



OPEN

## Spatial correlations in geographical spreading of COVID-19 in the United States

Troy McMahon<sup>1</sup>, Adrian Chan<sup>2</sup>, Shlomo Havlin<sup>2</sup> & Lazaros K. Gallos<sup>1</sup>✉

The global spread of the COVID-19 pandemic has followed complex pathways, largely attributed to the high virus infectivity, human travel patterns, and the implementation of multiple mitigation measures. The resulting geographic patterns describe the evolution of the epidemic and can indicate areas that are at risk of an outbreak. Here, we analyze the spatial correlations of new active cases in the USA at the county level and characterize the extent of these correlations at different times. We show that the epidemic did not progress uniformly and we identify various stages which are distinguished by significant differences in the correlation length. Our results indicate that the correlation length may be large even during periods when the number of cases declines. We find that correlations between urban centers were much more significant than between rural areas and this finding indicates that long-range spreading was mainly facilitated by travel between cities, especially at the first months of the epidemic. We also show the existence of a percolation transition in November 2020, when the largest part of the country was connected to a spanning cluster, and a smaller-scale transition in January 2021, with both times corresponding to the peak of the epidemic in the country.

The high infection rate of COVID-19<sup>1</sup>, combined with modern society behavioral patterns and a high volume of travel<sup>2</sup>, enabled the rapid spreading of the virus globally<sup>3</sup>. The COVID-19 pandemic is among the few examples of an infectious disease which spread widely in a relatively short amount of time and whose evolution was closely monitored and documented<sup>4,5</sup>. In the absence of restriction measures, the entire population would have been exposed to the virus resulting in a much higher number of active cases and consequently deaths. Many measures have been implemented in different countries and most of them aim to limit contacts between large numbers of people<sup>6</sup> and limiting travel over large distances<sup>7</sup>. These measures include travel restrictions, quarantines, social distancing, lockdowns, curfews, and others<sup>8</sup>. Such restrictions have led to varying degrees of success, which are difficult to estimate since individuals may not have followed state directives or the measures may have been inadequate<sup>9</sup>. However, it is clear that these mitigation measures have had a significant impact on the evolution of the pandemic and that the geographical spreading would have been very different if the virus was left to spread uncontrolled within the population<sup>10</sup>.

The special features of this virus and the unprecedented global response present potentially novel paths of disease transmission that have not been observed in modern times<sup>11</sup>. The combination of these transmission paths manifests itself as the number of new active cases over a region of interest, such as a county in the USA, and these numbers are reported daily. The geographical arrangement of variables such as the number of new cases gives rise to larger scale spatial patterns which span broader areas on the map. The analysis of these patterns and their pertinent features can provide important information on the extent of the epidemic at a given time and areas which may be at higher risk of an outbreak<sup>12–14</sup>. In practice, it is possible to identify geographical clusters whose connectivity is based on similar local levels of infections or similar trends in the local progress of infection<sup>15</sup>. We can then assess how these clusters evolve with time, for example in terms of their size and persistence.

Here, we suggest that the use of spatial correlation statistics and cluster analysis of spreading indicators can provide valuable information on the extent of spatial spreading and how spatial correlations arose within and between geographical areas. Similar approaches have been shown to be very successful in other contexts of spreading<sup>16,17</sup>, where they have provided important results and insight in problems related to spatial epidemics.

We find that during the one year of spreading from February 2020 to February 2021, when vaccinations started becoming widely available, there were three main phases in terms of spatial spreading, which can be roughly described as localized, dormant, and system-wide outbreak. Interestingly, if we consider the whole

<sup>1</sup>DIMACS, Rutgers University, Piscataway, NJ 08854, USA. <sup>2</sup>Department of Physics, Bar-Ilan University, 52900 Ramat Gan, Israel. ✉email: lgallos@gmail.com

country to represent one system, then these three phases can be compared to the progression of a disease in an individual, moving from an acute infection to false recovery to severe illness. In spring 2020 (the localized, acute phase), spreading was contained within small clusters and there were only a few local outbreaks mainly located in the Northeast. From May 2020 to October 2020, (the incubation phase) correlations in new active cases were weaker across the country but at the same time the underlying clusters started growing in size while still remaining mostly localized in space. November 2020 marks the beginning of the third stage (the system-wide outbreak phase), when correlations spanned the largest part of the country as these clusters merged. Even though this spanning cluster dissolved in less than a month, correlations remained strong for the rest of this time interval indicating that virus transmission could still increase at a fast pace.

It is also noteworthy that correlations among neighboring urban centers remained strong throughout this year. Conversely, until November 2020 there were only weak or no correlations between rural areas, even for counties which are geographically close to each other. After November 2020, these correlations increased in strength and became comparable to those between urban counties. This behavior can also explain the percolation transition to a cluster spanning the largest part of the country at that time.

## Methods

We analyze COVID-related data for the USA at the county level using the Johns Hopkins dataset<sup>5,18</sup>. The main quantity we study is the number of new COVID cases. We use a 7-day window in order to alleviate problems with inconsistent data reporting, such as weekend vs weekday testing patterns. Starting on February 1 2020, we aggregate the total number of newly infected cases in a given county over the previous 7 days (including the given day) and calculate the daily average during this time period. We then convert this number to the average daily fraction of the population in each county that was infected during this week by dividing with the county population. In short, if  $z_t(i)$  denotes the number of new cases in county  $i$  on day  $t$ , then for week  $T$  we calculate the fraction  $Z_T(i)$  as:

$$Z_T(i) \equiv \frac{1}{7p_i} \sum_{t=7T-6}^{7T} z_t(i), \quad (1)$$

where  $p_i$  is the county population. We remove 697 counties with a population less than 10,000 because a small change in the number of cases in a small population can lead to large fluctuations, which results in a total of 2411 counties in our calculations. In this way, we create weekly maps for the infection rates of each county  $Z_t(i)$  for the time period from February 1, 2020 to February 1, 2021. Some of the resulting maps are shown in Fig. 1a. As expected, these maps indicate that the spatial coverage of the virus is not uniform but incidents are geographically clustered. These clusters change significantly with time, both in terms of their size and location. Our main goal here is to quantify these clusters through a spatial correlation analysis<sup>19</sup>, so that we can detect the evolution of spreading and the potential impact of restriction measures.

