Disentangling the rhythms of human activity in the built environment for airborne transmission risk: an analysis of large-scale mobility data

Zachary Susswein[1,], Eva C. Rest[1], and Shweta Bansal[1,*]

[1]Department of Biology, Georgetown University, Washington, DC, USA
[*]Corresponding Author: shweta.bansal@georgetown.edu

December 2022

Abstract

Background Since the outset of the COVID-19 pandemic, substantial public attention has focused on the role of seasonality in impacting transmission. Misconceptions have relied on seasonal mediation of respiratory diseases driven solely by environmental variables. However, seasonality is expected to be driven by host social behavior, particularly in highly susceptible populations. A key gap in understanding the role of social behavior in respiratory disease seasonality is our incomplete understanding of the seasonality of indoor human activity.

Methods We leverage a novel data stream on human mobility to characterize activity in indoor versus outdoor environments in the United States. We use an observational mobile app-based location dataset encompassing over 5 million locations nationally. We classify locations as primarily indoor (e.g. stores, offices) or outdoor (e.g. playgrounds, farmers markets), disentangling location-specific visits into indoor and outdoor, to arrive at a fine-scale measure of indoor to outdoor human activity across time and space.

Results We find the proportion of indoor to outdoor activity during a baseline year is seasonal, peaking in winter months. The measure displays a latitudinal gradient with stronger seasonality at northern latitudes and an additional summer peak in southern latitudes. We statistically fit this baseline indoor-outdoor activity measure to inform the incorporation of this complex empirical pattern into infectious disease dynamic models. However, we find that the disruption of the COVID-19 pandemic caused these patterns to shift significantly from baseline, and the empirical patterns are necessary to predict spatiotemporal heterogeneity in disease dynamics.

Conclusions Our work empirically characterizes, for the first time, the seasonality of human social behavior at a large scale with high spatiotemporal resolution, and provides a parsimonious parameterization of seasonal behavior that can be included in infectious disease dynamics models. We provide critical evidence and methods necessary to inform the public health of seasonal and pandemic respiratory pathogens and improve our understanding of the relationship between the physical environment and infection risk in the context of global change.

Funding Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R01GM123007.

1 Introduction

The seasonality of infectious diseases is a widespread and familiar phenomenon. Although a number of potential mechanisms driving seasonality in directly transmitted infectious diseases have been proposed, the causal process behind seasonality is still largely an open question [1, 2, 3]. In the case of the influenza virus, seasonal changes in humidity have been identified as a potential mechanism, with drier winter months enhancing transmission [4, 5, 6]; similar patterns have been observed for respiratory syncytial virus and hand

foot and mouth disease [7, 8]. However, humidity is but one of many mechanisms contributing to seasonality in infectious disease transmission. Seasonal changes in temperature, human mixing patterns, and the immune landscape, among other factors, are thought to contribute to transmission dynamics [9, 10, 11, 12, 2]. The relative importance of these disparate mechanisms varies across directly-transmitted pathogens and is still largely unexplained [1, 3]. The influence of seasonal host behavior on respiratory disease seasonality remains particularly understudied [13, 11] except for a few notable examples [14, 15, 16].

For respiratory pathogens spread via the aerosol transmission route, in particular, seasonality may be mediated by multiple behaviorally-driven mechanisms. Aerosol transmission, a significant mode of transmission for a number of respiratory pathogens including tuberculosis, measles, and influenza [17], has become increasingly acknowledged during the COVID-19 pandemic [18, 19, 20, 21, 22]. The role of aerosols in respiratory disease transmission allows for transmission outside of the traditional 6 ft. radius and 5-minute duration for the droplet mode and implicates human mixing in indoor locations with poor ventilation as being a high risk for transmission, regardless of the intensity of the social contact. While more is known about the spatiotemporal variation in environmental factors such as temperature and humidity in the indoor environment (e.g. [23]) and about the impact these factors have on airborne pathogen transmission (e.g. [24, 25]), limited information is available on rates of human indoor activity and how this varies geographically and seasonally. In the US, most studies quantifying indoor and outdoor time are conducted in the context of air pollutants, suffer from small study sizes, lack spatiotemporal resolution, and are outdated. The most cited estimates originate from the 1980s-90s and estimate that Americans spend upwards of 90% of their time indoors [26]; and more recent data agree with these estimates [27, 28]. While it is well understood that seasonal differences and latitude likely affect time spent indoors, little is known of the spatiotemporal variation in indoor activity beyond this one monolithic estimate, vastly limiting our ability to comprehensively characterize the seasonality of airborne disease exposure risk.

Because our understanding of the drivers of seasonality for respiratory diseases has been limited, the modeling of seasonally-varying infectious disease dynamics has been traditionally done using environmental data-driven or phenomenological approaches. Environmental data-driven approaches incorporate seasonality into epidemiological models through environmental correlates of seasonality, such as solar exposure or outdoor temperature [12, 7, 29]. This approach to seasonal dynamics controls for inter-seasonal variation in transmission dynamics and measures the strength of correlations between proposed metrics and seasonal variation in force of infection – although the observed relationship is rarely causally relevant for respiratory disease transmission. In contrast, phenomenological models such as seasonal forcing approaches modulate transmissibility over time without specifying a particular mechanism for this modulation [30, 2]. By applying well-understood functions (such as sine functions), seasonal forcing allows for flexible specification and quantification of dynamics, such as periodicity or oscillation damping, and indirectly captures seasonal variation in non-environmental factors such as school mixing. A significant remaining gap in seasonal infectious disease modeling is thus the ability to empirically incorporate spatiotemporal variation in behavioral mechanisms driving seasonality of disease exposure and transmission.

Thus, despite the role of the indoor built environment in exposure to the airborne transmission route, seasonal variation in indoor human mixing has not yet been systemically characterized nor integrated into mathematical models of seasonal respiratory pathogens. To address this gap, we construct a novel metric quantifying the relative propensity for human mixing to be indoors at a fine spatiotemporal scale across the United States. We derive this metric using anonymized mobile GPS panel data of visits of over 45 million mobile devices to approximately 5 million public locations across the United States. We find a systematic latitudinal gradient, with indoor activity patterns in the northern and southern United States following distinct temporal trends at baseline. However, we find that the COVID-19 pandemic disrupted this structure. Lastly, we fit simple parametric models to incorporate these seasonal activity dynamics into models of infectious disease transmission when indoor activity is expected to be at baseline. Our work provides the evidence and methods necessary to inform the epidemiology of seasonal and pandemic respiratory pathogens and improve our understanding of the relationship between the physical environment and infection risk in light of global change.

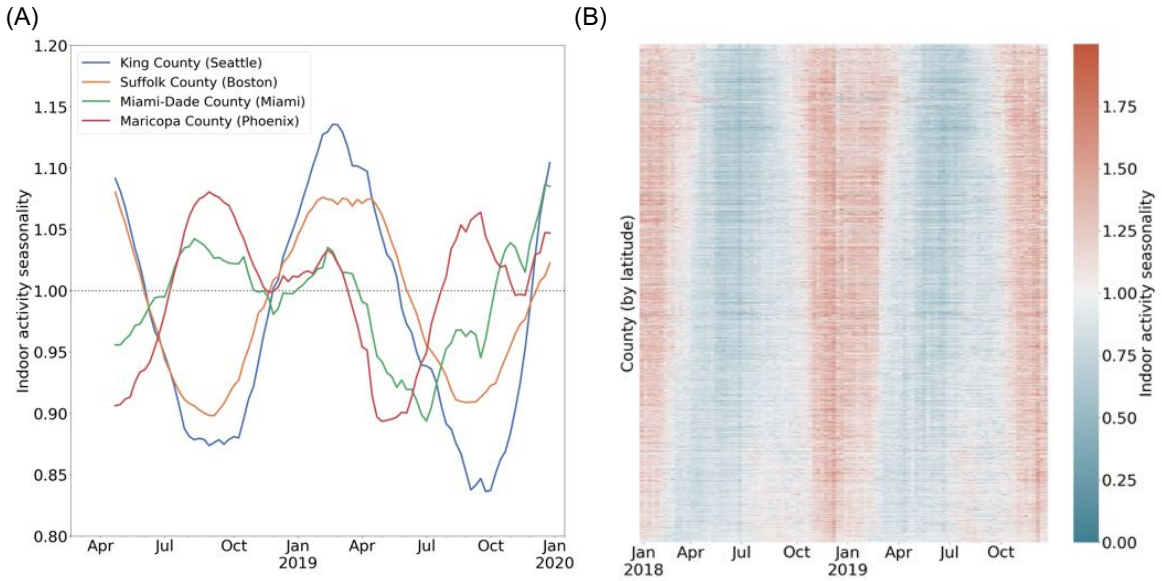


Figure 1: (A) Case studies to highlight varying trends in indoor activity seasonality during 2018 and 2019: King County and Suffolk County (in the northern US) have high indoor activity in the winter months and a trough in indoor activity in the summer months. Miami-Dade and Maricopa County (in the southern US) see moderate indoor activity in the winter and may have an additional peak in indoor activity during the summer. We apply a rolling window mean for visualization purposes. (B) A heatmap of the indoor activity seasonality metric for all US counties by week for 2018 and 2019. Counties are ordered by latitude. We see significant spatiotemporal heterogeneity with distinct trends in the summer versus winter seasons.

Methods

Data Source

We use the SafeGraph Weekly Patterns data, which provides foot traffic at public locations ("points of interest", referred to as POIs from here on) across the US based on the usage of mobile apps with GPS [31]. The data are from 2018 to 2020, and 4.6 million POIs are sampled in all years of our study. The data is anonymized by applying noise, omitting data associated with a single mobile device, and is provided at the weekly temporal scale. Data are sampled from over 45 million smartphone devices (of approximately 275-290 million smartphone devices in the US during 2018-2021 [32]), and does not include devices that are out of service, powered off, or ones that opt out of location services on their devices.

This is secondary data analysis, so no informed consent or consent to publish was necessary. Ethical review for this study (STUDY00003041) was sought from the Institutional Review Board at Georgetown University and was approved on October 14, 2020.

Defining indoor activity seasonality

Safegraph Points of Interest (POIs) are locations where consumers can spend money and/or time and include schools, hospitals, parks, grocery stores, and restaurants, etc, but do not include home locations. (In Figure 1—figure supplement 1, we show that time at home does not display significant seasonal variation). Each POI is assigned a six-digit North American Industry Classification System (NAICS) code in the SafeGraph Core Places dataset to classify each location into a business category. We classify each 6-digit NAICS codes (363 unique codes in total) as primarily *indoor* (e.g. schools, hospitals, grocery stores), primarily *outdoor* (e.g. parks, cemeteries, zoos). We classify some locations as *unclear* if the location is a potentially mixed indoor and outdoor setting (e.g. gas stations with convenience stores, automobile dealerships). Approximately 90% of POIs were classified as indoors, 6.5% were classified as outdoors, and 3.5% were classified as unclear. In Figure 1—figure supplement 2, we illustrate the robustness of our metric to the classification of unclear locations.

