Accepted March 28 2020 to Applied Geography

Michael Desjardins, Alexander Hohl and Eric Delmelle
John Hopkins University, University of Utah and University of North Carolina at Charlotte

Rapid surveillance of COVID-19 in the United States using a prospective space-time scan statistic: detecting and evaluating emerging clusters

Abstract.

Coronavirus disease 2019 (COVID-19) was first identified in Wuhan, China in December 2019, and is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). COVID-19 is a pandemic with an estimated death rate between 1% and 5%; and an estimated R_0 between 2.2 and 6.7 according to various sources. As of March 28th, 2020, there were over 649,000 confirmed cases and 30,249 total deaths, globally. In the United States, there were over 115,500 cases and 1,891 deaths and this number is likely to increase rapidly. It is critical to detect clusters of COVID-19 to better allocate resources and improve decision-making as the outbreaks continue to grow. Using daily case data at the county level provided by Johns Hopkins University, we conducted a prospective spatial-temporal analysis with SaTScan. We detect statistically significant space-time clusters of COVID-19 at the county level in the U.S. between January 22nd-March 9th, 2020, and January 22nd-March 27th, 2020. The space-time prospective scan statistic detected "active" and emerging clusters that are present at the end of our study periods – notably, 18 more clusters were detected when adding the updated case data. These timely results can inform public health officials and decision makers about where to improve the allocation of resources, testing sites; also, where to implement stricter quarantines and travel bans. As more data becomes available, the statistic can be rerun to support timely surveillance of COVID-19, demonstrated here. Our research is the first geographic study that utilizes space-time statistics to monitor COVID-19 in the U.S.

Keywords. COVID-19; SaTScan; Space-time clusters; pandemic; disease surveillance

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgements

The authors would like to acknowledge Johns Hopkins University Center for Systems Science and Engineering for providing open access to the COVID-19 data. Doing so supports timely research to facilitate surveillance and decision-making in this ongoing pandemic. Next, we thank Dr. Frank Curriero from Johns Hopkins Bloomberg School of Public Health, for his insightful and constructive feedback that improved the quality of this paper. Finally, we would especially like to thank the timely and constructive feedback of the editor and the anonymous reviewers that ultimately improved the quality of this manuscript.

1. Introduction

Coronavirus disease 2019 (COVID-19) was first identified in Wuhan city, Hubei province, China in December of 2019 (Huang et al. 2020; Li et al. 2020) and is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). COVID-19 is a pandemic (cases confirmed in more than 140 territories) with an estimated death rate between 1% and 5% (Roser et al. 2020); and an estimated R_0 between 2.2 and 6.7 (Liu et al. 2020; Sanche et al. 2020). As of March 28th, 2020, there were over 649,000 confirmed cases and 115,500 total deaths, globally. In the United States (U.S.), there were over 115,000 cases and 1,891 deaths (Dong et al. 2020). Approximately 80% of confirmed cases are mild, with symptoms including fever, cough, and shortness of breath (Ruan et al. 2020). Severe cases may experience pneumonia, multi-organ failure, and death (Mahase 2020). The vast majority of deaths from COVID-19 are those with preexisting conditions (e.g. hypertension and heart disease), are immunocompromised, or above 60 years old (Wu and McGoogan 2020).

During an emerging infectious disease like COVID-19, it is critical to implement space-time surveillance that can prioritize locations for targeted interventions, rapid testing, and resource allocation. One such method is the space-time scan statistic (Kulldorff 1997), which is widely used to identify significant clusters of disease. Space-time scan statistics supplement and can study basic rate maps of disease by relying on a variety of data models to determine whether the observed space-time patterns of a disease are due to chance or randomly distributed. In other words, scan statistics detect clusters that are outliers (e.g. unexpected clustering given baseline conditions). The statistic utilizes circles or ellipses (scanning window) that are centered on grid points and move (scan) systematically across a study area to identify clusters of cases (each window counts number of aggregated cases per geographic unit). In its space-extension, the location, size, and duration of statistically significant clusters of disease cases are subsequently reported (Desjardins et al. 2018; Owusu et al. 2019; Whiteman et al. 2019; Desjardins et al. 2020).

