Epidemic time series similarity is related to geographic distance and age structure

Tad A Dallas[a,*], Grant Foster[a], Robert L Richards[b] and Bret D Elderd[c]

[a]*Department of Biological Sciences, University of South Carolina, Columbia, SC, 29208*

[b]*Odum School of Ecology, University of Georgia, Athens, GA, 30609*

[c]*Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70802*

*Corresponding author: tad.a.dallas@gmail.com

Running title: Space, age, and epidemic similarity

Author contributions:TAD performed the analysis. All authors contributed to manuscript writing.

Acknowledgements: This work has been supported by the U.S. National Science Foundation RAPID grant (NSF-DEB-2031196).

Data accessibility: R code is available on figshare at `https://doi.org/10.6084/m9.figshare.19782406.v1` .

Keywords: Epidemic similarity, SARS-CoV2, Age structure, Distance decay

Epidemic time series similarity is related to geographic distance and age structure

Abstract

More similar locations may have similar infectious disease dynamics. There is clear overlap in putative causes for epidemic similarity, such as geographic distance, age structure, and population size. We compare the effects of these potential drivers on epidemic similarity compared to a baseline assumption that differences in the basic reproductive number (R_0) will translate to differences in epidemic trajectories. Using COVID-19 case counts from United States counties, we explore the importance of geographic distance, population size differences, and age structure dissimilarity on resulting epidemic similarity. We find clear effects of geographic space, age structure, population size, and R_0 on epidemic similarity, but notably the effect of age structure was stronger than the baseline assumption that differences in R_0 would be most related to epidemic similarity. Together, this highlights the role of spatial and demographic processes on SARS-CoV2 epidemics in the United States.

Introduction

The most recent pandemic of severe acute respiratory syndrome coronavirus 2 (SARS-CoV2) has highlighted the pressing need to understand how epidemics emerge and spread, and how epidemic models may be used for control and mitigation efforts. Models are used to estimate parameters of interest, which are then used to calculate composite properties (e.g., basic reproduction number R_0; Brauner *et al.* (2021); Ives & Bozzuto (2021)) and to simulate epidemics under different mitigation scenarios (e.g., Baker *et al.* (2020); Hinch *et al.* (2021); Sun *et al.* (2020)). However, these composite pathogen properties are not properties of the pathogen alone, but are conditional on the host population. Differences in susceptibility and contact patterns among individuals is critical to pathogen transmission and epidemic trajectories (Yin *et al.*, 2017). Measures of R_0 – quantifying the approximated number of secondary cases from a single case in a wholly susceptible host population – based on temporal case counts can hint at these differences in individual contact and transmission, but could also suggest differences in pathogen strain diversity and numerous other factors contributing to epidemic dynamics (Corcoran *et al.*, 2020; Ives & Bozzuto, 2021). Understanding the processes that lead to differing epidemic dynamics is a pressing research need, as many of these underlying drivers of estimated R_0 may potentially change over time or with different intervention strategies (Islam *et al.*, 2021).

The SARS-CoV-2 pandemic has created a situation where it may be possible to start to disentangle the role of different factors on resulting epidemic trajectories. For one, county-level data on infectious case counts provide a means to compare how epidemics progressed at the county scale, and to compare epidemic trajectories between counties. At a basic level, this allows for the comparison of epidemic trajectories to differences in R_0, as the larger difference in R_0 would suggest that the epidemics should be quite dissimilar in their trajectories. For one, R_0 may be estimated from the epidemic time series itself, such that epidemics with similar R_0 would naturally have similar dynamics. However, R_0 is a simple composite measure estimated from a time series that may belie the influence of mitigation efforts and fluctuating epidemic dynamics (e.g., COVID-19 case counts appeared in distinct waves, while R_0 estimates do not use all waves; Ives & Bozzuto (2021)). Apart from similarity in R_0 leading to similar epidemics, differences in epidemic trajectories may be driven simply by geographic space between two epidemics. That is, epidemics should be more similar in nearby counties than in distant counties. This could be driven by several interwoven drivers, which may not be reflected in differences in estimated R_0, including spatial autocorrelation in demographics, climatic effects on transmission, differences in mitigation efforts, or the movement of infectious individuals.

But there is an inherent circularity here, in that estimates of R_0 are based on the epidemic trajectories, such that pairwise differences in R_0 between counties should inherently be related to differences in epidemic trajectories. This creates an interesting baseline for comparison. That is, differences in R_0 should hypothetically relate to differences in epidemic trajectory – barring time-varying R_0 and assuming R_0 can be estimated accurately – simply because R_0 is estimated from a portion of the epidemic time series. Here, we explore how epidemic trajectories are related to differences in R_0, and how other important differences between counties may further influence epidemic trajectories. Specifically, epidemic trajectories may differ as a function of geographic distance between counties, and differences in age structure and population size. We find that there is a clear signal of geographic distance and demographic (population size and age structure) dissimilarity on resulting epidemic trajectory differences for a set of 3139 US counties. We compare the strength of these relationships to the potentially circular relationship between epidemic trajectory differences and differences in R_0, finding that age structure dissimilarity is more strongly related to epidemic trajectory similarity compared to differences in R_0. Together, this suggests an important role for age structure to epidemic emergence and progression, and highlights the importance of considering the spatial landscape of infectious disease.

Methods

COVID-19 epidemic time series data Time series case data for SARS-CoV-2 were compiled by the Center for Systems Science and Engineering at Johns Hopkins University Dong, Du & Gardner (2020) for a set of 3139 United States counties, with recorded case counts every day for the period between January 22, 2020, and May 9, 2022. These data were then rescaled to cases per 100,000 residents based on county population estimates from the United States Census Bureau from 2019 Loftin (2019). County age structure data was also obtained from the US Census Bureau Loftin (2019), and standardized to sum to one within a given county. Age structure dissimilarity was estimated as the Euclidean distance between two counties in their age structure distributions. Estimates of R_0 were obtained from Ives & Bozzuto (2021), which were estimated from the epidemic time series directly.