We define $\widetilde{\sigma}_{it}$, equation (1), as the propensity for visits to be to indoor locations relative to outdoor locations. We aggregated raw visit counts, defined when a device is present at a non-home POI for longer than one minute, to all indoor POIs and all outdoor POIs in a given week (t) at the U.S. county level (i). Visit counts are normalized by the maximum visit counts for indoor or outdoor locations in each county during the year 2019. (In Figure 1—figure supplement 3, we show that the max visit count is comparable in 2018 and 2019).

$$\widetilde{\sigma}_{it} = \frac{N_{it}^{indoor}/max_t\{N_{it}^{indoor}\}}{N_{it}^{outdoor}/max_t\{N_{it}^{outdoor}\}} \tag{1}$$

This metric is then mean-centered to arrive at a relative measure of indoor activity seasonality, σ_{it}, which is comparable across all counties:

$$\sigma_{it} = \frac{\widetilde{\sigma}_{it}}{\mu_{\tilde{\sigma}}} \tag{2}$$

We note that $\mu_{\tilde{\sigma}}$ is not spatially structured (see Figure 2—figure supplement 1).

As a data cleaning step, we use spatial imputation for any county-weeks where sample sizes are small. For location-weeks in which the total visit count is less than 100, we impute the indoor activity seasonality using an average of σ in the neighboring locations (where neighbors are defined based on shared county borders). This affects 0.6% of all county-weeks and a total of 79 (out of 3143) counties.

Time series clustering analysis

To characterize groups of US counties with similar indoor activity dynamics, we use a complex networks-based time series clustering approach. We first calculate the pairwise similarity between z-normalized indoor activity time series for each pair of counties, i and j using the Pearson correlation coefficient (ρ_{ij}). For pairs

of locations where ρ_{ij} is in the top 10% of all correlations, we represent the pairwise time series similarities as a weighted network where nodes are US counties and edges represent strong time series similarity. (In Figure 1—figure supplement 4, we show the robustness of our clustering results to this choice of correlation threshold.)

We then cluster the time series similarity network using community structure detection. This method effectively clusters nodes (counties) into groups of nodes that are more connected within than between. The resulting clustering thus represents a regionalization of the U.S. in which regions consist of counties that have more similar indoor activity dynamics to each other than to other regions. One benefit of the network-based community detection approach over other clustering methods is that community detection does not require user specification of the number of clusters (regions, in this case); instead the number of clusters emerge organically from the data connectivity [33]. For community detection, we use the Louvain method [34], a multiscale method in which modularity is first optimized using a greedy local algorithm, on the similarity network with edge weights (i.e. time series correlations) using a `igraph` implementation in *Python* [35].

We performed a robustness assessment of the community structure using a set of 25 "bootstrap networks", B_i. For each bootstrap network, the edge weight (i.e. the time series correlation) for each edge of the network was perturbed by $\epsilon\ N(0, 0.05)$. The community structure algorithm was performed on each bootstrap network. A consensus value was then calculated as the sum of the normalized mutual information between the community structure partition of bootstrap network B_i and all other bootstrap networks. The partition with the largest consensus value was defined as the robust community structure partition.

Given some known limitations to the time series correlation network-based approach to clustering [36], we validated our network-based clustering results with another common clustering method. In particular, we used hierarchical clustering with Ward linkage and Euclidean distance on z-normalized indoor activity time series, implemented using `scipy` in *Python*. (We note that Euclidean distance is equivalent to Pearson's correlation on normalized time series [37]). The results of this comparison are summarized in Figure 1—figure supplement 5.

Disruptions to indoor activity due to pandemic response

We investigate the COVID-19 pandemic's impact on indoor activity seasonality by comparing pre-pandemic mobility patterns in 2018 and 2019 with mobility patterns during the COVID-19 pandemic in 2020. We compared the proportion of indoor visits at the county level, σ_{it}, across 2018, 2019, and 2020 to examine changes in indoor activity seasonality during the COVID-19 pandemic. We also examined total activity, aggregating visits to all indoor, outdoor, and unclear POIs by week and mean-centering them for each US county during the COVID-19 pandemic in 2020.

Incorporating indoor activity into infectious disease models

We seek to illustrate the impact of incorporating seasonality into an infectious disease model using a phenomenological model versus empirical data. To achieve this, we parameterize a simple compartmental disease model with a seasonality term, using either our empirically-derived indoor activity seasonality metric or an analytical phenomenological model of seasonality fit to this metric.

Phenomenological model of seasonality

We first fit our empirically-derived indoor activity seasonality metric using a time-varying non-linear model. We specify the time-varying effect as a sinusoidal function as is commonly done to incorporate seasonality into infectious disease models phenomenologically. The indoor activity seasonality, σ_{it} for cluster i at week t is specified as: $\sigma_{it} = 1 + \alpha_i \sin(\omega_i t + \phi_i)$, where α_i is the sine wave amplitude, ω_i is the frequency and ϕ_i is the phase. We fit a model for locations in the northern cluster separately from those in the southern cluster, as identified above. We fit the parameters for this model using the `nlme`, a standard package in `R` for fitting Gaussian nonlinear models.

Disease model

We model infectious disease dynamics through a simple SIR model of disease spread:

$$\frac{dS}{dt} = -\beta_0 \beta(t) SI$$
$$\frac{dI}{dt} = \beta_0 \beta(t) SI - \gamma I$$
$$\frac{dR}{dt} = \gamma I$$

We incorporate alternative seasonality terms to consider the impact of heterogeneity in indoor seasonality on disease dynamics. For the northern and southern cluster separately, we define modeled seasonality as $\beta(t) = 1 + \alpha \sin(\omega t + \phi)$, with the fitted parameters for each cluster (Figure 4—figure supplement 1 and Figure 4—figure supplement 2). We also consider two exemplar locations for empirical estimates of seasonality, where $\beta(t) = \sigma_t$ after rolling window smoothing: Cook County for an example county from the northern cluster, and Maricopa County for an example location from the southern cluster. We also compare against a null expectation where $\beta(t) = 1$. (All seasonality functions are illustrated in Figure 4—figure supplement 3). We assume that $\beta_0 = 0.0025$ and $\gamma = 2$ (on a weekly time scale).

Results

Based on anonymized location data from mobile devices, we construct a novel metric that measures the relative propensity for human activity to be indoors at a fine geographic (US county) and temporal (weekly) scale. Activity is measured as number of visits to unique physical, public (non-residential) locations across the United States. Locations are classified as indoors if they are enclosed environments (i.e. buildings and transportation services). We characterize the systematic spatiotemporal structure in this metric of indoor activity seasonality with a time series clustering analysis. We also characterize the shift that occurred in the baseline patterns of indoor activity seasonality during the COVID-19 pandemic. We note that this seasonal variation in the propensity of human activity to be indoors differs from the variation in overall rates of contact or mobility, which does not appear to be highly seasonal (Figure 1—figure supplement 1, [38]). Lastly, we fit non-linear models to the indoor activity metric at baseline, comparing the ability of a simple model to capture seasonal variation in transmission risk.

Quantifying empirical dynamics in indoor activity

The indoor activity seasonality metric, σ, captures the relative frequency of visits to indoor versus outdoor locations within an area. The components of σ capture the degree to which indoor and outdoor locations are occupied; when $\sigma = 1$, a given county is at its county-specific average propensity (over time) for indoor activity relative to outdoor. When $\sigma < 1$, activity within the county is more frequently outdoor and less frequently indoor than average, while $\sigma > 1$ indicates that activity is more frequently indoor and less frequently outdoor than average. Thus, a σ of 1.2 indicates that the county's activity is 20% more indoor than average and a σ of 0.80 indicates that the county's activity is 20% less indoor than average (additional details in methods).

Through this metric, we measure the relative propensity for human activity to be indoors for every community (i.e. US county) across time (at a weekly timescale), finding significant heterogeneity between counties (Figure 1A). The representative examples of Cook County, Illinois (home of the city of Chicago in the midwestern US) and Maricopa County, Arizona (home of the city of Phoenix in the southwestern US) highlight systematic spatial and temporal heterogeneity in indoor activity dynamics. In Cook County, indoor activity varies over time, at its peak in the winter, with the relative odds of an indoor visit well above average. During the summer, σ in Cook County reaches its trough, with activity systematically more outdoors on average. On the other hand, the variation of σ across time in Maricopa County is characterized by a smaller winter peak in indoor activity, and an additional peak in the summer (i.e. July and August); this peak occurs concurrently with the trough in Cook County. Unlike in Cook County, σ in Maricopa County is lowest in the spring and fall. These representative counties illustrate the systematic within-county variation in indoor activity over time, as well as the between-county variation in temporal trends as represented in Figure 1B for all US communities.

To identify systematic geographic structure, we cluster the heterogeneous time series of county-level, weekly indoor activity. We find three geographic clusters corresponding to groups of locations that experience similar indoor activity dynamics (Figure 2). These clusters primarily split the country into two clusters: a northern cluster and a southern cluster. Among the communities in the northern cluster, activity is more commonly outdoor over the summer months, trending toward indoor during fall, with a peak in the winter months, as observed in Cook County. Comparatively, the southern cluster has a larger winter peak (i.e. between December and February) and a smaller summer peak (i.e. between July and August); most summer peaks are less extreme than that of Maricopa County (shown). We hypothesize that these two clusters are consistent with climate zones. While there is a moderate association between indoor activity seasonality and environmental variables such as temperature and humidity (Figure 2—figure supplement 2), we expect that the northern and southern indoor activity clusters will be more consistent with climate zones defined for the construction of the indoor built environment and find that there is indeed substantial consistency between the two (Figure 2—figure supplement 3). The third cluster differs substantially: it is geographically discontiguous and its two annual peaks occur during the spring (close to April) and fall (closer to November) seasons. Thus, the counties in this cluster have outdoor activity more frequently than average during both

