasons to get a single value for each state which gives a sense of the relative diversity per local outbreak (Figure 4.4). We observe that local transmission clusters in some states tend to be less diverse than seasonal averages, e.g., Georgia and New York, while those in other states tend to be more diverse, e.g., West Virginia and Oklahoma (Figure 4.4). When comparing the distributions of MPD values across all subtypes within each season, the MPD values seemingly exhibit some negative / antagonistic interactions / interference. For example, in the 2016-2017 season, MPD values for H1 transmission clusters are lower than those of IBV and H3 transmission clusters within the same season and H1 transmission clusters from surrounding seasons. Similarly, for the 2018-2019 season, MPD values for IBV transmission clusters are distributed at lower values compared to those of IAV transmission clusters. To further investigate associations, we quantified the correlations of these measures of diversity across space, time, and influenza virus subtypes.

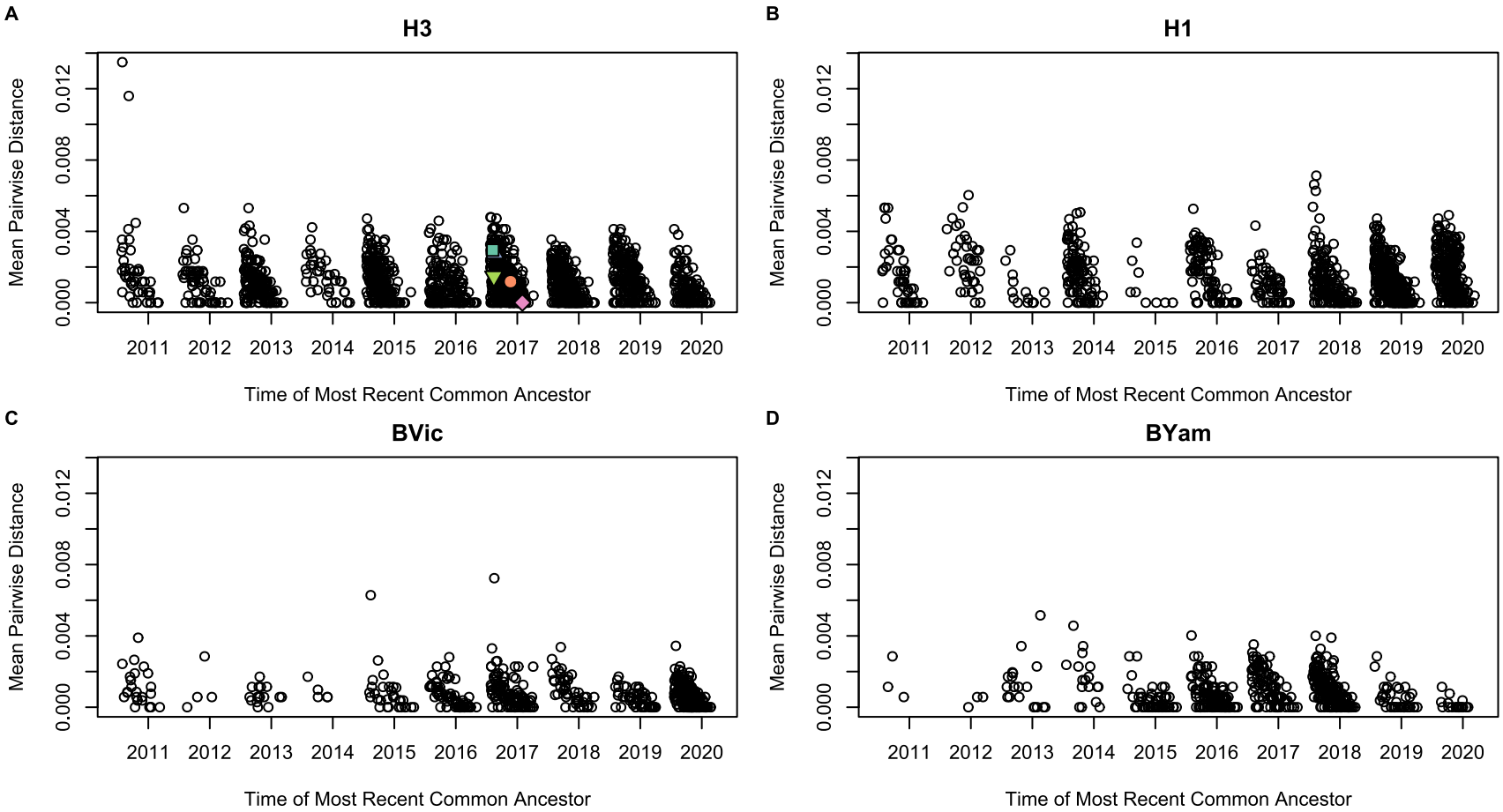


Figure 4.2. Profiles of Mean Pairwise Distances of Local Transmission Clusters of Seasonal Influenza. Influenza A Subtype H3 with highlighted examples from Figure 1 (A), Influenza A Subtype H1 (B), Influenza B Lineage Victoria (C), Influenza B Lineage Yamagata (D). Each data point represents a single local transmission cluster. The times of the most recent common ancestor (TMRCA) on the x-axis corresponds to the estimated time of the single internal node from which all tips in a local transmission cluster descended.

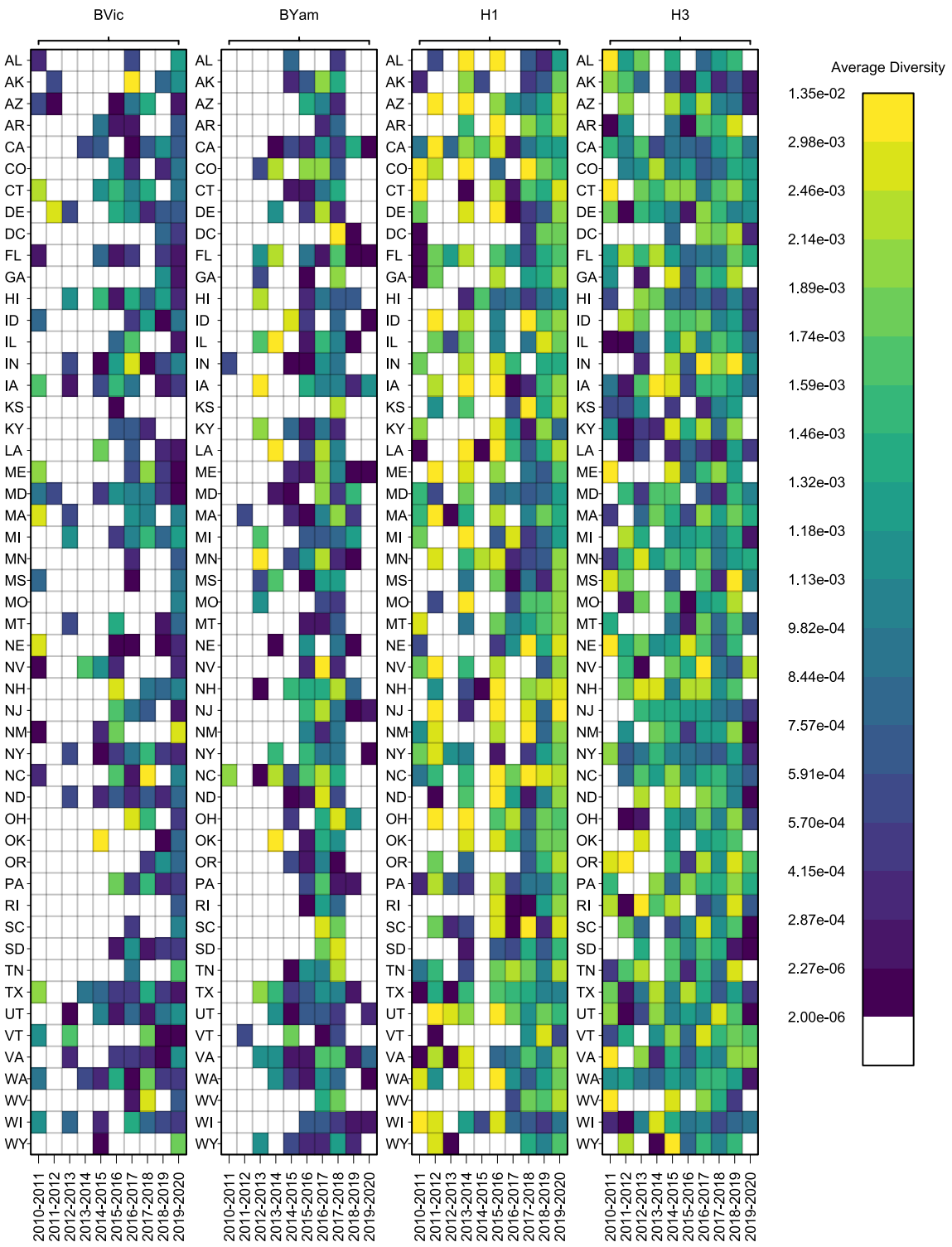


Figure 4.3. Average Diversity Realized in Local Transmission Clusters. All local transmission clusters for a given subtype-season-state were averaged to generate the shown values. White spots represent combinations of subtype-season-state where there is no data on local transmission clusters. Diversity here is the mean pairwise phylogenetic distance and has units of substitutions per site.

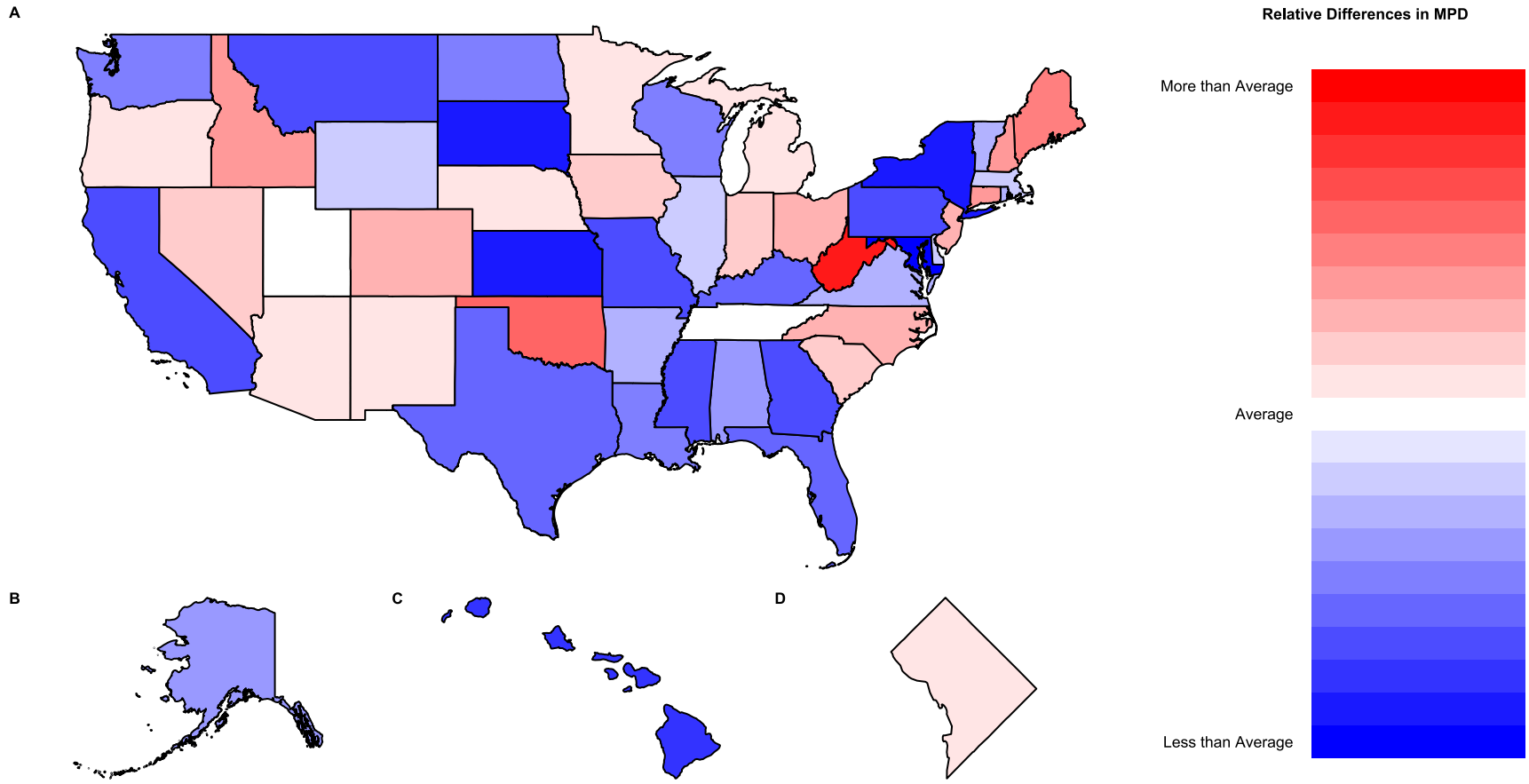


Figure 4.4. Relative Differences of Local Transmission Clusters' Diversity. Values of mean pairwise phylogenetic distance were standardized for each subtype-season combination. These standardized values were then summarized with a simple arithmetic average for each state yielding the values shown on the graph. Lower, blue values correspond to marginally lesser isolate diversity within local transmission clusters, and higher, red values correspond to marginally greater isolate diversity within local transmission clusters. Of note, there seems to be a general pattern with more populous states having marginally lesser isolate diversity within local transmission clusters.

Spatiotemporal autocorrelations within each subtype were relatively weak with the maximum significant correlation coefficient of 0.369 for contemporary (temporal lag = 0) for BYam trees for spatial neighbors (spatial lag = 1; Figure 4.5d). Significant positive correlations were observed within each subtype when comparing contemporary transmission clusters across spatial neighbors (Figure 4.5); that is, increasing values of MPD in local transmission clusters are observed with increasing values of MPD in local transmission clusters from neighboring locations. As the spatial lag increases, there is a relative decrease in the correlation coefficients for each subtype. For example, MPD values for local transmission clusters of H3 are slightly more correlated with transmission clusters in neighboring locations (rho = 0.141) than those clusters found at more distant locations (three spatial lags, rho = 0.126, Figure 4.5a). This decreasing trend seems to slightly vary among subtypes; the correlations decrease most for BVic and least for H3 at increasing spatial distances.

There are no consistently significant correlations across temporal lags for any of the spatial lags. Contemporary correlations across all spatial lags for each subtype are positive; that is, regardless of spatial separation, increasing values of MPD in transmission clusters for a given subtype and season are observed with increasing values of MPD in other transmission clusters of the same subtype and season. However, when comparing the correlations across temporal lags, there is an alternating pattern in the direction of the association which seems to vary between IAVs and IBVs. The correlations observed for a single temporal lag (i.e., comparing a season to the season before) are negative for IAVs; that is, increasing values of MPD in transmission clusters for a given IAV subtype in this influenza season are observed with decreasing values of

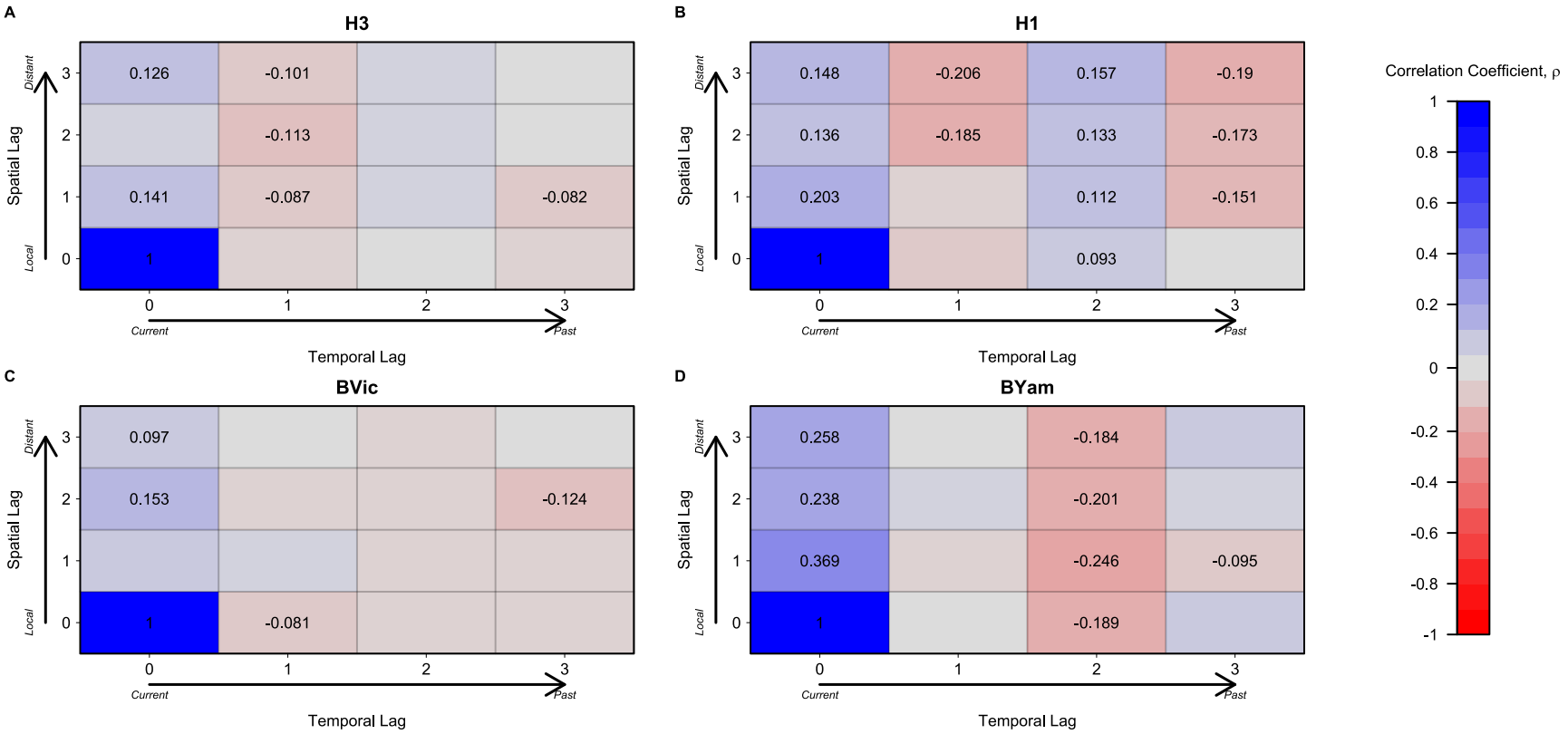


Figure 4.5. Spatiotemporal Autocorrelations of Viral Diversification / Tree Shape (MPD). Matrices have a focal point in the bottom left with both spatial and temporal lags of zero. Increasing up the y-axis, each increasing spatial lag corresponds to the order of separation among border-sharing states; that is, spatial lag of 1 compares a focal state to the neighboring states which share a border with the focal state, spatial lag of 2 compares a focal state the neighbors of border-sharing neighbors, and so on. Increasing along the x-axis, each increasing temporal lag corresponds to a previous influenza season; that is, temporal lag of 1 compares a focal influenza season to the previous season, temporal lag of 2 compares a focal influenza season to two seasons previous, and so on. Spatiotemporal lags are represented by the boxes not immediately adjacent to either axis. Median Pearson correlation coefficients are shown if 95% bootstrap intervals did not cross zero.

MPD in transmission clusters for the same respective subtype in the previous influenza season. The correlations continue this alternating pattern at increasing temporal lags for both IAV subtypes, though it is much more pronounced in H1 transmission clusters (Figure 4.5a-b). On the other hand, for IBVs, a similar alternating pattern is not observed (Figure 4.5c-d). For BVic, the temporal signal is least pronounced with most correlations with past seasons being negative and scantly significant. For BYam, the correlations at a single temporal lag are null-valued and insignificant. However, at two temporal lags, the signal is very pronounced and consistently negative across all spatial lags; that is, increasing values of MPD in BYam transmission clusters this season are observed with decreasing values of MPD in BYam transmission clusters from two seasons ago.

Spatiotemporal cross correlations between influenza subtypes are generally weaker than the autocorrelations within each subtype; the maximum magnitude of significant correlation coefficients is 0.189 (Supplementary Figures C.8-C.11).

For H3 transmission clusters, we observe weakly negative correlations with contemporary H1 transmission clusters while there are somewhat stronger positive correlations with contemporary BYam transmission clusters across all spatial lags (Supplementary Figure C.8). The correlations with contemporary BVic transmission clusters are not as consistent and shift from a weakly negative association at two spatial lags (rho = -0.068) to a stronger positive association at three spatial lags (rho = 0.162). Similar to the spatiotemporal autocorrelations within each subtype, the cross correlations between MPD values of H3 and H1 transmission clusters exhibit some alternating signals across temporal lags. That is, while there is a negative correlation between contemporary H3 and H1 transmission clusters, when comparing H3 transmission clusters to H1

transmission clusters from the previous season, we observe positive correlations. In other words, increasing values of MPD in H1 transmission clusters from last season are observed with decreasing values of MPD in H3 transmission clusters this season. This alternating pattern extends across all temporal lags. The temporal signal is less clear when comparing H3 transmission clusters to those of either IBV, though there may be some consistently negative correlations at two temporal lags for both IBV.

For H1 transmission clusters, the relationships with increasingly distant and past H3 transmission clusters are similar to those comparing H3 transmission clusters to distant and past H1 transmission clusters, with the alternating temporal signal being most prominent (Supplementary Figure C.9). The correlation between H1 and BYam transmission clusters is weak and inconsistent, save a strong positive correlation between H1 transmission clusters and local BYam transmission clusters from three seasons ago, three temporal lags (rho = 0.141). We observe weakly positive correlations between H1 transmission clusters and contemporary BVic transmission clusters, and these correlations decrease in magnitude with increasing spatial lags.

For IBV transmission clusters, we see consistent patterns relating each IBV to each IAV. For example, BVic transmission clusters are positively correlated with contemporary and local H1 transmission clusters (rho = 0.106) and positively correlated with contemporary H3 transmission clusters at three spatial lags (rho = 0.185; Supplementary Figure C.10). However, instead of antagonism or interference between IBV, we observe positive correlations between contemporary and local BVic and BYam transmission clusters (rho = 0.136; Supplementary Figure C.10-C.11). For BVic, this positive relationship extends to local BYam transmission clusters from the previous

season (rho = 0.127; Supplementary Figure C.10); but for BYam, the relationship is observed as antagonistic at larger spatial lags (rho = -0.091 at two spatial lags; rho = -0.102 at three spatial lags; Supplementary Figure C.11). Even though, MPD values for IAV transmission clusters did not seem to depend on BYam transmission clusters, the converse does not seem true; that is, we observe more significant correlations between BYam transmission clusters and IAV transmission clusters across space and time. Particularly, we see negative correlations between BYam transmission clusters and H3 transmission clusters from the previous season (at one spatial lag rho = -0.135), positive correlations with H1 transmission clusters from two seasons ago (at one spatial lag rho = 0.17), and positive correlations with H3 transmission clusters from three seasons ago (at two spatial lags rho = 0.133).

DISCUSSION

The evolution of seasonal influenza viruses contributes to the recurrence of outbreaks around the world. As viruses spread along chains of transmission within an outbreak, they diversify and the extent of genetic and antigenic change realized can impact viral fitness and, consequently, the likelihood of survival and persistence. Moreover, the overlapping nature of co-circulating pathogens allows for ecological interactions which may impact chains of transmission and, consequently, viral evolution. However, the spatial variation in transmission patterns of co-circulating influenza viruses within the US has not been thoroughly characterized. In this study, we characterize spatial variation in the phylogenetic signal of local influenza virus outbreaks in the US. We find evidence of the dependence of transmission patterns across space, time, and

influenza subtypes, and that these patterns of dependence vary among seasonal influenza subtypes. Particularly, we note that the spatial scale of dependence is seemingly greater for H1 and BYam than H3 and BVic and that temporal dependence patterns are more consistent in IAVs than IBVs. Furthermore, we describe ecological interactions between seasonal influenza viruses and note that the nature of these interactions is greatly dependent on the spatial and temporal scope of analysis.

IAVs have been shown to circulate on larger geospatial scales, with regular emergence of variants causing outbreaks around the world(5). Comparatively, IBVs tend to circulate in more local geographies (3,5). Our findings support this notion of smaller spatial scales for IBVs compared to IAVs as we observed the strength of spatial interaction to decrease more rapidly for IBVs than IAVs at increasing spatial distances. IAVs are thought to be more virulent pathogens that evolve on quicker time scales than IBVs (6). This aspect of the infectious disease systems manifests evidently in the epidemiology of influenza illnesses. As IAVs experience greater extents of genetic and antigenic drift, they are less limited by the standing population adaptive immunity and able to infect a general population (6,7). Conversely, IBVs change less and less quickly, and, as such, induced population adaptive immunity can considerably impact their ability to spread within populations (5). For these reasons, IBVs are relatively more prevalent in children who have relatively naïve immunities compared to adults who have had opportunities to develop protective if not sterilizing immunity to IBVs (27). Given the potential differences in the population demography of susceptible individuals, it is not surprising that IBVs have a more limited spatial scale than IAVs as children have a more limited range of mobility than adults (27,28).

The relative differences of evolutionary potential between IAVs and IBVs may also help to explain the differences we observed between the temporal patterns of dependence in diversification of transmission clusters. Temporal patterns for the IAVs were more regular than those of IBVs. As IAV variants regularly arise around the world and invade the United States each year, marginal population immunity may be similar at the onset of seasonal forcing. On the other hand, the local persistence of IBVs may allude to a more dynamic landscape of population immunity which changes along different time scales than that related to IAVs. Often, the incidence of influenza illness within a seasonal outbreak is seemingly dominated by a single subtype; in other seasons, influenza illnesses are more well distributed among subtypes. Predicting which subtype will predominate in upcoming seasons has proven as challenging as it would be rewarding. This may be, in part, due to complex interactions among the co-circulating viruses.