Dynamic time warping Dynamic time warping (DTW) is an approach to measure the similarity between two time series based on the notion that there is not an inherent 1:1 matching between values in each time series (Berndt & Clifford, 1994), largely applied to problems in speech (Amin & Mahmood, 2008) and gait (Boulgouris, Plataniotis & Hatzinakos, 2004) recognition and comparison. The underlying idea is that the speed of speech or gait could be different, while the actual underlying pattern is the same (e.g., the same words can be spoken more quickly or with differing amounts of pauses). In our application to infectious disease, there is no reason to believe that the pairwise difference in Covid-19 case counts between two counties is *actually* a measure of how similar the epidemics are, given that the epidemics may have started at different times. This fundamental disconnect means that perhaps it is more suitable to attempt to match the time series data based on the start of the epidemic or to use an approach which is flexible to different epidemic start times, as we do here. By allowing an *elastic* transformation of the time series, DTW attempts to minimize the difference between the two trajectories while accounting for phase shifts in epidemic dynamics (Figure 1).

$$DTW(x,y) = min_{\pi \in \mathbf{A(x,y)}} \left(\sum_{(i,j) \in \pi} d(x_i, y_j)^q \right)^{1/q} \tag{1}$$

Here, we want to compare two epidemic time series (x and y), considering an alignment path π of all possible paths ($A_{x,y}$), where i and j correspond to the position in the time series mapping onto the potential alignments, where q is a normalization constant. The goal is to find an alignment which minimizes the overall dissimilarity between the two time series. We use the `dtw` R package (Giorgino, 2009), and consider the dissimilarity between the time series to be the normalized cumulative dissimilarity between the two time series. There is a possibility that the results could be sensitive to the inclusion of many leading or trailing zero counts, where epidemics were on a fundamentally different timescale across US counties. While this approach should account for this, we explore the effect of truncating the epidemic time series to include 5 leading and 5 trailing zero values before the calculation of the DTW values. Trimming the time series to remove these zero-values did not affect our findings (see Supplementary Material).

What is related to epidemic similarity? Epidemic similarity was measured by comparing epidemic time series for every pair of US counties. This creates a pairwise dissimilarity matrix. To project this high-dimensional matrix into lower dimensions for analysis, we used t-distributed stochastic neighbor embedding (t-SNE), a method that offers a low-dimensional projection of high-dimensional data (Gisbrecht, Schulz & Hammer, 2015). The result of this embedding is the production of two t-SNE axes, in which each axis contains one value per US county, and the distance along each axis relates to epidemic dissimilarity, mapping counties out along the two axes. This allows us to relate these low-dimensional axes representing epidemic trajectory similarity to differences between counties in terms of spatial distance, demographics (e.g., age structure and population size), and estimated epidemic properties (R_0 (Ives & Bozzuto, 2021)).

We used Moran's I to quantify the effects of geographic distance and age structure dissimilarity on resulting epidemic similarity. That is, how similar are epidemics in different counties as a function of geographic distance between counties or differences in age structure between counties? Originally designed as a measure of spatial autocorrelation, Moran's I is essentially a distance-weighted Pearson's correlation, allowing the relationship between a distance matrix (e.g., pairwise geographic distance between all US counties) and a county-level trait (e.g., t-SNE axis values). We related each t-SNE axis – representing the projected epidemic dissimilarity between two US counties – to pairwise matrices of 1) geographic distance between US counties, 2) age structure dissimilarity, 3) absolute difference in population size, and 4) absolute difference in R_0. The underlying idea being that counties that are closer to one another, with similar age structure, and not differing greatly in population size or estimated R_0 (Ives & Bozzuto, 2021) would also be closer together along t-SNE axes. All distance and dissimilarity matrices – describing the relative difference in geographic distance, age structure, population size, and R_0 among US county pairs – were standardized to be bound between 0 and 1, and inverted, such that the largest distances corresponded to the smallest values. This allows us to calculate z-scores based on the null distributions, and to compare these scores across the different distance/dissimilarity matrices.

However, we are fundamentally limited by the almost inherent collinearity between some of these measures. For instance, geographic distance and age structure dissimilarity were positively related, based on a Mantel test ($z = 247$, $p = 0.001$), suggesting that more distant counties also have more dissimilar age structure. We explore this further in the Supplemental Materials, where we use Mantel tests on the pairwise epidemic dissimilarity matrix directly, instead of attempting to project the dissimilarity into two axes using t-SNE. However, regressions of distance matrices are notoriously error-prone (Legendre, Fortin & Borcard, 2015), which is why we present the analyses of the t-SNE axes here. By compressing epidemic similarity into a low-dimensional space, more traditional regression techniques can be used. The results of both analyses are qualitatively similar (see Supplementary Materials for further discussion).

Reproducibility R code and data to reproduce the analyses is provided at `https://doi.org/10.6084/m9.figshare.19782406.v1`

Results

Pairwise epidemic time series similarity was calculated using dynamic time warping (DTW), which was weakly related to Euclidean distance in epidemic time series, suggesting that this approach was able to capture additional information relative to a more simple distance measure (see Supplemental Materials). The matrix of pairwise DTW values were reduced to two axes using t-SNE (Gisbrecht, Schulz & Hammer, 2015). This low-dimensional representation of site-level epidemic similarity showed clear spatial patterns for the first two t-SNE axes (Figure 2). Interestingly, the spatial patterns adhere to geopolitical (i.e., US state) boundaries in some instances, a phenomenon which may be due to differences between states in case reporting standards and practices (Sen-Crowe *et al.*, 2021), but is worthy of future investigation. The extent to which geographic distance is related to epidemic similarity is difficult to discern, as we observed spatial structure in population age structure differences (Figure S3), as well as clear relationships between R_0 and population size (Figure 3).

What is related to epidemic dissimilarity? Despite these difficulties, we find a clear relationship between epidemic similarity and geographic distance, age structure dissimilarity, and differences in population size and R_0 between counties (Table 1). These relationships were estimated using Moran's I, relating the two axes of epidemic similarity to pairwise matrices describing differences in age structure, geographic distance, R_0, and population size. Moran's I is scaled between -1 and 1, where a value of 0 represents a lack of distance-based (or dissimilarity-based) autocorrelation (either negative or positive). All estimated Moran's I values in the current analysis were positive, suggesting positive spatial autocorrelation for all dissimilarity and distance matrices examined here. Both t-SNE axes – representing epidemic dissimilarity – were positively related to 1) geographic distance between US counties, 2) age structure dissimilarity, 3) absolute difference in population size, and 4) absolute difference in R_0 (Table 1). Geographic distance was more related to both t-SNE axes relative to age structure, population size, and R_0 based on both the raw observed value and the corresponding standardized z-score (Table 1). Differences in R_0 between counties showed the next strongest signal in the t-SNE axes, followed by age structure dissimilarity (Table 1).

Discussion

Here, we explored how geographic space, demographics, and R_0 influence differences in epidemic trajectories for over 3000 United States counties. We expected – and found – that counties with similar R_0 values tended to have similar epidemics. Independent of this, we found clear effects of geographic distance between counties and dissimilarities in county age structure on resulting epidemic trajectories, suggesting that R_0 estimated from case or mortality data (Ives & Bozzuto, 2021) may not capture the full potential of the epidemic in a given location. Together, we highlight the importance of considering population demographics, age-specific contact network structure, and geographic distance when attempting to estimate epidemic trajectories. While we approach the problem as one of pairwise dissimilarity in epidemics, it may be possible to use similar approaches to recreate an expected epidemic time series for an unsampled location given information on geography and demography.