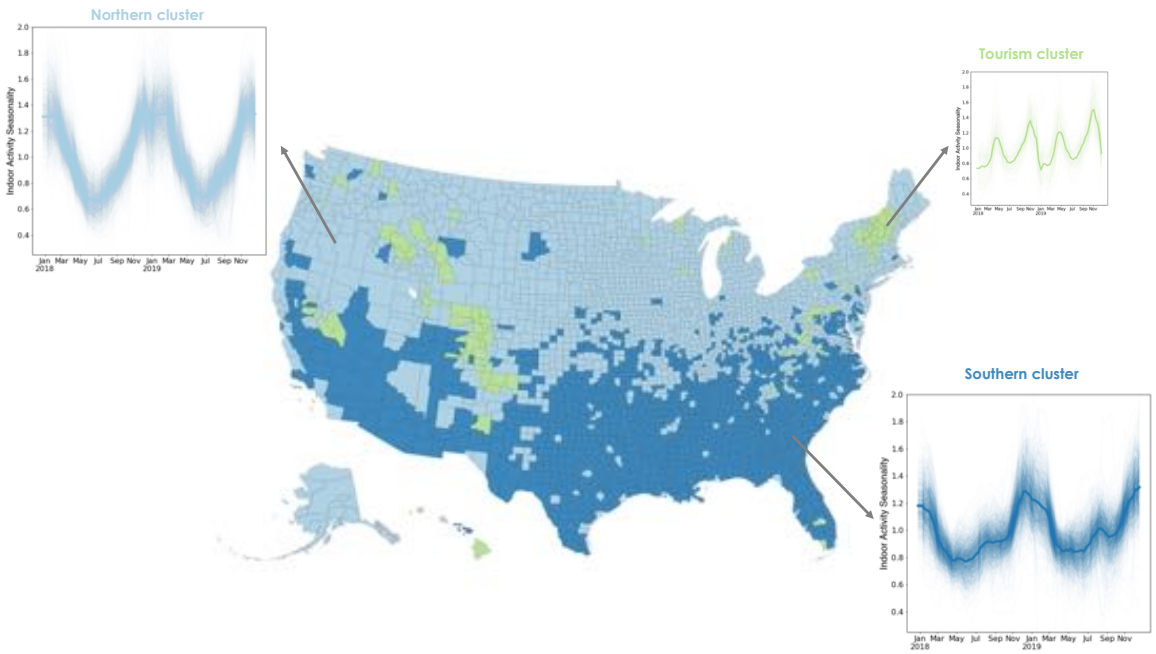


Figure 2: Using a time series clustering approach on the indoor activity time series for each US county, we identify groups of counties that experience similar trends in indoor activity. Locations in the northern cluster (light blue) follow a single peak pattern with the highest indoor activity occurring every winter. Locations in the southern cluster (dark blue) experience two peaks in indoor activity each year, one in the winter and a second, smaller one in the summer. The third cluster also experiences two peaks not matching environmental conditions, but potentially corresponding to winter or other tourism areas. We apply a rolling window mean to the time series for visualization purposes.

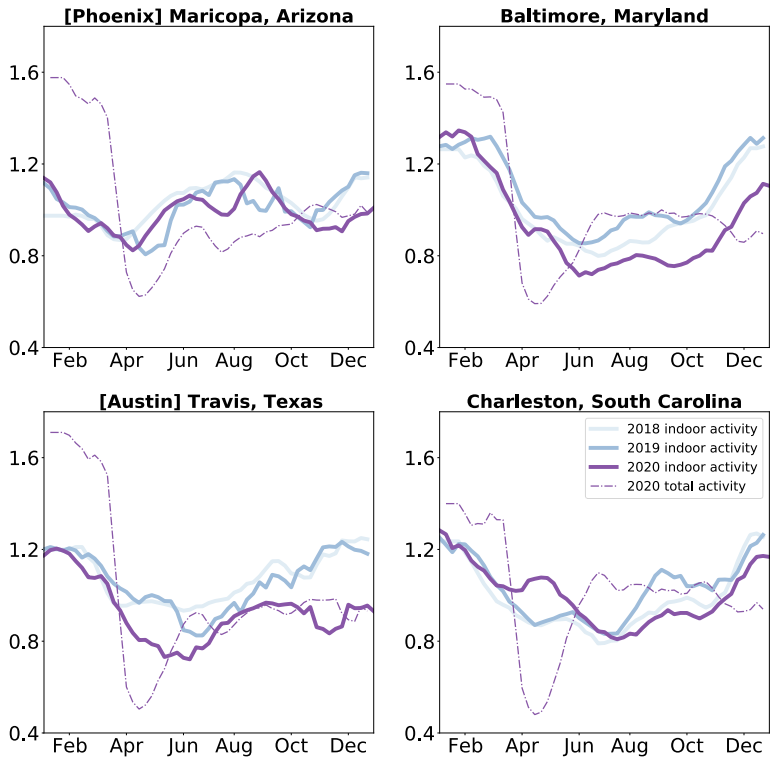


Figure 3: Indoor activity during the COVID-19 pandemic was shifted: We compare indoor activity trends in the baseline years of 2018 and 2019 to the pandemic year 2020 in four case study locations. We find that most locations saw a shift in their indoor activity patterns, while others (such as Maricopa County) did not. We also find that while overall activity was diminished uniformly during the Spring of 2020, indoor activity decreased in some locations (Travis County, Texas and Baltimore County, Maryland) and increased in others (Charleston County, South Carolina). We apply a 3-week rolling window mean to the time series for visualization purposes.

the winter and the summer. The counties in this cluster correspond to locations that are hubs for winter or other tourism, which we speculate is driving their unique dynamics (Figure 2—figure supplement 4).

Characterizing pandemic disruption to baseline indoor activity seasonality

In addition to the description of indoor activity seasonality at baseline, we examine the impact of a large-scale disruption – the COVID-19 pandemic – to these patterns. We compare indoor activity seasonality during the COVID-19 pandemic in 2020 to the baseline patterns of 2018 and 2019. We find that the temporal trends in indoor activity are less geographically structured in 2020 than those of previous years (see Figure 3–figure supplement 2 for a characterization of the time series patterns). We find that indoor activity deviated from pre-pandemic trends beyond interannual deviations (Figure 3—figure supplement 1). We focus on four case studies to highlight the varying impacts on indoor activity of the pandemic disruption (Figure 3). In all four communities, 2020 indoor activity trends shift from 2018 and 2019 patterns, with Maricopa County (home of the city of Phoenix, AZ) showing the least perturbation relative to prior years. We also find that in early

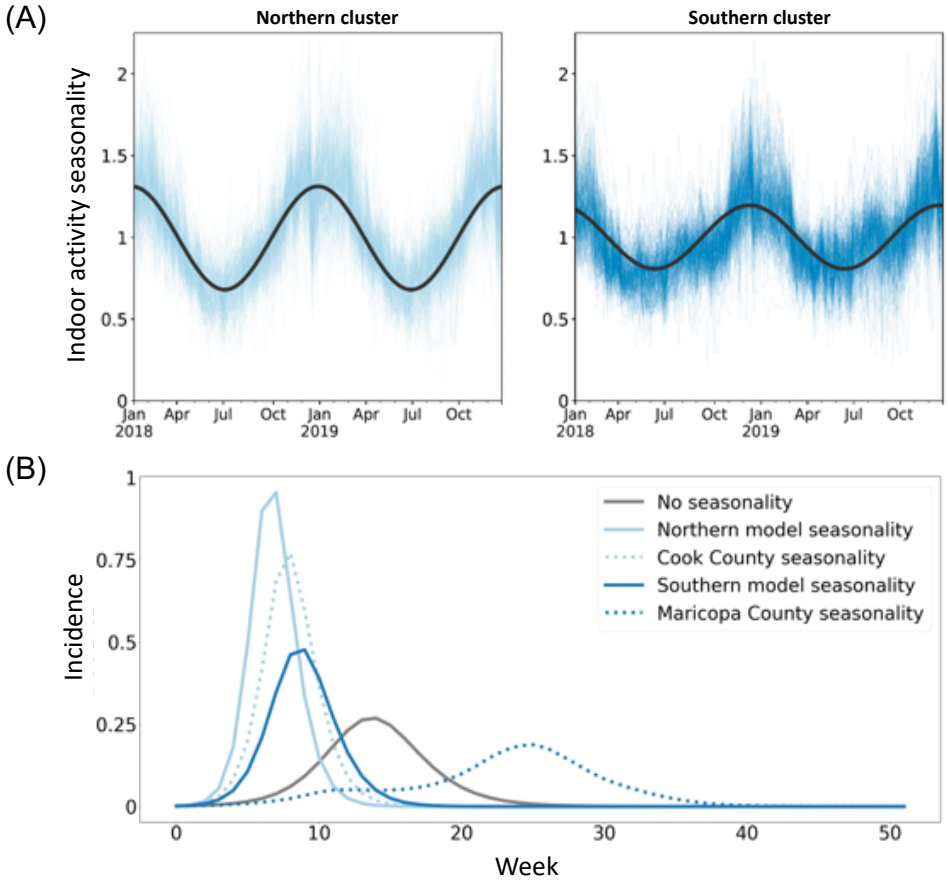


Figure 4: (A) Sine curves fit to the 2018 and 2019 time series data (analogous to seasonal forcing model components) fit the northern cluster better than the southern cluster, with a markedly poorer fit for the southern cluster's second summer peak. (B) Regional seasonal forcing models display variation in patterns of disease incidence omitted by a non-seasonal model, but even region-level seasonal forcing does not fully capture within-cluster county-level variation.

2020, when there was substantial social distancing in the United States (e.g. school closures, remote work), activity was more likely to be outdoor than in prior years, independent of changes in overall activity levels. With our case studies, we highlight that social distancing policies can have different impacts on airborne exposure risk in different locations: while some locations, such as Travis County (home of Austin, Texas), shifted activities outdoors during this period, reducing their overall risk further, other locations, such as Charleston County, South Carolina (home of Charleston, South Carolina) increased indoor activity above the seasonal average during this period, potentially diminishing the effect of reducing overall mobility. The trends in Charleston are representative of those in the southeastern United States during the spring of 2020 (Figure 3—figure supplement 1). By the end of 2020 (and the first winter wave of SARS-CoV2), many parts of the country were shifting activity more outdoors than seasonally expected (Figure 3—figure supplement 1).

Implications for modeling seasonal disease dynamics

We use this finely-grained spatiotemporal information on indoor activity to incorporate airborne exposure risk seasonality into compartmental models of disease dynamics using common, coarser seasonal forcing ap-

proaches. To investigate the impact of heterogeneity in σ on the estimation of seasonal forcing for infectious disease models, we fit a sinusoidal model to the time series of indoor activity for each of the primary clusters (Figure 4—figure supplement 4A). We note that because σ is defined as deviation from baseline indoor activity, the sinusoidal parameters (amplitude, frequency, phase) should be interpreted as a measure of seasonality in indoor activity, relative to each location's baseline. We find that the parameters of seasonality vary across clusters: the amplitude is higher, and the phase is lower in the northern cluster compared to the southern cluster, indicating a difference in the variability of indoor and outdoor activity seasonality in each cluster (Figure 4—figure supplement 1). While the fits are comparable for both clusters (Figure 4—figure supplement 2), the sinusoidal model does not capture the second peak of indoor activity during the summer months in the southern cluster. These differences in best fit indicate that sinusoidal models may have an overly restrictive functional form, limiting the accuracy of the approximation, and may underestimate the impacts of seasonality on transmission, obscuring systemic differences between regions. Furthermore, differences in seasonal activity of the observed magnitude can have important implications for disease modeling; applying region-level and county-level forcing to a simple disease model alters incidence patterns (igure 4—figure supplement 4B). Although region-level seasonality changes incidence timing and peak size relative to a non-seasonal model, it does not fully capture the changes produced by county-level seasonality. These differences indicate that while coarser geographic approximations of seasonality can be appropriate, these approximations can also oversimplify, reducing the accuracy of disease models. Additionally, while simple models of baseline indoor activity can capture seasonality in exposure risk, disruptions such as pandemics can alter this baseline structure and increase heterogeneity.