Antagonistic interactions or interference between IAVs has been noted in infectious disease modeling studies (29) as well as in molecular epidemiology studies (7). We add to this body of literature of antagonistic interaction between H3 and H1 with evidence negatively correlated diversification in transmission clusters occurring within the same influenza season. This finding is likely simply due to the competition between pathogens for susceptible hosts. The disruption of chains of transmission of one subtype from the other could be the result of changes to host behavior, e.g., sick behavior, impacting epidemiologic contact rates or changes to host innate immunity, e.g., induced antiviral states, which may impact secondary host susceptibility. At the molecular scale, conserved or convergent epitopes in the proteomes of influenza viruses may allow for cross-reactive adaptive immunity (30). Interaction as this may be more plausible/probable

for more closely related influenza viruses, e.g., homologous strains. So, while potentially less impactful across influenza subtypes or even contemporary strains, the population adaptive immunity induced from previous influenza outbreaks may have lasting impressions on the future incidence of infections of similar viruses. Even without overlap in protein epitope / antigen profiles, co-circulating influenza viruses may still interact across scales. For example, the landscape of an outbreak can also be drastically altered when co-circulating, contemporary and collocated, pathogens aggregately act to disrupt chains of transmission. This could be through moderation of either contact patterns between hosts, e.g., via sick behavior, or the susceptibility of secondary hosts, e.g., induced antiviral states. Either of these fundamental interactions could explain the antagonistic interactions between the IAVs, but there remains the possibility of antigenic overlap between subtypes as well. Ecological interactions as these have been suggested in explanations to observed patterns of branching in influenza virus phylogenies and incidence of influenza illness (7,31). However, study of these interactions has been conducted using rather coarse spatial resolutions, and, as such, the spatial extent, or scale, of interaction is still unclear. Delineating the root cause of such interaction through comparisons of epitope profiles could be a fruitful and challenging avenue for future research. The synergistic interactions between IBVs and between IBVs and IAVs, though, is not so simply explained. Aside from explanations towards increased susceptibility and co-infections, it may be possible that the observed synergy is more due to chance and the alignment of unrelated outbreaks within an influenza season. To this point, the limited spatial scale of IBV outbreaks may offer support. Given the somewhat coarse, state-level resolution of our analysis combined with differences in population

demography of susceptible peoples, it may be possible that the identified transmission clusters for IBVs have little to no spatial nor demographic overlap with each other or with IAV outbreaks. As such, the coinciding seasonal forcing could be driving the observe relationships.

A study such as this has only been recently made possible due to the accumulation of relevant pathogen genomic sequence data. Still, we have found that the coverage of suitable data is flimsy and there remain many gaps that can impact the scope of analytical inferences. Centers for Disease Control and Prevention (CDC) have developed a "Right Size Roadmap" in which they promote increasing sample sizes for pathogen molecular epidemiological surveillance (32). This initiative has seemingly borne fruit as the coverage of sequence data drastically improved within our study scope from 2010 through 2020, particularly from 2015 onwards. However, data coverage is the greatest limitation in this study. In addition to the coverage with respect to subtypes and seasons, the granularity of spatial data in pathogen molecular sequences greatly limits analytical scope. Despite this limitation, we still identify spatial dependence, though, the relatively weak correlations that we did identify could be indicative of a poorly aligned spatial scale between influenza ecology and data. This is not an easily overcome barrier in data collection as increasingly precise spatial resolutions in data come with threats to privacy of infected people who have deserved rights for their health information to be protected. However, there may be a middle ground between state-level resolutions and those at which patients are potentially identifiable such as with core-based statistical areas (33) or labor markets (34,35). Yet, the transition to finer spatial resolutions in data would bring about novel challenges as well, e.g., regions that cross administrative borders.

Another limitation to this study lies within the methodology, particularly the phylogenetic reconstructions and dating procedures. The choice to generate series of "small" phylogenetic trees is both a strength and a limitation of this research. This relatively simple approach allowed us to use the entirety of available influenza isolate sequence data without necessarily subsampling, a key advantage to our approach to those found elsewhere. Additionally, this relatively simple approach could be conducted in a decentralized manner in near real time during influenza season, potentially augmenting molecular surveillance efforts from public health labs around the country. Furthermore, another strength is in the relative simplicity of phylogenetic reconstruction with fewer taxa included. Phylogenetic complexity does not scale linearly, and, at big data scales, the degree of phylogenetic uncertainty can become extensive and computation complexity related to phylogenetic reconstruction can become intractable. By limiting the scope of each phylogenetic reconstruction, these issues of complexity are somewhat mitigated, and, as such, this methodology can be performed with limited expertise in phylogenetics and by using relatively simple techniques, e.g., maximum likelihood versus Bayesian frameworks.

Still, the limited scope of included isolate sequences inherently limits the temporal signal in the data. This temporal signal is essential for accurate phylogenetic dating. We acknowledge this limitation to be present in our study evidenced by extreme estimates of tMRCAs in some phylogenies. However, we mitigated this limitation in two ways. First, we included outgroup isolates to aid in the identification of valid roots for each phylogenetic tree; though, the use of outgroups did not guarantee accurate dating for the entire phylogeny, particularly at the roots. Second, we included influenza isolates

collected over an entire two-year period, both well before and well after the focal influenza season. Having these bookends in each phylogeny can help to constrain the phylogenetic dating procedure to identify more plausible tMRCAs. Still, it is likely that we missed some local transmission clusters. This limitation is also found in our choices for the start dates and end dates of an influenza season. Though influenza viruses in the United States have the potential to arise from introductions from outside the US, there remains the possibility of chains of transmission beginning before our designated start date. As such, we have potentially missed some of those transmission clusters that originated before calendar week 30 each year. However, this limitation in the identification of transmission clusters may act more to contributed to missingness in our data rather than bias our estimates.

Altogether, we have described the spatial variation in the phylogenetic signal of seasonal influenza viruses and characterized the scales at which these pathogens interact within their respective subtypes and across subtypes. Epidemic and evolutionary dynamics are co-dependent, yet little research has gone into characterizing the extent of evolutionary change realized within an outbreak. Rather, the focus in the literature has been directed towards using evolutionary change to predict epidemics, such as in (7). Here, we take a first step to assess the flow of information in the other direction with the quantification of diversity realized over the course of an influenza outbreak. We employ a simple metric in the mean pairwise patristic/tree distance to quantify evolution, but there remain a multitude of tree shape statistics with which similar analyses can be performed and each of which may divulge a different aspect of influenza transmission. We note that the spatial resolution of our study is limited, but we have also limited the spatial scope of

the analysis to the United States. Future works may also expand the analytical scope to the global scale as well as to incorporate additional infectious disease systems, e.g., RSV and COVID-19. The efforts of this work may culminate in improving efforts towards the prediction of evolutionary trajectories and epidemic dynamics.

REFERENCES

1. Grubaugh ND, Ladner JT, Lemey P, et al. Tracking virus outbreaks in the twenty-first century. *Nat Microbiol*. 2019;4(1):10–19.

2. Ladner JT, Grubaugh ND, Pybus OG, et al. Precision epidemiology for infectious disease control. *Nat Med*. 2019;25(2):206–211.

3. Bedford T, Cobey S, Beerli P, et al. Global Migration Dynamics Underlie Evolution and Persistence of Human Influenza A (H3N2). *PLOS Pathogens*. 2010;6(5):e1000918.

4. Bahl J, Nelson MI, Chan KH, et al. Temporally structured metapopulation dynamics and persistence of influenza A H3N2 virus in humans. *Proceedings of the National Academy of Sciences*. 2011;108(48):19359–19364.

5. Bedford T, Riley S, Barr IG, et al. Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature*. 2015;523(7559):217–220.

6. Bedford T, Suchard MA, Lemey P, et al. Integrating influenza antigenic dynamics with molecular evolution. *eLife*. 2014;3:e01914.

7. Perofsky AC, Huddleston J, Hansen C, et al. Antigenic drift and subtype interference shape A(H3N2) epidemic dynamics in the United States. *eLife* [electronic article]. 2024;13. (https://elifesciences.org/reviewed-preprints/91849). (Accessed May 2, 2024)

8. Magee D, Suchard MA, Scotch M. Bayesian phylogeography of influenza A/H3N2 for the 2014-15 season in the United States using three frameworks of ancestral state reconstruction. *PLOS Computational Biology*. 2017;13(2):e1005389.

9. Damodaran L. MOLECULAR EPIDEMIOLOGY OF CONTEMPORARY SEASONAL INFLUENZA EPIDEMICS. University of Georgia; 2023 (Accessed October 12, 2024).(https://esploro.libs.uga.edu/esploro/outputs/doctoral/MOLECULAR-EPIDEMIOLOGY-OF-CONTEMPORARY-SEASONAL-INFLUENZA/9949558923602959). (Accessed October 12, 2024)

10. Colijn C, Gardy J. Phylogenetic tree shapes resolve disease transmission patterns. *Evolution, Medicine, and Public Health*. 2014;2014(1):96–108.

11. Hayati M. Tree shape statistics and their applications. 2019;(https://summit.sfu.ca/item/19964). (Accessed July 11, 2024)

12. Núñez RC, Hart GR, Famulare M, et al. Using phylogenetic summary statistics for epidemiological inference. 2024;2024.08.07.607080.

(https://www.biorxiv.org/content/10.1101/2024.08.07.607080v1). (Accessed October 12, 2024)

13. CDC. Key Facts About Influenza (Flu). *Centers for Disease Control and Prevention*. 2024;(https://www.cdc.gov/flu/about/keyfacts.htm). (Accessed August 13, 2024)

14. Viboud C, Bjørnstad ON, Smith DL, et al. Synchrony, Waves, and Spatial Hierarchies in the Spread of Influenza. *Science*. 2006;312(5772):447–451.

15. Charu V, Zeger S, Gog J, et al. Human mobility and the spatial transmission of influenza in the United States. *PLOS Computational Biology*. 2017;13(2):e1005382.

16. Chen C, Jiang D, Yan D, et al. The global region-specific epidemiologic characteristics of influenza: World Health Organization FluNet data from 1996 to 2021. *Int J Infect Dis*. 2023;129:118–124.

17. Zanobini P, Bonaccorsi G, Lorini C, et al. Global patterns of seasonal influenza activity, duration of activity and virus (sub)type circulation from 2010 to 2020. *Influenza and Other Respiratory Viruses*. 2022;16(4):696–706.

18. Caini S, Alonso WJ, Séblain CE-G, et al. The spatiotemporal characteristics of influenza A and B in the WHO European Region: can one define influenza transmission zones in Europe? *Eurosurveillance*. 2017;22(35):30606.

19. Dalziel BD, Kissler S, Gog JR, et al. Urbanization and humidity shape the intensity of influenza epidemics in U.S. cities. *Science*. 2018;362(6410):75–79.

20. Re3data.Org. GISAID. 2012;(https://www.re3data.org/repository/r3d100010126). (Accessed September 2, 2024)

21. Bureau UC. Cartographic Boundary Files. *Census.gov*. (https://www.census.gov/geographies/mapping-files/time-series/geo/cartographic-boundary.html). (Accessed July 11, 2024)

22. Bureau UC. Commuting Flows. *Census.gov*. (https://www.census.gov/topics/employment/commuting/guidance/flows.html). (Accessed July 11, 2024)

23. Posit team. RStudio: Integrated development environment for R. Boston, MA: Posit Software, PBC; 2023.(http://www.posit.co/)

24. Manpage of MAFFT. (https://mafft.cbrc.jp/alignment/software/manual/manual.html). (Accessed April 28, 2024)

25. Minh BQ, Schmidt HA, Chernomor O, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*. 2020;37(5):1530–1534.

26. Bivand R, Altman M, Anselin L, et al. spdep: Spatial Dependence: Weighting Schemes, Statistics. 2024;(https://cran.r-project.org/web/packages/spdep/index.html). (Accessed October 13, 2024)

27. Petrova VN, Russell CA. The evolution of seasonal influenza viruses. *Nat Rev Microbiol*. 2018;16(1):47–60.

28. Charaudeau S, Pakdaman K, Boëlle P-Y. Commuter Mobility and the Spread of Infectious Diseases: Application to Influenza in France. *PLOS ONE*. 2014;9(1):e83002.

29. Chen J, Gokhale DV, Liu L, et al. Characterizing Potential Interaction Between Respiratory Syncytial Virus and Seasonal Influenza in the U.S. 2023;2023.10.04.23296424. (https://www.medrxiv.org/content/10.1101/2023.10.04.23296424v1). (Accessed July 11, 2024)

30. Nickbakhsh S, Mair C, Matthews L, et al. Virus–virus interactions impact the population dynamics of influenza and the common cold. *Proceedings of the National Academy of Sciences*. 2019;116(52):27142–27150.

31. Chen J, Gokhale DV, Liu L, et al. Characterizing Potential Interaction Between Respiratory Syncytial Virus and Seasonal Influenza in the U.S. 2023;2023.10.04.23296424. (https://www.medrxiv.org/content/10.1101/2023.10.04.23296424v1). (Accessed May 2, 2024)

32. APHL. *APHL*. (http://www.aphl.org). (Accessed November 10, 2022)

33. Bureau UC. Statistical Areas. *Census.gov*. (https://www.census.gov/programs-surveys/metro-micro/about.html). (Accessed September 2, 2024)

34. Fowler CS, Jensen L. Bridging the gap between geographic concept and the data we have: The case of labor markets in the USA. *Environ Plan A*. 2020;52(7):1395–1414.

35. Nelson GD, Rae A. An Economic Geography of the United States: From Commutes to Megaregions. *PLOS ONE*. 2016;11(11):e0166083.

CHAPTER 5

CONCLUSIONS

In this dissertation, we characterize spatial structuring in the United States (US) and its impact on the epidemiological and evolutionary patterns observed in seasonal influenza. We do this in recognition of challenges arising from the integration of data and theory across ecological scales. Specifically, at larger scales, important spatial constructs lack concrete definitions which inhibits integrative approaches to studying seasonal influenza across scales. To address this in the context of seasonal influenza in the US, we carefully analyzed publicly available data related to seasonal influenza, including commuting flows, influenza-like illness incidence, and genetic sequences of seasonal influenza viruses. We integrated these disparate data streams using several different methodologies (incorporating elements of geography, human mobility, disease incidence, and molecular evolution). Each of the aims / chapters offers evidence suggestive of important spatial structuring, and, when taken together, this work constitutes a holistic characterization of subnational, regional scale within the US.

In Aim 1 / Chapter 2, we analyzed patterns in commuting flows, and, by using human mobility models, we were able to identify a critical point in the distance distributions of commuting flows. This potentially region-defining distance, ~146km or ~91mi, has been found in others' work, both with similar approaches as ours outside of the context of defining spatial scales [~119km in US commutes by Viboud et al (2004)(1); ~300km in global commutes by Balcan et al (2009)(2); and ~146km or

~293km, depending on the formulation, in US commutes by Truscott & Ferguson (2012)(3)] and as a result of specific inquiry to scales in human mobility using more complex human mobility models [~161km in high-resolution displacements in Denmark by Alessandretti et al (2020)(4)].

In Aim 2 / Chapter 3, we focused more intently on delineating regions within the US. Using a network science approach, we were able to generate an array of candidate regional delineations which were then evaluated for their ability to capture patterns in commuting flows, influenza-like illness incidence, and viral ancestry. Our results suggest that the US may be well-characterized with ~8 multistate agglomerations / regions, or influenza transmission zones. Due to differences in the resolution of data and results, it is more difficult to directly compare our findings with those found in the literature [cf., the epidemic invasive tree from Figure 4b of Balcan et al (2009)(2) or the epidemiological geographies from Figure 4 of Rosensteel et al (2021)(5). This challenge of reconciling spatial units across scales remains unmitigated in the absence of more standardized, concrete definitions of larger scale spatial constructs. We are still able to interpret our regionalization results from Aim 2 / Chapter 3 with the results from Aim 1 / Chapter 2, i.e., the region defining distance. Many of our delineated regions span substantially longer distances than what would be dictated by a region-defining distance of <100mi. Rather than interpreting this as suggestive of another, separate, larger regional scale within the US, we feel this observed pattern relates more to heterogeneity in the spatial distribution and organization of populations. In other words, regions may be influenced by scale-related distance, but they are ultimately defined by local metapopulation structures.

This notion of local metapopulation dynamics may be further supported by our findings in Aims 1 & 3. In Aim 1, we explored the association between commuting summaries and influenza-like illness epidemic intensity. Viboud et al (2004) showed that the synchrony of influenza outbreaks among states in the US is influenced by shared commuting flows, i.e., metapopulation dynamics (1). Consider that the aggregation of data across spatial scales can explain an inherent relationship between outbreak synchrony and epidemic intensity. For example, asynchronous outbreaks within a region may be interpreted as an overall diffuse outbreak when aggregated to the regional scale, while simultaneous, synchronized outbreaks within a region would be interpreted as an intense outbreak when similarly aggregated. Extending the implications of these more localized spatial contexts, we may interpret this to mean that state-level resolution is too coarse to properly characterize the spatial epidemiology of seasonal influenza, potentially supported by our findings in Aim 3 / Chapter 4.

In Aim 3 / Chapter 4, we took a molecular epidemiology approach using phylogenetics to explore the similarities of local outbreaks across space and time. Ultimately, we found weak that local outbreaks are rather dissimilar when compared to outbreaks in neighboring states (gauged via correlations in the phylogenetic signals (i.e., mean pairwise tree distances) of each seasonal influenza virus). While at least partly founded in the inefficiency of our approach to comparisons, our results from this analysis could be explained by a poor alignment of spatial scales between the research questions and data. That is, the state- and multistate-/regional-level resolution may be too broad / coarse to capture meaningful relationships in the spatial contexts of local outbreaks.

With the continued emphasis on molecular surveillance, we will continually observe the epidemiology, ecology, and evolution of influenza with increasingly finer detail. Consequently, this will allow us to uncover and resolve more intricate patterns. This point encapsulates both potential future direction and an inherent limitation of this study: availability of high-resolution data. We were able to overcome some aspects of the limited resolution in the data, e.g., summarizing county-level commuting patterns to the state-level without complete loss of nested information. However, ultimately, we were limited to a state-level resolution for all analyses. While we acknowledge that state-level characterizations have practical advantages, e.g., working within defined constructs in data and public health practice, important relationships found at smaller spatial scales are likely obfuscated by such coarse resolution. As influenza incidence and molecular sequence data become increasingly available and with improvements in spatial coverage, we may be able to leverage this high-resolution observations to better characterize local transmission dynamics, a fundamental linkage between spatial epidemiology and viral evolution.

We find that our work is most immediately applicable to computational modeling, both mathematical simulations and phylogeographic analyses. There is a precedence of increased model accuracy with more explicit consideration / parameterization of spatial relationships / hierarchies; Turtle et al (2021) showed that models fit to subpopulations with multi-county clusters outperformed those fit to aggregated cluster data (6). Also discussed by Turtle et al (2021)(6), Centers for Disease Control and Prevention (CDC) in the US host an annual influenza forecasting challenge where forecasts are generated for the aggregate US and individual Health and Human Services (HHS) regions (7,8). If we

consider that modeling at finer resolutions may improve forecast accuracy, then state-level predictions would yield similarly accurate predictions regardless of whether the predictions are aggregated to a regional level as defined in our Aim 2 or that defined by the HHS regions. However, continuing up spatial hierarchies, we may speculate that different subnational geographic representations could impact larger scale model estimations, e.g., international spread, especially in phylogeographic contexts.

Characteristics of spread (local metapopulation dynamics & phylogenetic signal in local outbreaks) may also be useful signals to consider as indicators for anomaly detection in influenza surveillance. The emergence of antigenic variants is a continual concern for influenza, whether from zoonotic origins following antigenic shift events or a more subtle change arising somewhere in human populations following antigenic drift (9,10). The early detection of these emergent strains is paramount to control efforts. Antigenic novelty has been shown to allow for more widespread transmission, both with respect to geographical diffusion as well as demographic segments of a population (9,11,12). In characterizing typical spatial patterns in the spread of seasonal influenza, e.g., delineating specific influenza transmission zones, we may also be thereby creating a useful case for anomaly detection. For example, when spread patterns substantially diverge from those expected given established spatial constructs, this could be indicative of novel variant emergence. We feel this similarly applies to the characterization of local outbreaks using phylogenetic approaches; Perofsky et al (2021) corroborate this with a suggestion that quantifications of patterns in phylogenies [via local branching index] can indicate “selective sweeps” which are characteristic in the emergence of antigenic variants (11).

REFERENCES

1. Viboud C, Bjørnstad ON, Smith DL, et al. Synchrony, Waves, and Spatial Hierarchies in the Spread of Influenza. *Science*. 2006;312(5772):447–451.

2. Balcan D, Colizza V, Gonçalves B, et al. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*. 2009;106(51):21484–21489.

3. Truscott J, Ferguson NM. Evaluating the Adequacy of Gravity Models as a Description of Human Mobility for Epidemic Modelling. *PLOS Computational Biology*. 2012;8(10):e1002699.

4. Alessandretti L, Aslak U, Lehmann S. The scales of human mobility. *Nature*. 2020;587(7834):402–407.

5. Rosensteel GE, Lee EC, Colizza V, et al. Characterizing an epidemiological geography of the United States: influenza as a case study. 2021;2021.02.24.21252361. (https://www.medrxiv.org/content/10.1101/2021.02.24.21252361v1). (Accessed July 11, 2024)

6. Turtle J, Riley P, Ben-Nun M, et al. Accurate influenza forecasts using type-specific incidence data for small geographic units. *PLOS Computational Biology*. 2021;17(7):e1009230.

7. Results from the second year of a collaborative effort to forecast influenza seasons in the United States. *Epidemics*. 2018;24:26–33.

8. McGowan CJ, Biggerstaff M, Johansson M, et al. Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. *Sci Rep*. 2019;9(1):683.

9. Krammer F, Smith GJD, Fouchier RAM, et al. Influenza. *Nat Rev Dis Primers*. 2018;4(1):1–21.

10. Petrova VN, Russell CA. The evolution of seasonal influenza viruses. *Nat Rev Microbiol*. 2018;16(1):47–60.

11. Perofsky AC, Huddleston J, Hansen C, et al. Antigenic drift and subtype interference shape A(H3N2) epidemic dynamics in the United States. *eLife* [electronic article]. 2024;13. (https://elifesciences.org/reviewed-preprints/91849). (Accessed July 11, 2024)

12. Bedford T, Cobey S, Beerli P, et al. Global Migration Dynamics Underlie Evolution and Persistence of Human Influenza A (H3N2). *PLOS Pathogens*. 2010;6(5):e1000918.

BIBLIOGRAPHY

Affairs (IEA) O of I and E. HHS Regional Offices. 2006;(https://www.hhs.gov/about/agencies/iea/regional-offices/index.html). (Accessed October 7, 2024)

Alessandretti L, Aslak U, Lehmann S. The scales of human mobility. Nature. 2020;587(7834):402–407.

APHL. APHL. (http://www.aphl.org). (Accessed November 10, 2022)

Bahl J, Nelson MI, Chan KH, et al. Temporally structured metapopulation dynamics and persistence of influenza A H3N2 virus in humans. Proceedings of the National Academy of Sciences. 2011;108(48):19359–19364.

Balcan D, Colizza V, Gonçalves B, et al. Multiscale mobility networks and the spatial spreading of infectious diseases. Proceedings of the National Academy of Sciences. 2009;106(51):21484–21489.

Barbosa H, Barthelemy M, Ghoshal G, et al. Human mobility: Models and applications. Physics Reports. 2018;734:1–74.

Barbosa H, Hazarie S, Dickinson B, et al. Uncovering the socioeconomic facets of human mobility. Sci Rep. 2021;11(1):8616.

BaTS – Bayesian Tip-association Significance testing. (https://online.fliphtml5.com/yyhu/fvkl/). (Accessed May 15, 2024)

Bedford T, Cobey S, Beerli P, et al. Global Migration Dynamics Underlie Evolution and Persistence of Human Influenza A (H3N2). PLOS Pathogens. 2010;6(5):e1000918.

Bedford T, Riley S, Barr IG, et al. Global circulation patterns of seasonal influenza viruses vary with antigenic drift. Nature. 2015;523(7559):217–220.

Bedford T, Suchard MA, Lemey P, et al. Integrating influenza antigenic dynamics with molecular evolution. eLife. 2014;3:e01914.

Biggerstaff M, Cauchemez S, Reed C, et al. Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature. BMC Infect Dis. 2014;14(1):480.

Billings WZ, Cleven A, Dworaczyk J, et al. Use of Patient-Reported Symptom Data in Clinical Decision Rules for Predicting Influenza in a Telemedicine Setting. J Am Board Fam Med. 2023;36(5):766–776.

Bivand R, Altman M, Anselin L, et al. spdep: Spatial Dependence: Weighting Schemes, Statistics. 2024;(https://cran.r-project.org/web/packages/spdep/index.html). (Accessed October 13, 2024)

Boucherie L, Maier BF, Lehmann S. Decomposing geographical and universal aspects of human mobility. 2024;(http://arxiv.org/abs/2405.08746). (Accessed September 2, 2024)

Brockmann D, Hufnagel L, Geisel T. The scaling laws of human travel. Nature. 2006;439(7075):462–465.

Brownstein JS, Wolfe CJ, Mandl KD. Empirical Evidence for the Effect of Airline Travel on Inter-Regional Influenza Spread in the United States. PLOS Medicine. 2006;3(10):e401.

Bureau UC. Cartographic Boundary Files - Shapefile. Census.gov. (https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html). (Accessed May 11, 2024)

Bureau UC. Centers of Population. Census.gov. (https://www.census.gov/geographies/reference-files/time-series/geo/centers-population.html). (Accessed May 11, 2024)

Bureau UC. Commuting Flows. Census.gov. (https://www.census.gov/topics/employment/commuting/guidance/flows.html). (Accessed May 11, 2024)

Bureau UC. County Population Totals and Components of Change: 2020-2023. Census.gov. (https://www.census.gov/data/datasets/time-series/demo/popest/2020s-counties-total.html). (Accessed September 2, 2024)

Bureau UC. Metropolitan and Micropolitan Statistical Areas Map (March 2020). Census.gov. (https://www.census.gov/geographies/reference-maps/2020/geo/cbsa.html). (Accessed May 17, 2024)

Bureau UC. Statistical Areas. Census.gov. (https://www.census.gov/programs-surveys/metro-micro/about.html). (Accessed September 2, 2024)

Caini S, Alonso WJ, Séblain CE-G, et al. The spatiotemporal characteristics of influenza A and B in the WHO European Region: can one define influenza transmission zones in Europe? Euro Surveill. 2017;22(35):30606.