We start by detrending the data for the correlation analysis. We use the differences method<sup>20</sup> where we consider the difference of  $Z$  between two consecutive weeks  $T$  and  $T - 1$ , so that

$$X_T(i) = \Delta Z_T(i) = Z_T(i) - Z_{T-1}(i). \quad (2)$$

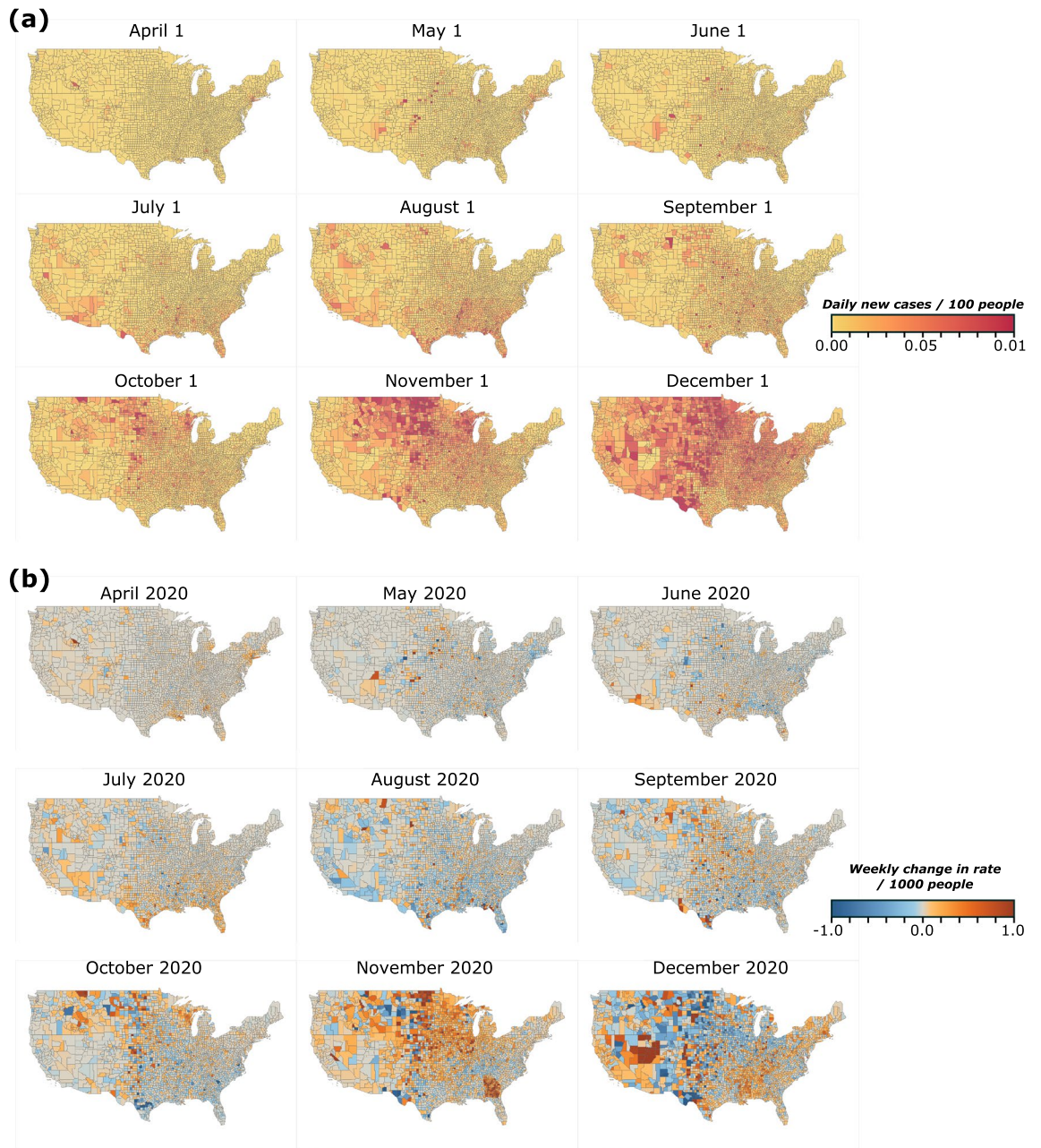
In practice,  $X_T(i)$  measures the extent to which the relative number of cases increased or decreased on a given week compared to the previous week. Figure 1b shows maps of the values of  $X_T(i)$  at different times.

We start by constructing a set  $N(r)$  which includes all pairs of counties which are at a distance  $r$  from each other, as measured between the centroids of two counties. Because of the natural inhomogeneity in the geographical distribution of the counties,  $N(r)$  includes pairs of counties within a distance  $[r, r + dr]$  from each other. In practice, we bin the distances logarithmically, so that  $dr$  increases by a factor 1.05 between successive bins after fixing the first bin in the range  $[0 \text{ km}, 75 \text{ km}]$ . The pairs  $(I, J)$  in this set represent a vector from county  $I$ , at the tail of the vector, to county  $J$ , at the head of the vector. When we average over the entire country, the set includes vectors in both directions, so that both  $(I, J)$  and  $(J, I)$  appear in the set. When we study correlations at a local level, such as in Northeast, then we restrict the set to those vectors whose tail originates in this area. For example, if counties  $I$  and  $J$  are in the Northeast and county  $K$  is in the South, then the set would include pairs  $(I, J)$ ,  $(J, I)$ ,  $(I, K)$ , but not  $(K, I)$ . In the following, for each pair of counties  $(s_1, s_2)$  in  $N(r)$  we denote the tail county as  $s_1$  and the head county as  $s_2$ . The number of pairs in the set is denoted by  $|N(r)|$ . There is no explicit dependence on time in our calculations and we use information only of a given snapshot in time, typically corresponding to the duration of a week  $T$ . We then repeat the correlation calculations at different times. We define the average value of  $X_T$  for all counties  $s_1$  that appear at the head of pairs in  $N(r)$  and for those counties  $s_2$  that appear at the tail of those pairs as follows:

$$m_{T,s_1} = \frac{\sum_{(s_1,s_2) \in N(r)} X_T(s_1)}{|N(r)|}, \quad m_{T,s_2} = \frac{\sum_{(s_1,s_2) \in N(r)} X_T(s_2)}{|N(r)|}. \quad (3)$$

These two values coincide when we consider the entire country, but they are different when we focus on a smaller geographical area, as described above. We can define the corresponding variances as:

$$\sigma_{T,s_1}^2 = \frac{\sum_{(s_1,s_2) \in N(r)} (X_T(s_1) - m_{T,s_1})^2}{|N(r)|}, \quad \sigma_{T,s_2}^2 = \frac{\sum_{(s_1,s_2) \in N(r)} (X_T(s_2) - m_{T,s_2})^2}{|N(r)|}. \quad (4)$$



**Figure 1.** Evolution of COVID-19 spreading in the continental US. (a) Average daily rate of new infections from April to December 2020 for the first week of the month at the county level. (b) Maps of the quantity  $X_T(i)$ , which corresponds to the change in the daily rate of infections between two consecutive weeks in the beginning of each month. Red color indicates an increase compared to the previous week and blue color indicates that the rate decreased.