Discussion

The seasonality of influenza, SARS-CoV-2, and other respiratory pathogens depends not only on environmental variables but also on the social behavior of hosts. In settings with little prior immunity – such as a pandemic – host social behavior (generating contacts during which transmission may occur) primarily drives heterogeneity in disease dynamics, and seasonality is dwarfed by susceptibility [39]. In settings with higher rates of immunity, contact remains critically important, and seasonal changes in contacts (both direct and indirect) can contribute to the movement of R_t above and below 1 – providing noticeable changes in incidence. Although environmental variables play a role in the seasonality of respiratory pathogens, the role of host social behavior in pathogen seasonality is poorly understood, driven by a poor understanding of indoor versus outdoor social interactions and interactions between behavior and the environment. In this study, we propose a fine-grain measure of indoor activity seasonality across time and space. This metric is a relative quantity of behavior, comparable across locations, and thus intended to be a measure of seasonality beyond a baseline. We determine that indoor activity seasonality displays significant spatiotemporal heterogeneity and that this variability is highly geographically structured. We also find that while indoor activity seasonality may be highly predictable under baseline conditions, disruptions such as the COVID-19 pandemic can alter these patterns. Finally, we provide an illustration of how our findings can be incorporated into classical infectious disease models using parsimonious models of exposure seasonality.

The indoor activity seasonality that we quantify may reflect heterogeneity in transmission risk via a number of mechanisms including those affecting host contact, susceptibility, or transmissibility. Increased indoor activity may indicate longer-duration airborne contact (e.g., co-location without direct interaction) between susceptible and infected individuals, elevating respiratory transmission risk. Increased indoor density may also suggest increased droplet contact (e.g., a conversation in close proximity), under homogeneous mixing. Additionally, indoor activity may suggest increased susceptibility as poor ventilation, increased pollutants, reduced solar exposure, and low humidity of the indoor environment have been shown to weaken immune response [40]. Finally, increased indoor activity may indicate an increase in transmissibility due to higher exposure as low humidity caused by climate control (heating, ventilation, and cooling, HVAC) in indoor environments has been shown to increase viral survival and HVAC re-circulation has been shown to increase viral dispersion [41, 42]. While our new measure does not disentangle these component mechanisms, it represents an integrated seasonality in exposure risk due to all of these factors and can help lead us to a more complete understanding of the heterogeneity and seasonality in disease dynamics and outcomes.

We find that spatiotemporal heterogeneity in the indoor activity metric can be decomposed into two large geographically-contiguous groups in the northern and southern United States representing distinct temporal dynamics in indoor activity. These groups closely correspond to built environment climate zones, potentially explaining this systematic variability. We note, however, that while these clusters overlap with climate classifications, this correspondence does not suggest that environmental variables such as temperature and humidity should be used to represent behavioral heterogeneity. Climatic factors within these climate zones may be related to, but not necessarily correlated with, the seasonality of human mixing within these zones. Additionally, even in the case that environmental factor variability drives behavioral variability, it would be critical to capture the effect of behavior on disease directly so as to not obscure any direct effects of climatic factors on disease.

We illustrate how to incorporate seasonality in exposure risk to future models of disease dynamics using a simple phenomenological model. We use this traditional model of infectious disease dynamics to evaluate the implications of the spatial coarseness of seasonal forcing. Our results suggest that the substantial local heterogeneity in the dynamics of indoor activity across time and space could be large enough to alter seasonality in infectious disease dynamics. While our work does not consider observed transmission patterns, we suggest that researchers carefully consider the spatial scale on which they model seasonality in theoretical models, commonly used for scenario analysis and model-based intervention design (e.g., [43]). We additionally highlight that the use of simple or complex functional forms of seasonality requires statistical fits to baseline data, and in the case of disruptions, these fitted models may no longer be appropriate. Although indoor activity is moderately anticorrelated with temperature and humidity (Figure 1. Consequently, weather-derived covariates may have some statistical power to reflect impacts of human movement, but is not able to completely reflect this phenomenon. As we show, patterns of human mobility changed substantially during the COVID-19 pandemic, potentially contributing to changes in infectious disease seasonality.

Recent work during the COVID-19 pandemic demonstrates the impact of reduced occupancy in indoor locations and increasing outdoor activity on the likelihood of disease transmission. In particular, behavioral interventions or nudges that reduce occupancy are more impactful than reducing overall mobility as they reduce visitor density and the likelihood of density-dependent airborne transmission [44]. Similarly, the availability of outdoor areas in urban settings, such as public parks, has been demonstrated to reduce case rates when population mobility becomes less restricted [45]. Our results suggest that such public health strategies should be implemented in a targeted manner, informed by real-time data and with clear communication of the goals. We found notable changes occurred in indoor activity seasonality at the start of the COVID-19 pandemic, despite relatively consistent patterns during the spring season in prior years. Designing a behavioral strategy and measuring its effectiveness without real-time data could thus be misleading. Our finding of two distinct geographic clusters of indoor activity suggests the need for geographical targeting of strategies to reduce indoor transmission risk. While northern latitudes might benefit from decreased indoor occupancy and increased outdoor activity in Northern Hemisphere winters, southern latitudes should be additionally targeted for such interventions in the summer months. Lastly, our findings highlight the need to communicate the goals of behavioral interventions clearly. While all communities universally reduced overall activity during the early days of the COVID-19 pandemic, some increased indoor activity during this time, potentially diminishing the positive effects of the social distancing policies put into place. A public health education campaign to clarify the role of indoor interactions in transmission risk may have ameliorated this.

Our study leverages a novel data stream made available to researchers due to the COVID-19 pandemic. Similar datasets are available globally, part of a $12 billion location intelligence industry [46]. Such novel data streams offer many opportunities to address long-unanswered questions in infectious disease and climate change behavior dynamics, but these data must be interpreted carefully. Safegraph's mobile-app-based location data does not include data on individuals less than 16 years of age [47]. While we may expect that children under 12 may be accompanied by adults that may be represented in the dataset, our metric likely does not capture the activity dynamics of older children (children 12-15 make up 5% of the US population). For those included in the Safegraph database, representation is dependent on smartphone usage and a number of business processes not transparent to users of the data, thus we expect that there is geographic variation in the representativeness of the data. Smartphone ownership has increased in recent years, with 85% of US adults reporting smartphone ownership; however, smartphone usage does vary significantly by age, with only 61% of adults over 65 reporting smartphone use [48]. Additionally, data shows that location sharing among

mobile users is not significantly biased by age, gender, race/ethnicity, income, or education (with 40-65% of all demographic groups participating in location sharing) [49]. Based on an analysis done by Safegraph, the panel is representative of race, educational attainment, and income [50]. On the other hand, a recent independent analysis shows that older and non-white individuals are less likely to be captured in the panel for POI-specific analyses [51]. It is important to note that both studies are associative in nature as the devices in the panel are fully anonymized, so no device-level demographic data exists. Continued work to understand the sampling biases of such datasets will be needed so that improved bias correction approaches can be developed [51]. Additionally, we limit our scope in this study to consider only the number of visits and do not incorporate information about visit duration. The dataset counts all visits of one minute or longer. For disease transmission, there may be a threshold duration required for an interaction between an infected and susceptible individual for infection to be propagated. These thresholds are not well-understood for all respiratory diseases, but evidence that SARS-CoV-2 transmission can occur with brief encounters has emerged [52]. While the Safegraph dataset does provide median dwell times for POIs, the likely significant heterogeneity in the distribution of dwell times remains unknown and is difficult to capture in an aggregated manner.

Our metric and analysis also focus on the US county scale to reflect the finest scale generally used for infectious disease modeling as well as public health decision-making. This choice is likely to ignore some within-county heterogeneity and means that our metric does not represent the experience of all groups, particularly by socioeconomic status. For example, low-income and racially marginalized communities have systematically less access to outdoor, natural spaces and spend more time indoors due to structural inequities including lack of paid leave [28, 53, 54]. Such socio-economic disparities have been further exacerbated during the pandemic, which potentially affects our indoor activity estimates during 2021. Thus, our estimate of a county's indoor transmission risk may represent an underestimate of the risk experienced by individuals in these communities. We commit to continued work to better characterize the transmission risk experienced by vulnerable populations. Lastly, we acknowledge that data modeling work that can influence public health policy decisions, particularly during an ongoing crisis, must be done with care to prevent misconceptions from having adverse effects on risk perception and policies [55]. We thus strongly note that while our measure of indoor behavioral seasonality provides a potential driver of respiratory disease seasonality, it remains one among many complex factors which integrate to predict the transmission potential of an ongoing epidemic or pandemic [56]. Thus we cannot rely on behavioral seasonality to diminish transmission naturally, and pandemic intervention strategies should not be planned around behavioral seasonality while population susceptibility remains high in so many locations.

Ongoing global change events highlight the importance of this work, as it informs how widespread disruptions may shift patterns of indoor activity, potentially altering traditional infectious disease seasonality. Climate change events will continue to cause significant disruption to normal behavior patterns; mechanistic understanding of infectious disease seasonality and real-time data collection will be crucial components of future disease control efforts. While other global change events may impact indoor activity in different ways than the COVID-19 pandemic, a rigorous understanding of the impact of host behavior on infectious disease allows policymakers and emergency preparedness experts to effectively address future disruptions.

Data Availability

We make available on Github the data and code needed to reproduce all figures and analyses in this manuscript: `https://github.com/bansallab/indoor_outdoor`. This dataset is of the metric used in all our analyses and figures ("indoor activity"). This dataset can be regenerated using the Safegraph Patterns and Places datasets found at `https://www.safegraph.com/covid-19-data-consortium` and code in the

Github repository.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R01GM123007. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

We gratefully acknowledge data sharing by Safegraph which made this study possible. We thank Alexes Merritt for her data processing efforts.