CDC. Burden of Influenza. Centers for Disease Control and Prevention. 2024;(https://www.cdc.gov/flu/about/burden/index.html). (Accessed July 11, 2024)

CDC. Flu Symptoms & Complications. Centers for Disease Control and Prevention. 2022;(https://www.cdc.gov/flu/symptoms/symptoms.htm). (Accessed August 13, 2024)

CDC. Key Facts About Influenza (Flu). Centers for Disease Control and Prevention. 2024;(https://www.cdc.gov/flu/about/keyfacts.htm). (Accessed August 13, 2024)

CDC. Learn more about the flu season. Centers for Disease Control and Prevention. 2022;(https://t.cdc.gov/C03). (Accessed July 11, 2024)

Charaudeau S, Pakdaman K, Boëlle P-Y. Commuter Mobility and the Spread of Infectious Diseases: Application to Influenza in France. PLOS ONE. 2014;9(1):e83002.

Charu V, Zeger S, Gog J, et al. Human mobility and the spatial transmission of influenza in the United States. PLOS Computational Biology. 2017;13(2):e1005382.

Chastagner A, Hervé S, Bonin E, et al. Spatiotemporal Distribution and Evolution of the A/H1N1 2009 Pandemic Influenza Virus in Pigs in France from 2009 to 2017: Identification of a Potential Swine-Specific Lineage. Journal of Virology. 2018;92(24):10.1128/jvi.00988-18.

Chen C, Jiang D, Yan D, et al. The global region-specific epidemiologic characteristics of influenza: World Health Organization FluNet data from 1996 to 2021. Int J Infect Dis. 2023;129:118–124.

Chen J, Gokhale DV, Liu L, et al. Characterizing Potential Interaction Between Respiratory Syncytial Virus and Seasonal Influenza in the U.S. 2023;2023.10.04.23296424. (https://www.medrxiv.org/content/10.1101/2023.10.04.23296424v1). (Accessed May 2, 2024)

Chen X, Liu S, Goraya MU, et al. Host Immune Response to Influenza A Virus Infection. Front. Immunol. [electronic article]. 2018;9. (https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2018.00320/full). (Accessed September 2, 2024)

Chen Y, Tang F, Cao Z, et al. Global pattern and determinant for interaction of seasonal influenza viruses. Journal of Infection and Public Health. 2024;17(6):1086–1094.

Cheng C, Li J, Liu W, et al. Modeling analysis revealed the distinct global transmission patterns of influenza A viruses and their influencing factors. Integrative Zoology. 2021;16(6):788–797.

Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. Phys. Rev. E. 2004;70(6):066111.

Colijn C, Gardy J. Phylogenetic tree shapes resolve disease transmission patterns. Evolution, Medicine, and Public Health. 2014;2014(1):96–108.

Csárdi G, Nepusz T, Traag V, et al. igraph: Network Analysis and Visualization. 2024;(https://cran.r-project.org/web/packages/igraph/index.html). (Accessed October 13, 2024)

Dalziel BD, Kissler S, Gog JR, et al. Urbanization and humidity shape the intensity of influenza epidemics in U.S. cities. Science. 2018;362(6410):75–79.

Damodaran L. MOLECULAR EPIDEMIOLOGY OF CONTEMPORARY SEASONAL INFLUENZA EPIDEMICS. University of Georgia; 2023 (Accessed October 12, 2024).(https://esploro.libs.uga.edu/esploro/outputs/doctoral/MOLECULAR-EPIDEMIOLOGY-OF-CONTEMPORARY-SEASONAL-INFLUENZA/9949558923602959). (Accessed October 12, 2024)

Delamater PL, Street EJ, Leslie TF, et al. Complexity of the Basic Reproduction Number (R0). Emerg Infect Dis. 2019;25(1):1–4.

Diagnosing Flu | CDC. 2022;(https://www.cdc.gov/flu/symptoms/testing.htm). (Accessed August 13, 2024)

Disease Ecology | Learn Science at Scitable. (https://www.nature.com/scitable/knowledge/library/disease-ecology-15947677/). (Accessed October 7, 2024)

Ebell MH, Rahmatullah I, Cai X, et al. A Systematic Review of Clinical Prediction Rules for the Diagnosis of Influenza. J Am Board Fam Med. 2021;34(6):1123–1140.

Ferguson NM, Galvani AP, Bush RM. Ecological and immunological determinants of influenza evolution. Nature. 2003;422(6930):428–433.

FluView Interactive | CDC. 2023;(https://www.cdc.gov/flu/weekly/fluviewinteractive.htm). (Accessed September 2, 2024)

Fortunato S, Barthélemy M. Resolution limit in community detection. Proceedings of the National Academy of Sciences. 2007;104(1):36–41.

Fowler CS, Jensen L. Bridging the gap between geographic concept and the data we have: The case of labor markets in the USA. Environ Plan A. 2020;52(7):1395–1414.

Gerrymandering | Definition, Litigation, & Facts | Britannica. 2024;(https://www.britannica.com/topic/gerrymandering). (Accessed October 7, 2024)

Gong LI, Suchard MA, Bloom JD. Stability-mediated epistasis constrains the evolution of an influenza protein. eLife. 2013;2:e00631.

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

Grenfell BT, Pybus OG, Gog JR, et al. Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. Science. 2004;303(5656):327–332.

Grubaugh ND, Ladner JT, Lemey P, et al. Tracking virus outbreaks in the twenty-first century. Nat Microbiol. 2019;4(1):10–19.

Hanski I. Metapopulation dynamics. Nature. 1998;396(6706):41–49.

Hayati M. Tree shape statistics and their applications. 2019;(https://summit.sfu.ca/item/19964). (Accessed July 11, 2024)

Heikkinen T, Ikonen N, Ziegler T. Impact of Influenza B Lineage-Level Mismatch Between Trivalent Seasonal Influenza Vaccines and Circulating Viruses, 1999–2012. Clinical Infectious Diseases. 2014;59(11):1519–1524.

Identification of Shared Populations of Human Immunodeficiency Virus Type 1 Infecting Microglia and Tissue Macrophages outside the Central Nervous System | Journal of Virology. (https://journals.asm.org/doi/10.1128/JVI.75.23.11686-11699.2001). (Accessed November 10, 2022)

Influenza Transmission Zones. (https://www.who.int/publications/m/item/influenza_transmission_zones). (Accessed April 18, 2024)

Information for Clinicians on Influenza Virus Testing | CDC. 2023;(https://www.cdc.gov/flu/professionals/diagnosis/index.htm). (Accessed August 13, 2024)

Johnson PTJ, de Roode JC, Fenton A. Why infectious disease research needs community ecology. Science. 2015;349(6252):1259504.

Krammer F, Smith GJD, Fouchier RAM, et al. Influenza. Nat Rev Dis Primers. 2018;4(1):1–21.

Ladner JT, Grubaugh ND, Pybus OG, et al. Precision epidemiology for infectious disease control. Nat Med. 2019;25(2):206–211.

Lemey P, Rambaut A, Bedford T, et al. Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2. PLOS Pathogens. 2014;10(2):e1003932.

Luo W, Liu Q, Zhou Y, et al. Spatiotemporal variations of "triple-demic" outbreaks of respiratory infections in the United States in the post-COVID-19 era. BMC Public Health. 2023;23(1):2452.

Magee D, Suchard MA, Scotch M. Bayesian phylogeography of influenza A/H3N2 for the 2014-15 season in the United States using three frameworks of ancestral state reconstruction. PLOS Computational Biology. 2017;13(2):e1005389.

Manpage of MAFFT. (https://mafft.cbrc.jp/alignment/software/manual/manual.html). (Accessed April 28, 2024)

McGowan CJ, Biggerstaff M, Johansson M, et al. Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. Sci Rep. 2019;9(1):683.

Minh BQ, Schmidt HA, Chernomor O, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution. 2020;37(5):1530–1534.

Morgan OW, Abdelmalik P, Perez-Gutierrez E, et al. How better pandemic and epidemic intelligence will prepare the world for future threats. Nat Med. 2022;28(8):1526–1528.

Morgenstern H. Ecologic Studies in Epidemiology: Concepts, Principles, and Methods. Annual Review of Public Health. 1995;16(Volume 16, 1995):61–81.

Moriyama M, Hugentobler WJ, Iwasaki A. Seasonality of Respiratory Viral Infections. Annual Review of Virology. 2020;7(Volume 7, 2020):83–101.

Morris SE, Blasio BF de, Viboud C, et al. Analysis of multi-level spatial data reveals strong synchrony in seasonal influenza epidemics across Norway, Sweden, and Denmark. PLOS ONE. 2018;13(5):e0197519.

Morris SE, Nguyen HQ, Grijalva CG, et al. Influenza virus shedding and symptoms: Dynamics and implications from a multi-season household transmission study. 2024;2024.03.04.24303692. (https://www.medrxiv.org/content/10.1101/2024.03.04.24303692v2). (Accessed July 11, 2024)

National, Regional, and State Level Outpatient Illness and Viral Surveillance. (https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html). (Accessed May 11, 2024)

Nelson GD, Rae A. An Economic Geography of the United States: From Commutes to Megaregions. PLOS ONE. 2016;11(11):e0166083.

Nelson MI, Simonsen L, Viboud C, et al. Phylogenetic Analysis Reveals the Global Migration of Seasonal Influenza A Viruses. PLOS Pathogens. 2007;3(9):e131.

Neumann G, Kawaoka Y. Seasonality of influenza and other respiratory viruses. EMBO Molecular Medicine. 2022;14(4):e15352.

Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Phys. Rev. E. 2004;69(2):026113.

Nickbakhsh S, Mair C, Matthews L, et al. Virus–virus interactions impact the population dynamics of influenza and the common cold. Proceedings of the National Academy of Sciences. 2019;116(52):27142–27150.

Núñez RC, Hart GR, Famulare M, et al. Using phylogenetic summary statistics for epidemiological inference. 2024;2024.08.07.607080. (https://www.biorxiv.org/content/10.1101/2024.08.07.607080v1). (Accessed October 12, 2024)

Orthomyxoviruses: Structure of Antigens. 2016;(https://www.sciencedirect.com/science/article/pii/B9780128012383957210). (Accessed August 21, 2024)

Parino F, Gustani-Buss E, Bedford T, et al. Integrating dynamical modeling and phylogeographic inference to characterize global influenza circulation. 2024;2024.03.14.24303719. (https://www.medrxiv.org/content/10.1101/2024.03.14.24303719v1). (Accessed July 11, 2024)

Parker J, Rambaut A, Pybus OG. Correlating viral phenotypes with phylogeny: Accounting for phylogenetic uncertainty. Infection, Genetics and Evolution. 2008;8(3):239–246.

Paul Glezen W. Editorial Commentary: Changing Epidemiology of Influenza B Virus. Clinical Infectious Diseases. 2014;59(11):1525–1526.

Perofsky AC, Huddleston J, Hansen C, et al. Antigenic drift and subtype interference shape A(H3N2) epidemic dynamics in the United States. eLife [electronic article]. 2024;13. (https://elifesciences.org/reviewed-preprints/91849). (Accessed May 2, 2024)

Petrova VN, Russell CA. The evolution of seasonal influenza viruses. Nat Rev Microbiol. 2018;16(1):47–60.

Posit team. RStudio: Integrated development environment for R. Boston, MA: Posit Software, PBC; 2023.(http://www.posit.co/)

Prachanronarong KL, Canale AS, Liu P, et al. Mutations in Influenza A Virus Neuraminidase and Hemagglutinin Confer Resistance against a Broadly Neutralizing Hemagglutinin Stem Antibody. Journal of Virology. 2019;93(2):10.1128/jvi.01639-18.

Re3data.Org. GISAID. 2012;(https://www.re3data.org/repository/r3d100010126). (Accessed September 2, 2024)

Results from the second year of a collaborative effort to forecast influenza seasons in the United States. Epidemics. 2018;24:26–33.

Rhee I, Shin M, Hong S, et al. On the Levy-Walk Nature of Human Mobility. IEEE/ACM Transactions on Networking. 2011;19(3):630–643.

Rosensteel GE, Lee EC, Colizza V, et al. Characterizing an epidemiological geography of the United States: influenza as a case study. 2021;2021.02.24.21252361. (https://www.medrxiv.org/content/10.1101/2021.02.24.21252361v1). (Accessed July 11, 2024)

SaTScan - Software for the spatial, temporal, and space-time scan statistics. (https://www.satscan.org/). (Accessed May 11, 2024)

Shin GY, Manuel R. Letter to the editor: Sampling bias should be minimised when analysing influenza transmission zones involving very large countries. Eurosurveillance. 2017;22(40):17.

Shrestha S, Foxman B, Weinberger DM, et al. Identifying the Interaction Between Influenza and Pneumococcal Pneumonia Using Incidence Data. Science Translational Medicine. 2013;5(191):191ra84-191ra84.

Smith AM, McCullers JA. Secondary Bacterial Infections in Influenza Virus Infection Pathogenesis. Influenza Pathogenesis and Control - Volume I. 2014;385:327–356.

Smith DJ, Lapedes AS, de Jong JC, et al. Mapping the Antigenic and Genetic Evolution of Influenza Virus. Science. 2004;305(5682):371–376.

Smith NR, Zivich PN, Frerichs L, et al. A guide for choosing community detection algorithms in social network studies: The Question-Alignment approach. Am J Prev Med. 2020;59(4):597–605.

Southworth M. Designing the Walkable City. Journal of Urban Planning and Development. 2005;131(4):246–257.

Sreenivasan CC, Sheng Z, Wang D, et al. Host Range, Biology, and Species Specificity of Seven-Segmented Influenza Viruses—A Comparative Review on Influenza C and D. Pathogens. 2021;10(12):1583.

Stark JH, Cummings DAT, Ermentrout B, et al. Local Variations in Spatial Synchrony of Influenza Epidemics. PLOS ONE. 2012;7(8):e43528.

Suzuki A, Mizumoto K, Akhmetzhanov AR, et al. Interaction Among Influenza Viruses A/H1N1, A/H3N2, and B in Japan. International Journal of Environmental Research and Public Health. 2019;16(21):4179.

The global region-specific epidemiologic characteristics of influenza: World Health Organization FluNet data from 1996 to 2021 - ClinicalKey. (https://www.clinicalkey.com/#!/content/playContent/1-s2.0-S1201971223000528?returnurl=https:%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS1201971223000528%3Fshowall%3Dtrue&referrer=https:%2F%2Fpubmed.ncbi.nlm.nih.gov%2F). (Accessed May 7, 2024)

Thyagarajan B, Bloom JD. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. eLife. 2014;3:e03300.

Tokars JI, Olsen SJ, Reed C. Seasonal Incidence of Symptomatic Influenza in the United States. Clinical Infectious Diseases. 2018;66(10):1511–1518.

Topham DJ, DeDiego ML, Nogales A, et al. Immunity to Influenza Infection in Humans. Cold Spring Harb Perspect Med. 2021;11(3):a038729.

Truscott J, Ferguson NM. Evaluating the Adequacy of Gravity Models as a Description of Human Mobility for Epidemic Modelling. PLOS Computational Biology. 2012;8(10):e1002699.

Turtle J, Riley P, Ben-Nun M, et al. Accurate influenza forecasts using type-specific incidence data for small geographic units. PLOS Computational Biology. 2021;17(7):e1009230.

U.S. Influenza Surveillance: Purpose and Methods | CDC. 2023;(https://www.cdc.gov/flu/weekly/overview.htm). (Accessed August 13, 2024)

United States. The World Factbook. 2024;(https://www.cia.gov/the-world-factbook/countries/united-states/). (Accessed May 7, 2024)

US Census Bureau CHS. Regions and Divisions - History - U.S. Census Bureau. (https://www.census.gov/history/www/programs/geography/regions_and_divisions.html). (Accessed April 27, 2022)

Uyeki TM, Hui DS, Zambon M, et al. Influenza. The Lancet. 2022;400(10353):693–706.

Vaccines. (https://www.who.int/teams/global-influenza-programme/vaccines). (Accessed October 7, 2024)

Viboud C, Bjørnstad ON, Smith DL, et al. Synchrony, Waves, and Spatial Hierarchies in the Spread of Influenza. Science. 2006;312(5772):447–451.

Wu NC, Otwinowski J, Thompson AJ, et al. Major antigenic site B of human influenza H3N2 viruses has an evolving local fitness landscape. Nat Commun. 2020;11(1):1233.

Yang W, Lau EHY, Cowling BJ. Dynamic interactions of influenza viruses in Hong Kong during 1998-2018. 2019;19008987. (https://www.medrxiv.org/content/10.1101/19008987v1). (Accessed September 2, 2024)

Zanobini P, Bonaccorsi G, Lorini C, et al. Global patterns of seasonal influenza activity, duration of activity and virus (sub)type circulation from 2010 to 2020. Influenza and Other Respiratory Viruses. 2022;16(4):696–706.

Zost SJ, Wu NC, Hensley SE, et al. Immunodominance and Antigenic Variation of Influenza Virus Hemagglutinin: Implications for Design of Universal Vaccine Immunogens. The Journal of Infectious Diseases. 2019;219(Supplement_1):S38–S45.

APPENDIX A

SUPPLEMENTAL MATERIALS FOR AIM 1

DATA MANAGEMENT

The commuting data in its raw format was already a dataframe of adjacency, containing records for location pairs of origins and destinations (nodes) and the commuter volume between those locations (edges with weight). This data was used to estimate the gravity model parameters. To fit the gravity models, data for the population sizes of the locations and the distances between locations were also required.

It was important to first consider the granularity of the data. The commuting data were at the county level for spatial scale and were 5-year period estimates for the temporal scale (2011-2015 and 2016-2020). Data on the location of population centers for US counties for 2010 and 2020 were found; these data included population size and coordinates for the population centers. While population estimates for each year were available, the coordinates for population centers were not readily available for each year. Interpolation between 2010 and 2020 coordinates was a potential solution, but it was likely overly complex, especially considering the commuting data were 5-year estimates.

Therefore, it was necessary to consider how to best align the data to combine across sources. The commuting data were used as the base dataset and augmented with population data and spatial data (i.e., the coordinates of population centers). The population data consisted of the midpoint year for the commuting data time periods; for the 2011-2015 commuting data, 2013 population data were used, and for the 2016-2020

commuting data, 2018 population data were used. For the spatial data, the 2010 population center coordinates were used for the 2011-2015 commuting data, and the 2020 population center coordinates were used for the 2016-2020 commuting data.

Supplementary Table A.1. Commuting and Population Data Summaries by US Census Region

Period	Variable	Class	Parameter	All US	Midwest	Northeast	South	West
2011-2015	Total Observations			N = 137 806	N = 43 414	N = 18 479	N = 60 328	N = 15 585
	Total Workers			N = 144 550 912	N = 31 597 818	N = 27 387 863	N = 52 401 175	N = 33 164 056
	Extent of Workers' Commute	Intracounty	n (%)	104 478 675 (72.3)	22 310 313 (70.6)	17 729 877 (64.7)	36 942 071 (70.5)	27 496 414 (82.9)
		Intrastate	n (%)	34 738 165 (24)	8 123 392 (25.7)	8 074 883 (29.5)	13 284 204 (25.4)	5 255 686 (15.8)
		Intraregion	n (%)	4 349 729 (3)	896 122 (2.8)	1 380 498 (5)	1 786 716 (3.4)	286 393 (0.9)
		Interregion	n (%)	984 343 (0.7)	267 991 (0.8)	202 605 (0.7)	388 184 (0.7)	125 563 (0.4)
	Workers in Commuting Flow		Mean (SD)	1 048.9 (20 024.4)	727.8 (13 247)	1 482.1 (16 371.1)	868.6 (14 263.5)	2 127.9 (44 155.9)
			Median [IQR]	17 [8, 53]	13 [5, 42]	17 [8, 61]	20 [9, 61]	17 [8, 48]
			[Min, Max]	[1, 4 181 968]	[1, 2 095 117]	[1, 730 763]	[1, 1 886 175]	[1, 4 181 968]
	Distance (km)		Mean (SD)	544.2 (835)	399.6 (597.5)	558.2 (909.8)	487.6 (731.2)	1 149.2 (1 297.2)
			Median [IQR]	197.38 [82.47, 608.56]	165.67 [78.72, 426.28]	202.49 [89.05, 492.53]	182.6 [75.49, 579.21]	547.68 [165.64, 1 834.72]
			[Min, Max]	[0.000, 9 415.947]	[0.000, 7 422.532]	[0.000, 9 415.947]	[0.000, 7 887.425]	[0.000, 8 176.692]
	Surrounding Population		Mean (SD)	43 741 678 (74 384 947)	34 708 218 (67 817 479)	52 615 236 (75 835 385)	42 142 969 (72 807 674)	64 582 343 (89 566 987)
			Median [IQR]	6 550 510 [849 815.8, 46 147 049.8]	3 704 492 [517 450.5, 25 536 715.5]	18 642 837 [2 616 856, 60 764 239]	6 060 403 [769 861, 42 024 103]	14 935 436 [1 746 573, 98 340 837]
			[Min, Max]	[0, 319 489 617]	[0, 318 585 761]	[0, 318 553 808]	[0, 319 428 840]	[0, 319 489 617]
2013	Total Counties			N = 3 220	N = 1 055	N = 295	N = 1 422	N = 448
	Total Population			N = 319 653 024	N = 67 576 524	N = 59 505 852	N = 118 397 213	N = 74 173 435
	Population		Mean (SD)	99 271.1 (318 499.2)	64 053.6 (212 727.6)	201 714.8 (341 719.5)	83 261 (221 953.3)	165 565.7 (612 641.2)
			Median [IQR]	26 023 [11 193.0, 66 204.5]	20 117 [8 301.5, 44 788.0]	64 545 [37 414.5, 182 953.5]	26 277.5 [13 523.50, 62 592.25]	21 286 [7 274.75, 83 254.75]
			[Min, Max]	[89, 9 987 189]	[454, 5 252 513]	[1 815, 2 587 759]	[106, 4 352 419]	[89, 9 987 189]

Period	Variable	Class	Parameter	All US	Midwest	Northeast	South	West
2016-2020	Total Observations			N = 121 034	N = 37 860	N = 15 897	N = 53 707	N = 13 570
	Total Workers			N = 154 581 044	N = 32 926 166	N = 28 172 851	N = 57 101 842	N = 36 380 185
	Extent of Workers' Commute	Intracounty	n (%)	112 280 334 (72.6)	23 321 016 (70.8)	18 290 061 (64.9)	40 380 024 (70.7)	30 289 233 (83.3)
		Intrastate	n (%)	36 851 534 (23.8)	8 416 523 (25.6)	8 284 190 (29.4)	14 485 335 (25.4)	5 665 486 (15.6)
		Intraregion	n (%)	4 425 438 (2.9)	919 948 (2.8)	1 392 442 (4.9)	1 810 530 (3.2)	302 518 (0.8)
		Interregion	n (%)	1 023 738 (0.7)	268 679 (0.8)	206 158 (0.7)	425 953 (0.7)	122 948 (0.3)
	Workers in Commuting Flow		Mean (SD)	1 277.2 (23 065.9)	869.7 (14 931.4)	1 772.2 (18 232.7)	1 063.2 (16 597.4)	2 680.9 (51 390)
			Median [IQR]	21 [9, 74]	15 [6, 57]	23 [9, 86]	25 [11.0, 85.5]	21 [9, 64]
			[Min, Max]	[1, 4 429 523]	[1, 2 192 398]	[1, 748 979]	[1, 2 014 104]	[1, 4 429 523]
	Distance (km)		Mean (SD)	520 (820.2)	377.7 (590.7)	531.1 (893.3)	470.2 (717.5)	1 101.1 (1 284.9)
			Median [IQR]	173.25 [75.58, 564.14]	143.23 [72.43, 375.55]	178.5 [79.07, 447.95]	162.03 [70.07, 556.95]	484.32 [147.08, 1 685.97]
			[Min, Max]	[0.000, 8 385.497]	[0.000, 7 354.188]	[0.000, 8 385.497]	[0.000, 7 901.971]	[0.000, 8 177.124]
	Surrounding Population		Mean (SD)	43 102 871 (75 849 047)	33 670 220 (69 439 275)	50 852 875 (77 049 161)	42 199 502 (74 342 561)	63 946 833 (91 057 534)
			Median [IQR]	5 473 778 [695 817.8, 42 555 217.8]	2 949 276 [409 620, 20 157 979]	15 738 528 [2 140 863, 56 818 646]	5 335 326 [650 389, 41 049 150]	13 796 610 [1 507 259, 92 929 858]
			[Min, Max]	[0, 329 899 075]	[0, 328 952 132]	[0, 328 938 728]	[0, 329 899 075]	[0, 329 543 803]
2018	Total Counties			N = 3 222	N = 1 055	N = 296	N = 1 422	N = 449
	Total Population			N = 330 023 248	N = 68 263 019	N = 59 269 592	N = 124 649 156	N = 77 841 481
	Population		Mean (SD)	102 428.1 (328 388.5)	64 704.3 (212 404.5)	200 235.1 (341 022.4)	87 657.6 (237 931.5)	173 366.3 (630 587.1)
			Median [IQR]	26 080.5 [11 094.5, 67 012.5]	19 847 [8 170.5, 45 055.0]	63 028.5 [35 641.0, 180 890.2]	26 280.5 [13 301.50, 64 494.75]	22 286 [7 409, 86 761]
			[Min, Max]	[87, 10 061 533]	[463, 5 171 007]	[1 713, 2 580 088]	[149, 4 676 913]	[87, 10 061 533]

There were additional alignment issues with Alaska and Connecticut. While exploring the data, data sources, and relevant documentation, it was observed that there were slight changes in the "counties" for both of these states around 2019.

Alaska had a county-equivalent called Valdez-Cordova Census Area that was split into Chugach Census Area and Copper River Census Area. The population data for 2010-2020 already included the population estimates for these separated "counties," but the commuting data for 2011-2015 only had records for the combined Valdez-Cordova Census Area. As a simple fix, the data for Chugach and Copper River were summed for the 2010-2015 years to recreate population estimates for Valdez-Cordova. Downloading and using the 2010-2015 census population estimates that contained Valdez-Cordova records was considered, but it was noted that estimates can change from year to year and that newer estimates supersede older ones. Although it likely would not matter significantly, the decision was made to use the newer estimates.