Finally, we consider the equal-time two-point correlation function on week  $T$ ,  $C_T(r)$ , which is the average of the correlation of  $X_T$  over all counties at distance  $r$ . For the calculation of  $C_T(r)$  we use the following formula

$$C_T(r) \equiv \frac{\frac{1}{|N(r)|} \sum_{(s_1, s_2) \in N(r)} X_T(s_1) X_T(s_2) - m_{T, s_1} m_{T, s_2}}{\sqrt{\sigma_{T, s_1}^2 \sigma_{T, s_2}^2}}, \quad (5)$$

where the average values  $m$  and variances  $\sigma$  have been defined in Eqs. (3) and (4).

The function  $C_T(r)$  in Eq. (5) is an indicator of how correlation decays with distance on week  $T$  as we move away from a given point in space<sup>21</sup>. The correlation length,  $\xi$ , is then defined as the minimum distance where this function assumes a value of 0, i.e.  $C_T(\xi) = 0$ <sup>22</sup>. Long-range correlations are manifested by a large correlation length, while  $\xi$  vanishes for a random distribution. In the context of an epidemic process, long-range correlations are a hallmark of virus transmission through travel between distant places<sup>23</sup>. If travel was severely limited and

people could only interact locally then the evolution of the disease at different areas would be largely independent of each other which would be manifested by weak correlations and small correlation lengths.

## Results

**Use of correlation functions to describe the spatial extent and intensity of epidemics.** The evolution of the COVID epidemic has been highly inhomogeneous and did not follow spatial and temporal patterns of typical infectious diseases<sup>24</sup>. The origin of the outbreak in large cities, which receive a lot of travel from other countries, combined with the implementation of quarantine and other travel restriction measures at relatively early stages of the outbreak resulted in the general absence of infected clusters persistently spanning a significant fraction of the country<sup>25</sup>. To study the extent of these areas we calculate the correlation function  $C(r)$  as a function of the distance  $r$  at different times, where we include the 2411 counties with a population greater than 10,000 people. As expected, correlation values start at  $C(0) = 1$  and decay to 0, which also defines the correlation length  $\xi$  as the shortest distance where  $C(r) = 0$ . In Fig. 2a we calculate  $C(r)$  for the first week of each month from March 2020 to February 2021. It is obvious that correlation varies significantly over time, both in terms of the local correlation strength at small distances (typically less than 100 km) and in terms of the correlation length. In Fig. 2b we focus on some of the strongest correlations and use double-logarithmic axes to highlight their different behavior. For example, during July 2020 the local correlations are weak but persist over a distance of 800 km, while in April 2020 the local correlations were much stronger and decayed faster with a correlation length of 400 km.

We found that the epidemic patterns changed significantly over time, as shown in Fig. 2c where we isolate four  $C(r)$  curves per plot for clarity. In the first phase of the epidemic, from March to June 2020, correlations fluctuated around zero with the striking exception of April when local correlations were strong and decayed fast. Local correlations were higher in the next time interval, from July to October 2020, but remained relatively weak. However, the correlation length increased significantly during this time. While the relatively low infection numbers during summer seemingly suggest that the country is “recovering” from the pandemic, the increasing correlation length tells us the opposite—that the virus is silently taking root everywhere, which will lead to the eventual outbreak in the next phase. This indicates that the correlation length is not just a different way to look at the raw numbers but can reveal underlying phenomena that can serve as warning signs to policymakers not to relax restrictions too early. Indeed, during the next phase of the epidemic, from November 2020 to February 2021, the local correlations increased significantly and the correlation length remained relatively large. In a fast-spreading epidemic, the expectation is that correlations become strong and extend over long distances, especially if they are facilitated by long-range travel<sup>26</sup>. Here, we see that after an initial peak in April 2020, it took many months for correlations to increase and to remain strong over a long period of time.