Spatial structure in both age structure and population sizes precludes the attribution of any form of causal link between age structure or geographic distance and resulting epidemic trajectories. However, our findings, based on the entire epidemic time, broadly agree with similar studies which focused on components of the transmission process or summary statistics such as R_0. Further, the analyses can be updated as the epidemic progresses, or using different time windows to explore how time series clustering changes temporally. It is recognized that both parts of the transmission process – encounter and susceptibility – vary with individual age

(Covid *et al.*, 2020; Jones *et al.*, 2021; Kerr *et al.*, 2021; Magpantay, King & Rohani, 2019), suggesting that for some pathogens including SARS-CoV-2, considering the age structure is quite important to epidemic forecasting (Kerr *et al.*, 2021). Additionally, geographic patterns in R_0 (Ives & Bozzuto, 2021), non-pharmaceutical interventions initiation and compliance (Amuedo-Dorantes, Kaushal & Muchow, 2021; Yang *et al.*, 2021), and vaccine hesitancy (Zuzek, Zipfel & Bansal, 2022) have emerged as potential drivers for spatial variation in epidemic progression (Richards *et al.*, 2022). By comparing epidemic trajectories directly, using a flexible framework which allows epidemics to be sampled at different timescales, we have found that these similarity patterns in summary values, transmission components, and intervention uptake scale up directly to the similarity between entire epidemics.

One major result is the marked state-level clustering of epidemic similarity (Figure 2). Previous clustering of US states was observed early in the pandemic at the state-level (Rojas, Valenzuela & Rojas, 2020), potentially reflecting large scale differences in mitigation protocols (e.g., closing bars and restaurants) or differences in testing regimes across US states. The consistent clustering at US state level when considering counties as the unit of study suggests that state-level variation in reporting, testing, or mitigation may manifest to influence epidemic similarity. Understanding the cause of this clustering may help to inform mitigation efforts, and help to uncover differences in testing or reporting that may be important to understand spatial patterns of infectious disease.

It is interesting that epidemic similarity showed clear signals of geographic distance, age structure, and county-level differences in population size and R_0, given that counties also varied in other marked ways. For instance, differences in non-pharmaceutical interventions, vaccination rate variation, and other demographic factors which we recognize are important to pathogen spread (Abedi *et al.*, 2021; Ge *et al.*, 2022; Zuzek, Zipfel & Bansal, 2022) did not mask the effect of age structure. One reason for this may be that age structure is serving as a surrogate for other measures of population demography not inherently related to age-structured transmission. That is, differences in vaccination hesitancy (Zuzek, Zipfel & Bansal, 2022) and risk perception (Bruine de Bruin, 2021) may differ across age groups. One way to parse this out would be to examine epidemic trajectory similarity in other geopolitical locations and at different spatial scales, where the relative influence of geographic connectivity, population demographics, and pathogen strain diversity may be quite different. The incorporation of temporal information on mitigation efforts, strain diversity, and availability of health care infrastructure is a clear next step to understanding and forecasting epidemic time series. This effort is obviously not aimed at forecasting directly, but could potentially be used to infer approximate epidemic dynamics for future epidemics or to explore how deviations from epidemic trajectories between neighboring counties (or those with similar age structure) may be driven by other critical variables.

The COVID-19 pandemic will not be the last pandemic (Medicine, 2022), and understanding the factors which influence epidemic dynamics are intrinsically important to public health measures. Perhaps this current pandemic is a special case, as comparisons in R_0 between SARS-CoV2 and 1918 pandemic influenza revealed little consensus in heavily impacted cities (Foster *et al.*, 2022). But it seems relevant to use approaches such as the one we do here to understand how epidemic trajectories differ, both within the same pandemic and potentially for different pathogens (e.g., how dissimilar are temporal patterns in seasonal flu epidemics in a given location?). The comparison of epidemic trajectories – especially along moving windows as the epidemic progresses – can provide insight into the relative effects of different mitigation and control efforts. Finally, while many approaches to forecasting epidemics rely on a single time series, this work alludes to the possibility of incorporating information on nearby or similar time series, creating the possibility of joint epidemic forecasts.

Conflict of interest: The authors have no conflicts of interest to declare.

Funding source: This work has been supported by the U.S. National Science Foundation RAPID grant (NSF-DEB-2031196).

Ethical Approval statement: The present study used publicly available data compiled by the Center for Systems Science and Engineering at Johns Hopkins University Dong, Du & Gardner (2020).

References

Abedi, V., Olulana, O., Avula, V., Chaudhary, D., Khan, A., Shahjouei, S., Li, J. & Zand, R. (2021) Racial, economic, and health inequality and COVID-19 infection in the United States. *Journal of racial and ethnic health disparities*, **8**, 732–742.

Amin, T.B. & Mahmood, I. (2008) Speech recognition using dynamic time warping. *2008 2nd international conference on advances in space technologies*, pp. 74–79. IEEE.

Amuedo-Dorantes, C., Kaushal, N. & Muchow, A.N. (2021) Timing of social distancing policies and COVID-19 mortality: county-level evidence from the US. *Journal of Population Economics*, **34**, 1445–1472.

Baker, R.E., Park, S.W., Yang, W., Vecchi, G.A., Metcalf, C.J.E. & Grenfell, B.T. (2020) The impact of COVID-19 nonpharmaceutical interventions on the future dynamics of endemic infections. *Proceedings of the National Academy of Sciences*, **117**, 30547–30553.

Berndt, D.J. & Clifford, J. (1994) Using dynamic time warping to find patterns in time series. *KDD workshop*, vol. 10, pp. 359–370. Seattle, WA, USA:.

Boulgouris, N.V., Plataniotis, K.N. & Hatzinakos, D. (2004) Gait recognition using dynamic time warping. *IEEE 6th Workshop on Multimedia Signal Processing, 2004.*, pp. 263–266. IEEE.

Brauner, J.M., Mindermann, S., Sharma, M., Johnston, D., Salvatier, J., Gavenčiak, T., Stephenson, A.B., Leech, G., Altman, G., Mikulik, V. *et al.* (2021) Inferring the effectiveness of government interventions against COVID-19. *Science*, **371**.

Bruine de Bruin, W. (2021) Age differences in COVID-19 risk perceptions and mental health: Evidence from a national US survey conducted in March 2020. *The Journals of Gerontology: Series B*, **76**, e24–e29.

Corcoran, D., Urban, M.C., Wegrzyn, J. & Merow, C. (2020) Virus evolution affected early COVID-19 spread. *medRxiv*.