Similarly but distinctly, Connecticut recently requested that the census use newly designated "Planning Regions" rather than their former counties. The population estimates for 2010-2020 still had records for the counties, not planning regions. However, the commuting data for 2016-2020 had records for the planning regions and not the counties. Since the planning regions were not directly aligned with county borders, it was not possible to simply aggregate county-level data to reflect the planning regions. A census notice was found that provided a simple table giving 2010 and 2019 population estimates for the planning regions. Additionally, the census population estimates data at the county level for 2020-2022 included data for the planning regions. Therefore, the

simple table of planning region population estimates for 2010 and 2019 was downloaded, and linear regression was used to interpolate the values for 2011-2018.

GRAVITY MODELS & DISTANCE THRESHOLD OPTIMIZATION

Three gravity models were fit: one including only the distance threshold, another incorporating a large-population-assortative component similar to Truscott & Ferguson (2012) (1), and a more extensive population-size-assortative model, wherein the model includes separate terms for all pairwise combinations of population size tertiles, i.e., not only large-to-large population flows.

Population sizes for each location were categorized using tertiles (e.g., small, medium, large population sizes). These categories were used to fit two variations of the gravity models. First, similar to Truscott and Ferguson (2012), an interaction term was introduced via an indicator variable identifying commuting flows between two large populations; Truscott and Ferguson (2012) suggest this assortative component is important to maintain epidemiological relationships in simulations using synthetic data. Continuing with this notion, another gravity model variation uses an interaction term for all nine, unique pairings of population size categories (e.g., (small | medium | large) + (small | medium | large)). Additionally, the gravity models were further extended by compounding interaction terms for time period and census region. These increasingly complex formulations were fit to the commuting data and compared using analyses of variance.

We use these additional gravity models for comparisons in the distance threshold optimization, but we also use them to formally test whether the increases in complexity are warranted via fit improvements to the data.

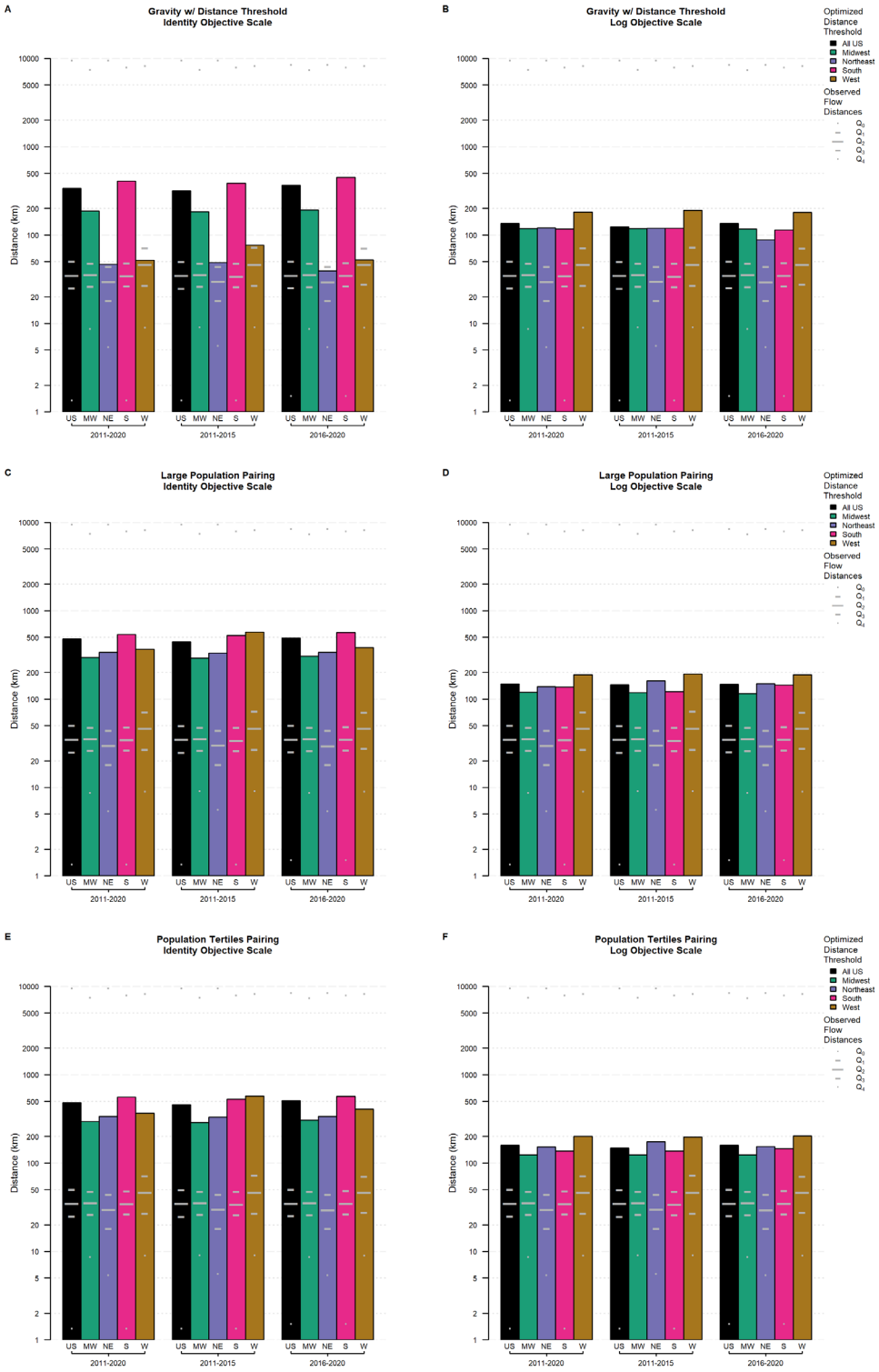
Supplementary Table A.2. Gravity Model Analyses of Variance

Model	Terms	Df	Sum of Sq	F	Pr(>F)	AIC	BIC	RMSE
1	~log(Origin Population) +log(Destination Population) +log(Distance)					868,871.0	868,923.2	1.353
2	*I(Long Distance)	4	143,355.776	32,658.726	<0.001	775,114.6	775,208.5	1.124
3	*(Large-Large Population Size Pairings)	8	8,849.267	1,008.002	<0.001	768,021.8	768,199.3	1.108
4	*(Population Size Category Pairings)	56	4,806.450	78.213	<0.001	764,187.1	764,949.1	1.099
5	*(Origin Census Region)	216	27,707.115	116.891	<0.001	740,580.0	743,596.8	1.048
6	*(Time Period)	284	928.451	2.979	<0.001	740,301.4	746,282.8	1.046

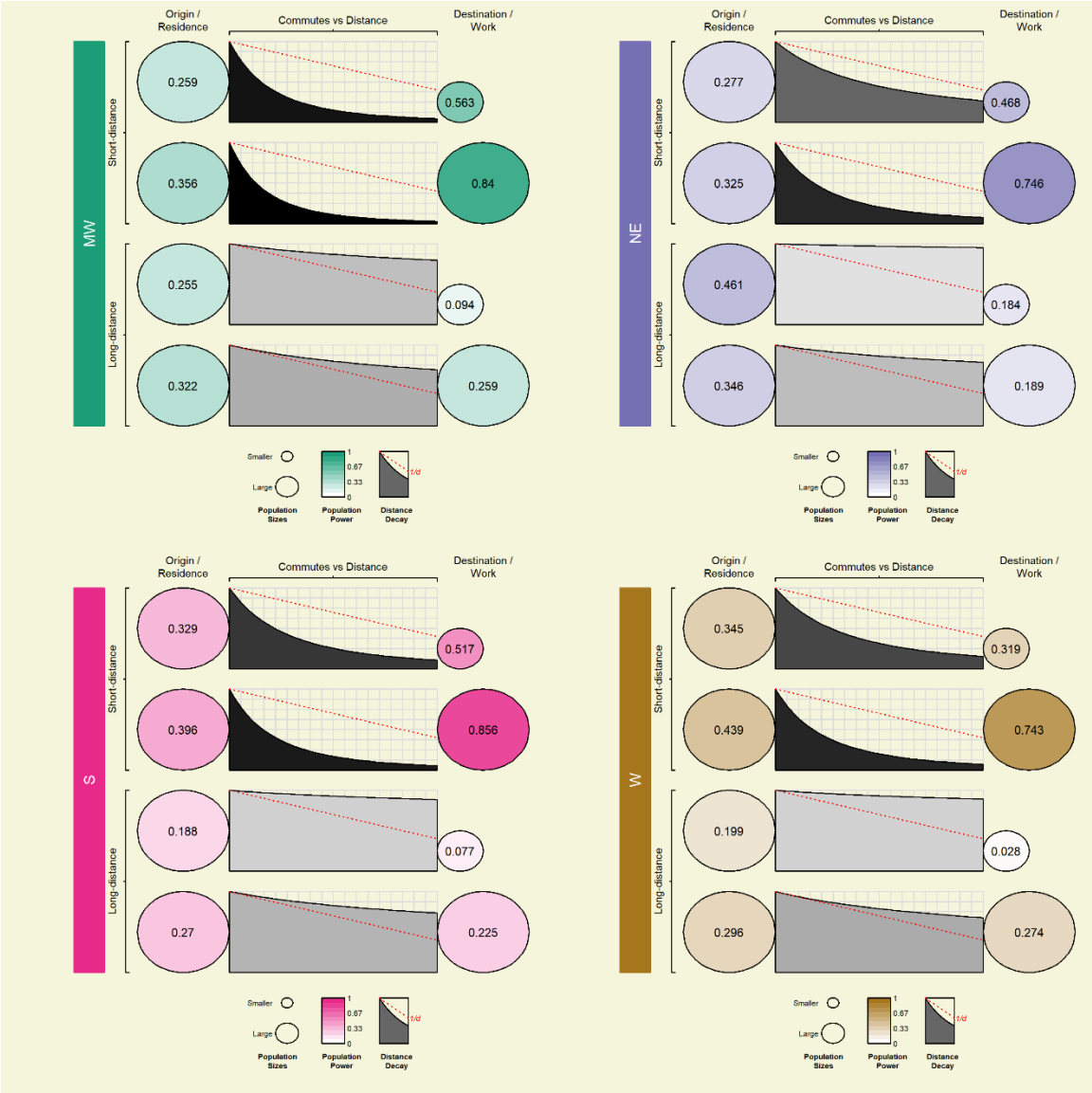
The distance threshold parameter was chosen to minimize the root mean square error (RMSE) of the gravity model predictions compared to the observed commuter volume. The gravity models were fitted using log-linear models; thus, the modeled outcome was the log-transformed count of workers in the commuting flow. For the sake of comparisons, the impact of the choice of data scale (i.e., either raw counts referred to as “identity” or log-transformed counts) for use in the objective function during

optimization was explored. Furthermore, tuning results were compared between a gravity model including only the distance threshold interaction, a gravity model including distance threshold interaction and pairings between two large populations, and a gravity model including distance threshold and population size category pairings interactions (i.e., the full 18-group gravity model discussed previously).

The distance threshold was tuned across different subsets of the total data based on combinations of the time period and the census region of the origin location. Three time periods were considered, corresponding to the time frames of the commuting data collection: 2011-2015, 2016-2020, or both time periods 2011-2020. Five "regions" were considered, corresponding to the four US census regions (Midwest, Northeast, South, West) plus the entire aggregate US. Thus, there were 15 combinations of time periods and regions for which the distance threshold was calibrated.



Supplementary Figure A.1. Sensitivity Analysis of Distance Threshold Optimization



Supplementary Figure A.2. Gravity Model Power Parameter Estimates

Collectively, much lower values for the distance power were estimated for short-distance commutes than long-distance commutes; that is, the frequencies of commutes at short-distances decay much more quickly along increasing distances, while long-distance

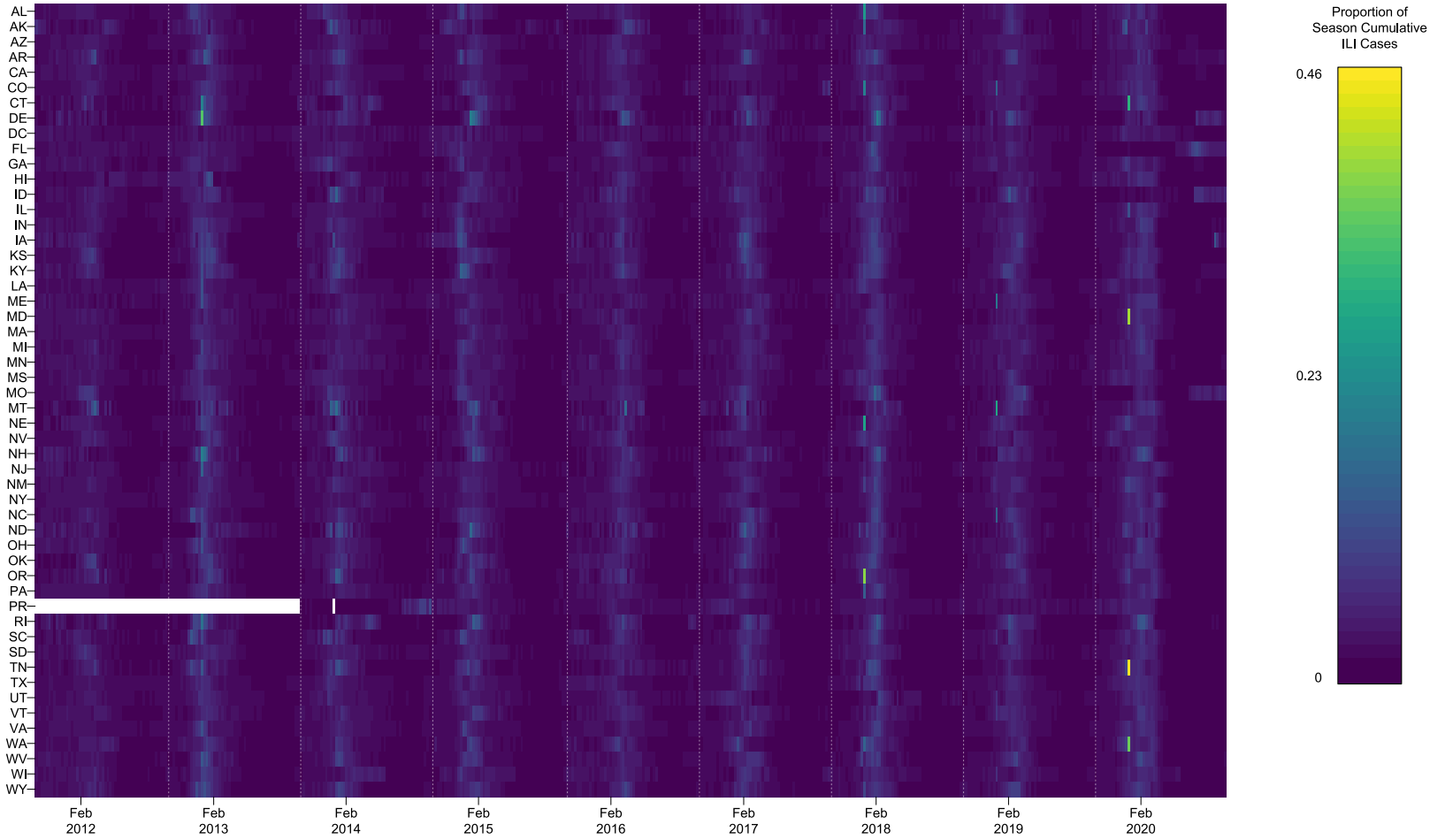
commutes are less affected by increasing distance experiencing only gradual declines. The distance power parameters seemed most variable among regions for short-distance commutes. For short-distance commutes in both population size pairing subgroups, the distance power is estimated to be much lower for the MW region. Additionally, the distance power for short-distance commutes between other population size pairings in the NE is markedly greater than those of the other regions. Overall, this suggests that commuting frequencies decay more steeply over increasing distances in the MW and more gradually in the NE.

Population power estimates across regions were comparably more consistent, though we do find some differences among regions. For all short-distance commutes, the destination population power parameter estimates are larger than those of the origin population, except for the W region. For non-assortative short-distance commutes originating in the W, the destination population power parameter is slightly less than the origin population power parameter. Additionally, the relative difference in population power parameter estimates for commutes between two large populations is also lower in magnitude for the W region compared to the others. This suggests that comparatively the commutes originating in the W are less impacted by destination population sizes.

EPIDEMIC INTENSITY CALCULATIONS

We partition the ILI time series according to the timing of influenza season, set to begin on the 40th calendar / epi week and end the following 39th calendar / epi week. Using each season's cumulative ILI case count, we calculate the relative frequencies of cases observed each week which is then summarized using Shannon's entropy. These

values are transformed and scaled so that zero corresponds to a diffuse epidemic with cases more evenly distributed among weeks and one corresponds to an intense epidemic with cases more concentrated / distributed among fewer weeks. We calculated the ILI epidemic intensity for each of the 52 locations in each of the 9 influenza seasons from 2011-2020, data permitting. Within the ILI data, we also determine the week at which ILI cases were greatest for each state in each season. These ILI data summaries were merged with population data for each state. The population estimates roughly correspond to the estimated population size at the midpoint, 1 July, for each year. To better align with the influenza season, population data were averaged in a two-year rolling window and then joined with the data on epidemic intensities.

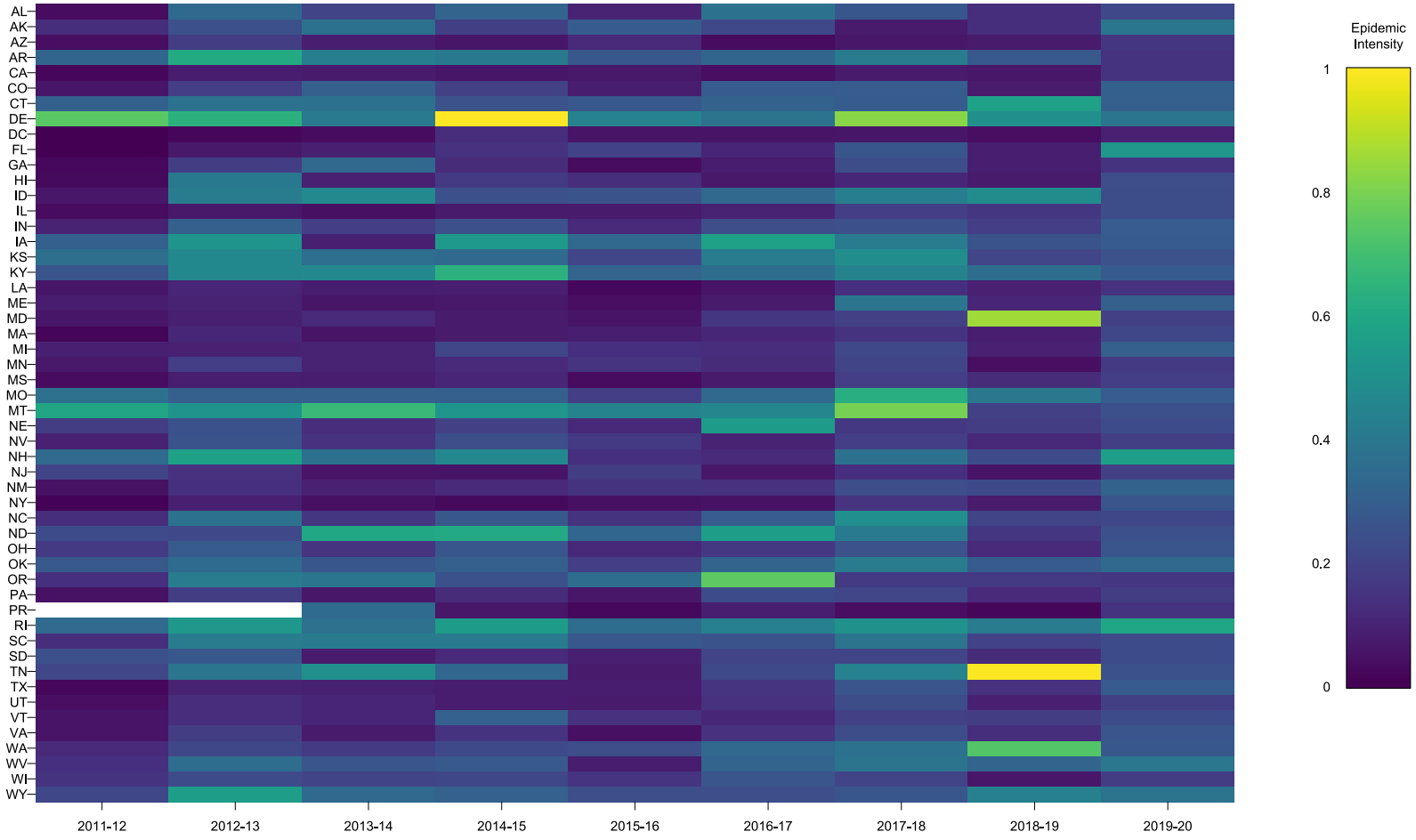


Supplementary Figure A.3. Influenza-like Illness Incidence

Supplementary Table A.3. Epidemic Intensities by Season and State

Region	2011-2012	2012-2013	2013-2014	2014-2015	2015-2016	2016-2017	2017-2018	2018-2019	2019-2020	*Average*
Alabama	0.031	0.338	0.204	0.315	0.098	0.376	0.268	0.132	0.215	*0.220*
Alaska	0.135	0.242	0.375	0.188	0.285	0.217	0.075	0.134	0.391	*0.227*
Arizona	0.036	0.179	0.082	0.057	0.118	0.031	0.062	0.070	0.157	*0.088*
Arkansas	0.322	0.609	0.428	0.404	0.275	0.325	0.414	0.278	0.146	*0.356*
California	0.019	0.078	0.074	0.060	0.066	0.038	0.075	0.065	0.146	*0.069*
Colorado	0.056	0.181	0.299	0.193	0.080	0.284	0.292	0.072	0.302	*0.196*
Connecticut	0.305	0.379	0.376	0.252	0.270	0.313	0.286	0.571	0.297	*0.339*
Delaware	0.743	0.640	0.405	1	0.436	0.381	0.821	0.496	0.390	*0.590*
District of Columbia	0.001	0.019	0.034	0.126	0.054	0.053	0.056	0.040	0.090	*0.053*
Florida	0	0.067	0.086	0.140	0.196	0.102	0.265	0.084	0.532	*0.164*
Georgia	0.025	0.175	0.332	0.122	0.033	0.078	0.232	0.082	0.147	*0.136*
Hawaii	0.030	0.401	0.091	0.171	0.133	0.068	0.106	0.074	0.239	*0.146*
Idaho	0.064	0.411	0.473	0.244	0.257	0.330	0.423	0.484	0.234	*0.324*
Illinois	0.035	0.069	0.045	0.072	0.075	0.095	0.184	0.158	0.231	*0.107*
Indiana	0.096	0.296	0.182	0.244	0.118	0.236	0.244	0.183	0.293	*0.210*
Iowa	0.300	0.515	0.090	0.541	0.334	0.571	0.411	0.257	0.290	*0.368*
Kansas	0.364	0.468	0.359	0.337	0.205	0.417	0.485	0.208	0.254	*0.344*
Kentucky	0.261	0.468	0.472	0.635	0.318	0.353	0.450	0.352	0.282	*0.399*
Louisiana	0.055	0.101	0.076	0.084	0.022	0.053	0.135	0.092	0.142	*0.085*
Maine	0.079	0.098	0.059	0.069	0.038	0.074	0.385	0.103	0.296	*0.133*
Maryland	0.058	0.086	0.116	0.074	0.052	0.159	0.195	0.859	0.193	*0.199*
Massachusetts	0.012	0.106	0.055	0.072	0.081	0.109	0.147	0.080	0.211	*0.097*
Michigan	0.087	0.091	0.098	0.208	0.139	0.128	0.218	0.095	0.299	*0.152*
Minnesota	0.066	0.184	0.100	0.115	0.151	0.124	0.204	0.038	0.170	*0.128*
Mississippi	0.035	0.084	0.077	0.103	0.034	0.071	0.183	0.121	0.183	*0.099*
Missouri	0.367	0.297	0.295	0.297	0.182	0.333	0.628	0.397	0.293	*0.343*
Montana	0.586	0.513	0.673	0.525	0.449	0.463	0.791	0.195	0.244	*0.493*
Nebraska	0.178	0.250	0.125	0.191	0.112	0.550	0.160	0.175	0.230	*0.219*
Nevada	0.093	0.259	0.141	0.238	0.167	0.099	0.185	0.110	0.187	*0.164*
New Hampshire	0.344	0.566	0.374	0.466	0.139	0.117	0.379	0.226	0.562	*0.353*
New Jersey	0.205	0.142	0.053	0.054	0.178	0.062	0.137	0.051	0.193	*0.120*
New Mexico	0.047	0.136	0.093	0.115	0.148	0.143	0.237	0.224	0.312	*0.162*
New York	0.009	0.089	0.044	0.026	0.044	0.047	0.154	0.072	0.264	*0.083*

Region	2011-2012	2012-2013	2013-2014	2014-2015	2015-2016	2016-2017	2017-2018	2018-2019	2019-2020	*Average*
North Carolina	0.135	0.377	0.158	0.274	0.147	0.291	0.493	0.200	0.212	*0.254*
North Dakota	0.227	0.215	0.598	0.600	0.327	0.551	0.410	0.158	0.242	*0.370*
Ohio	0.165	0.281	0.152	0.269	0.110	0.160	0.250	0.120	0.262	*0.197*
Oklahoma	0.276	0.349	0.269	0.297	0.185	0.319	0.414	0.290	0.338	*0.304*
Oregon	0.138	0.413	0.389	0.246	0.353	0.753	0.167	0.168	0.160	*0.310*
Pennsylvania	0.050	0.177	0.060	0.123	0.064	0.229	0.210	0.117	0.176	*0.134*
Puerto Rico			0.346	0.061	0.021	0.083	0.038	0.019	0.148	*0.102*
Rhode Island	0.342	0.531	0.365	0.549	0.351	0.433	0.503	0.414	0.595	*0.454*
South Carolina	0.134	0.411	0.408	0.409	0.272	0.260	0.382	0.196	0.227	*0.300*
South Dakota	0.242	0.273	0.075	0.120	0.089	0.200	0.204	0.124	0.229	*0.173*
Tennessee	0.207	0.389	0.497	0.326	0.075	0.222	0.444	0.991	0.247	*0.378*
Texas	0.020	0.099	0.087	0.079	0.078	0.150	0.265	0.144	0.282	*0.134*
Utah	0.036	0.129	0.100	0.076	0.075	0.145	0.233	0.088	0.186	*0.119*
Vermont	0.054	0.126	0.102	0.309	0.145	0.108	0.194	0.176	0.226	*0.160*
Virginia	0.055	0.178	0.073	0.146	0.042	0.145	0.233	0.131	0.265	*0.141*
Washington	0.118	0.214	0.167	0.225	0.239	0.333	0.380	0.726	0.274	*0.297*
West Virginia	0.139	0.352	0.261	0.276	0.076	0.318	0.381	0.315	0.397	*0.279*
Wisconsin	0.145	0.230	0.202	0.213	0.151	0.263	0.205	0.062	0.178	*0.183*
Wyoming	0.214	0.555	0.339	0.308	0.239	0.237	0.269	0.432	0.381	*0.330*
Average	*0.152*	*0.271*	*0.220*	*0.243*	*0.160*	*0.231*	*0.288*	*0.220*	*0.258*	*0.227*