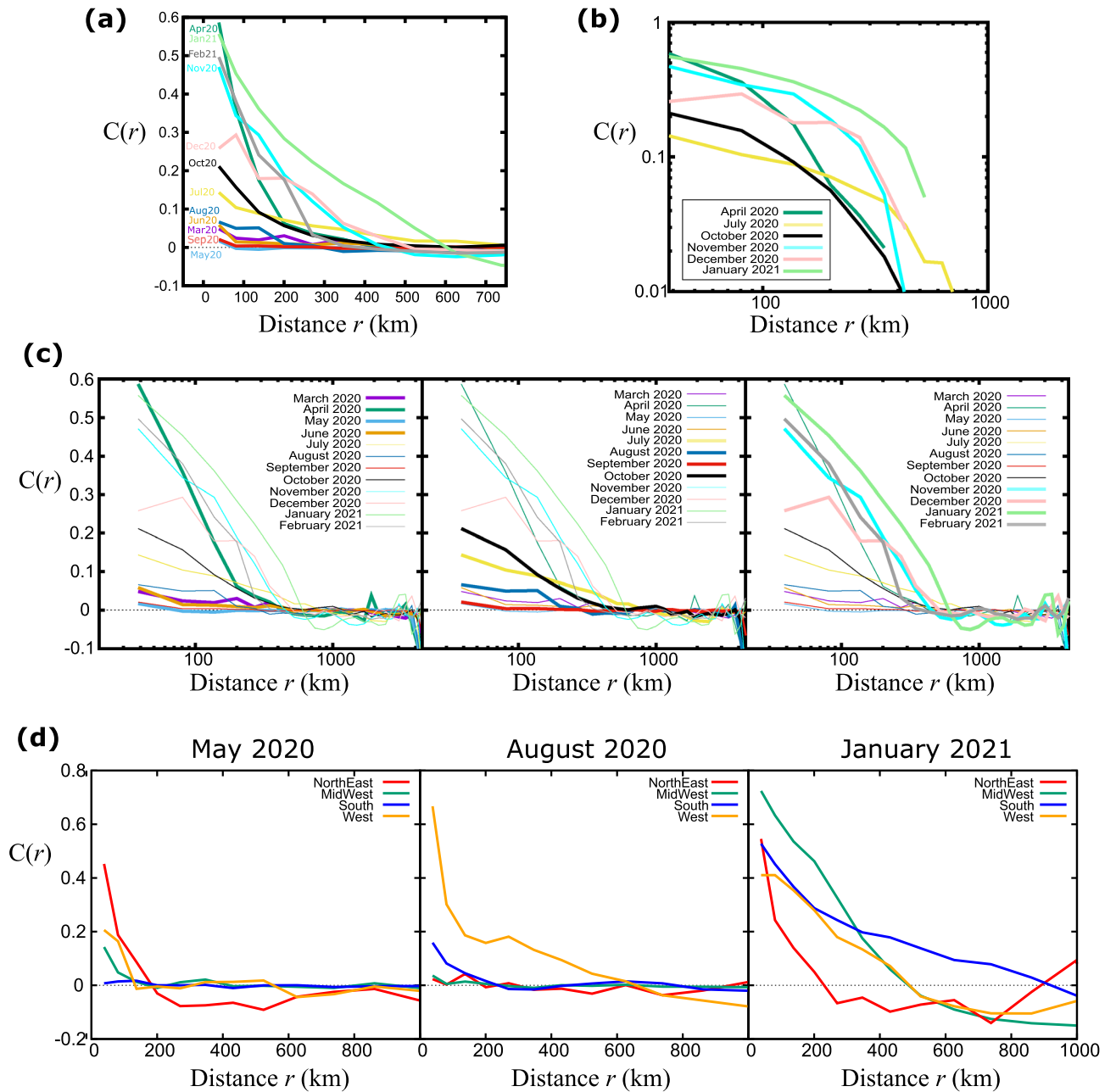
Importantly, these curves are averages over all US counties but the evolution of the epidemic may be different at different regions of the country. In Fig. 2d we calculate correlations in each of the four Census Bureau-designated regions, i.e. Northeast, Midwest, South, and West<sup>27</sup>. In the examples shown, local correlations and correlation lengths vary quite significantly in different regions and while the Northeast dominates in May, the West exhibits much stronger correlations in August.

The evolution of the  $C(r)$  curves carries a lot of information and we need some way to probe this information and compare different instances in time. If the functional form remained the same throughout this process then we could compare any parameters that would appear in a model that could describe the evolution (for example, if these curves could be described by a power-law we could compare the power-law exponents). As can be seen in the various plots of Fig. 2, the dependence of  $C(r)$  on  $r$  cannot be described by the same form. For example, in Fig. 2b some curves can be described by a modified power-law while others are fitted better by an exponential form, but there is not a uniform description over the entire time interval. Therefore, we choose to characterize this behavior by comparing both the strength of ‘local’ correlations, i.e. the value of  $C(r)$  at small distances around 50 km, and the spatial extent of correlations through the value of the correlation length. In a sense, these two parameters capture the basic trends that we are interested in these plots: all curves decay from a given value at  $C(r < 50 \text{ km})$  down to  $C(\xi) = 0$ . Even though we lose the exact form of the decay, these two points define in very broad terms the extent and strength of the correlation function and here we are not concerned about the curvature of the line in the intermediate regime. The combination of these two parameters can already inform on whether counties tend to belong to large clusters and whether the influence within these clusters is strong or weak.

**Correlation function encapsulates the temporal evolution of the disease.** How do we characterize the COVID evolution to determine if spreading expands geographically or if it shrinks? As discussed above, the correlation length itself is important but is not fully adequate. Even if  $\xi$  is large, it is possible that correlations in shorter distances could be weak and, as a result, less influential. In Fig. 3a we plot the correlation function continuously from Feb 1, 2020 to Feb 1, 2021 in weekly increments. The correlation length is marked as the distance where  $C(r)$  becomes zero on each week. The correlation length fluctuates significantly but we can distinguish two peaks in April and July, followed by a rather constant high value from October to January. Time periods with stronger local correlations are identified by red color. Even though many of the correlation values may seem small, these are averages over the entire country where spreading may be very inhomogeneous.

The consideration of these two parameters allows us to plot the ‘trajectory’ of the epidemic in a phase space of  $C(r < 50 \text{ km})$  and  $\xi$  (Fig. 3b), where we plot these two quantities continuously for different points in time. To conceptualize this plot we can split the phase space into four quadrants. When  $C(r)$  is small, i.e. at the bottom quadrants, correlations are weak. The large correlation lengths found at the quadrants at the right, indicate that epidemic can spread easily across large distances. From the diagram in Fig. 3a we can see that in April 2020 there was a significant increase in local correlation strength but correlation length was still relatively small. The