References

[1] Micaela Elvira Martinez. The calendar of epidemics: Seasonal cycles of infectious diseases. *PLoS Pathogens*, 14(11):e1007327, 2018.

[2] Sonia Altizer, Andrew Dobson, Parviez Hosseini, Peter Hudson, Mercedes Pascual, and Pejman Rohani. Seasonality and the dynamics of infectious diseases. *Ecology Letters*, 9(4):467–484, 2006.

[3] Nicholas C Grassly and Christophe Fraser. Seasonal infectious disease epidemiology. *Proceedings of the Royal Society B: Biological Sciences*, 273(1600):2541–2550, 2006.

[4] Jeffrey Shaman and Melvin Kohn. Absolute humidity modulates influenza survival, transmission, and seasonality. *Proceedings of the National Academy of Sciences*, 106(9):3243–3248, 2009.

[5] Jeffrey Shaman, Virginia E Pitzer, Cécile Viboud, Bryan T Grenfell, and Marc Lipsitch. Absolute humidity and the seasonal onset of influenza in the continental United States. *PLoS Biology*, 8(2):e1000316, 2010.

[6] Benjamin D Dalziel, Stephen Kissler, Julia R Gog, Cecile Viboud, Ottar N Bjørnstad, C Jessica E Metcalf, and Bryan T Grenfell. Urbanization and humidity shape the intensity of influenza epidemics in US cities. *Science*, 362(6410):75–79, 2018.

[7] Rachel E Baker, Ayesha S Mahmud, Caroline E Wagner, Wenchang Yang, Virginia E Pitzer, Cecile Viboud, Gabriel A Vecchi, C Jessica E Metcalf, and Bryan T Grenfell. Epidemic dynamics of respiratory syncytial virus in current and future climates. *Nature Communications*, 10(1):1–8, 2019.

[8] Daisuke Onozuka and Masahiro Hashizume. The influence of temperature and humidity on the incidence of hand, foot, and mouth disease in Japan. *Science of the Total Environment*, 410:119–125, 2011.

[9] C Jessica E Metcalf, Ottar N Bjørnstad, Bryan T Grenfell, and Viggo Andreasen. Seasonality and comparative dynamics of six childhood infections in pre-vaccination Copenhagen. *Proceedings of the Royal Society B: Biological Sciences*, 276(1676):4111–4118, 2009.

[10] Joël Mossong, Niel Hens, Mark Jit, Philippe Beutels, Kari Auranen, Rafael Mikolajczyk, Marco Massari, Stefania Salmaso, Gianpaolo Scalia Tomba, Jacco Wallinga, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine*, 5(3):e74, 2008.

[11] Noga Kronfeld-Schor, Tamara J Stevenson, Sema Nickbakhsh, Eva S Schernhammer, Xaquin C Dopico, Tamar Dayan, Maria Martinez, and Barbara Helm. Drivers of infectious disease seasonality: potential implications for COVID-19. *Journal of Biological Rhythms*, 36(1):35–54, 2021.

[12] Kevin M Bakker, Marisa C Eisenberg, Robert Woods, and Micaela E Martinez. Exploring the seasonal drivers of varicella zoster transmission and reactivation. *American Journal of Epidemiology*, 2021.

[13] D Fisman. Seasonality of viral infections: mechanisms and unknowns. *Clinical Microbiology and Infection*, 18(10):946–954, 2012.

[14] Nita Bharti, Andrew J Tatem, Matthew J Ferrari, Rebecca F Grais, Ali Djibo, and Bryan T Grenfell. Explaining seasonal fluctuations of measles in Niger using nighttime lights imagery. *Science*, 334(6061):1424–1427, 2011.

[15] Roger Few, Iain Lake, Paul R Hunter, and Pham Gia Tran. Seasonality, disease and behavior: Using multiple methods to explore socio-environmental health risks in the Mekong Delta. *Social Science & Medicine*, 80:1–9, 2013.

[16] Allisandra G Kummer, Juanjuan Zhang, Maria Litvinova, Alessandro Vespignani, Hongjie Yu, and Marco Ajelli. Measuring the seasonality of human contact patterns and its implications for the spread of respiratory infectious diseases. *medRxiv*, 2022.

[17] Raymond Tellier, Yuguo Li, Benjamin J Cowling, and Julian W Tang. Recognition of aerosol transmission of infectious agents: a commentary. *BMC infectious diseases*, 19(1):1–9, 2019.

[18] Trisha Greenhalgh, Jose L Jimenez, Kimberly A Prather, Zeynep Tufekci, David Fisman, and Robert Schooley. Ten scientific reasons in support of airborne transmission of SARS-CoV-2. *The Lancet*, 397(10285):1603–1605, 2021.

[19] Chia C Wang, Kimberly A Prather, Josué Sznitman, Jose L Jimenez, Seema S Lakdawala, Zeynep Tufekci, and Linsey C Marr. Airborne transmission of respiratory viruses. *Science*, 373(6558):eabd9149, 2021.

[20] Mahesh Jayaweera, Hasini Perera, Buddhika Gunawardana, and Jagath Manatunge. Transmission of COVID-19 virus by droplets and aerosols: A critical review on the unresolved dichotomy. *Environmental Research*, 188:109819, 2020.

[21] Michael Klompas, Meghan A Baker, and Chanu Rhee. Airborne transmission of SARS-CoV-2: theoretical considerations and available evidence. *JAMA*, 2020.

[22] Lidia Morawska and Donald K Milton. It is time to address airborne transmission of coronavirus disease 2019 (COVID-19). *Clinical Infectious Diseases*, 71(9):2311–2313, 2020.

[23] Jennifer L Nguyen and Douglas W Dockery. Daily indoor-to-outdoor temperature and humidity relationships: a sample across seasons and diverse climatic regions. *International Journal of Biometeorology*, 60(2):221–229, 2016.

[24] Alison J Robey and Laura Fierce. Sensitivity of airborne transmission of enveloped viruses to seasonal variation in indoor relative humidity. *International Communications in Heat and Mass Transfer*, 130:105747, 2022.

[25] Wan Yang and Linsey C Marr. Dynamics of airborne influenza A viruses indoors and dependence on humidity. *PloS one*, 6(6):e21481, 2011.

[26] Wayne R Ott. *Human activity patterns: a review of the literature for estimating time spent indoors, outdoors, and in transit*. US Environmental Protection Agency, 1988.

[27] Neil E Klepeis, William C Nelson, Wayne R Ott, John P Robinson, Andy M Tsang, Paul Switzer, Joseph V Behar, Stephen C Hern, and William H Engelmann. The National Human Activity Pattern Survey (NHAPS): a resource for assessing exposure to environmental pollutants. *Journal of Exposure Science & Environmental Epidemiology*, 11(3):231–252, 2001.

[28] Elizabeth W Spalt, Cynthia L Curl, Ryan W Allen, Martin Cohen, Sara D Adar, Karen H Stukovsky, Ed Avol, Cecilia Castro-Diehl, Cathy Nunn, Karen Mancera-Cuevas, et al. Time–location patterns of a diverse population of older adults: the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). *Journal of exposure science & environmental epidemiology*, 26(4):349–355, 2016.

[29] Pietro Coletti, Chiara Poletto, Clément Turbelin, Thierry Blanchon, and Vittoria Colizza. Shifting patterns of seasonal influenza epidemics. *Scientific Reports*, 8(1):1–12, 2018.

[30] Matt J Keeling, Pejman Rohani, and Bryan T Grenfell. Seasonally forced disease dynamics explored as switching between attractors. *Physica D: Nonlinear Phenomena*, 148(3-4):317–335, 2001.

[31] Safegraph. Safegraph Patterns. https://safegraph.com/, 2021 (Last accessed February 14, 2022).

[32] Statista Digital Market Outlook. Individuals of any age who own at least one smartphone and use the smartphone(s) at least once per month. https://www.statista.com/statistics/201182/forecast-of-smartphone-users-in-the-us/, 2022 (Last accessed Feb 17, 2022).

[33] Charu C. Aggarwal and Chandan K. Reddy. *Data Clustering: Algorithms and Applications*. Chapman & Hall/CRC, 1st edition, 2013.

[34] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[35] Vincent Traag. Louvain-igraph. https://louvain-igraph.readthedocs.io/en/latest/reference.html, 2018 (Last accessed Feb 17, 2019).

[36] Till Hoffmann, Leto Peel, Renaud Lambiotte, and Nick S Jones. Community detection in networks without observing edges. *Science advances*, 6(4):eaav1478, 2020.

[37] Michael R Berthold and Frank Höppner. On clustering time series using euclidean distance and pearson correlation. *arXiv preprint arXiv:1601.02213*, 2016.

[38] Brennan Klein, Timothy LaRock, Stefan McCabe, Leo Torres, Lisa Friedland, Maciej Kos, Filippo Privitera, Brennan Lake, Moritz UG Kraemer, John S Brownstein, et al. Characterizing collective physical distancing in the us during the first nine months of the covid-19 pandemic. *arXiv preprint arXiv:2212.08873*, 2022.

[39] Rachel E Baker, Wenchang Yang, Gabriel A Vecchi, C Jessica E Metcalf, and Bryan T Grenfell. Susceptible supply limits the role of climate in the early SARS-CoV-2 pandemic. *Science*, 369(6501):315–319, 2020.

[40] Miyu Moriyama, Walter J Hugentobler, and Akiko Iwasaki. Seasonality of respiratory viral infections. *Annual review of virology*, 7:83–101, 2020.

[41] Jianyun Lu, Jieni Gu, Kuibiao Li, Conghui Xu, Wenzhe Su, Zhisheng Lai, Deqian Zhou, Chao Yu, Bin Xu, and Zhicong Yang. COVID-19 outbreak associated with air conditioning in restaurant, Guangzhou, China, 2020. *Emerging infectious diseases*, 26(7):1628, 2020.

[42] Chung-Min Liao, Chao-Fang Chang, and Huang-Min Liang. A probabilistic transmission dynamic model to assess indoor airborne infection risks. *Risk Analysis: An International Journal*, 25(5):1097–1107, 2005.

[43] Rebecca K Borchering, Cécile Viboud, Emily Howerton, Claire P Smith, Shaun Truelove, Michael C Runge, Nicholas G Reich, Lucie Contamin, John Levander, Jessica Salerno, et al. Modeling of future covid-19 cases, hospitalizations, and deaths, by vaccination rates and nonpharmaceutical intervention scenarios—united states, april–september 2021. *Morbidity and Mortality Weekly Report*, 70(19):719, 2021.

[44] Serina Chang, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, and Jure Leskovec. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature*, 589(7840):82–87, 2021.

[45] Thomas F Johnson, Lisbeth A Hordley, Matthew P Greenwell, and Luke C Evans. Associations between COVID-19 transmission rates, park use, and landscape structure. *Science of the Total Environment*, 789:148123, 2021.