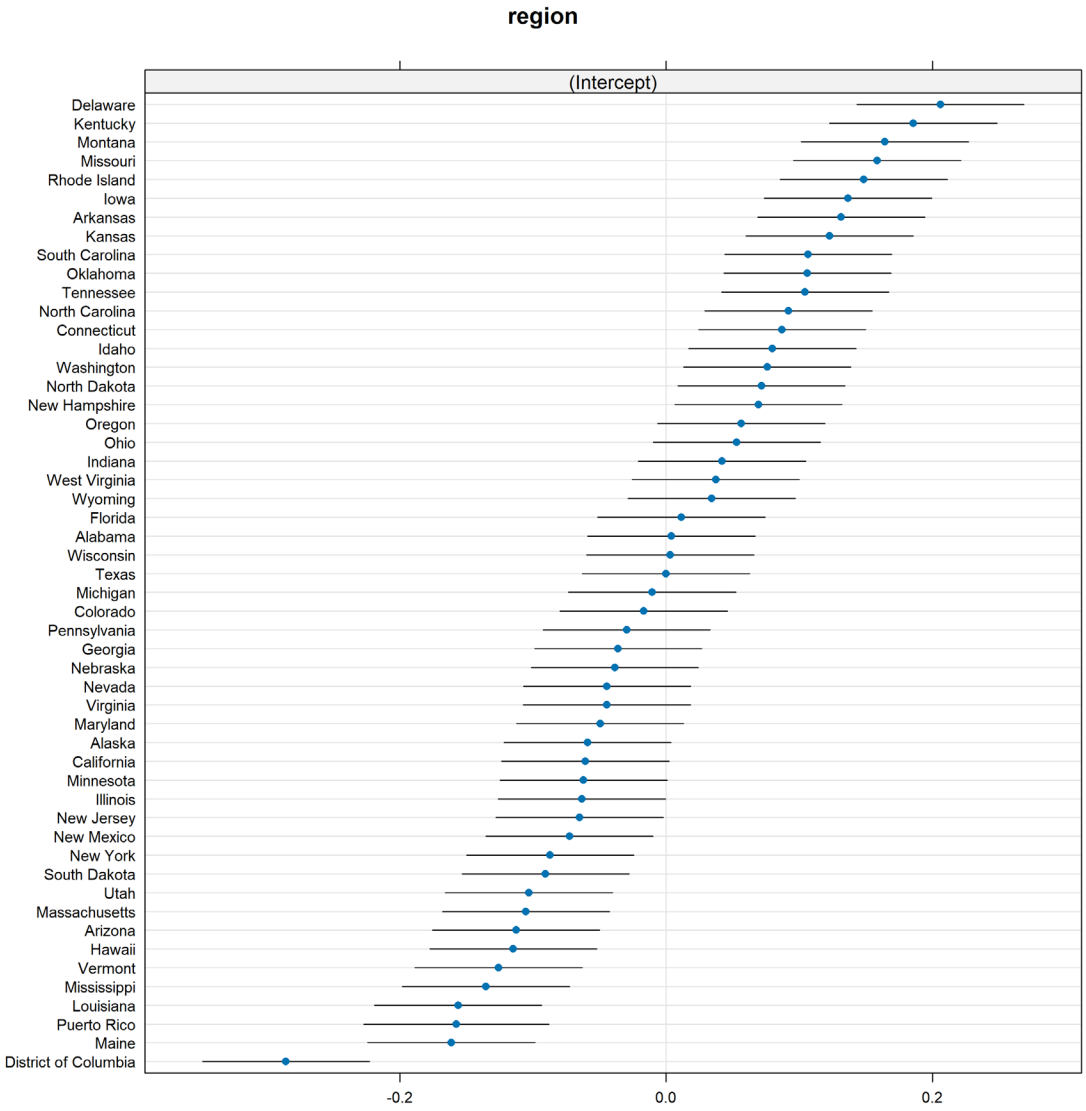
Supplementary Figure A.4. Influenza-like Illness Seasonal Epidemic Intensity for US States and the District of Columbia, 2011-2020

FEATURE ENGINEERING & INCLUDED PREDICTORS

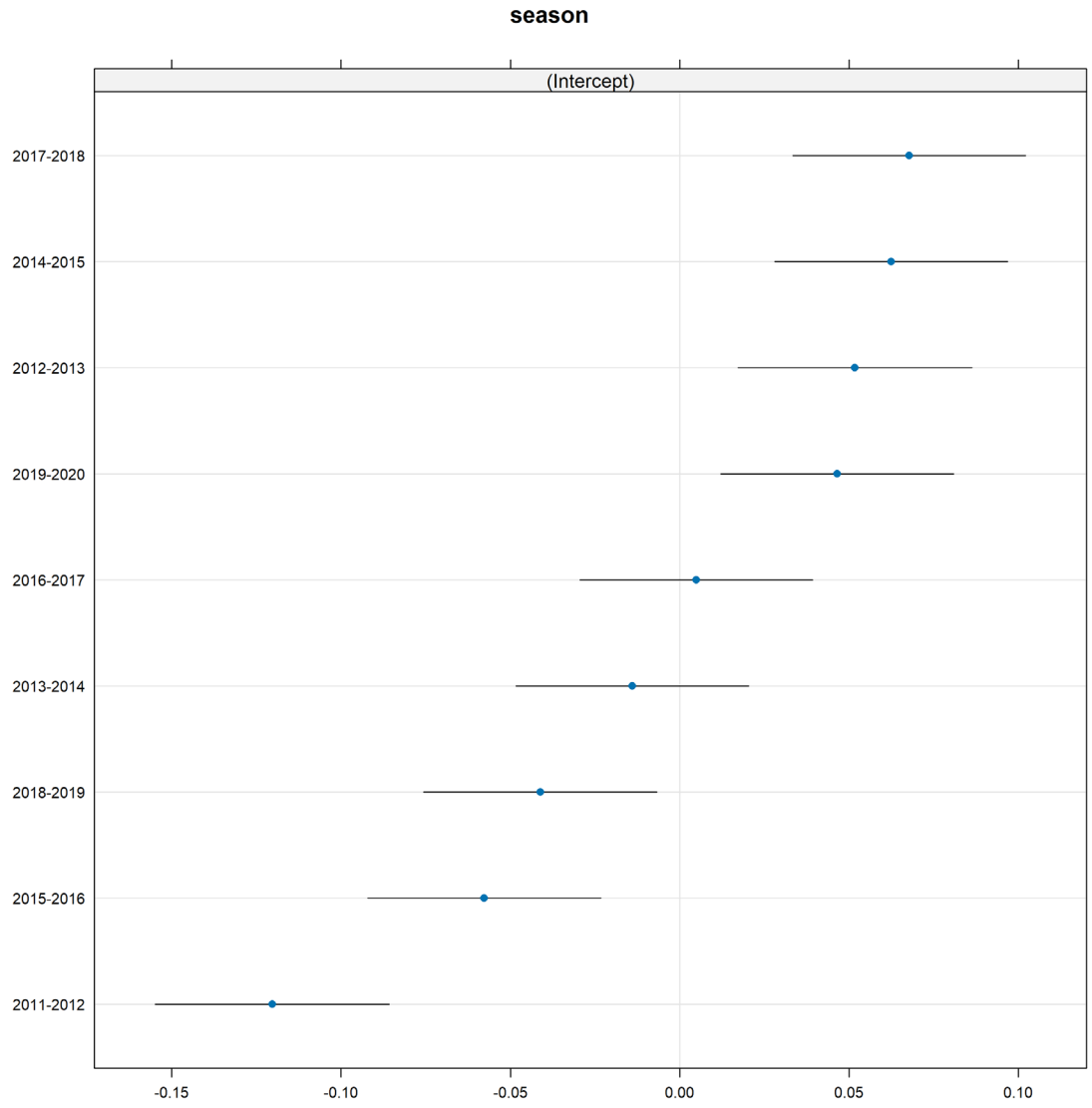
In addition to summarizing only the commuting data, we combine it with the population data to calculate summary metrics of the spatial distribution of population within each state. We use the county-level population data to calculate state-level mean crowding and patchiness, similar to Dalziel et al (2018) (2,3). Finally, using population and commuting data combined, we approximate "daytime" population distributions by shifting population counts according to the net change of population due to commutes (influx and outflux). These shifted population counts allowed us to calculate "workday/daytime" mean crowding and patchiness as well as the changes in these quantities due to commuting frequencies.

Supplementary Table A.4. State-level Summary Statistics of Epidemic Intensity and Predictor Variables

var	Mean (SD)	Median [Min,Max]	Missing n
ei	0.203 (0.147)	0.172 [0, 1]	2
peak_wk	5.425 (5.467)	5 [-12, 39]	2
county.pop.mean	149034.335	80872 [12557.568, 710534.5]	
population	6247251.358	4304076.25 [572073.5,	
Total.Workers	2886984.761	1962038 [284566, 18225448]	
Internal	2092556.346	1338846.5 [247142, 15116254]	
Short_Distance	759457.744	532604 [1458, 3144479]	
Long_Distance	34970.671 (40148.385)	23271 [1200, 234728]	
ratio.si	0.372 (0.226)	0.348 [0.002, 1.146]	
ratio.ls	0.132 (0.343)	0.051 [0.002, 2.816]	
ratio.li	0.018 (0.007)	0.018 [0.002, 0.045]	
Internal_prop	73.667 (11.059)	73.438 [46.486, 99.179]	
Short_Distance_prop	25.004 (11.253)	25.498 [0.215, 53.273]	
Long_Distance_prop	1.329 (0.543)	1.318 [0.117, 3.153]	
county.popday.mean	150681.895	80770.25 [12566.698, 797724.1]	
populationday	6247246.113	4311613.35 [573475.3,	
state.crowding	583346.798	324203.881 [46704.579,	
state.patchiness	4.637 (3.242)	3.904 [1, 18.832]	
state.crowding.day	591048.782	334840.209 [46996.232,	
state.patchiness.day	4.709 (3.293)	3.962 [1, 19.069]	
state.crowding.dailychange	7701.984 (13652.366)	4354.886 [-4733.437, 96496.2]	
state.patchiness.dailychange	0.072 (0.075)	0.044 [-0.005, 0.299]	
state.crowding.dailychangeratio	1.017 (0.021)	1.013 [0.992, 1.154]	
state.patchiness.dailychangeratio	1.014 (0.011)	1.013 [0.998, 1.048]	
distance_mean_km	19.888 (4.415)	19.455 [11.249, 41.091]	
distance_mean_km_nozeros	147.877 (348.839)	70.665 [24.349, 2543.536]	
county_count	61.923 (45.959)	63 [1, 254]	
county_area_km2_mean	4186.737 (7564.209)	1872.696 [113.704, 50994.472]	
state_area_km2_land	176121.104 (218805.58)	136859.004 [158.34,	
state_area_km2_water	13271.67 (35954.609)	3898.064 [18.687, 245481.577]	
pop_density	16394.05 (58676.313)	4139.118 [49.132, 448738.629]	
pop_density2	140.916 (523.059)	36.682 [0.421, 4013.694]	
state_area_km2	189392.775	143520.797 [177.028,	



Supplementary Figure A.5. Random Effect Estimates for Region



Supplementary Figure A.6. Random Effect Estimates for Season

Supplementary Table A.5. Coefficient Estimates from Linear Mixed-effects Regression Models

Population Size	Parameter	Linear	Quadratic	Cubic	Model.P
-0.051 (-0.082,-0.019)*	Peak Week	0 (-0.004,0.004)	0 (-0.001,0)	0 (0,0)*	0.316
-0.07 (-0.109,-0.031)*	Population Density (Land Area)	0.016 (-0.037,0.069)	**-0.02 (-0.04,-0.001)***	-0.007 (-0.016,0.002)	**0.047**
-0.071 (-0.109,-0.032)*	Population Density (Total Area)	0.018 (-0.035,0.071)	**-0.02 (-0.037,-0.002)***	-0.007 (-0.015,0.002)	**0.044**
-0.067 (-0.104,-0.031)*	County Count	0.001 (-0.003,0.005)	0 (0,0)*	0 (0,0)*	0.247
-0.065 (-0.095,-0.035)*	Average County Area	-0.029 (-0.08,0.021)	**-0.024 (-0.04,-0.009)***	0.008 (-0.001,0.016)	**0.009**
-0.036 (-0.072,-0.001)*	Average County Population	0.003 (-0.061,0.067)	-0.004 (-0.032,0.024)	-0.016 (-0.04,0.009)	0.124
-0.038 (-0.073,-0.003)*	Average County Population during Day	0.007 (-0.056,0.07)	-0.005 (-0.033,0.023)	-0.017 (-0.04,0.006)	*0.088*
-0.027 (-0.068,0.013)	State Crowding	-0.031 (-0.095,0.032)	-0.011 (-0.036,0.015)	-0.001 (-0.021,0.02)	0.246
-0.062 (-0.095,-0.029)*	State Patchiness	-0.009 (-0.063,0.046)	**-0.032 (-0.056,-0.007)***	0.01 (-0.006,0.026)	*0.074*
-0.027 (-0.067,0.013)	State Crowding during Day	-0.032 (-0.096,0.031)	-0.011 (-0.037,0.015)	-0.001 (-0.021,0.02)	0.225
-0.062 (-0.094,-0.029)*	State Patchiness during Day	-0.007 (-0.062,0.047)	**-0.031 (-0.055,-0.007)***	0.009 (-0.007,0.026)	*0.081*
-0.059 (-0.093,-0.025)*	Daily Change in State Crowding	0.011 (-0.033,0.055)	**-0.017 (-0.027,-0.008)***	-0.002 (-0.005,0.001)	**0.007**
-0.058 (-0.089,-0.027)*	Daily Change in State Patchiness	**0.055 (0.016,0.095)***	-0.017 (-0.041,0.007)	**-0.011 (-0.02,-0.003)***	**0.034**
-0.06 (-0.089,-0.031)*	Daily Change Ratio in State Crowding	0.021 (-0.015,0.056)	**-0.018 (-0.029,-0.006)***	**-0.005 (-0.008,-0.002)***	**0.008**
-0.057 (-0.087,-0.027)*	Daily Change Ratio in State Patchiness	**0.04 (0.007,0.073)***	-0.012 (-0.039,0.014)	-0.005 (-0.01,0)	*0.055*
-0.06 (-0.092,-0.027)*	State Area Total	-0.014 (-0.058,0.03)	-0.003 (-0.037,0.03)	0.003 (-0.005,0.011)	*0.078*
-0.061 (-0.093,-0.028)*	State Land Area	-0.011 (-0.055,0.033)	-0.004 (-0.04,0.032)	0.003 (-0.006,0.012)	*0.094*
-0.041 (-0.075,-0.008)*	State Water Area	**-0.061 (-0.112,-0.009)***	-0.007 (-0.024,0.01)	**0.009 (0.002,0.015)***	**0.002**
-0.223 (-0.542,0.099)	Total Workers	0.131 (-0.18,0.441)	-0.02 (-0.051,0.011)	0.018 (-0.006,0.043)	0.400
-0.049 (-0.082,-0.016)*	Average Commute Distance	0.002 (-0.034,0.038)	0.001 (-0.016,0.016)	0 (-0.006,0.006)	0.999
-0.053 (-0.085,-0.021)*	Average Commute Distance (no zeros)	0.016 (-0.033,0.065)	-0.018 (-0.061,0.023)	0.002 (-0.008,0.011)	0.455
-0.136 (-0.295,0.024)	Total Commutes Internal/Intracounty	0.05 (-0.117,0.216)	-0.025 (-0.057,0.007)	0.017 (-0.006,0.038)	0.323
-0.055 (-0.087,-0.023)*	Proportion Commutes Internal/Intracounty	-0.015 (-0.074,0.045)	**-0.028 (-0.049,-0.007)***	0.006 (-0.01,0.021)	**0.035**
-0.053 (-0.132,0.027)	Total Commutes Short Distance	0.007 (-0.092,0.105)	-0.024 (-0.07,0.021)	-0.005 (-0.019,0.008)	0.595
-0.055 (-0.087,-0.022)*	Proportion Commutes Short Distance	0.016 (-0.046,0.076)	**-0.026 (-0.047,-0.004)***	-0.006 (-0.023,0.01)	*0.051*
-0.118 (-0.193,-0.043)*	Total Commutes Long Distance	**0.105 (0.009,0.199)***	-0.004 (-0.028,0.019)	-0.012 (-0.025,0.002)	0.153
-0.041 (-0.073,-0.008)*	Proportion Commutes Long Distance	**0.044 (0.01,0.079)***	0.007 (-0.017,0.03)	-0.005 (-0.012,0.003)	*0.097*
-0.054 (-0.087,-0.022)*	Ratio of Short Distance to Internal Commutes	0.004 (-0.045,0.053)	**-0.037 (-0.063,-0.01)***	-0.008 (-0.018,0.002)	**0.047**
-0.051 (-0.084,-0.018)*	Ratio of Long Distance to Short Distance Commutes	0.014 (-0.033,0.061)	-0.017 (-0.05,0.015)	0.002 (-0.003,0.008)	0.701
-0.045 (-0.075,-0.015)*	Ratio of Long Distance to Internal Commutes	**0.06 (0.026,0.095)***	-0.005 (-0.021,0.011)	**-0.009 (-0.014,-0.003)***	**0.002**

APPENDIX B

SUPPLEMENTARY MATERIALS FOR AIM 2

CLUSTERING OF INFLUENZA-LIKE ILLNESS

To characterize coupling/linkages between outbreaks in separate locations, we investigate clustering patterns in the incidence of ILI. For the response variable, weekly reported ILI counts were transformed to a bi-weekly rate of change. We define this rate of change (RoC) as the percent change of the current week's ILI case count from the previous week's count; that is,

$$RoC_t = \frac{ILI_t - ILI_{t-1}}{ILI_{t-1}},$$

where RoC_t is the rate of change observed for week t, ILI_t is the count of ILI cases for week t, and ILI_{t-1} is the count of ILI cases for week $t-1$, the week prior. These rate of change values were calculated for each state across the study period of 2011-2020 and served as the outcome of interest, or response variable, for the clustering analyses.

Clusters were identified using scan statistics via the SaTScan software (16). Briefly, clusters are searched in an iterative fashion using each state as a focal point from which circles of increasing radii extend to define potential cluster constituents. Values of ILI RoC are compared between locations within the cluster and those outside of the cluster using the scan statistics. Cluster significance is determined via permutation tests

with a significance level of $\alpha = 0.05$. A model for a normal distribution was specified, and total patient counts were used as population weights.

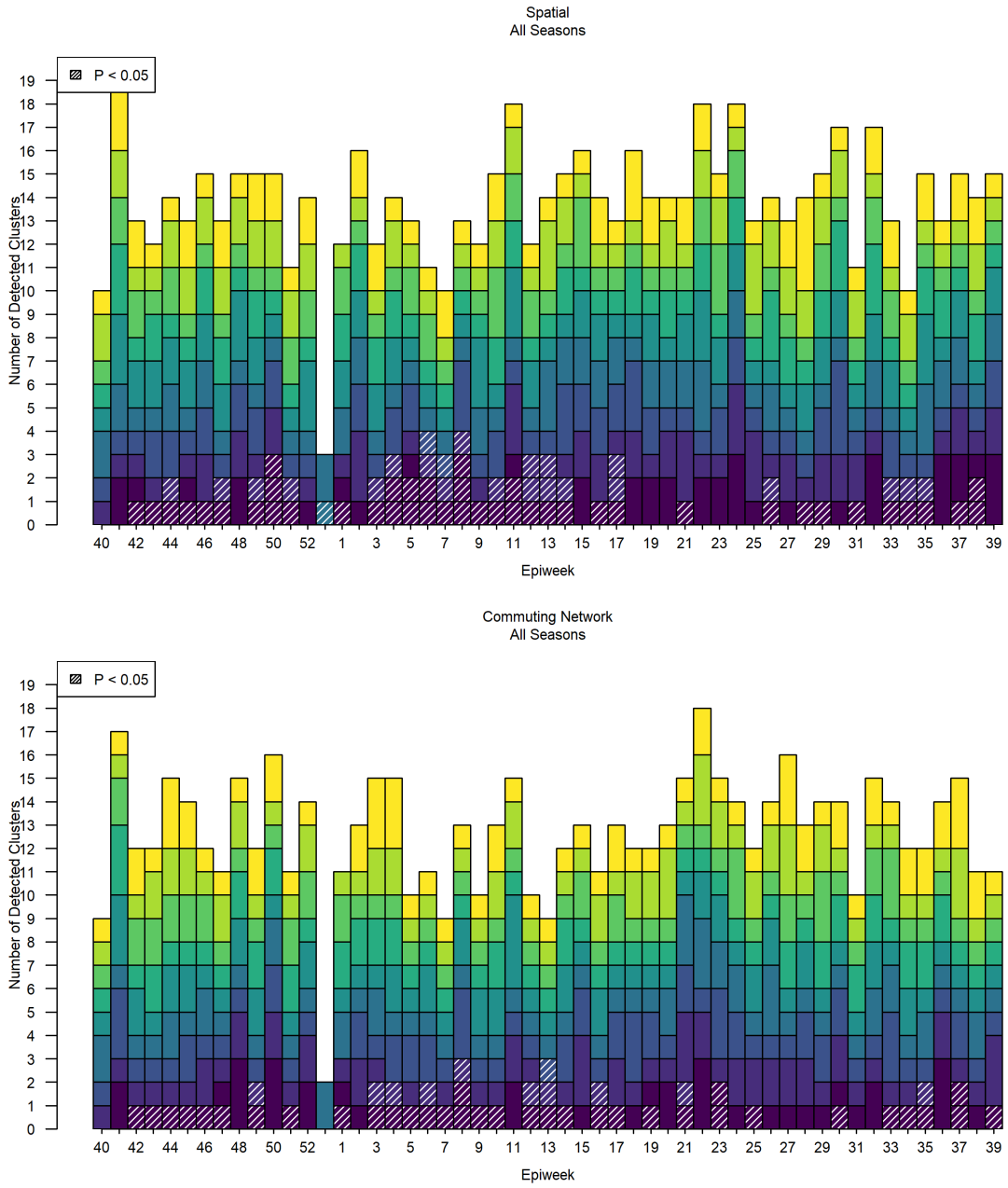
The relationship between locations is necessary to identify potential clusters, i.e., plausible groupings/agglomerations of locations. We define this relationship using two distance metrics. First, we use a simple geographic distance, i.e., Haversine / Great Circle distance. Each state is represented as a point location with latitude and longitude coordinates of the spatial center of population. These coordinates are used by the SaTScan software to calculate distances for generating candidate clusters. We refer to the clusters generated using this distance metric as spatial clusters. Additionally, we investigate clustering using a distance metric related to how many people commute between locations. Data on commuting flows were aggregated to the state level for each unique pair of locations; that is, the data were transformed into undirected, state-level, commuting networks with edges weighted by the number of people commuting between two states, separately for 2011-2015 and 2016-2020. As the edge weights in this commuting network represent the strength of the coupling between nodes, we need to transform the edge weights so that they can be interpreted as distances. We take the reciprocal of the commuting totals to represent the network distance between two locations. Clusters identified using commuting network distance, or network distance, are referred to as commuting clusters or network clusters. We briefly compare the spatial and network clusters in their effect sizes, overall cluster size, and timing relative to influenza season.