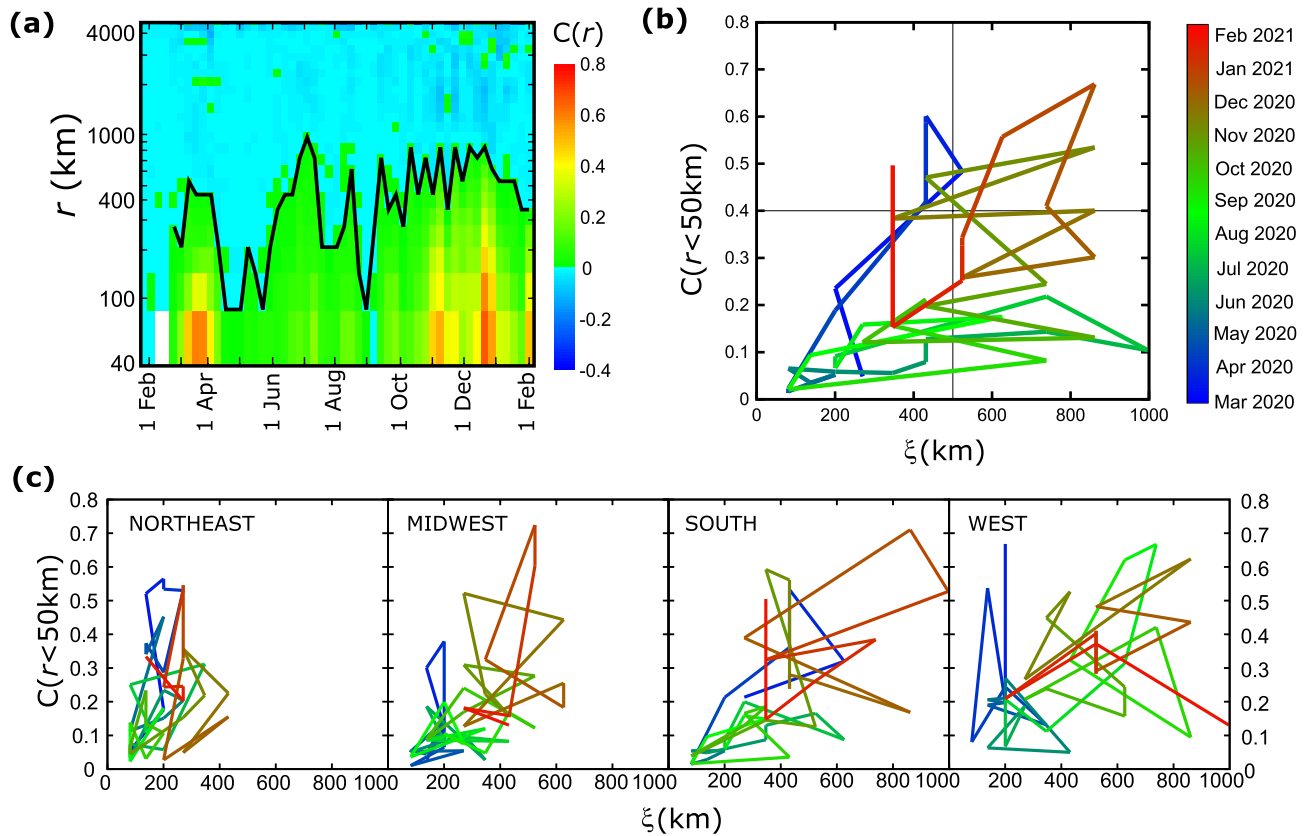




**Figure 2.** Decay of the correlation function with distance. **(a)** The correlation function as a function of the distance, averaged over all counties in the United States. Different curves correspond to different instances of time, for the first two weeks of the month from March 2020 to February 2020. **(b)** Same plot as in **(a)** in double logarithmic axes for select months. **(c)** For clarity, we split the plot in **(a)** into three different time intervals and focus on four months within these intervals. **(d)** Comparison of the correlation functions, averaged over all counties within the four US regions, for May 2020, August 2020, and January 2021.

trajectory returned close to the origin (indicating a random distribution of cases) until the end of summer, when both the intensity and correlation length increased significantly. The trajectory has remained at the area of the upper right quadrant from October 2020 until the end of the year, which indicates extended correlations over space. Starting in January 2021, both the correlation length and the correlation strength decreased considerably, with an abrupt increase of  $C(r < 50 \text{ km})$  at the beginning of February 2021.

These trajectories can shed light on the different evolution of spreading in different areas of the country (Fig. 3c). For example, the correlation length in the Northeast has remained relatively short (this region is also the smallest) but the local correlation strength increased significantly in April 2020 and January 2021. This shows that at the early stages of the epidemic there was a strong cluster in this area but it remained localized. In contrast to that, the correlation length in the West was consistently large since June 2020, with spikes of the intensity between September 2020 and January 2021. The behavior in the Midwest was also different. The trajectory there