Covid, C., Team, R., COVID, C., Team, R., COVID, C., Team, R., Bialek, S., Boundy, E., Bowen, V., Chow, N. *et al.* (2020) Severe outcomes among patients with coronavirus disease

2019 (COVID-19) - United States, February 12 - March 16, 2020. *Morbidity and mortality weekly report*, **69**, 343.

Dong, E., Du, H. & Gardner, L. (2020) An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases*, **20**, 533–534. Publisher: Elsevier.

Foster, G., Elderd, B., Richards, R. & Dallas, T. (2022) Estimating R_0 from Early Exponential Growth: Parallels between 1918 influenza and 2020 SARS-CoV-2 Pandemics. *in review*.

Ge, Y., Zhang, W.B., Liu, H., Ruktanonchai, C.W., Hu, M., Wu, X., Song, Y., Ruktanonchai, N.W., Yan, W., Cleary, E. *et al.* (2022) Impacts of worldwide individual non-pharmaceutical interventions on COVID-19 transmission across waves and space. *International Journal of Applied Earth Observation and Geoinformation*, **106**, 102649.

Giorgino, T. (2009) Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package. *Journal of Statistical Software*, **31**, 1–24.

Gisbrecht, A., Schulz, A. & Hammer, B. (2015) Parametric nonlinear dimensionality reduction using kernel t-SNE. *Neurocomputing*, **147**, 71–82.

Hinch, R., Probert, W.J., Nurtay, A., Kendall, M., Wymant, C., Hall, M., Lythgoe, K., Bulas Cruz, A., Zhao, L., Stewart, A. *et al.* (2021) OpenABM-Covid19An agent-based model for non-pharmaceutical interventions against COVID-19 including contact tracing. *PLoS computational biology*, **17**, e1009146.

Islam, A., Sayeed, M.A., Rahman, M.K., Zamil, S., Abedin, J., Saha, O. & Hassan, M.M. (2021) Assessment of basic reproduction number (R_0), spatial and temporal epidemiological determinants, and genetic characterization of SARS-CoV-2 in Bangladesh. *Infection, Genetics and Evolution*, **92**, 104884.

Ives, A.R. & Bozzuto, C. (2021) Estimating and explaining the spread of COVID-19 at the county level in the USA. *Communications Biology*, **4**, 1–9.

Jones, J.M., Stone, M., Sulaeman, H., Fink, R.V., Dave, H., Levy, M.E., Di Germanio, C., Green, V., Notari, E., Saa, P. *et al.* (2021) Estimated US infection-and vaccine-induced SARS-CoV-2 seroprevalence based on blood donations, July 2020-May 2021. *JAMA*, **326**, 1400–1409.

Kerr, C.C., Stuart, R.M., Mistry, D., Abeysuriya, R.G., Rosenfeld, K., Hart, G.R., Núñez, R.C., Cohen, J.A., Selvaraj, P., Hagedorn, B. *et al.* (2021) Covasim: an agent-based model of COVID-19 dynamics and interventions. *PLOS Computational Biology*, **17**, e1009149.

Legendre, P., Fortin, M.J. & Borcard, D. (2015) Should the Mantel test be used in spatial analysis? *Methods in Ecology and Evolution*, **6**, 1239–1247.

Loftin, L.E. (2019) Overview of the Census Bureau: The Other Nine Years. *The Geography Teacher*, **16**, 103–106.

Magpantay, F., King, A. & Rohani, P. (2019) Age-structure and transient dynamics in epidemiological systems. *Journal of the Royal Society Interface*, **16**, 20190151.

Medicine, T.L.R. (2022) Future pandemics: failing to prepare means preparing to fail. *The Lancet. Respiratory Medicine*, **10**, 221.

Richards, R.L., Foster, G., Elderd, B.D. & Dallas, T.A. (2022) Comparing Waves of COVID-19 in the US: Scale of response changes over time. *medRxiv*.

Rojas, F., Valenzuela, O. & Rojas, I. (2020) Estimation of covid-19 dynamics in the different states of the united states using time-series clustering. *medRxiv*.

Sen-Crowe, B., Sutherland, M., McKenney, M. & Elkbuli, A. (2021) The Florida COVID-19 mystery: Lessons to be learned. *The American Journal of Emergency Medicine*, **46**, 661.

Sun, J., Chen, X., Zhang, Z., Lai, S., Zhao, B., Liu, H., Wang, S., Huan, W., Zhao, R., Ng, M.T.A. *et al.* (2020) Forecasting the long-term trend of COVID-19 epidemic using a dynamic model. *Scientific reports*, **10**, 1–10.

Yang, B., Huang, A.T., Garcia-Carreras, B., Hart, W.E., Staid, A., Hitchings, M.D., Lee, E.C., Howe, C.J., Grantz, K.H., Wesolowksi, A. *et al.* (2021) Effect of specific non-pharmaceutical intervention policies on SARS-CoV-2 transmission in the counties of the United States. *Nature communications*, **12**, 1–10.

Yin, Q., Shi, T., Dong, C. & Yan, Z. (2017) The impact of contact patterns on epidemic dynamics. *PLoS One*, **12**, e0173411.

Zuzek, L.G.A., Zipfel, C.M. & Bansal, S. (2022) Spatial clustering in vaccination hesitancy: the role of social influence and social selection. *medRxiv*.

327 Tables

Table 1: Moran's I analysis exploring how t-SNE axes are related to geographic distance, age structure dissimilarity, difference in population size, and difference in R_0. Mantel tests use a randomization approach to generate null distributions to compare observed (`obs`) to null (`exp` and `sd`) distributions. Z-scores estimate the divergence of the test statistic from the null distribution.

covariate	t-SNE axis	obs	exp	sd	p-value	z-score
geography	1	0.02963	-0.00032	0.00014	$<$ **0.0001**	216.3
	2	0.01930	-0.00032	0.00014	$<$ **0.0001**	141.7
age structure	1	0.00043	-0.00032	0.00001	$<$ **0.0001**	60.5
	2	0.00017	-0.00032	0.00001	$<$ **0.0001**	39.4
population size	1	0.00002	-0.00032	0.00003	$<$ **0.0001**	11.7
	2	0.00004	-0.00032	0.00003	$<$ **0.0001**	12.3
R_0	1	0.00339	-0.00032	0.00003	$<$ **0.0001**	110.7
	2	0.00135	-0.00032	0.00003	$<$ **0.0001**	49.8