Spatial clusters had a median 11 [Q1: 4, Q3: 18.5] constituent states; commuting clusters were similarly sized (p=0.47) with a median 8.5 [4, 15.75] states comprising each

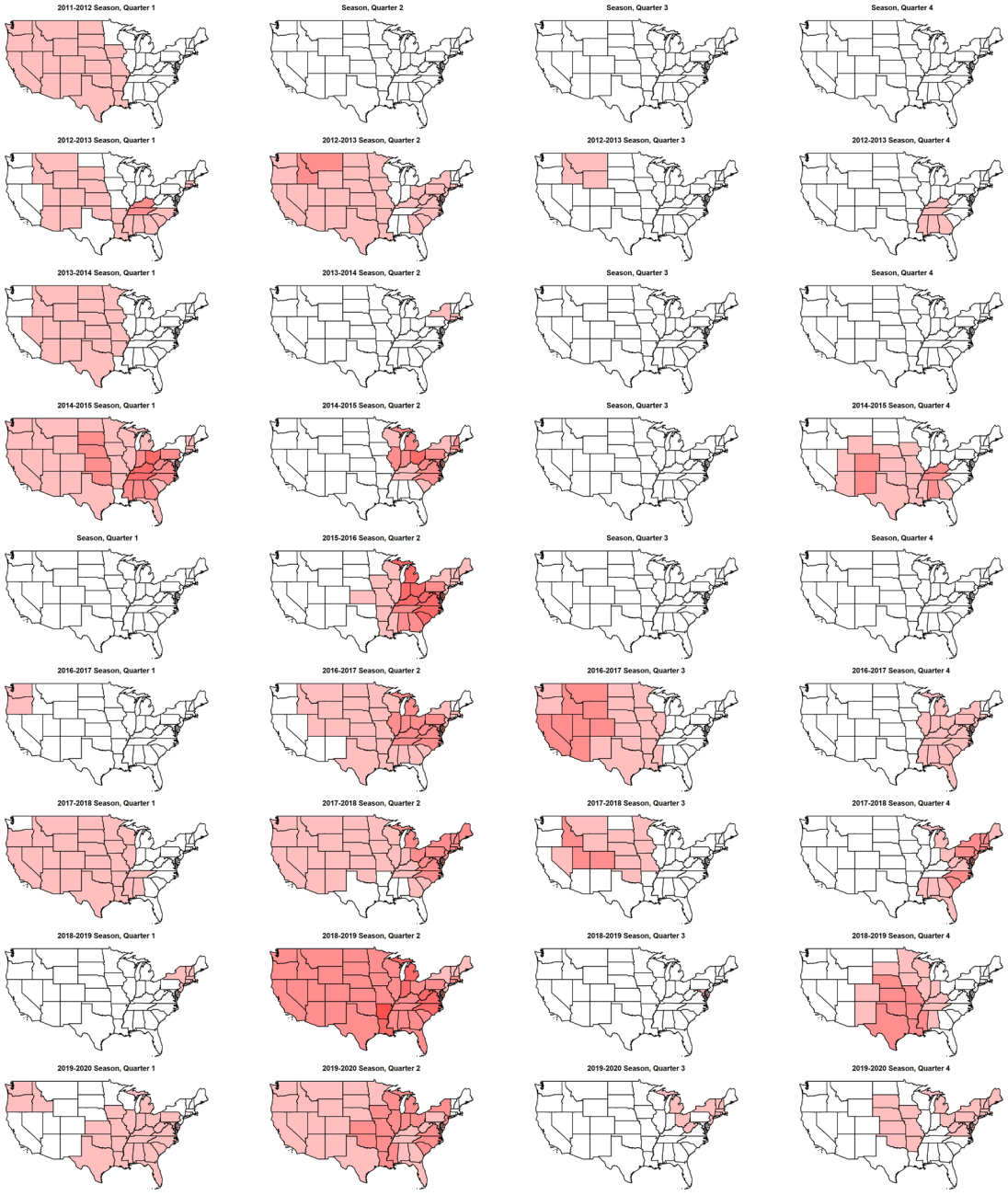
cluster. ILI RoC within spatial and commuting clusters were also similar with an average RoC of +38.6 and +34% change, respectively; weighted means within clusters were also similar at +36.1 and +32.1% change for spatial and commuting clusters, respectively. The magnitudes of ILI RoC outside of clusters were also similar (p=0.33) between spatial and commuting clusters, with an average ILI RoC of -1.8% change outside of spatial clusters and -4.9% change outside of commuting clusters. Consequently, the differences in ILI RoC between clustered and non-clustered states are similar for both types of clusters. At the margin, both types of clusters occurred at similar times with an average occurrence on the 23rd epidemiological/calendar week, i.e., the end of May. When examining the spatial distribution of clusters over the course of an influenza season, there is considerable overlap between spatial and commuting clusters (Figure 1. However, from January-March, there seem to be a higher concentration of spatial clusters, and, from approximately April-June each year, there seem to be relatively few spatial clusters (Figures #, Supplemental Figure #). Comparatively, commuting clusters seem more evenly spread across the year.

Supplementary Table B.1. Identified Clusters of ILI Incidence

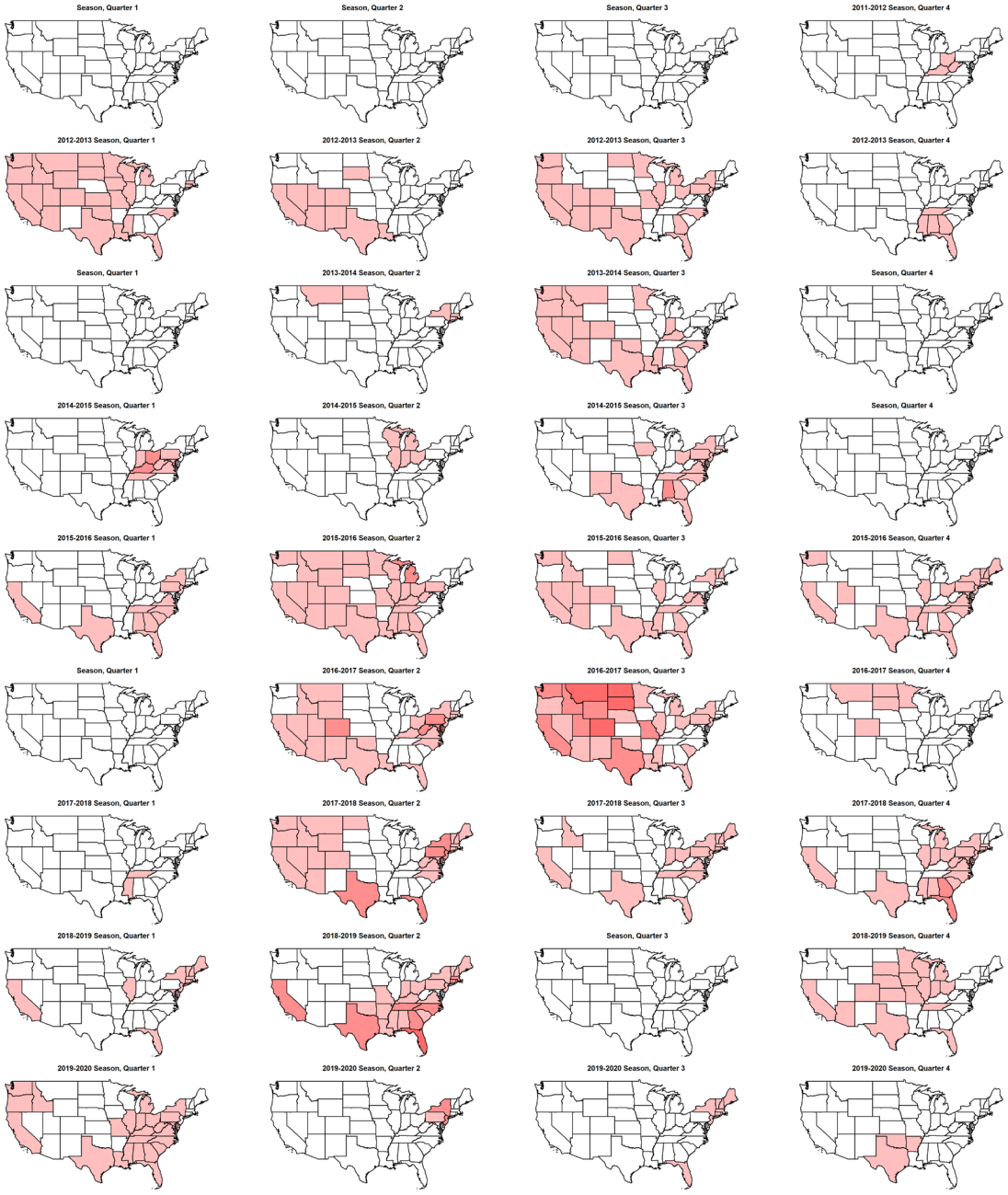
Variable	Levels	Commuting	Spatial	Statistic	Parameter	p
		n = 46 (40.7)	n = 67 (59.3)			
Cluster Size		8.5 [4, 15.75]	11 [4, 18.5]	1,416.00		0.47
	Missing	0 (0)	0 (0)			
Mean inside		0.34235 (0.32614)	0.38601	-0.72	91.470.47	
	Missing	0 (0)	0 (0)			
Mean outside		-0.04947	-0.01809	-0.99	85.660.33	
	Missing	0 (0)	0 (0)			
meandiff		0.39182 (0.21314)	0.4041 (0.24543)	-0.28	104.930.78	
	Missing	0 (0)	0 (0)			
Weighted mean		0.32087 (0.33735)	0.36103	-0.66	87.350.51	
	Missing	0 (0)	0 (0)			
Weighted mean		-0.06417	-0.02984	-1.20	83.520.23	
	Missing	0 (0)	0 (0)			
wtdmeandiff		0.38504 (0.23461)	0.39087	-0.13	96.890.9	
	Missing	0 (0)	0 (0)			
ew		22.52174	23.37313	-0.27	101.480.78	
	Missing	0 (0)	0 (0)			



Supplementary Figure B.1. Frequencies of Spatial and Commuting Clusters by Calendar/Epi Week



Supplementary Figure B.2. Spatial Clusters



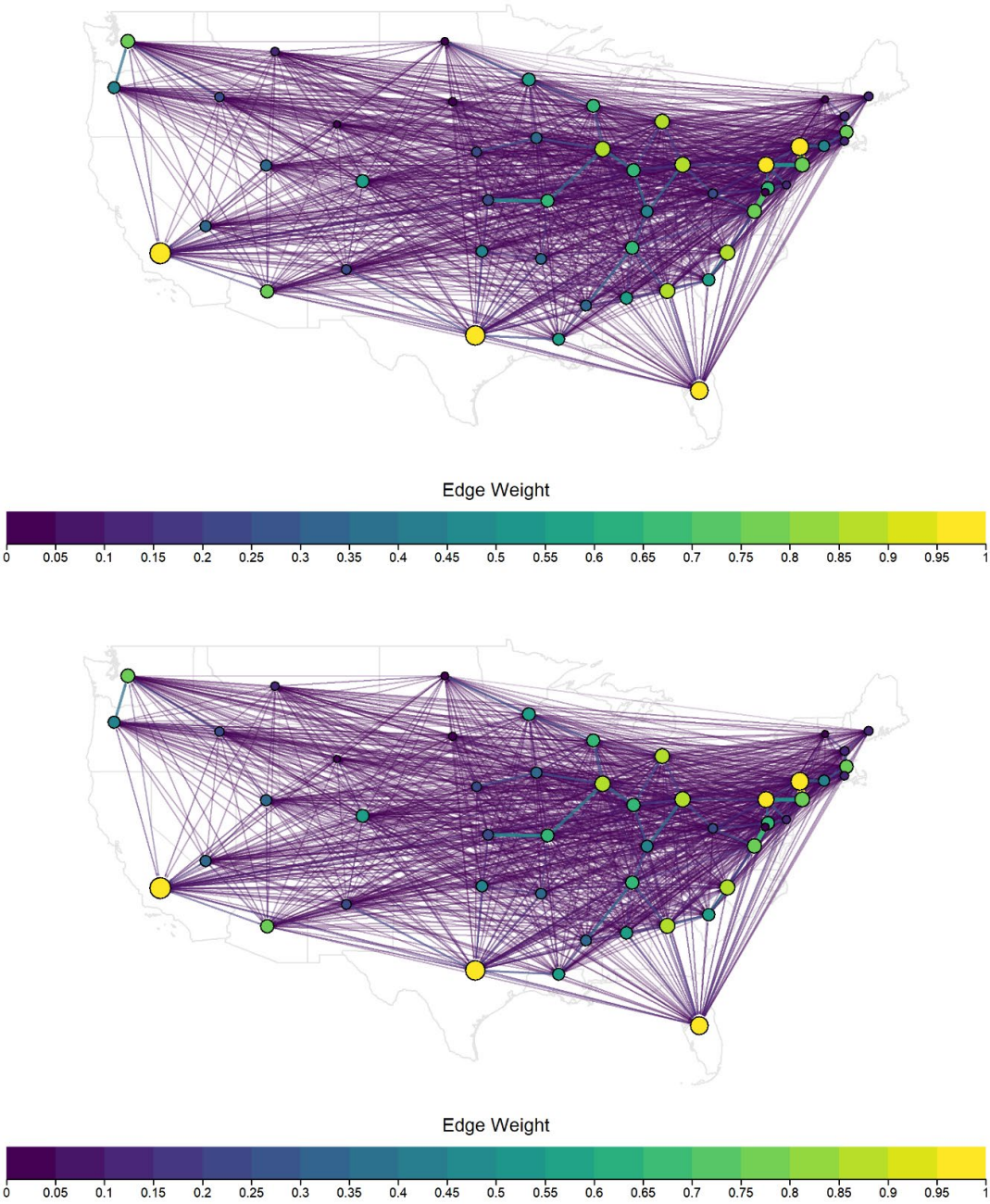
Supplementary Figure B.3. Commuting Network Clusters

Supplementary Table B.2. Summary Statistics of Networks included in Community Detection Analyses

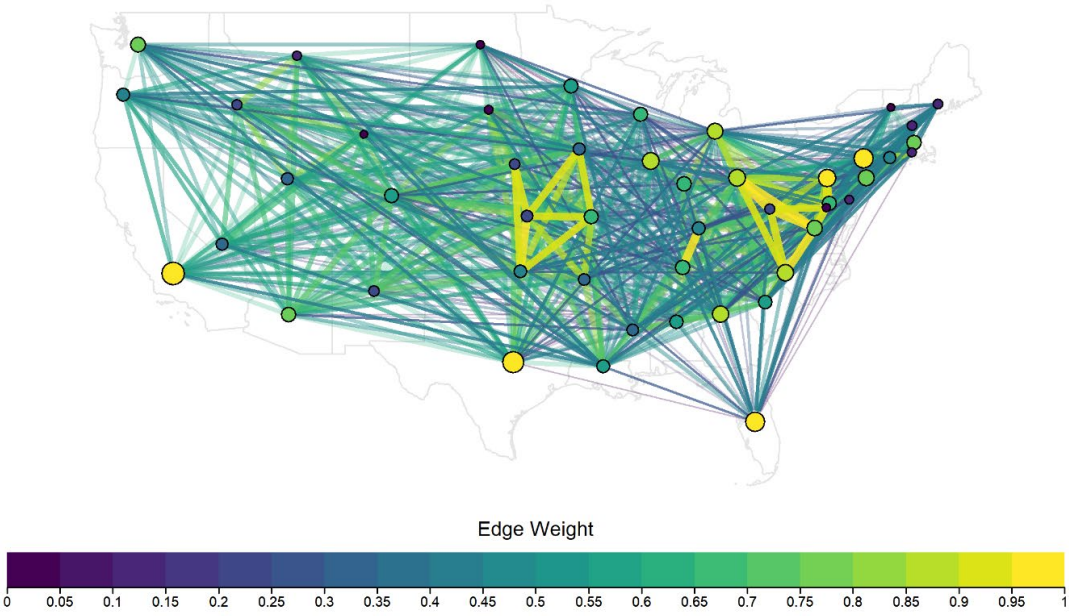
Variable	ACS 2011-2015	ACS 2011-2015 (UD)	ACS 2011-2015 (Scaled)	ACS 2011-2015 (Scaled; UD)
is.directed	1	0	1	0
is.connected	1	1	1	1
any loop	0	0	0	0
adhesion	22	31	22	31
cohesion	22	31	22	31
diameter	2	2	2.017	2.001
reciprocity	0.933	1	0.933	1
alpha centrality.mean	-0.022	-0.02	-0.021	-0.02
authority score.mean	0.886	0.939	0.881	0.943
hub score.mean	0.901	0.939	0.912	0.943
betweenness.mean	5.346	1.135	5.346	1.135
closeness.mean	0.018	0.019	0.018	0.018
degree.mean	91.308	48.731	91.308	48.731
diversity.mean	--	1	--	0.999
eccentricity.mean	1.558	1.558	1.558	1.558
edge betweenness.mean	1.234	1.093	1.234	1.093
meansd.edge.weight	0.005 (0.035)	0.008 (0.046)	0.025 (0.093)	0.027 (0.096)
iqr.edge.weight	0 (0, 0.001)	0 (0, 0.002)	0.003 (0.001, 0.008)	0.003 (0.001, 0.01)
n.edges	2374	1267	2374	1267

Variable	ACS 2016-2020	ACS 2016-2020 (UD)	ACS 2016-2020 (Scaled)	ACS 2016-2020 (Scaled; UD)
is.directed	1	0	1	0
is.connected	1	1	1	1
any loop	0	0	0	0
adhesion	26	32	26	32
cohesion	26	32	26	32
diameter	2	2	2.01	2.004
reciprocity	0.93	1	0.93	1
alpha centrality.mean	-0.017	-0.022	-0.018	-0.02
authority score.mean	0.872	0.925	0.861	0.927
hub score.mean	0.885	0.925	0.895	0.927
betweenness.mean	6.327	1.596	6.327	1.596
closeness.mean	0.018	0.018	0.017	0.018
degree.mean	89.346	47.808	89.346	47.808
diversity.mean	--	1	--	0.999
eccentricity.mean	1.654	1.654	1.654	1.654
edge betweenness.mean	1.283	1.134	1.283	1.134
meansd.edge.weight	0.005 (0.034)	0.008 (0.046)	0.026 (0.094)	0.028 (0.095)
iqr.edge.weight	0 (0, 0.001)	0 (0, 0.002)	0.003 (0.001, 0.009)	0.003 (0.001, 0.01)
n.edges	2323	1243	2323	1243

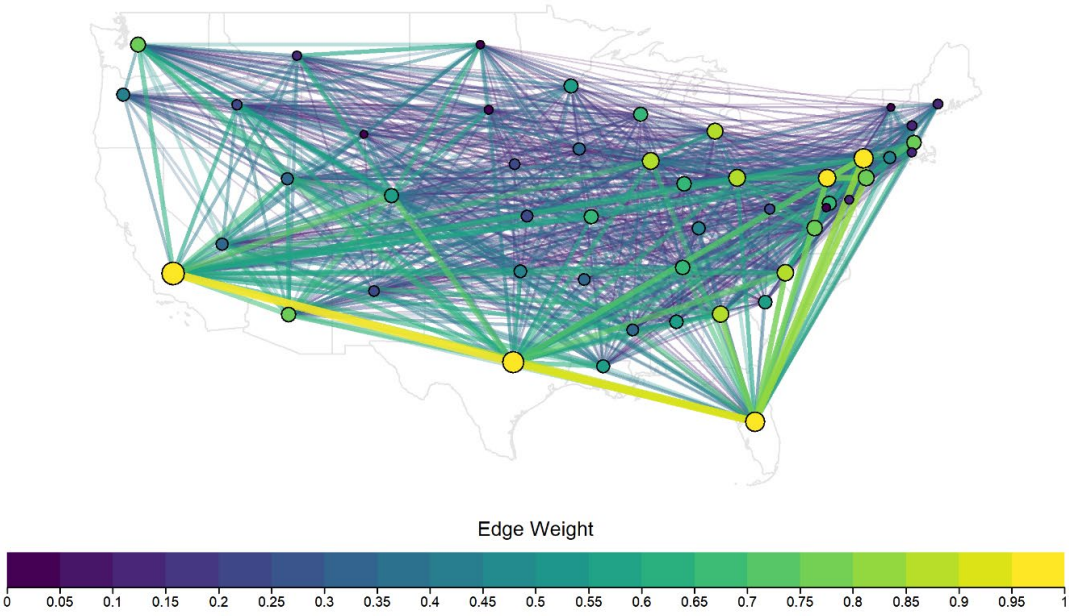
Variable	Clusters ACS	Clusters Spatial
is.directed	0	0
is.connected	1	0
any loop	0	0
adhesion	22	0
cohesion	22	0
diameter	2.118	2.45
reciprocity	1	1
alpha centrality.mean	-0.015	-0.009
authority score.mean	0.706	0.735
hub score.mean	0.706	0.735
betweenness.mean	5.115	7.378
closeness.mean	0.015	--
degree.mean	40.769	34.308
diversity.mean	0.998	--
eccentricity.mean	1.904	1.942
edge betweenness.mean	1.502	1.859
meansd.edge.weight	0.146 (0.152)	0.324 (0.225)
iqr.edge.weight	0.118 (0, 0.235)	0.3 (0.15, 0.5)
n.edges	1060	892



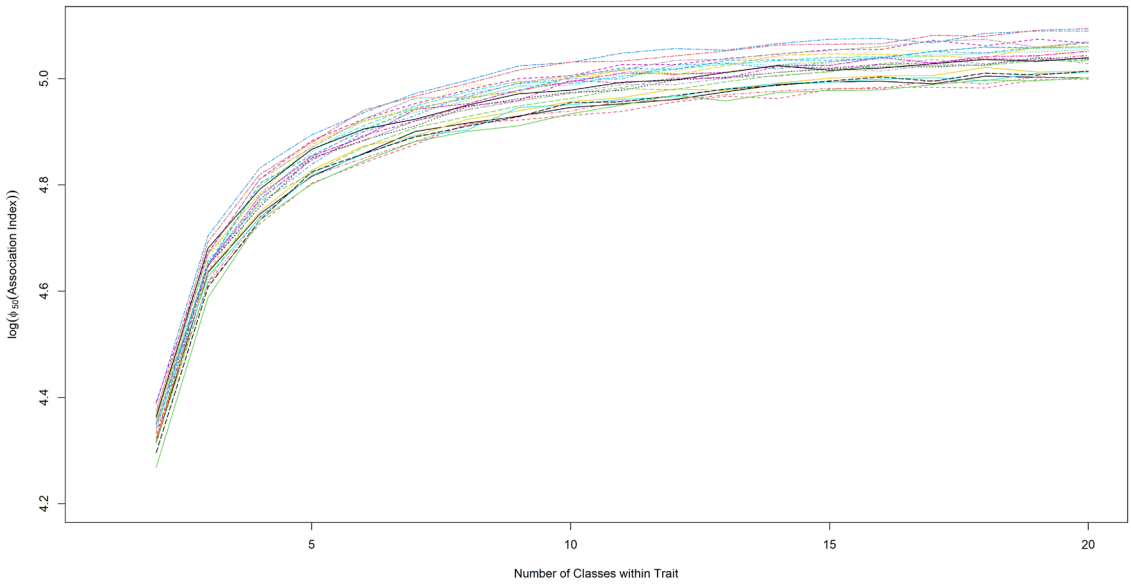
Supplementary Figure B.4. Commuting Networks for 2011-2015 (top) and 2016-2020 (bottom)



Supplementary Figure B.5. Spatial Clusters Network



Supplementary Figure B.6. Commuting Clusters Network

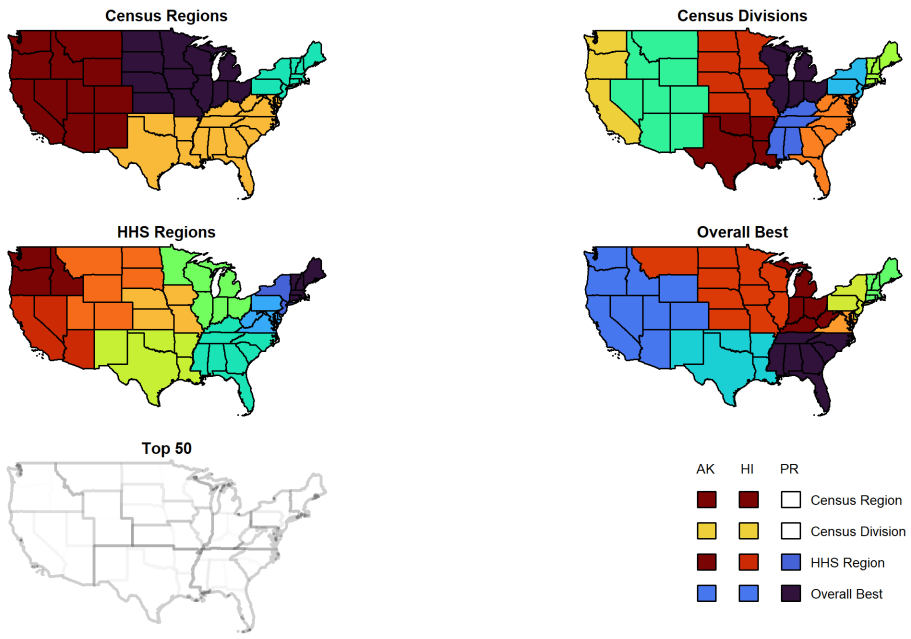


Supplementary Figure B.7. Relationship between Tip-Trait Association Index and the Number of Trait Classes

I devised a simulation to explore how the values of the tip-trait association index depended on the number of classes within a given trait. Color may be categorized three classes as red, blue, or green, or it could be broken down into many more categories; I was curious how the tip-trait association index statistic would differ when only the number of classes changed. I simulated phylogenetic trees and randomly assigned tips to some trait class. This was repeated for many iterations and many different traits which varied in the number of trait classes.

Each tree has 100 tips. Tips are classified into from 2 up through 20 separate classes for a "trait". Traits for a given class number are generated in 10 repetitions. There are 19 "traits" simulated for 10 repetitions for 1000 trees in sets of 25 trees.

Supplementary Figure B.7 shows that the value of the association index will depend on the number of classes within the trait. Therefore, it may be difficult to compare the fit of two separate traits which differ in the number of classes by direct comparison of the association index statistics.