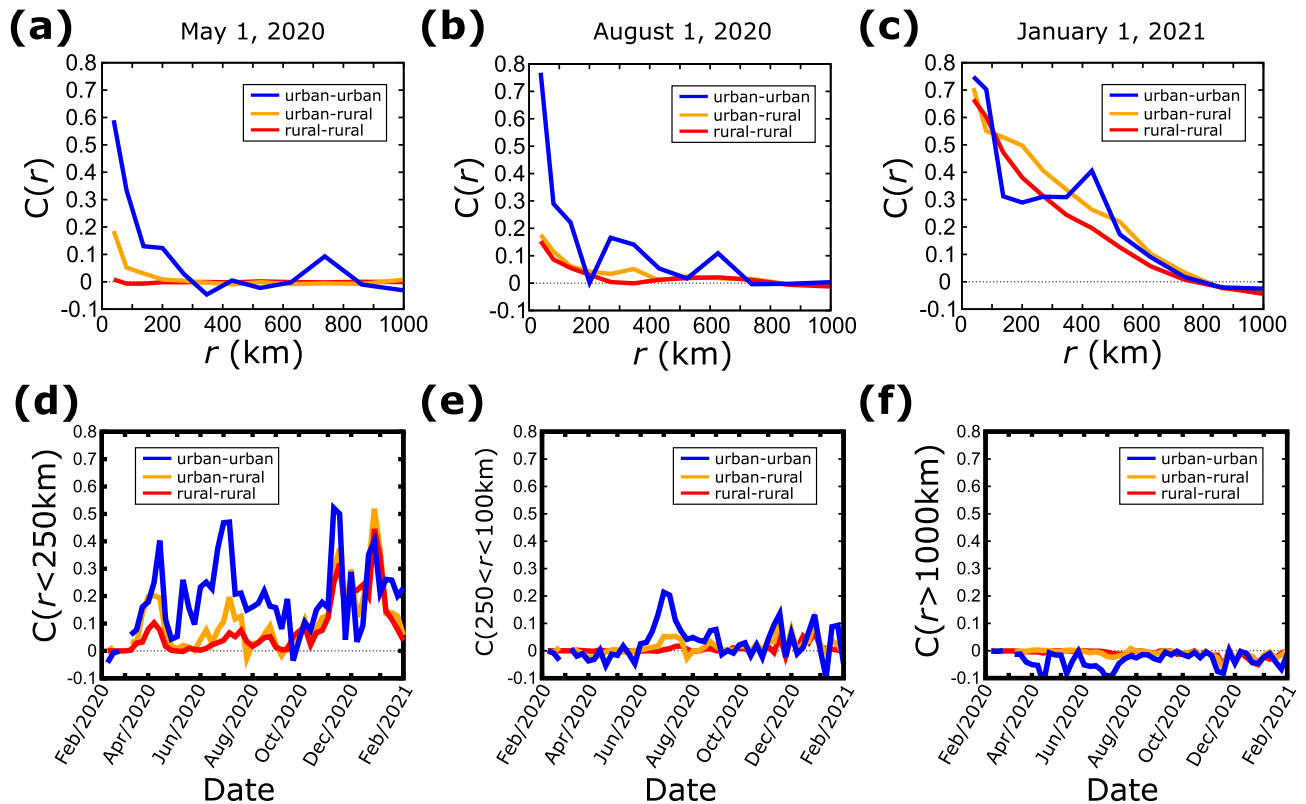


**Figure 3.** Time evolution of the correlation function. **(a)** The value of the correlation function is shown as a function of the date (x-axis) and distance (y-axis). Red and green colors correspond to positive values while blue colors indicate negative values. The line that separates the two regimes describes the correlation length as a function of time. **(b)** The ‘trajectory’ of the epidemic is plotted by the values of local correlations  $C(r < 50 \text{ km})$  vs the correlation length,  $\xi$ , at different times for the entire country. **(c)** Same trajectory plots as in **(b)** for the four regions of US.

remained close to the origin from the beginning until October 2020 when both the intensity and correlation length increased. Notice that in all regions the correlation length in November 2021 grew larger, which is an indication that local clusters joined into a larger cluster spanning the majority of the country.

*Cities are the main drivers of the epidemic.* In typical epidemic processes, it is expected that correlations are higher in areas geographically close to each other, since the main method of transmission is close contact between individuals<sup>28</sup>. Travel represents another important mechanism which contributes to long-range correlations, where now the underlying mechanism is the direct transfer of the virus over a large distance through air or ground travel<sup>29</sup>. To determine the contribution of travel we calculated the correlation as a function of the distance in the case of urban areas vs rural areas. Here, we set an arbitrary criterion for an urban county as one with a population larger than 250,000 people and a rural county with a population less than that. Using this threshold, there are 273 urban and 2138 rural counties. We then calculate the correlation function for the new active cases between urban counties only, between rural counties, or between rural and urban. In Fig. 4a, we plot  $C(r)$  for these three cases for the week of May 1, 2020. There are practically no correlations between rural areas even for short distances and there is only weak correlation between rural and urban areas, and this is true only for short distances. In contrast, correlations between urban places are much higher for distances up to roughly 300 km. This is an important observation because at that time air travel was severely limited and passengers were heavily screened, while car transportation was not restricted. Therefore, there are two possible explanations. Assuming that the first significant center of the epidemic was located in the urban NYC area, either virus transmission was facilitated through car transportation from city to city or the city lifestyle made the virus spread easier in an urban environment and cities presented similar behavior independently of each other. Interestingly, in Fig. 4a correlations between urban places that are farther than 300 km become largely uncorrelated, which favors the idea of local transmission through ground transportation.

Similar results were obtained at other times, such as in August 2020 shown in Fig. 4b, when the difference between cities and city-rural areas was more pronounced. Correlations between urban counties remained stronger. However, travel had increased during summer and the restriction measures were relaxed, which in the plot is manifested by smaller peaks at 300 km and 600 km. In January 2021 (Fig. 4c), the differences vanished and



**Figure 4.** Spatial correlations between areas of different population (a–c) Correlation function  $C(r)$  between urban–urban, urban–rural, and rural–rural counties, as a function of the distance at three different times. (d) Average correlation function between urban–urban, urban–rural, and rural–rural counties which are within 250 km from each other, as a function of time. (e) Same plot as in (d) for county distances between 250 and 1000 km. (f) Same plot as in (d) for distances longer than 1000 km.

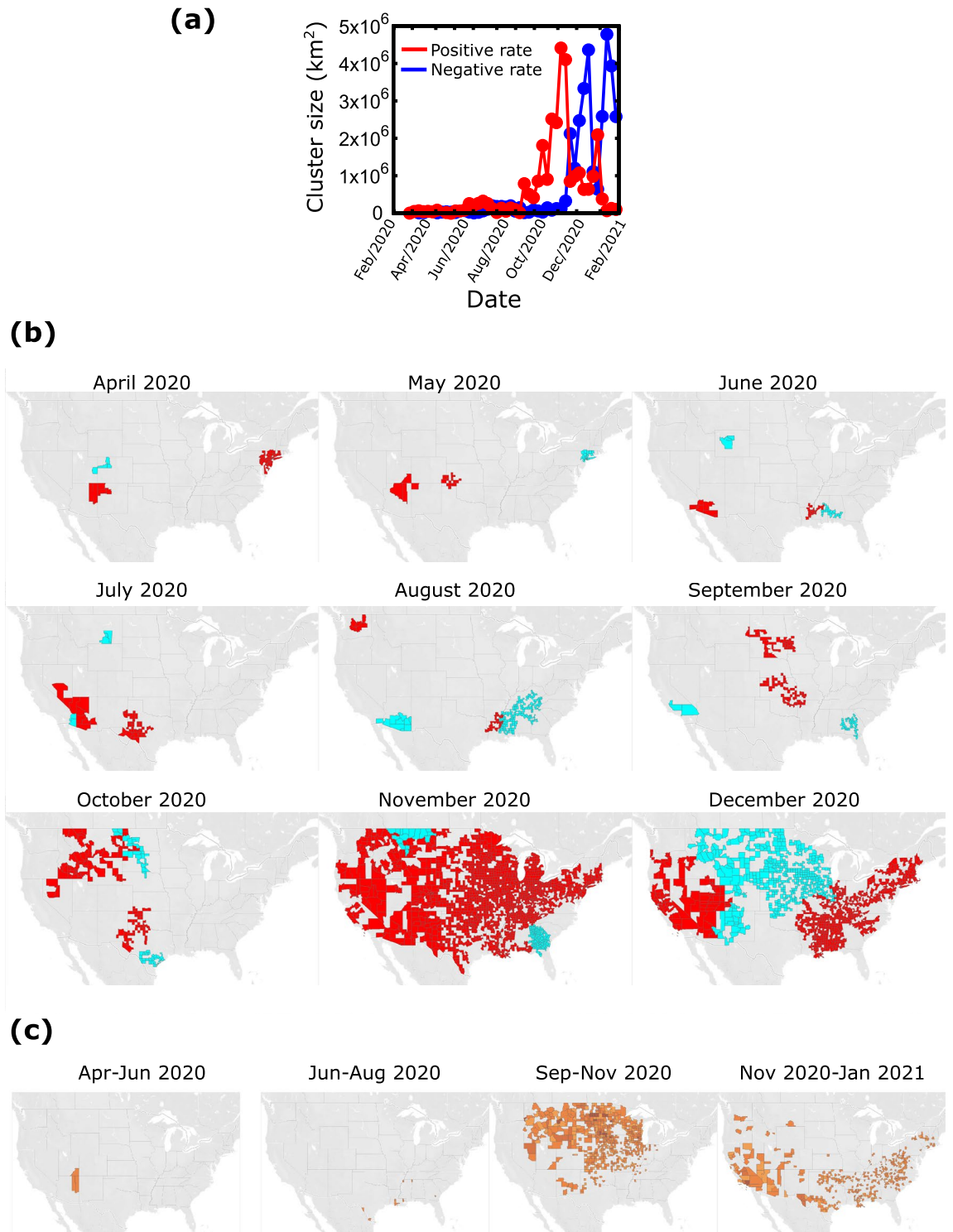
correlations were practically independent on the urban character of the counties. At that time, the correlation length increased to 800 km indicating again a global outbreak covering the largest part of the country.