Figures

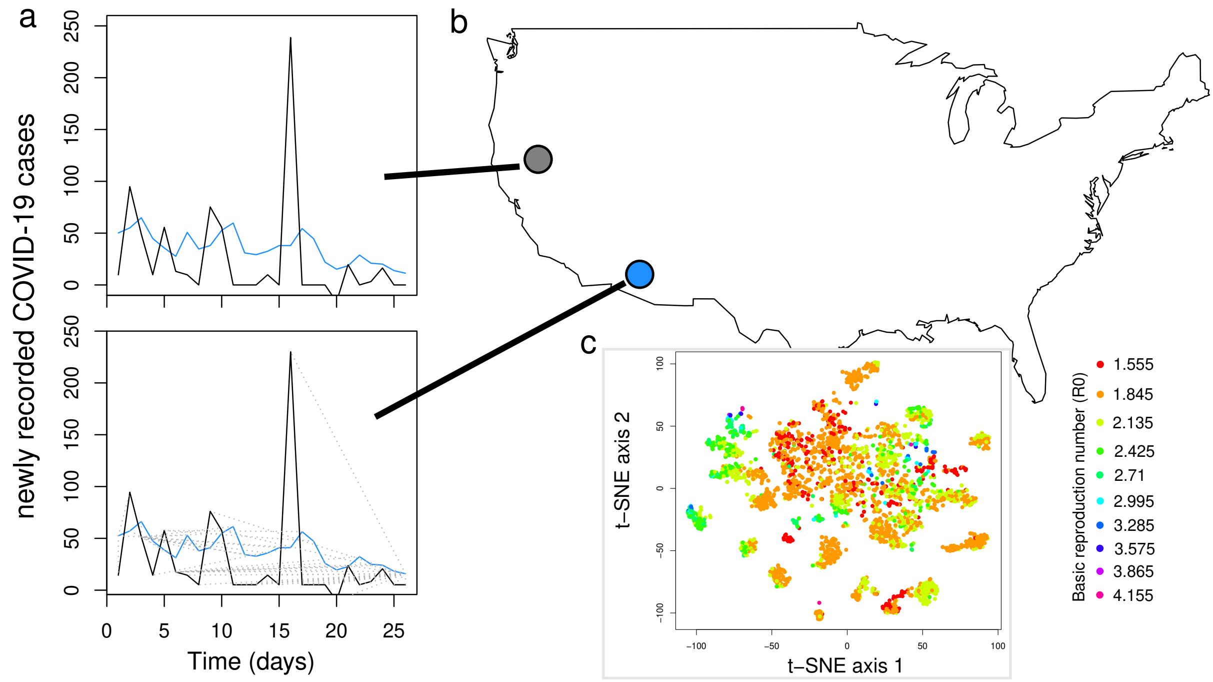

Figure 1: The similarity of epidemic time series was estimated using dynamic time warping, where two time series (in blue and black in panel *a*) are mapped onto one another (indicated by grey lines in panel *a*) to estimate epidemic dissimilarity. These time series are pairwise between every county in the United States (panel *b*). These pairwise values are then compressed to a low-dimensional space by using t-SNE (panel *c*), where point color corresponds to estimated R_0 for the given US county.

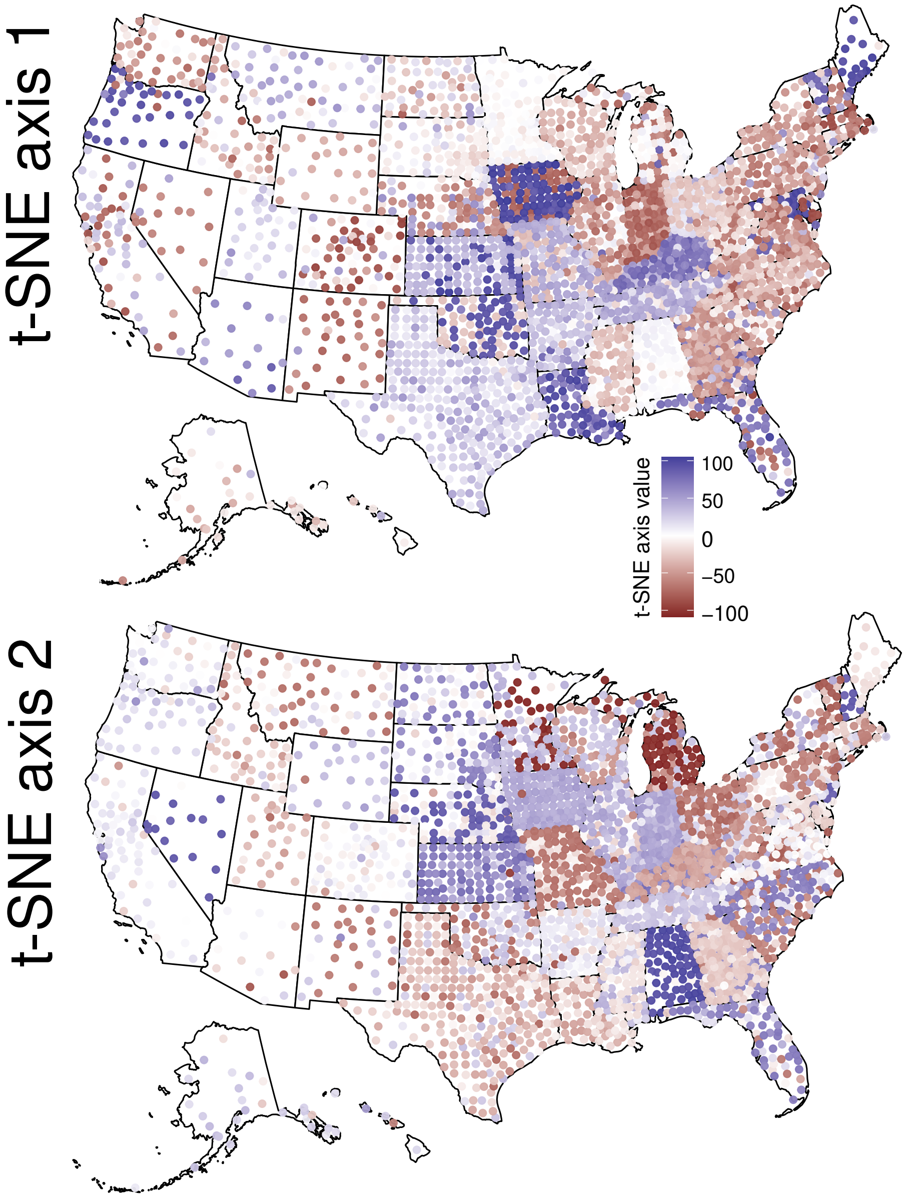

Figure 2: The spatial distribution of epidemic trajectory similarity (t-SNE decomposition of the pairwise dynamic time warping matrix). In this geographic projection of the t-SNE values, there are clearly some states which cluster, suggesting similar mitigation efforts, sampling/reporting biases, and/or epidemic trajectories.