Supplementary Figure B.8. Administrative and Generated Regions of the United States

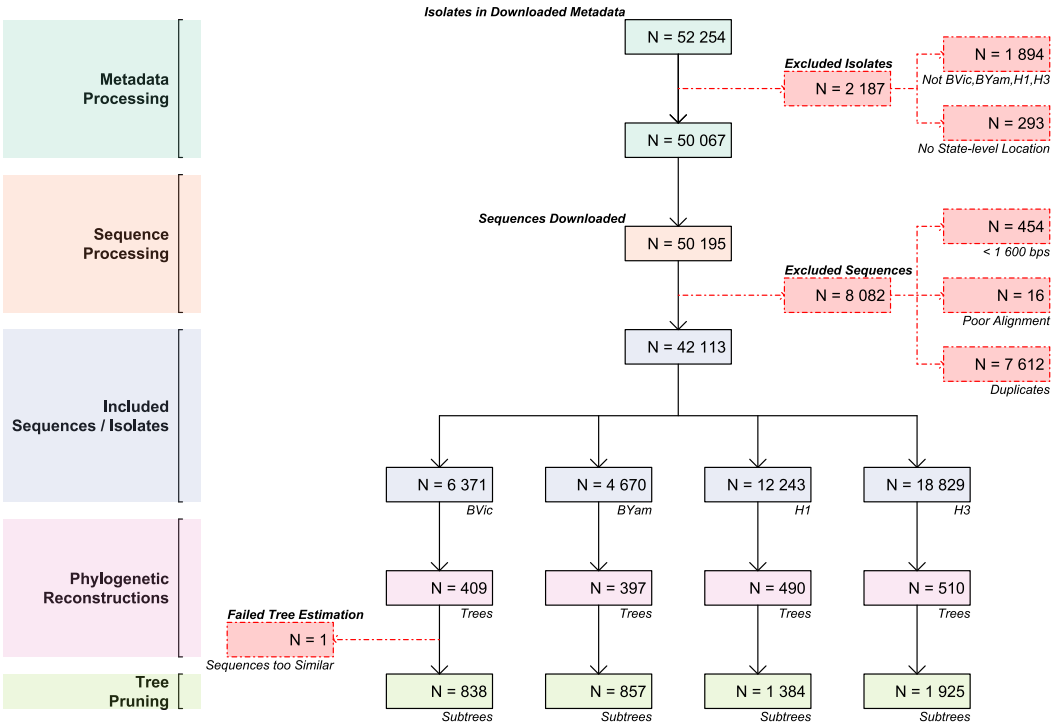
Supplementary Table B.3. Summary Statistics of Administrative and Generated Regions of the United States

Parameter	Census	Census	HHS	Overall	Top 50	All Others
groupedrank3	26	35	21	1	27 [2, 47]	112.5 [52, 173]
n_communities	10	5	10	8	9 [3, 19]	13 [1, 52]
modularity_network.acs.1115	0.492	0.506	0.508	0.602	0.577 [0.275,0	[-0.099, 0.608]
modularity_network.acs.1115.scaled	0.452	0.493	0.497	0.557	0.542 [0.318,0	[-0.054, 0.537]
modularity_network.acs.1115.scaled.undirected	0.45	0.491	0.496	0.555	0.541 [0.318,	-0.002 [-0.056,
modularity_network.acs.1115.undirected	0.491	0.506	0.508	0.601	0.577 [0.275,	-0.001 [-0.101,
modularity_network.acs.1620	0.489	0.504	0.511	0.602	0.577 [0.276,0	[-0.099, 0.606]
modularity_network.acs.1620.scaled	0.454	0.499	0.492	0.558	0.539 [0.316,0	[-0.055, 0.536]
modularity_network.acs.1620.scaled.undirected	0.451	0.497	0.491	0.557	0.538 [0.316,0	[-0.057, 0.535]
modularity_network.acs.1620.undirected	0.489	0.504	0.51	0.602	0.577 [0.276,0	[-0.101, 0.606]
modularity_network.clusters.acs	0.062	0.109	0.061	0.077	0.072 [0.038,	-0.008 [-0.029,
modularity_network.clusters.spatial	0.113	0.159	0.092	0.117	0.105 [0.057,	-0.003 [-0.025,
modularity_mean	0.394	0.427	0.417	0.483	0.457 [0.268,	-0.002 [-0.058,
modrank	58	46	57	9	28 [1, 65]	112 [20, 173]
modacsrank	54	53	52	10	27 [1, 65]	112.5 [18, 173]
modclustrank	32	9	39	21	30 [1, 64]	112 [3, 173]
ai_propp_grand	0.13	0.094	0.123	0.139	0.147 [0.034,	0.216 [0.026,
ps_propp_grand	0	0	0	0	0 [0, 0]	0 [0, 1]
ttairank	27	14	22	35	41 [2, 74]	112.5 [1, 173]
membership_entropy	2.196	1.43	2.266	1.995	2.044 [0.775,	1.84 [0, 3.951]
mcs_entropy	2.167	1.371	2.262	2.043	2.043 [0.692,	2.067 [0, 2.973]
entrank	55	102	46	71	68 [18, 119]	108 [1, 173]

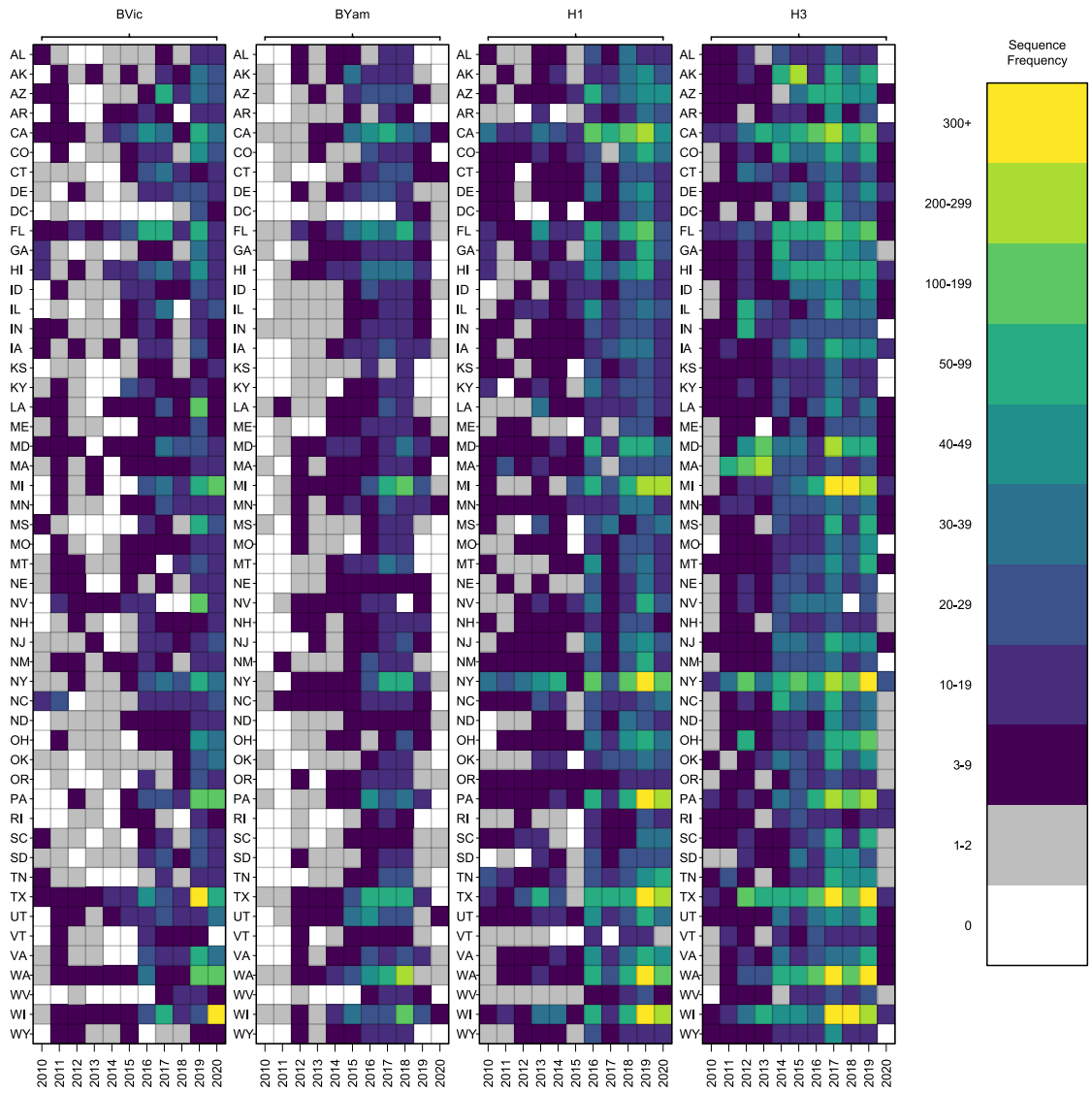
APPENDIX C

SUPPLEMENTAL MATERIALS FOR AIM 3

SUPPLEMENTARY TABLES & FIGURES



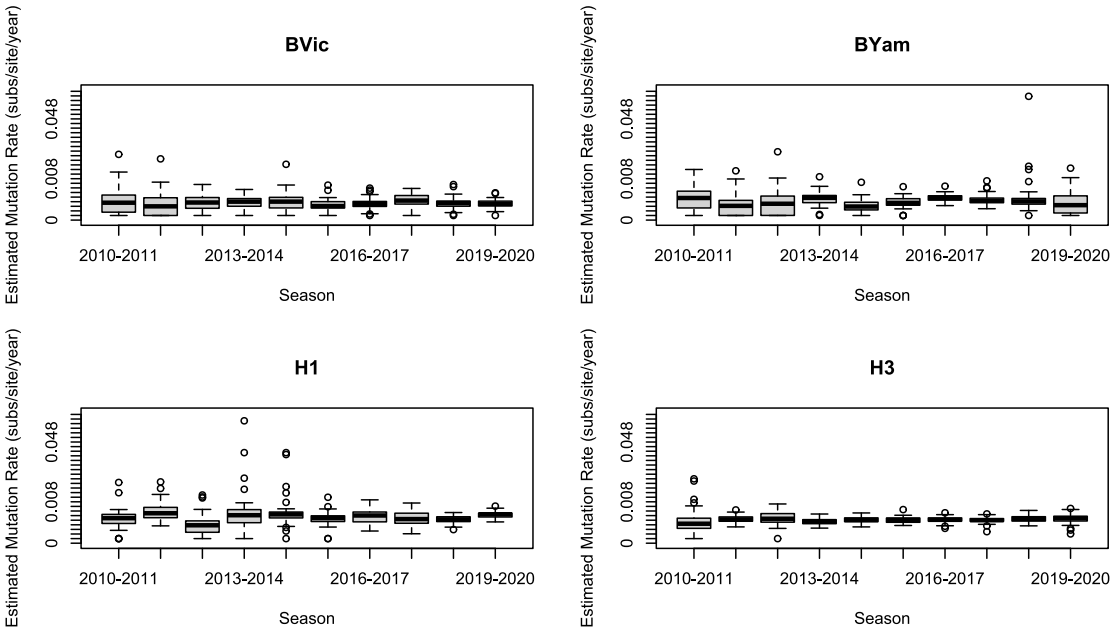
Supplementary Figure C.1. Data Processing



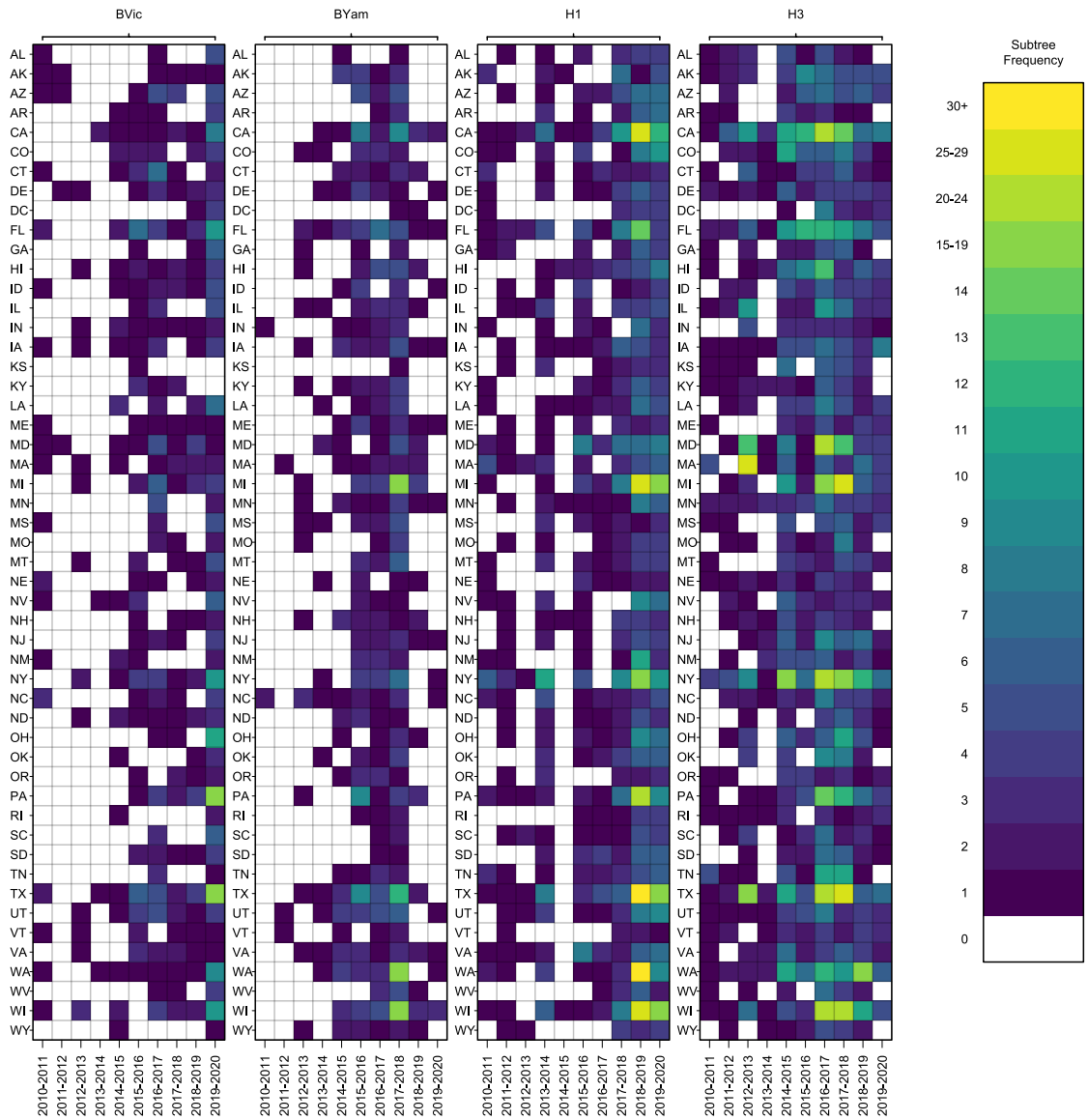
Supplementary Figure C.2. Coverage of Included Hemagglutinin Sequences

Supplementary Table C.1. Full Phylogenetic Tree Summaries

Variable	Levels	BVic	BYam	H1	H3	Statistic	Parameter	p	Test
		n = 510 (25)	n = 510 (25)	n = 510 (25)	n = 510 (25)				
ntips		14 [6, 32]	14 [7, 29]	25 [11, 48]	41.5 [20, 71]	255.40650	3	<0.001	Kruskal-
	Missing	102 (20)	113 (22.2)	20 (3.9)	0 (0)				
mpd		0.0104 (0.00529)	0.0145 (0.00723)	0.03407 (0.03194)	0.02029 (0.00917)	152.92612	3, 1801	<0.001	AoV
imbalance.collessnorm		0.51516 (0.21095)	0.47922 (0.20956)	0.46304 (0.22123)	0.33234 (0.1879)	70.44287	3, 1801	<0.001	AoV
avgladder		2.67982 (1.80769)	2.34488 (1.38364)	2.50771 (1.21182)	2.93548 (1.13976)	15.27091	3, 1801	<0.001	AoV
tmrca1		-3.12 [-	-3.34 [-	-2.1 [-	-4.96 [-	277.68576	3	<0.001	Kruskal-
rate		0.00175 (0.00201)	0.00221 (0.00412)	0.0038 (0.00471)	0.00272 (0.00146)	31.37160	3, 1801	<0.001	AoV



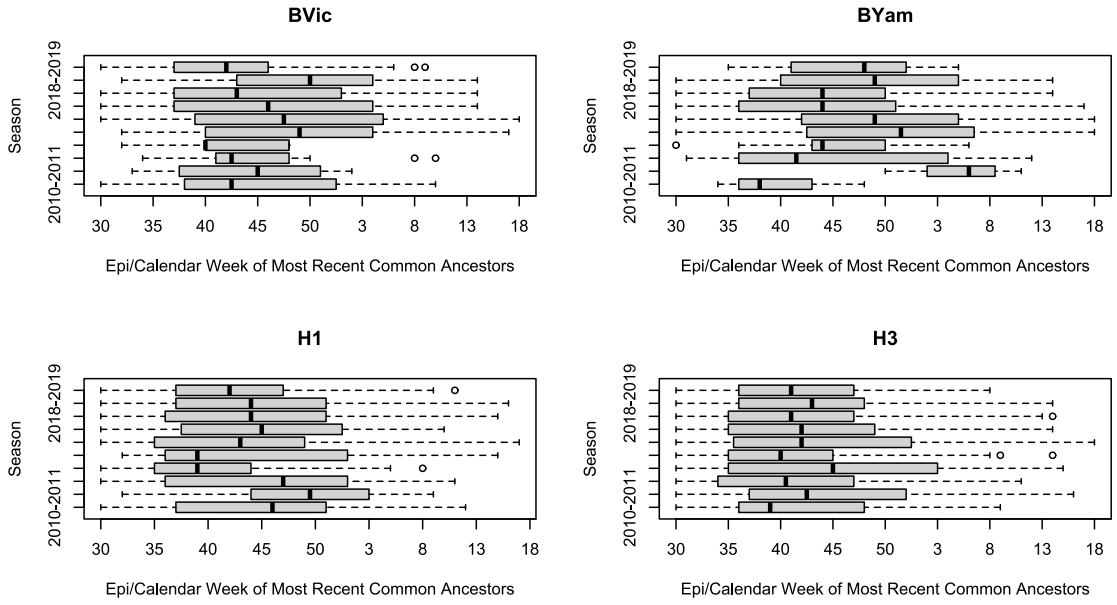
Supplementary Figure C.3. Mutation Rate of Full Phylogenetic Trees



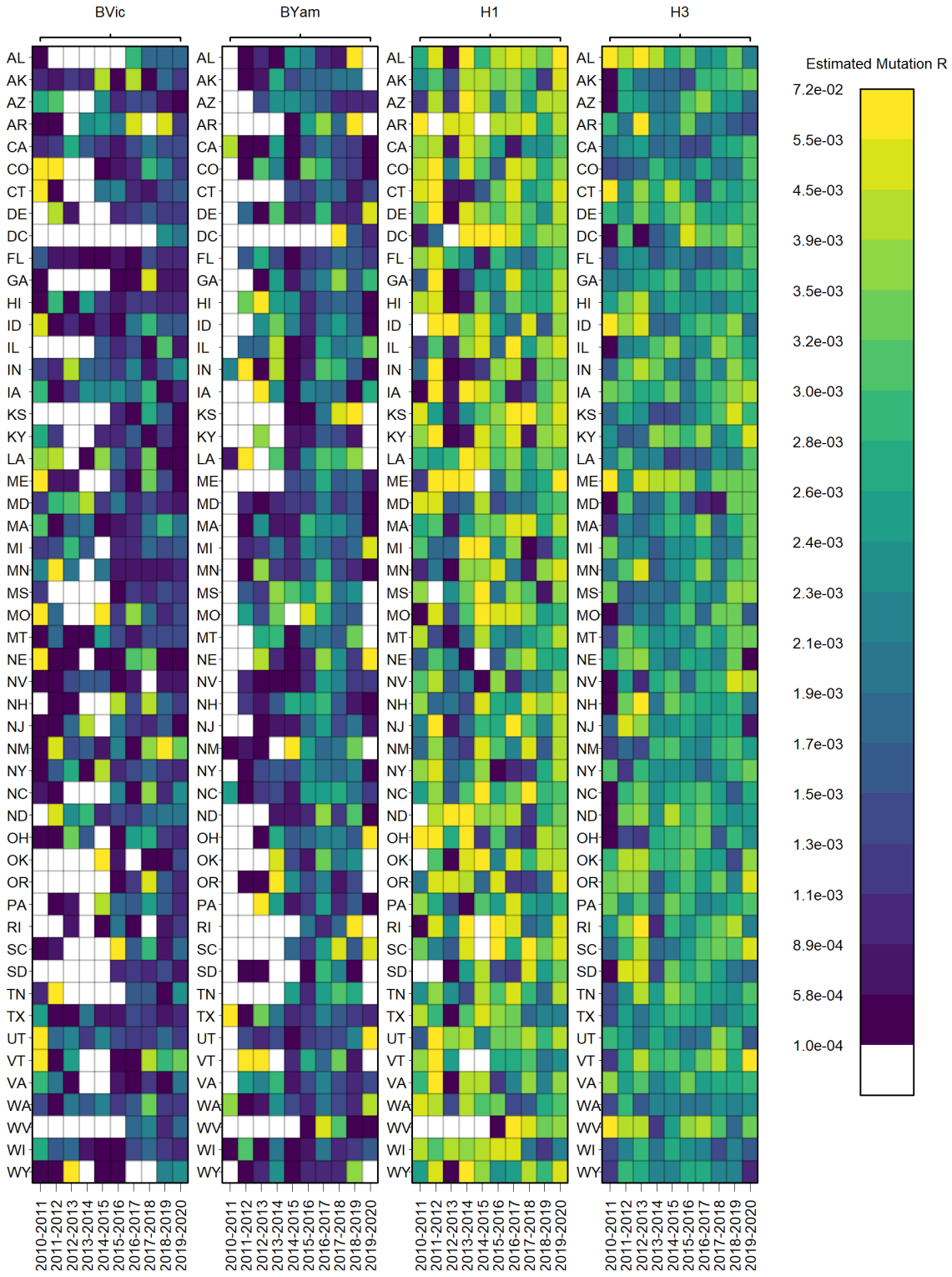
Supplementary Figure C.4. Coverage of Local Transmission Clusters

Supplementary Table C.2. Summaries of Local Transmission Clusters

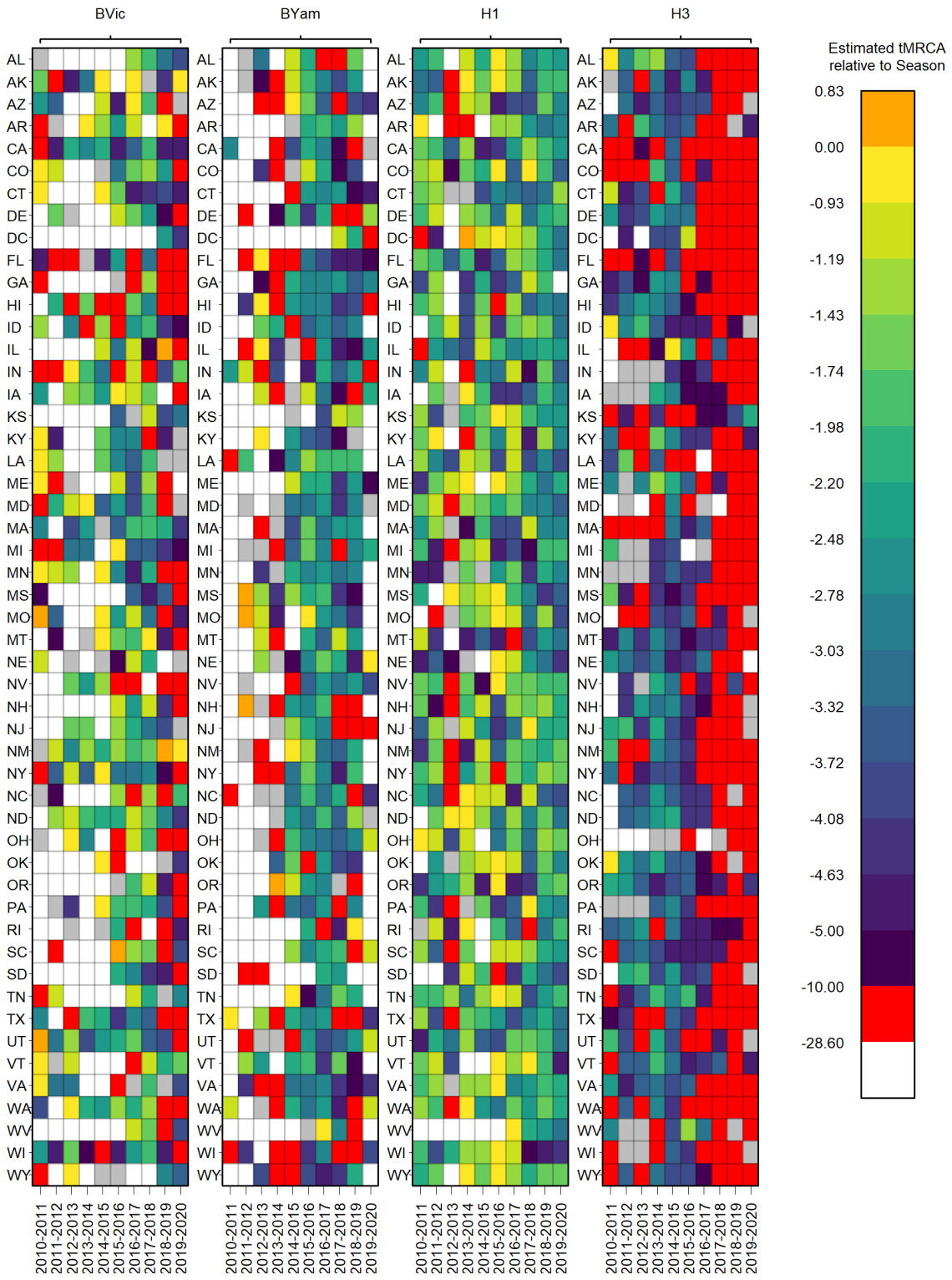
Variable	Levels	BVic	BYam	H1	H3	Statistic	Paramete	p	Test
		n = 831 (16.6)	n = 852 (17)	n = 1419 (28.4)	n = 1897 (37.9)				
ntips		3 [2, 6]	3 [2, 5]	3 [2, 7]	3 [2, 6]	27.863388	3	<0.00	Kruskal
	Missin	261 (31.4)	273 (32)	162 (11.4)	89 (4.7)				
mpd		7e-04 (0.00077)	0.00086 (0.00088)	0.00141 (0.00121)	0.0012 (0.00109)	74.223637	3, 4210	<0.00	AoV
imbalance.collessnor		0.74987 (0.30846)	0.79882 (0.29441)	0.72869 (0.29957)	0.73539 (0.31138)	4.342069	3, 2564	<0.00	AoV
		475 (57.2)	541 (63.5)	615 (43.3)	800 (42.2)				
avgladder		1.39375 (2.40215)	0.96765 (1.88494)	1.1379 (1.85129)	1.23945 (2.04366)	4.906241	3, 4210	<0.00	AoV
tmrca1		-0.17 [-	-0.13 [-	-0.19 [-	-0.22 [-	109.06572	3	<0.00	Kruskal