To explore the time evolution of these observations we calculated the average correlation  $\langle C(r) \rangle$  for the three cases (urban–urban, urban–rural, and rural–rural) over three distance intervals,  $r < 250$  km (Fig. 4d),  $250 \text{ km} < r < 1000$  km (Fig. 4e), and  $r > 1000$  km (Fig. 4f). When we consider short distances the correlation between urban areas is consistently higher than in the other two cases. In fact, from February to October 2020 the spreading in rural areas was uncorrelated when rural areas were involved which shows that the epidemic was largely contained in urban environments. After October 2020, correlations became stronger in all cases for short distances, and the behavior was consistent independently of the rural or urban character of the area.

In the case of intermediate distances (Fig. 4e) the value of  $\langle C(r) \rangle$  mostly fluctuates around 0 with a small peak of inter-city correlations from June to August 2020. This plot supports the idea of ground transportation transmission since long road trips are much more rare than local road trips and in these distances air travel is typically the preferred mode of transportation. For distances longer than 1000 km, the inter-city correlations are consistently zero or weakly anti-correlated. This shows that not only spreading in the west coast was different than in the east coast but in general they had opposite trends (notice that these are averages over the entire country and it is possible that locally correlations may be much stronger or weaker than these averages).

As shown above, in November the average correlation length for the entire country was of the order of 800–1000 km which is comparable to the size of the system. For reference, a 1000 km-radius circle centered in the middle of the continental US would cover roughly half of the country (the distance from New York to Los Angeles is around 4000 km). This is a strong indication of a widespread epidemic which approaches the percolation threshold.

**Percolation and cluster analysis of the epidemic.** To determine how close the epidemic came to percolating throughout the country and when this happened we performed a clustering analysis<sup>30</sup>. For a given point in time, we consider all clusters created by connecting counties whose weekly difference  $\Delta Z_T(i)$  exceeds a given threshold. We create two types of clusters, depending on whether these differences are positive or negative. Positive clusters are areas where the epidemic has increased significantly over the previous week and negative clusters are areas where it decreased. We determine the size of a cluster by the total geographic area covered by the counties that comprise this cluster. Figure 5a shows that the largest clusters were relatively small and localized until September 2020, which was followed by a rapid increase in the size of the positive cluster which within two months covered an area of around  $4.5 \times 10^6 \text{ km}^2$  (the total area of the 48 contiguous states is roughly  $8 \times 10^6 \text{ km}^2$ ).



**Figure 5.** Clustering analysis of the correlations (a) Total area of the largest clusters as a function of time. The red line corresponds to clusters of increasing number of cases and the blue line corresponds to clusters of decreasing cases. (b) The maps show the two largest clusters with positive rate (red) and the two largest clusters of negative rate (blue) at different times. (c) ‘Persistence’ maps. These maps show the counties which remained in positive clusters for more than half of the time period indicated on the map.

This cluster was dissolved within two weeks and its size remained small with an exception of a smaller peak during January 2021. The size of the negative cluster remained small until after the large positive cluster was formed in November 2020. It is interesting that the two processes of positive or negative change above a given threshold



are not completely synchronized and cannot fully explain each other, i.e. the extent and location of the negative clusters do not necessarily follow the positive cluster. This can be seen in the largest decrease of the positive cluster size compared to the smallest increase of the negative cluster in December 2020, as well as the second peak of the negative cluster in January 2021 which was not preceded by a positive cluster of comparable size.

In the maps of Fig. 5b we present the two largest positive and the two largest negative clusters at different times. In agreement with previous observations the clusters from April to October 2020 are relatively small. In October 2020, the clusters started getting bigger to the west and south of the country and in November 2020 there was a transition to a country-spanning positive rate cluster. This cluster started dissolving in December 2020, when the middle part of the country was connected through a negative rate cluster leaving connected areas of positive rates at the east and west parts of the country.

Clustering can also be used to identify areas of the country where the epidemic persisted the longest time as part of a large cluster. For this, we considered all the counties which belonged to a cluster of at least 10,000 km<sup>2</sup> for a minimum of 1.5 months during a three-month period. The results in Fig. 5c show that from April to August 2020 there were only few isolated counties involved in the largest clusters. Contrary to that, a large part at the north of the country was consistently included in large clusters during the period of September to November 2020. For the next time period, cases were consistently increasing throughout the southern part of the country and mainly in the southwest. In general, we see that the epidemic has been spreading quickly but has not persisted over extended areas for more than a few months, perhaps as a result of implementing restriction measures when active cases were increasing at a local level.

## Discussion

One loose interpretation of a correlation-based spatial cluster is that the virus may be transmitted more easily within counties in this cluster, compared to counties whose active cases are uncorrelated<sup>31</sup>. This is a phenomenological manifestation of the underlying transmission mechanisms, even though these mechanisms are not explicitly known<sup>32</sup>. This macroscopic evaluation of the epidemic footprint can identify areas of a country where spreading is highly correlated and since spatial correlations can exist almost independently of the current level of the number of active cases, they can potentially be used as warning signals. For example, a careful evaluation of Fig. 3 shows that the correlation length has grown significantly a few weeks before the number of cases has increased. However, this observation has not been verified in all cases and it is very difficult at this stage to determine whether correlations can be used as true predictors of the future trajectory of the epidemic, especially given the many and different local restriction measures.