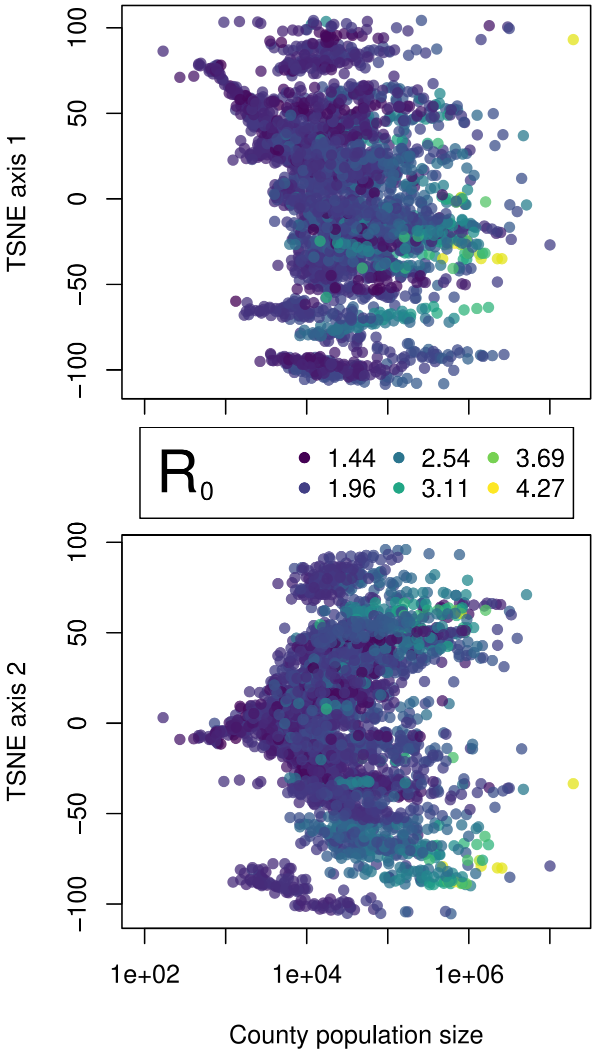

Figure 3: The relationship between t-SNE axes and county population size, with point color corresponding to R_0, highlighting the distribution of t-SNE values, the messy relationship between epidemic similarity and county population size, and the clear scaling of R_0 with county population size.

Supplementary materials

Title: Epidemic time series similarity is related to geographic distance and age structure

Authors: Tad A Dallas, Grant Foster, Robert Richards, & Bret D Elderd

Does time need to warped?

We use dynamic time warping as a flexible way to compare time series similarity. Here, we explore how much of this signal would be observed if we simply calculated the summed difference in pairwise epidemic trajectories. We found the two approaches are roughly similar, but that the dynamic time warping does result in different estimates of epidemic similarity (Figure S1), highlighting the application of such time series approaches to epidemic trajectory data.

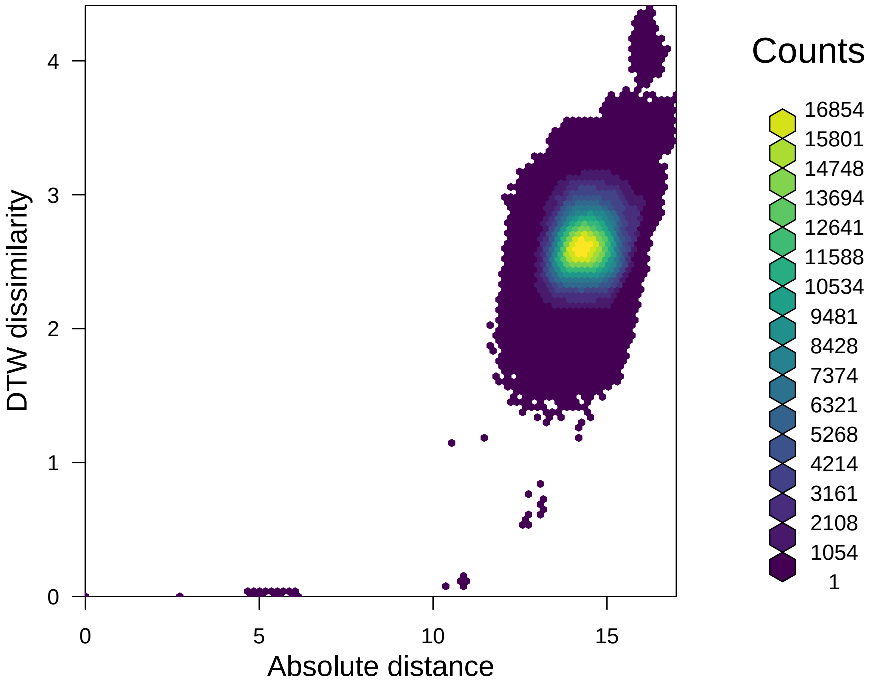

Figure S1: The sum of the absolute difference between the two time series is related to the dynamic time warp dissimilarity in this particular application. There are still clear differences between the two.

Truncating the epidemic time series

In the main text, we considered the full epidemic time series, including case counts in which case counts were zero-valued. Here, we explore to what extent this influences the dynamic time warping estimates, and our overall results. This does not influence our overall results (Table S1), and the two estimates of epidemic dissimilarity produced by truncating the epidemic time series versus keeping the entire time series are quite positively related (Figure S2).

Table S1: Moran's I analysis exploring how t-SNE axes are related to geographic distance, age structure dissimilarity, difference in population size, and difference in R_0. Mantel tests use a randomization approach to generate null distributions to compare observed (`obs`) to null (`exp` and `sd`) distributions. Z-scores estimate the divergence of the test statistic from the null distribution.

covariate	t-SNE axis	obs	exp	sd	p-value	z-score
geography	1	0.01832	-0.00032	0.00014	< **0.0001**	134.7
	2	0.02849	-0.00032	0.00014	< **0.0001**	208.1
age structure	1	0.00041	-0.00032	0.00001	< **0.0001**	58.9
	2	0.00024	-0.00032	0.00001	< **0.0001**	45.4
population size	1	0.00001	-0.00032	0.00003	< **0.0001**	11.2
	2	0.00008	-0.00032	0.00003	< **0.0001**	13.7
R_0	1	0.00231	-0.00032	0.00003	< **0.0001**	78.6
	2	0.00106	-0.00032	0.00003	< **0.0001**	41.2