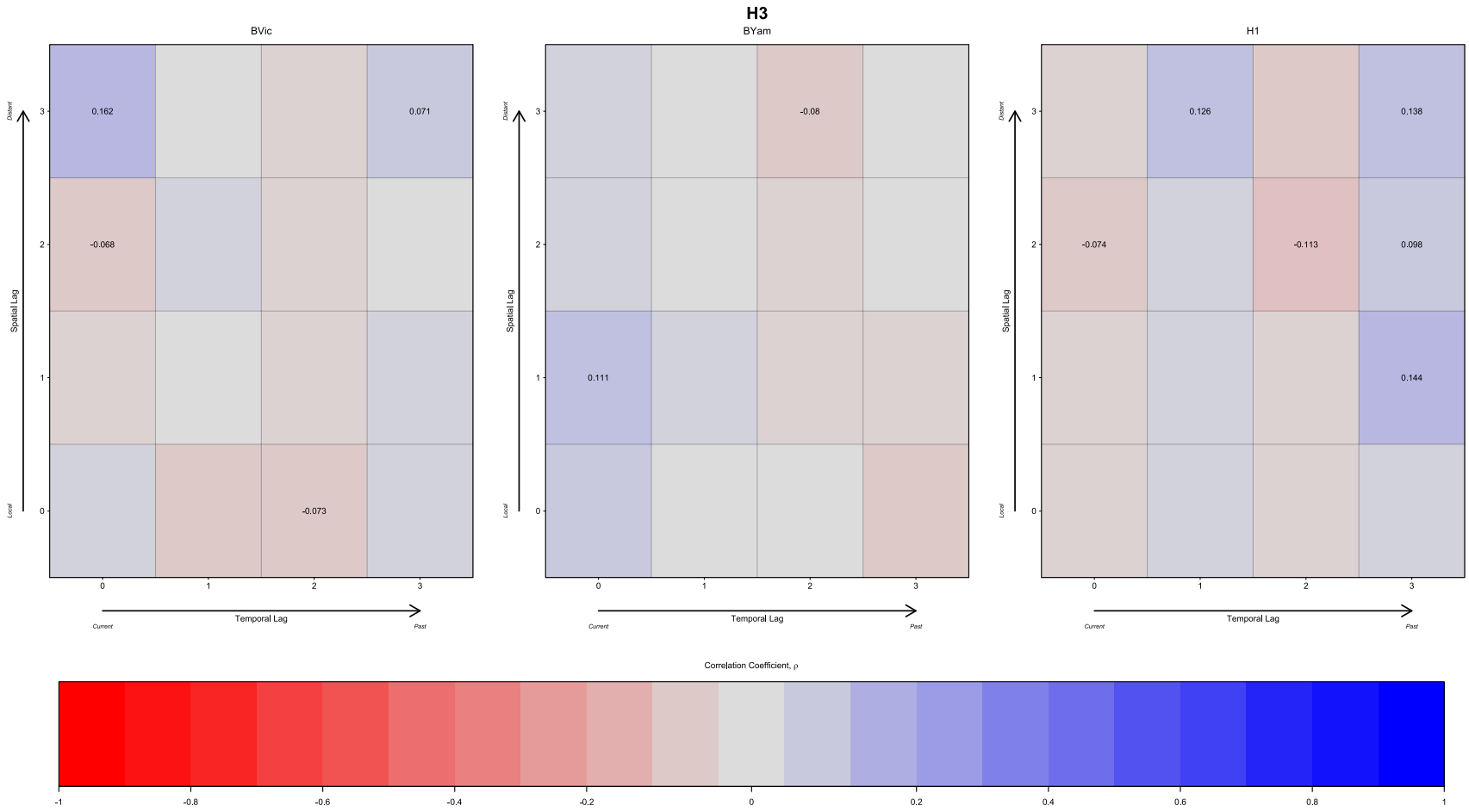
Supplementary Figure C.5. Time of the Most Recent Common Ancestor of Full Phylogenetic Trees



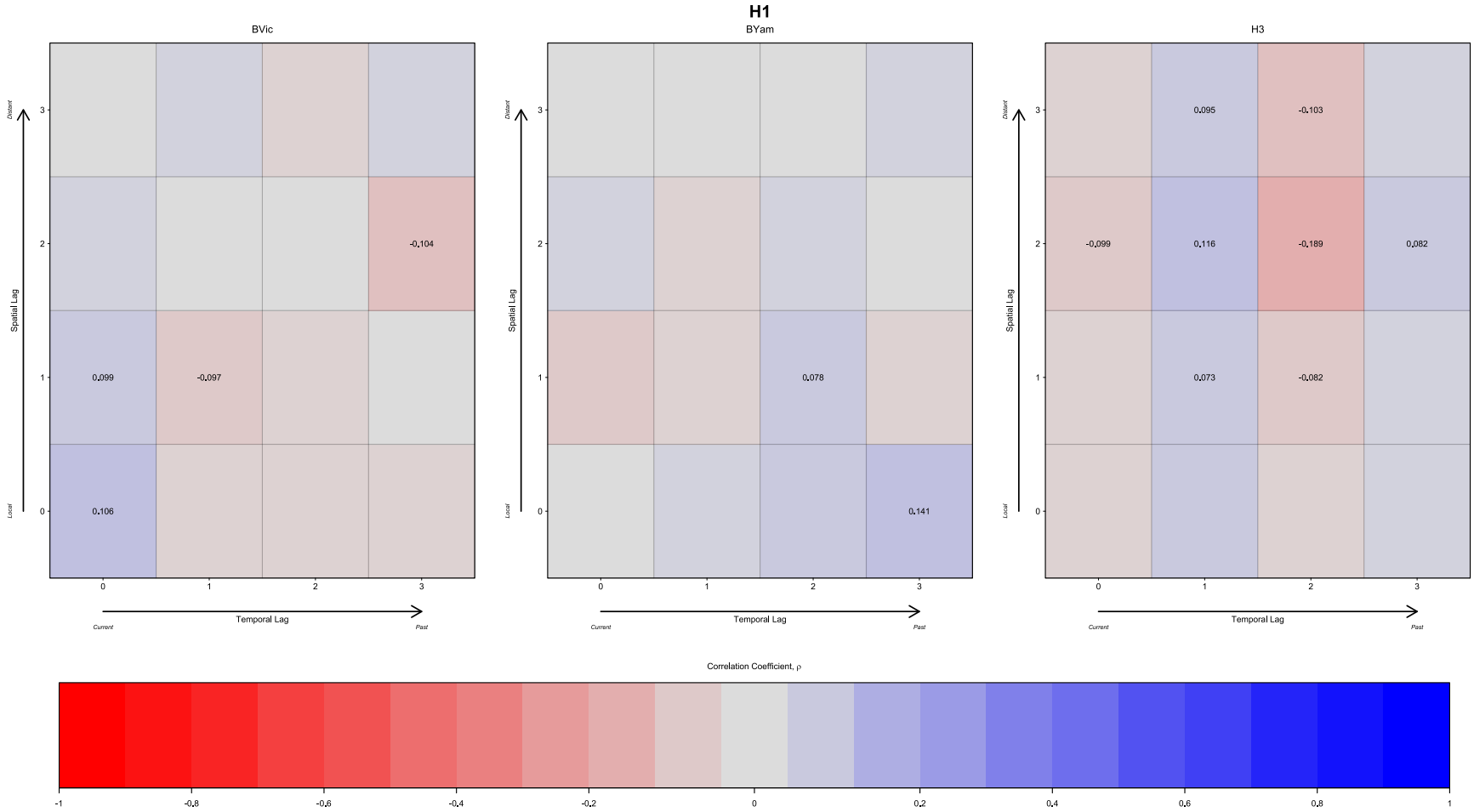
Supplementary Figure C.6. Estimated Mutation Rates from Full Phylogenetic Trees



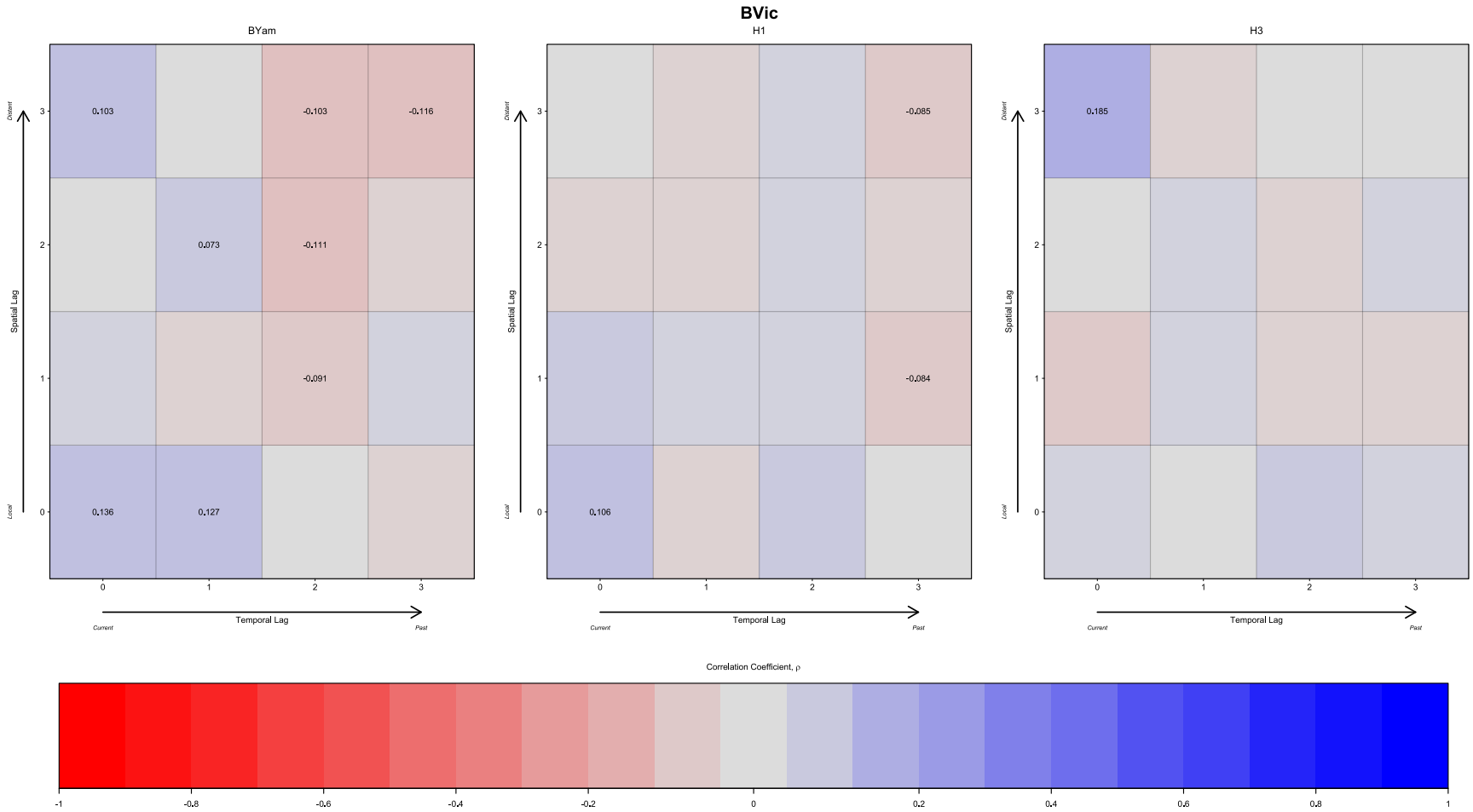
Supplementary Figure C.7. Time of the Most Recent Common Ancestor for Full Phylogenetic Trees



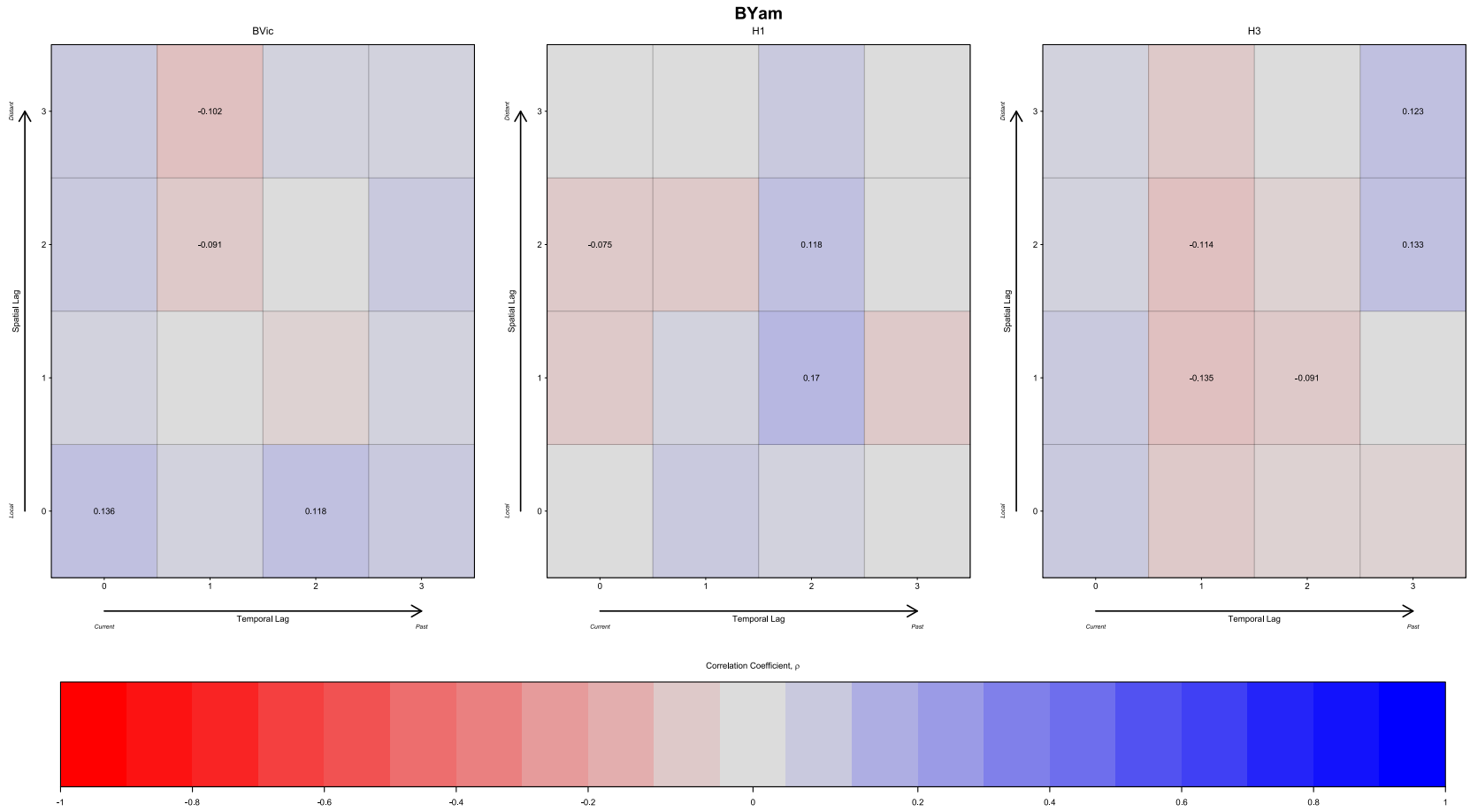
Supplementary Figure C.8. Cross-correlations of H3



Supplementary Figure C.9. Cross-correlations of H1



Supplementary Figure C.10. Cross-correlations of BVic



Supplementary Figure C.11. Cross-correlations of BYam

Supplementary Table C.3. Pearson Correlation Coefficients of Spatiotemporally Lagged Auto- and Cross-correlations of Seasonal Diversification of Influenza Subtypes / Lineages with BVictoria

Subtype	Spatial Lag	Spatial	BVic			
			Tlag = 0	1	2	3
BVic	Auto	0	1	-0.081 (-0.159, -	-0.033 (-0.112,	-0.021 (-0.113,
BVic	Network	1	0.057 (-0.006,	-0.017 (-0.066,	-0.025 (-0.085,	-0.074 (-0.132, -
BVic	Spatial	1	0.074 (-0.013,	0.036 (-0.02,	-0.046 (-0.128,	-0.031 (-0.128,
BVic	Spatial	2	0.153 (0.067,	-0.035 (-0.101,	-0.06 (-0.134,	-0.124 (-0.189, -
BVic	Spatial	3	0.097 (0.026,	0.018 (-0.038,	-0.061 (-0.123,	0.011 (-0.062,
BYam	Auto	0	0.136 (0.037,	0.036 (-0.048,	0.118 (0.04,	0.068 (-0.1,
BYam	Network	1	0.039 (-0.041,	-0.085 (-0.161, -	-0.062 (-0.162,	0.097 (0.023,
BYam	Spatial	1	0.047 (-0.04,	-0.004 (-0.098,	-0.036 (-0.136,	0.025 (-0.064,
BYam	Spatial	2	0.067 (-0.01,	-0.091 (-0.159, -	0.014 (-0.081,	0.094 (-0.019,
BYam	Spatial	3	0.076 (-0.005,	-0.102 (-0.161, -	0.029 (-0.05,	0.036 (-0.033,
H1	Auto	0	0.106 (0.012,	-0.031 (-0.099,	-0.033 (-0.118,	-0.04 (-0.127,
H1	Network	1	0.063 (-0.005,	-0.046 (-0.11,	-0.028 (-0.085,	-0.016 (-0.101,
H1	Spatial	1	0.099 (0.019,	-0.097 (-0.163, -	-0.049 (-0.116,	0.003 (-0.078,
H1	Spatial	2	0.024 (-0.044,	0.017 (-0.057,	0.013 (-0.055,	-0.104 (-0.173, -
H1	Spatial	3	0.01 (-0.059,	0.042 (-0.033,	-0.041 (-0.106,	0.053 (-0.02,
H3	Auto	0	0.056 (-0.03,	-0.065 (-0.145,	-0.073 (-0.143, -	0.041 (-0.025,
H3	Network	1	0.034 (-0.022,	0.003 (-0.056,	-0.024 (-0.088,	0.058 (-0.015,
H3	Spatial	1	-0.031 (-0.1,	-0.011 (-0.078,	-0.044 (-0.104,	0.05 (-0.035,
H3	Spatial	2	-0.068 (-0.127, -	0.057 (-0.017,	-0.06 (-0.124,	0.011 (-0.042,
H3	Spatial	3	0.162 (0.099,	0.019 (-0.052,	-0.053 (-0.119,	0.071 (0.008,

Supplementary Table C.4. Pearson Correlation Coefficients of Spatiotemporally Lagged Auto- and Cross-correlations of Seasonal Diversification of Influenza Subtypes / Lineages with BYamagata

Subtype	Spatial Lag	Spatial	BYam			
			0	1	2	3
BVic	Auto	0	0.136 (0.037,	0.127 (0.044,	-0.007 (-0.091,	-0.041 (-0.15,
BVic	Network	1	0.065 (-0.007,	-0.038 (-0.083,	-0.068 (-0.118, -	-0.005 (-0.061,
BVic	Spatial	1	0.034 (-0.052,	-0.043 (-0.099,	-0.091 (-0.149, -	0.021 (-0.061,
BVic	Spatial	2	0.018 (-0.042,	0.073 (0.003,	-0.111 (-0.159, -	-0.055 (-0.128,
BVic	Spatial	3	0.103 (0.02,	0.015 (-0.045,	-0.103 (-0.157, -	-0.116 (-0.193, -
BYam	Auto	0	1	0.005 (-0.1,	-0.189 (-0.306, -	0.073 (-0.07,
BYam	Network	1	0.356 (0.27,	-0.039 (-0.092,	-0.249 (-0.326, -	-0.034 (-0.113,
BYam	Spatial	1	0.369 (0.287,	-0.021 (-0.075,	-0.246 (-0.322, -	-0.095 (-0.184, -
BYam	Spatial	2	0.238 (0.158,	0.036 (-0.024,	-0.201 (-0.296, -	0.027 (-0.046,
BYam	Spatial	3	0.258 (0.175,	-0.008 (-0.048,	-0.184 (-0.289, -	0.088 (-0.004,
H1	Auto	0	-0.015 (-0.117,	0.052 (-0.034,	0.071 (-0.035,	0.141 (0.05,
H1	Network	1	-0.051 (-0.11,	-0.007 (-0.093,	0.011 (-0.058,	-0.022 (-0.076,
H1	Spatial	1	-0.07 (-0.148,	-0.033 (-0.101,	0.078 (0.01,	-0.028 (-0.082,
H1	Spatial	2	0.054 (-0.015,	-0.053 (-0.12,	0.04 (-0.028,	-0.016 (-0.089,
H1	Spatial	3	-0.01 (-0.082,	0.003 (-0.063,	0.014 (-0.056,	0.028 (-0.037,
H3	Auto	0	0.102 (-0.012,	0.002 (-0.079,	-0.013 (-0.128,	-0.081 (-0.184,
H3	Network	1	0.142 (0.049,	0.02 (-0.039,	-0.049 (-0.115,	-0.007 (-0.062,
H3	Spatial	1	0.111 (0.035,	0.036 (-0.034,	-0.035 (-0.113,	-0.021 (-0.087,
H3	Spatial	2	0.043 (-0.024,	0.011 (-0.051,	-0.048 (-0.122,	-0.008 (-0.073,
H3	Spatial	3	0.025 (-0.045,	0.008 (-0.074,	-0.08 (-0.141, -	0.002 (-0.065,

Supplementary Table C.5. Pearson Correlation Coefficients of Spatiotemporally Lagged Auto- and Cross-correlations of Seasonal Diversification of Influenza Subtypes / Lineages with H1

Subtype	Spatial Lag	Spatial	H1			
			0	1	2	3
BVic	Auto	0	0.106 (0.012,	-0.057 (-0.15,	0.069 (-0.015,	0.012 (-0.075,
BVic	Network	1	0.002 (-0.069,	0.009 (-0.063,	0.029 (-0.028,	-0.06 (-0.115, -
BVic	Spatial	1	0.08 (-0.02,	0.022 (-0.056,	0.024 (-0.043,	-0.084 (-0.159, -
BVic	Spatial	2	-0.027 (-0.087,	-0.026 (-0.086,	0.055 (-0.018,	-0.058 (-0.111,
BVic	Spatial	3	0.006 (-0.06,	-0.03 (-0.095,	0.049 (-0.019,	-0.085 (-0.143, -
BYam	Auto	0	-0.015 (-0.117,	0.078 (-0.01,	0.05 (-0.079,	-0.014 (-0.119,
BYam	Network	1	-0.14 (-0.225, -	0.05 (-0.017,	0.115 (0.053,	-0.027 (-0.109,
BYam	Spatial	1	-0.082 (-0.165,	0.047 (-0.02,	0.17 (0.104,	-0.066 (-0.146,
BYam	Spatial	2	-0.075 (-0.156, -	-0.074 (-0.164,	0.118 (0.046,	-0.015 (-0.108,
BYam	Spatial	3	-0.017 (-0.096,	0.006 (-0.071,	0.076 (0, 0.148)	-0.012 (-0.087,
H1	Auto	0	1	-0.086 (-0.172,	0.093 (0.004,	-0.005 (-0.098,
H1	Network	1	0.203 (0.108,	-0.066 (-0.129,	0.09 (0.014,	-0.173 (-0.25, -
H1	Spatial	1	0.203 (0.113,	-0.029 (-0.1,	0.112 (0.036,	-0.151 (-0.241, -
H1	Spatial	2	0.136 (0.048,	-0.185 (-0.251, -	0.133 (0.06,	-0.173 (-0.254, -
H1	Spatial	3	0.148 (0.068,	-0.206 (-0.275, -	0.157 (0.08,	-0.19 (-0.27, -
H3	Auto	0	-0.043 (-0.13,	0.043 (-0.04,	-0.041 (-0.122,	0.022 (-0.053,
H3	Network	1	-0.086 (-0.154, -	0.056 (-0.031,	-0.045 (-0.12,	0.105 (0.043,
H3	Spatial	1	-0.045 (-0.113,	0.024 (-0.055,	-0.043 (-0.112,	0.144 (0.082,
H3	Spatial	2	-0.074 (-0.156, -	0.047 (-0.025,	-0.113 (-0.185, -	0.098 (0.023,
H3	Spatial	3	-0.032 (-0.114,	0.126 (0.055,	-0.064 (-0.127,	0.138 (0.072,

Supplementary Table C.6. Pearson Correlation Coefficients of Spatiotemporally Lagged Auto- and Cross-correlations of Seasonal Diversification of Influenza Subtypes / Lineages with H3

Subtype	Spatial Lag	Spatial	H3			
			0	1	2	3
BVic	Auto	0	0.056 (-0.03,	-0.019 (-0.118,	0.061 (-0.034,	0.052 (-0.066,
BVic	Network	1	-0.01 (-0.095,	0.029 (-0.057,	-0.048 (-0.123,	-0.095 (-0.163, -
BVic	Spatial	1	-0.067 (-0.154,	0.048 (-0.047,	-0.059 (-0.146,	-0.045 (-0.12,
BVic	Spatial	2	0.02 (-0.054,	0.052 (-0.018,	-0.036 (-0.099,	0.041 (-0.036,
BVic	Spatial	3	0.185 (0.135,	-0.044 (-0.14,	0.003 (-0.059,	0.007 (-0.065,
BYam	Auto	0	0.102 (-0.012,	-0.066 (-0.171,	-0.036 (-0.142,	-0.026 (-0.124,
BYam	Network	1	0.093 (-0.013,	-0.133 (-0.244, -	-0.078 (-0.153, -	0.047 (-0.03,
BYam	Spatial	1	0.088 (-0.035,	-0.135 (-0.232, -	-0.091 (-0.164, -	0.006 (-0.072,
BYam	Spatial	2	0.05 (-0.063,	-0.114 (-0.193, -	-0.045 (-0.136,	0.133 (0.051,
BYam	Spatial	3	0.026 (-0.069,	-0.062 (-0.132,	-0.001 (-0.09,	0.123 (0.055,
H1	Auto	0	-0.043 (-0.13,	0.076 (-0.011,	-0.048 (-0.135,	0.046 (-0.032,
H1	Network	1	-0.082 (-0.162,	0.07 (-0.009,	-0.091 (-0.166, -	0.047 (-0.034,
H1	Spatial	1	-0.053 (-0.12,	0.073 (0.008,	-0.082 (-0.161, -	0.039 (-0.035,
H1	Spatial	2	-0.099 (-0.185, -	0.116 (0.052,	-0.189 (-0.254, -	0.082 (0.014,
H1	Spatial	3	-0.028 (-0.107,	0.095 (0.042,	-0.103 (-0.165, -	0.023 (-0.031,
H3	Auto	0	1	-0.023 (-0.094,	0.02 (-0.056,	-0.046 (-0.123,
H3	Network	1	0.13 (0.042,	-0.062 (-0.139,	0.048 (-0.036,	-0.074 (-0.141, -
H3	Spatial	1	0.141 (0.059,	-0.087 (-0.155, -	0.042 (-0.041,	-0.082 (-0.146, -
H3	Spatial	2	0.028 (-0.056,	-0.113 (-0.173, -	0.05 (-0.018,	-0.001 (-0.063,
H3	Spatial	3	0.126 (0.07,	-0.101 (-0.147, -	0.035 (-0.028,	-0.01 (-0.055,