[46] Jon Keegan and Alfred Ng. There's a multibillion-dollar market for your phone's location data. *The Markup*, 2021.

[47] Safegraph. Privacy Policy. https://www.safegraph.com/privacy-policy, 2021 (Last accessed Feb 17, 2022).

[48] Pew Resesarch Center. Mobile Fact Sheet. https://www.pewresearch.org/internet/fact-sheet/mobile/, 2021 (Last accessed Feb 17, 2022).

[49] Kathryn Zickuhr and Aaron Smith. 28% of american adults use mobile and social location-based services. 2011.

[50] Ryan Fox. "What about bias in your dataset?": Quantifying Sampling Bias in SafeGraph Patterns. https://colab.research.google.com/drive/1u15afRytJMsizySFqA2EPlXSh3KTmNTQ, 2019 (Last accessed Feb 17, 2022).

[51] Amanda Coston, Neel Guha, Derek Ouyang, Lisa Lu, Alexandra Chouldechova, and Daniel E Ho. Leveraging administrative data for bias audits: Assessing disparate coverage with mobility data for COVID-19 policy. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 173–184, 2021.

[52] Julia C Pringle, Jillian Leikauskas, Sue Ransom-Kelley, Benjamin Webster, Samuel Santos, Heidi Fox, Shannon Marcoux, Patsy Kelso, and Natalie Kwit. COVID-19 in a correctional facility employee following multiple brief exposures to persons with COVID-19—vermont, july–august 2020. *Morbidity and Mortality Weekly Report*, 69(43):1569, 2020.

[53] Lorien Nesbitt, Michael J Meitner, Cynthia Girling, Stephen RJ Sheppard, and Yuhao Lu. Who has access to urban vegetation? a spatial analysis of distributional green equity in 10 US cities. *Landscape and Urban Planning*, 181:51–79, 2019.

[54] Justine S Sefcik, Michelle C Kondo, Heather Klusaritz, Elisa Sarantschin, Sara Solomon, Abbey Roepke, Eugenia C South, and Sara F Jacoby. Perceptions of nature and access to green space in four urban neighborhoods. *International Journal of Environmental Research and Public Health*, 16(13):2313, 2019.

[55] Colin J Carlson, Ana CR Gomez, Shweta Bansal, and Sadie J Ryan. Misconceptions about weather and seasonality must not misguide COVID-19 response. *Nature Communications*, 11(1):1–4, 2020.

[56] Zachary Susswein, Eugenio Valdano, Tobias Brett, Pej Rohani, Vittoria Colizza, and Shweta Bansal. Ignoring spatial heterogeneity in drivers of SARS-CoV-2 transmission in the US will impede sustained elimination. *medRxiv*, 2021.

[57] Fedor Mesinger, Geoff DiMego, Eugenia Kalnay, Kenneth Mitchell, Perry C Shafran, Wesley Ebisuzaki, Dušan Jović, Jack Woollen, Eric Rogers, Ernesto H Berbery, et al. North american regional reanalysis. *Bulletin of the American Meteorological Society*, 87(3):343–360, 2006.

[58] International Code Council. 2012 International Energy Conservation Code , Printed 2015.

Supplementary Figures

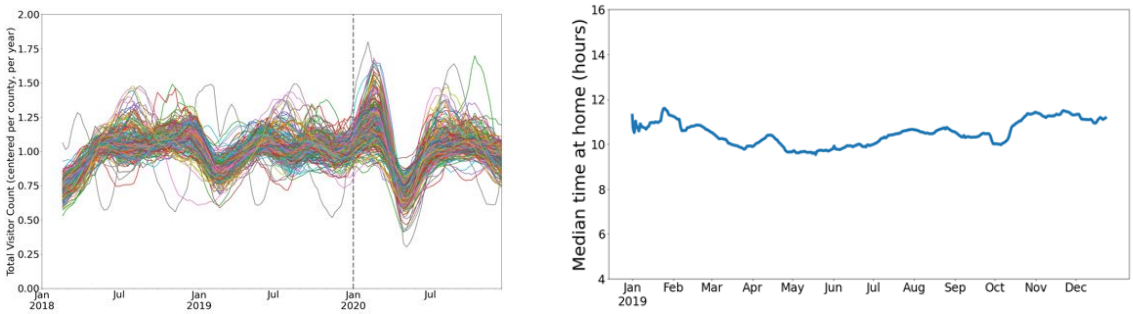


Figure 1—figure supplement 1: Left: Using the Safegraph Weekly Patterns dataset (https://docs.safegraph.com/docs/weekly-patterns), we show total (all non-home locations) visitor counts for a random sample of 310 counties (10% of all US counties). Overall mobility does not appear to be highly seasonal. Right: Using the Safegraph Social Distancing Metrics dataset (https://docs.safegraph.com/docs/social-distancing-metrics), we show time spent at home for a random sample of 310 counties (10% of all US counties). While home locations are not included in our indoor activity metric, time spent at home does not appear to be highly seasonal.

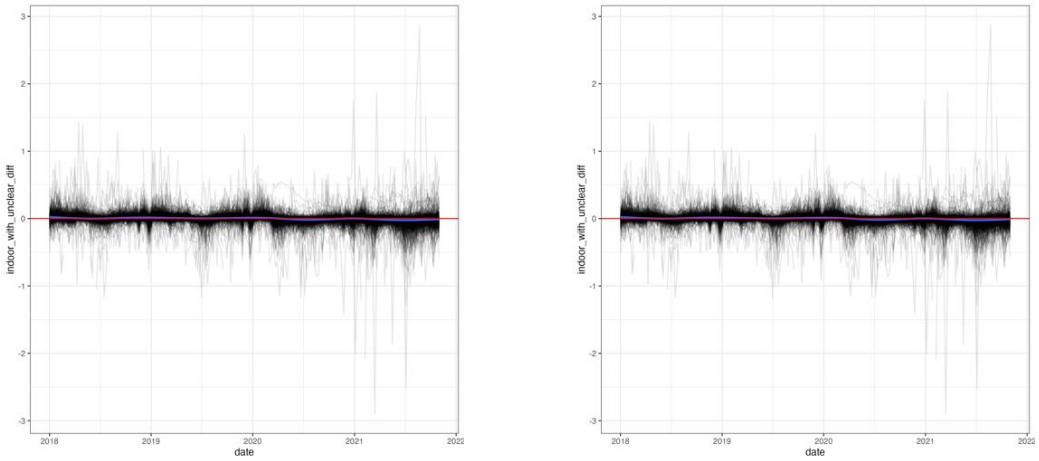


Figure 1—figure supplement 2: We demonstrate the effect of the "unclear" locations on the indoor activity seasonality. In the left panel, we show the difference in σ if all "unclear" locations were to be classified as indoor. In the right panel, we show the difference if σ if all "unclear" locations are classified as outdoor.

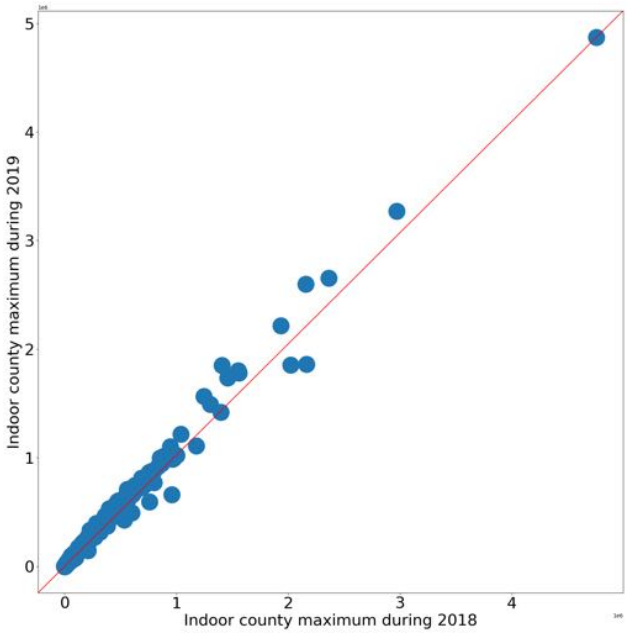
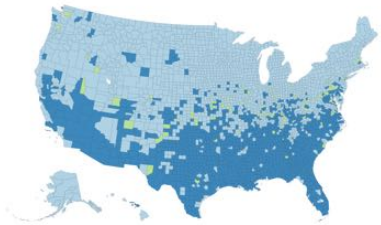
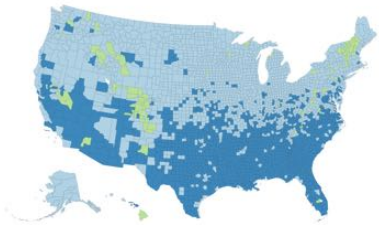


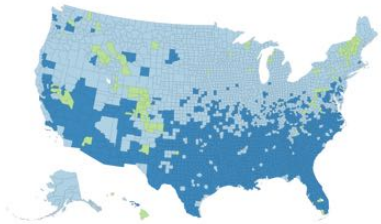
Figure 1—figure supplement 3: We show that the maximum number of visits used in the definition of the σ metric are highly comparable in 2018 and 2019.



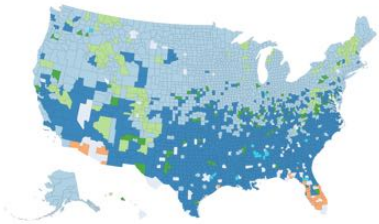
(a) $50th$ percentile, $\rho > 0.42$, NMI = 0.60



(b) $75th$ percentile, $\rho > 0.54$, NMI = 0.74



(c) $90th$ percentile, $\rho > 0.67$



(d) $99th$ percentile, $\rho > 0.82$, NMI = 0.67

Figure 1—figure supplement 4: We illustrate the impact of the correlation threshold on the clustering results (without post processing). For each panel, we list the percentile for time series correlations used as the threshold, the corresponding correlation value (ρ), and the normalized mutual information between each partition and the partition with the 90th percentile threshold (corresponding to the partition presented in Figure 2).