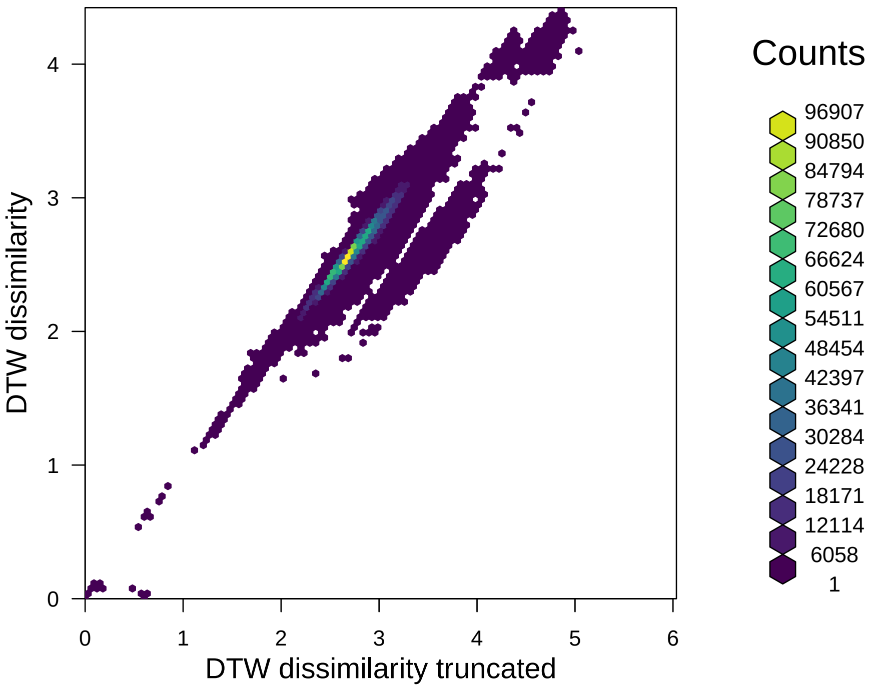

Figure S2: The relationship between dynamic time warping estimates when the time series was truncated to remove the majority of zero values (x-axis) compared to when the entire epidemic time series was used (y-axis). Small variations do exist, but this does not affect our overall findings.

R_0, population size, and epidemic similarity

While we can consider epidemics in US counties as being quasi-isolated, with travel restrictions and differing epidemic timing, it is not possible to control for the inherent link between R_0 (which is estimated from epidemic time series themselves) and population size (Figure S4) and the resulting epidemic trajectory similarity values obtained from the t-SNE decomposition of the pairwise dynamic time warping matrix of epidemic similarity.

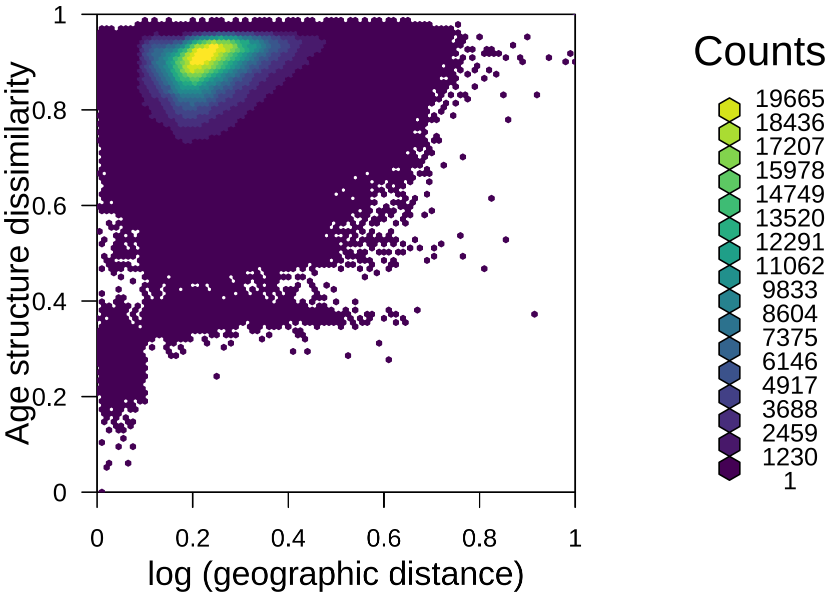

Figure S3: The relationship between geographic distance and age structure dissimilarity.