Using spatial statistical analysis, our results indicate that the COVID-19 epidemic in the continental USA went through different phases. The first localized clusters started dissolving quickly, which indicates that there was not any significant long-distance transmission during spring 2020. By the summer of 2020, many local clusters started emerging whose size was continuously increasing and by November 2020 they had merged into a country-spanning cluster. This formation was short-lived and even though local correlations remained strong, the global correlation length started decreasing.

A similar approach can be applied to smaller areas, such as regions or states. We have created an online tool, where the user can select individual states for further analysis. This analysis could provide additional information on how the epidemic evolved at a smaller geographic scale and a possible extension of this work is to relate differences in virus spreading with state-level mitigation measures.

## Data availability

The COVID-19 datasets are available under a CC by 4.0 license and were downloaded from the "COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University", <https://github.com/CSSEGISandData/COVID-19>.

Received: 27 July 2021; Accepted: 14 December 2021

Published online: 13 January 2022

## References

1. Levin, A. T. *et al.* Assessing the age specificity of infection fatality rates for COVID-19: Systematic review, meta-analysis, and public policy implications. *Eur. J. Epidemiol.* 1–16 (2020).
2. Chinazzi, M. *et al.* The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* **368**, 395–400 (2020).
3. Sohrabi, C. *et al.* World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). *Int. J. Surg.* **76**, 71–76 (2020).
4. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/>. Accessed 31 Dec 2021.
5. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **20**, 533–534 (2020).
6. Xiao, Y. & Torok, M. E. Taking the right measures to control COVID-19. *Lancet Infect. Dis.* **20**, 523–524 (2020).
7. Linka, K., Peirlinck, M., Sahli Costabal, F. & Kuhl, E. Outbreak dynamics of COVID-19 in Europe and the effect of travel restrictions. *Comput. Methods Biomech. Biomed. Eng.* **23**, 710–717 (2020).
8. Schlosser, F. *et al.* COVID-19 lockdown induces disease-mitigating structural changes in mobility networks. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 32883–32890 (2020).
9. Anderson, R. M., Heesterbeek, H., Klinkenberg, D. & Hollingsworth, T. D. How will country-based mitigation measures influence the course of the COVID-19 epidemic?. *Lancet* **395**, 931–934 (2020).
10. Lau, H. *et al.* The positive impact of lockdown in Wuhan on containing the COVID-19 outbreak in China. *J. Travel Med.* **27**, taaa037 (2020).
11. Franch-Pardo, I., Napoletano, B. M., Rosete-Verges, F. & Billa, L. Spatial analysis and GIS in the study of COVID-19. A review. *Sci. Total Environ.* **739**, 140033 (2020).

12. James, N., Menzies, M. & Bondell, H. Understanding spatial propagation using metric geometry with application to the spread of COVID-19 in the United States. *EPL (Europhysics Letters)* **135**, 48004 (2021).
13. Cuadros, D. F., Branscum, A. J., Mukandavire, Z., Miller, F. D. & MacKinnon, N. Dynamics of the COVID-19 epidemic in urban and rural areas in the United States. *Ann. Epidemiol.* **59**, 16–20 (2021).
14. Gross, B. *et al.* Spatio-temporal propagation of COVID-19 pandemics. *EPL (Europhysics Letters)* **131**, 58003 (2020).
15. Pfeiffer, D. *et al.* *Spatial Analysis in Epidemiology* (Oxford University Press, 2008).
16. Gallos, L. K., Barttfeld, P., Havlin, S., Sigman, M. & Makse, H. A. Collective behavior in the spatial spreading of obesity. *Sci. Rep.* **2**, 1–9 (2012).
17. Alves, L. G., Lenzi, E. K., Mendes, R. S. & Ribeiro, H. V. Spatial correlations, clustering and percolation-like transitions in homicide crimes. *EPL (Europhysics Letters)* **111**, 18002 (2015).
18. <https://github.com/CSSEGISandData/COVID-19>. Accessed 31 Dec 2021.
19. Schuenemeyer, J. H. & Drew, L. J. *Statistics for Earth and Environmental Scientists* (Wiley, 2011).
20. Brockwell, P. J. & Davis, R. A. *Time Series: Theory and Methods* (Springer Science & Business Media, 2009).
21. Stanley, H. Introduction to Phase Transitions and Critical Phenomena The International Series of Monographs on Physics Oxford University Press Inc. London, UK (1971).
22. Cavagna, A. *et al.* Scale-free correlations in starling flocks. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 11865–11870 (2010).
23. Crépey, P. & Barthélemy, M. Detecting robust patterns in the spread of epidemics: A case study of influenza in the United States and France. *Am. J. Epidemiol.* **166**, 1244–1251 (2007).
24. Flaxman, S. *et al.* Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584**, 257–261 (2020).
25. Gao, S., Rao, J., Kang, Y., Liang, Y. & Kruse, J. Mapping county-level mobility pattern changes in the United States in response to COVID-19. *SIGSpatial Spec.* **12**, 16–26 (2020).
26. Colizza, V., Barrat, A., Barthélemy, M. & Vespignani, A. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 2015–2020 (2006).
27. [https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us\\_regdiv.pdf](https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf). Accessed 31 Dec 2021.
28. Read, J. M. & Keeling, M. J. Disease evolution on networks: The role of contact structure. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* **270**, 699–708 (2003).
29. Riley, S. Large-scale spatial-transmission models of infectious disease. *Science* **316**, 1298–1301 (2007).
30. Havlin, S. & Nossal, R. Topological properties of percolation clusters. *J. Phys. A Math. Gen.* **17**, L427 (1984).
31. Bunde, A. & Havlin, S. *Fractals and Disordered Systems* (Springer Science & Business Media, 2012).
32. Balcan, D. *et al.* Modeling the spatial spread of infectious diseases: The GLOBAL Epidemic and Mobility computational model. *J. Comput. Sci.* **1**, 132–145 (2010).

## Acknowledgements

This work was supported by a joint NSF-BSF grant. TM and LKG were supported by NSF through DEB-2035297. AC and SH were supported by BSF through Grant 2020645.

## Author contributions

All authors contributed equally to the work presented in this paper. All authors reviewed the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to L.K.G.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022