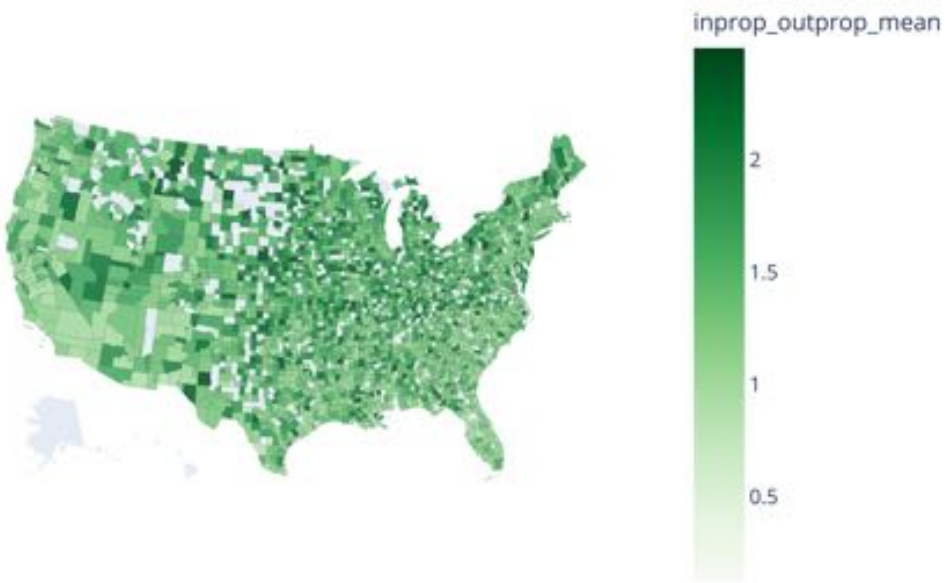


Figure 2—figure supplement 1: The mean proportion of indoor/outdoor activity ($\mu_{\tilde{\sigma}}$) in 2018 displays no latitudinal gradient and is relatively homogeneous across counties; outliers of mean ≥ 2.5 are removed

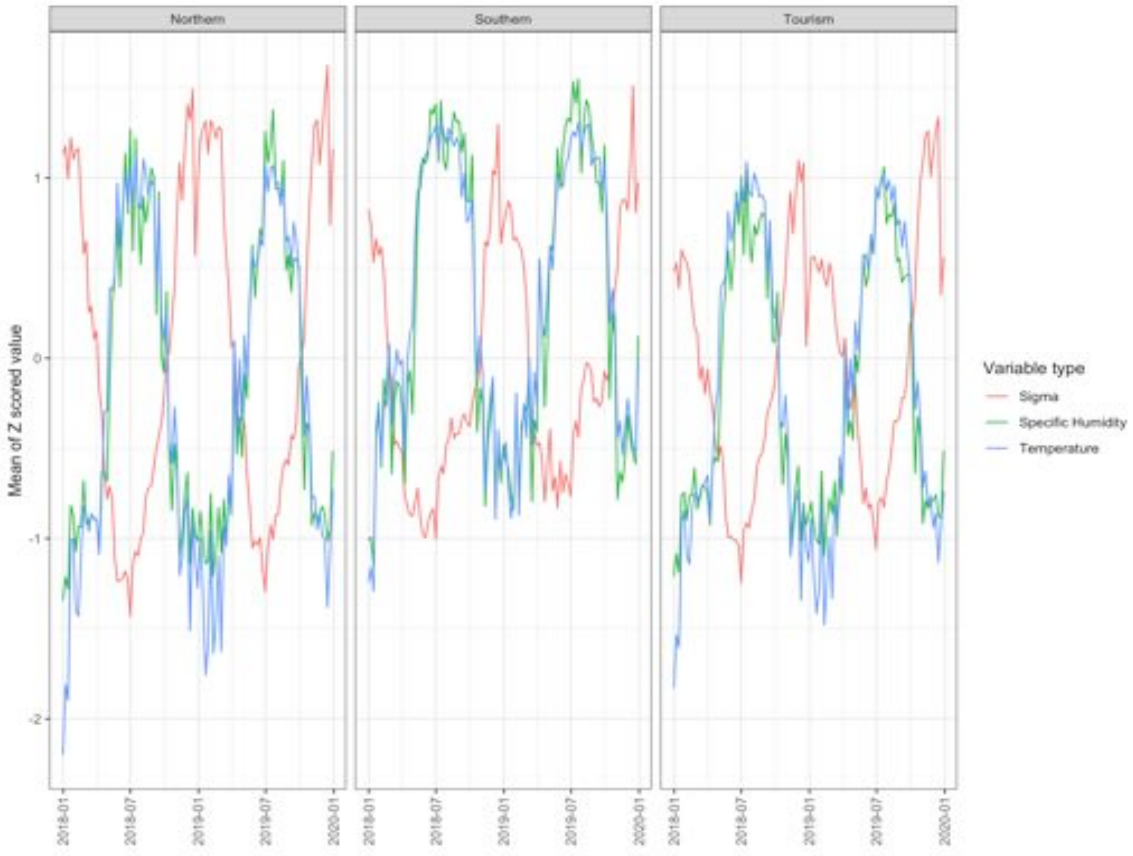


Figure 2—figure supplement 2: Using data on temperature and rainfall from NOAA's North American Regional Reanalysis [57], we find that indoor activity (sigma) is moderately anticorrelated with both temperature and humidity. Temperature and humidity are strongly correlated in all three clusters (pearson's $\rho \approx 0.87$). Across the three clusters, indoor activity is moderately associated with temperature ($\rho \approx -0.52$). Likewise, indoor activity is moderately anticorrelated with humidity ($\rho \approx -0.45$).

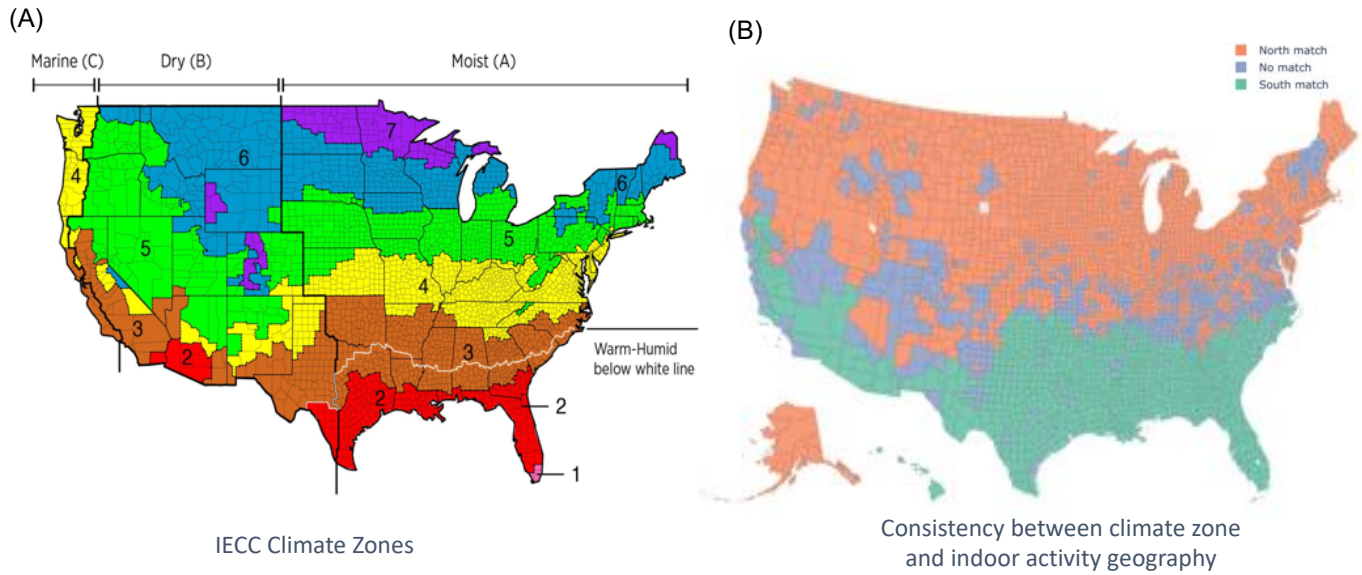


Figure 2—figure supplement 3: (A) The IECC climate zones are based on temperature, humidity, and rainfall in each county and govern the type building material and amount of ventilation required in a building [58]. (B) The consistency between the two primary clusters of indoor activity identified by our analysis and the IECC climate zones. Treating the IECC climate zones as "ground truth", we quantify the ability of our indoor activity clusters to predict the IECC climate zones We achieve this by collapsing the partitions into two clusters each (the tourism cluster is grouped with the northern cluster in the indoor activity clustering; and IECC climate zones 1/2/3 are grouped into one cluster and zones 4/5/6/7 into another cluster). Our indoor activity clusters have a 0.72 F1-score, with a precision of 0.92 and a recall score of 0.59 with the IECC zones.

Figure 2—figure supplement 4: The third indoor activity cluster displays some correlation with areas of increased tourism, including US ski areas in western and northeastern states, potentially contributing to off-season activity increases. Most areas in the cluster are either in a ski area or neighbor a ski area, with some parts of Hawaii and Florida being clear outliers of this pattern and suggests other types of tourism lead to similar behavioral seasonality.

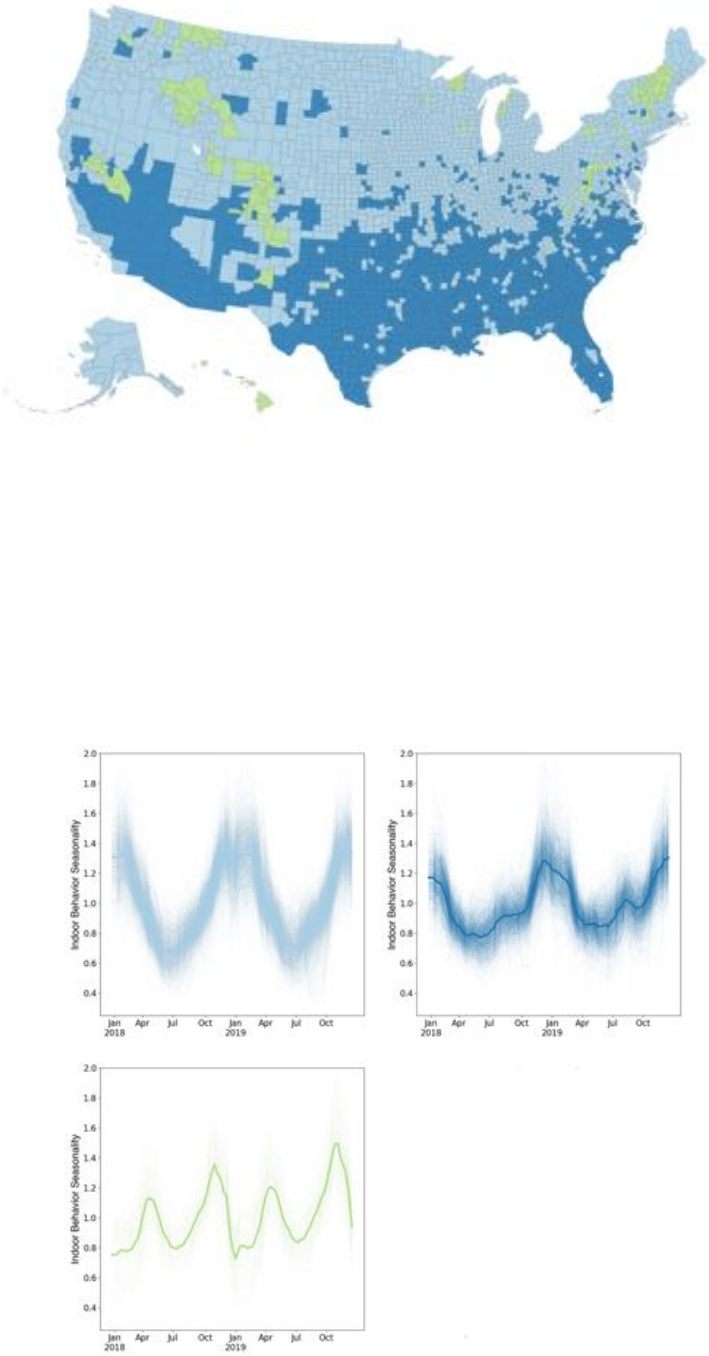
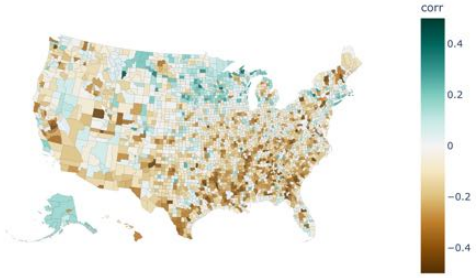