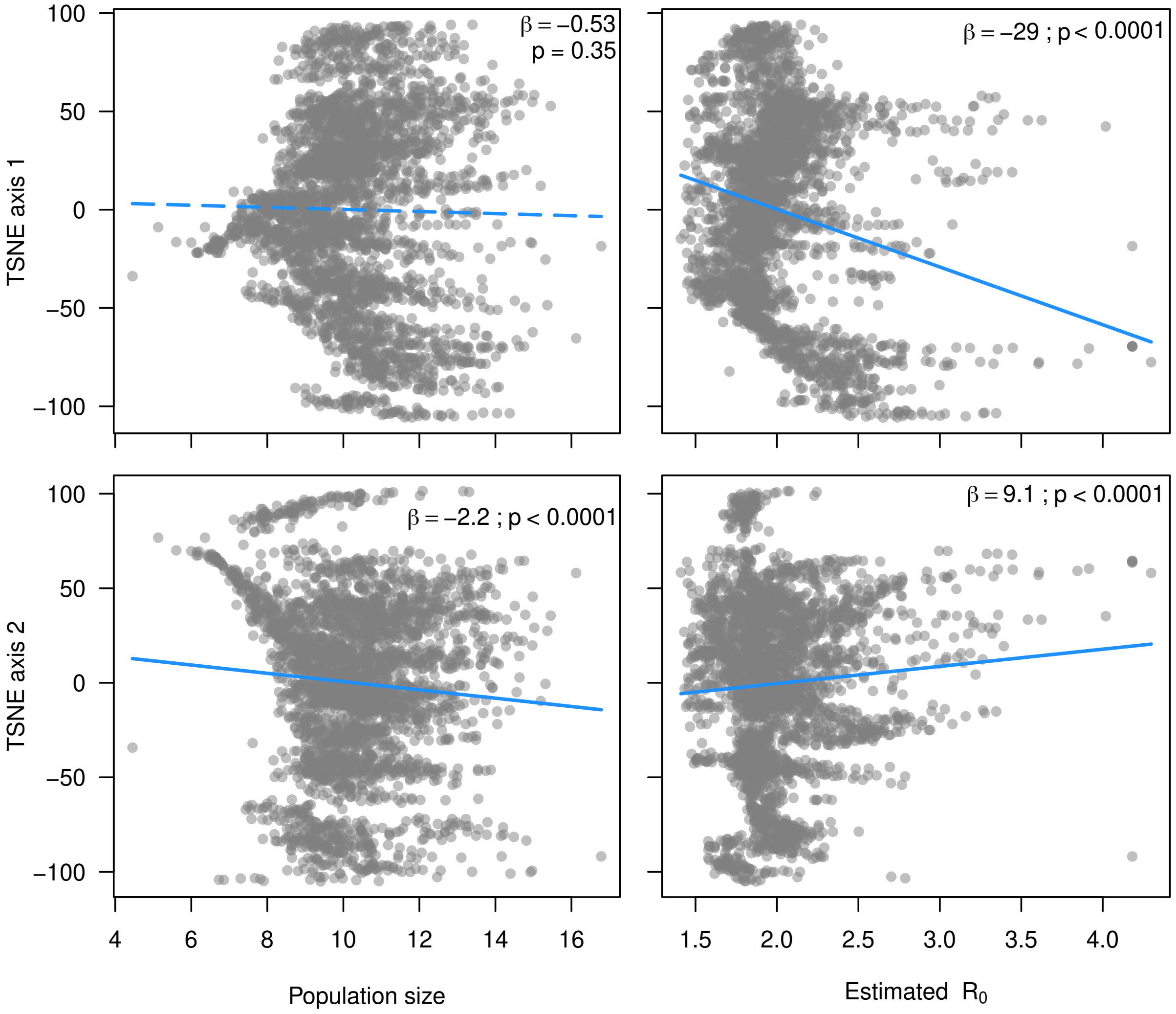

Figure S4: The relationship between epidemic dissimilarity (t-SNE axes as y-axes) and popularion size (first column) and estimated R_0 (second column). Blue lines are linear fits (with associated β and p-values in each panel), where significant lines are solid.

Epidemic similarity as a function of geopolitical boundaries

Epidemic similarity, when compressed to the two t-SNE axes, showed clear US state-level relationships. There are numerous potential reasons for this, including state-level implementations of lockdown orders, variation in state-level testing efforts, and variability in reporting. These are beyond the scope of the current work, but it seems prudent to highlight this variation in t-SNE space (Figure S5).

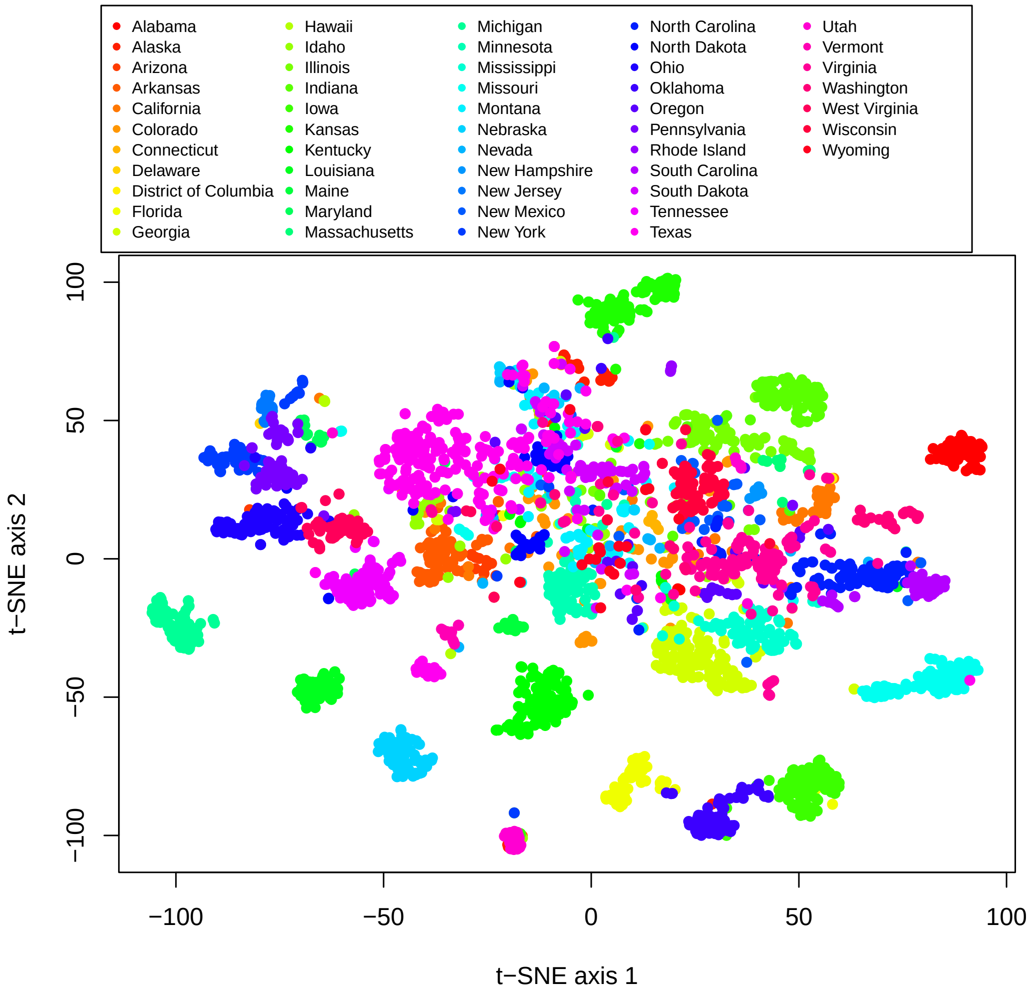

Figure S5: Epidemic similarity in t-SNE space shows clear state-level clustering, suggesting that epidemic similarity was related to some aspect of this geopolitical scale, such as variable mitigation, testing, and reporting efforts.

Mantel Tests

Here, we explore how the pairwise epidemic similarity is related to the distance (or dissimilarity) matrices related to demography and spatial processes. If we claim that z-score as a measure of association between epidemic trajectory similarity and geographic distance, age structure dissimilarity, population size difference, and R_0 difference, then we would conclude that geographic distance and R_0 difference between US counties are the *most* related to epidemic similarity. Each of the distance or dissimilarity matrices were significantly related to the pairwise epidemic dissimilarity matrix. Taking the estimated z-score from the Mantel tests as a measure of association would lead us to conclude that geographic distance was far less important than other matrices. Considering the inherent collinearity between many of these variables, the most salient aspect of this becomes that all of these demographic and spatial factors were significantly related to epidemic similarity.

Table S2: Mantel tests – permutation tests relating two pair-wise dissimilarities to one another – found that geographic distance, age structure dissimilarity, difference in population size, and difference in R_0 were all related to epidemic trajectory dissimilarity.

covariate	z	p
geography	3111220	$<$ **0.001**
age structure	11354792	$<$ **0.001**
population size	11218853	$<$ **0.001**
R_0	12819318	$<$ **0.001**