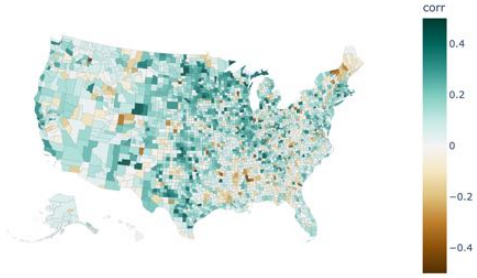
Figure 2—figure supplement 5: We show the results of time series clustering based on a hierarchical clustering method using Ward linkage and Euclidean distance, implemented using `scipy.cluster` in *Python*. This partition has high similarity to the network-based clustering algorithm results we illustrate in Figure 2: normalized mutual information = 0.56 with 89% of counties matching on cluster identity.

	2018/2019	2018/2020	2019/2020
Euclidean distance	1.1	1.5	1.5

(a) Interannual deviations



(b) Deviations during Spring 2020



(c) Deviations during Winter 2020

Figure 3—figure supplement 1: Top: Euclidean distance between indoor activity time series in corresponding years for each county, averaged over all counties. The 2020 time series show a higher deviation from each of the baseline years than the two baseline years do from each other. Bottom: We illustrate the mean difference in indoor activity at baseline (defined as the average of 2018 and 2019) and 2020 for two time periods: (a) Week 10 to Week 20 in spring 2020 during the initial lockdown period for COVID-19. (b) Week 44 to Week 52 in winter 2020 during the first winter surge of COVID-19. Positive mean differences suggest more outdoor activity in 2020 than at baseline and negative mean differences suggest more indoor activity in 2020 than at baseline.

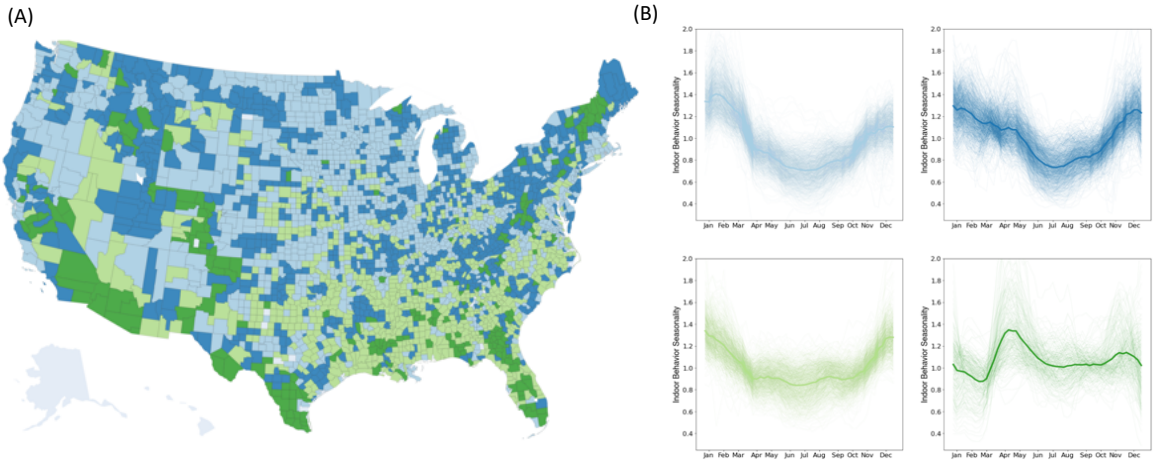
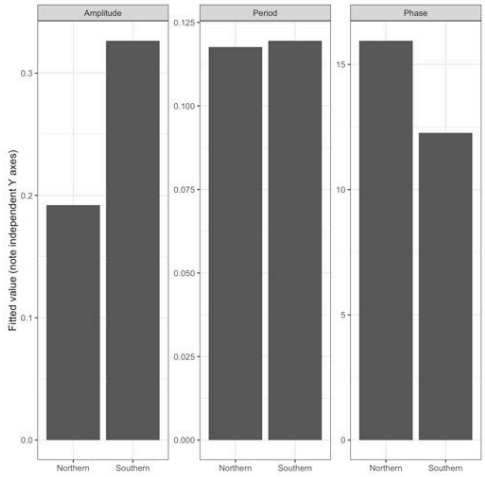


Figure 3—figure supplement 2: (A) Indoor seasonality during 2020 can be clustered into four groups, although clusters are more geographically fragmented than previous years. (B) Time series for 2020 indoor seasonality clusters display heterogeneous trends that were not apparent in previous years, with some clusters more variable than others.



Estimates and cluster differences			
term	Northern	Southern	diff
Amplitude	0.1921964	0.3264692	-0.134272737
Period	0.1176322	0.1195019	-0.001869687
Phase	15.9414221	12.2645018	3.676920333

Figure 4—figure supplement 1: Top: Inferred parameters for the sinusoidal model fits of the indoor activity data for the northern and southern clusters show a similar frequency, but greater amplitude and shorter phase in the southern cluster. Values displayed are mean parameter estimates. Standard errors for all parameters are smaller than 5e-3 and thus are not displayed. Bottom: We show the estimated parameters for the parameters of the sine curve fits to the Northern and Southern clusters as well as the difference between the parameter estimates. The period is in units of time (weeks). The amplitude matches the units of σ. The phase is in units of time (weeks).

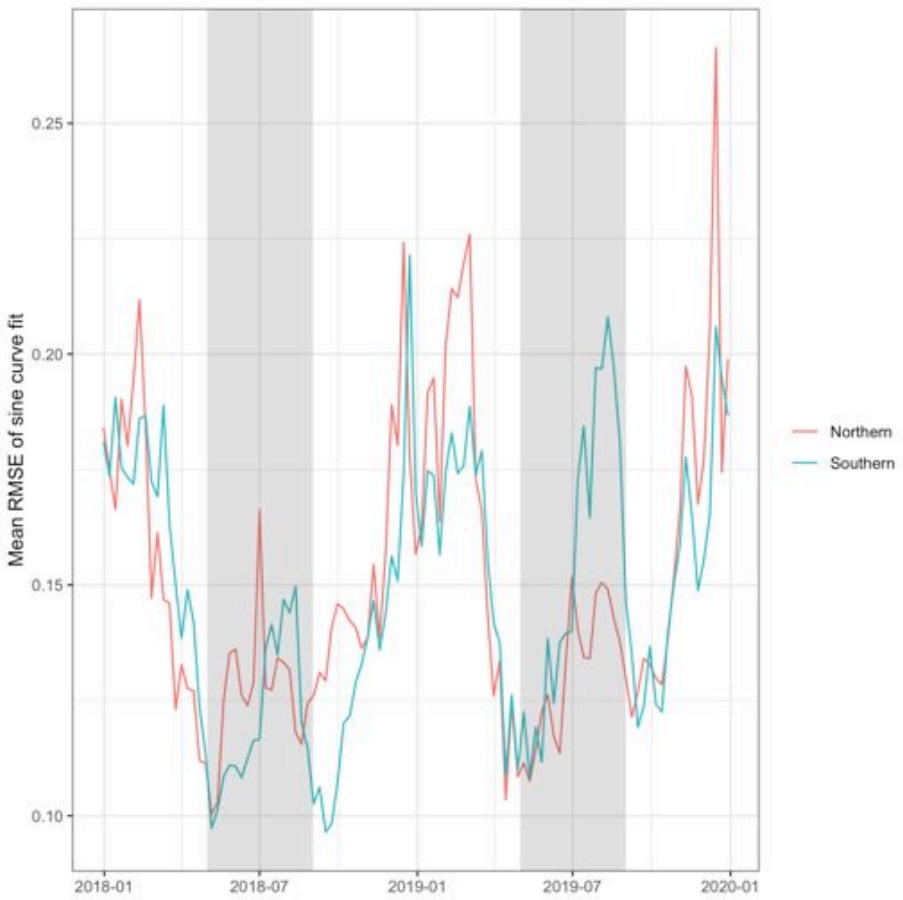


Figure 4—figure supplement 2: Model performance as measured by the root mean square error of the sine curve fit to the cluster averaged over counties within the cluster. The summer period between March and September is highlighted in light grey to emphasize the summer months.

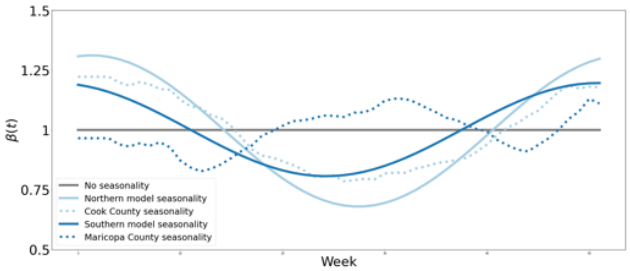


Figure 4—figure supplement 3: The seasonal forcing functions ($\beta 9t$)) we used in the epidemiological model. The non-seasonal model (grey) shows no variation in transmission risk over time. We model northern seasonality via a sinusoidal model fit to the northern indoor activity data (light blue solid) and via the empirically-measured indoor seasonality from a county in the northern cluster (Cook County, light blue dotted). We model southern seasonality via a sinusoidal model fit to the southern indoor activity data (dark blue solid) and via the empirically-measured indoor seasonality from a county in the northern cluster (Maricopa County, dark blue dotted).