To routinely monitor outbreaks, the prospective space-time scan statistic (Kulldorff 2001) is one method to detect "active" or emerging clusters of disease, which can be used for surveillance during an ongoing epidemic. The statistic will detect clusters that are "active" at the end of the study period; but as more data (e.g. confirmed cases) becomes available, the statistic can be rerun to confirm the presence and track the clusters in space and time, update relative risks for each location affected by a disease, and detect new emerging clusters. The main purpose of using a prospective statistic rather than retrospective is to only focus on significant clustering that is "active" or present at the time of the analysis; which disregards clusters that may have existed previously, and are no longer a public health threat (Kulldorff 2001). For example, the prospective space-time scan statistic has been utilized to detect emerging clusters of shigellosis (Jones et al. 2006), measles (Yin et al. 2007), thyroid cancer (Kulldorff 2001), and syndromic surveillance (Yih et al. 2010). Since COVID-19 data are updated daily, our approach can contribute to timely monitoring of the pandemic, focusing on the United States in this study.

This study contributes to ongoing COVID-19 surveillance efforts by detecting significant space-time clusters of reported cases at the county level in the U.S. The space-time prospective statistic is especially useful since it detects active and emerging clusters of COVID-19, which can inform public health officials and decision-makers where and when to improve targeted interventions, testing sites, and necessary isolation measures to mitigate further transmission. Our prospective analysis can be rerun each day as new data become available to detect new emerging clusters and identify areas where transmission is decreasing; suggesting where COVID-19 is potentially no longer a public health threat.

To demonstrate the notion of detecting new emerging clusters when adding updated case data using the prospective space-time scan statistic, we report results for two time periods: January 22nd-March 9th, 2020 and January 22nd-March 27th, 2020. Since COVID-19 is a highly infectious disease that can affect all segments of the population, we decided not to adjust for age. However, since the highest proportion of deaths occur among the elderly and those with preexisting conditions, an age-adjusted Bernoulli model accounting for cases and deaths could be conducted but is beyond the scope of this research.

2. Data & Methods

2.1 Data

We collected COVID-19 case and location data from Johns Hopkins University's Center for Systems Science and Engineering GIS dashboard (Dong et al. 2020). These data are freely available on their GitHub page (https://github.com/CSSEGISandData/COVID-19). Temporally, these data are currently updated daily, and we use available data between January 22nd and March 27th, 2020. Spatially, the daily confirmed cases if COVID-19 are aggregated at the county level.

Using the spatial location information in the COVID-19 dataset, we assigned the case counts to the appropriate counties in a geographic information systems compatible file we obtained from the U.S. Census. We focused our analysis on the contiguous 48 states and Washington D.C., excluded cases recorded at the state-level (no county-level information available) and cases diagnosed on the “Grand Princess” and “Diamond Princess” cruise ships. The infected passengers on the cruise ships were sent to various quarantine locations throughout the U.S. and their exact locations are not provided in the dataset. The COVID-19 dataset reports cumulative case counts (Figure 1). Therefore, for each day in the study period, we subtracted the previous day's count (n_{t-1}) from the current day's count (n_t) to obtain the number of new cases.

Figure 1: Cumulative number of COVID-19 cases in the contiguous United States between January 22nd and March 27th, 2020 (used for the statistical analysis).

2.2 Prospective Poisson space-time scan statistic

To identify space-time clusters that are still occurring or "active", we utilize the prospective version of the Poisson space-time scan statistic **(**Kulldorff 2001; Kulldorff et al. 1998) and implemented in SaTScan™ (Kulldorff 2018). As such, we can identify COVID-19 clusters that are still active (excess risk still present) during the last day in our dataset. In other words, we detect space-time clusters of COVID-19 that are emerging and "disregard" clusters in the study period that do not have a statistically significant excess relative risk (i.e. more observed than expected COVID-19 cases). In other words, the prospective statistic evaluates potential clusters that are still occurring at the end of the study period. The space-time scan statistic (STSS) employs moving cylinders that scan the U.S. for potential space-time clusters of COVID-

19 cases. The base of the cylinder is the spatial scanning window and the height reflects the temporal scanning window. The center of the cylinder is defined as the centroid of each U.S. county.

Next, each cylinder is expanded until a maximum spatial and temporal upper bound is reached, while each cylinder is a potential cluster. We set the upper bounds to have a maximum spatial and temporal scanning window size of 10% of the population at-risk to avoid extremely large clusters; and 50% of the study period, respectively. Each cluster's duration was set to a minimum of 2 days and a cluster must contain at least 5 confirmed cases of COVID-19. In other words, an unknown large number of cylinders of different spatial and temporal sizes are generated around each centroid until the maximum spatial and temporal thresholds are reached; the observed and expected case counts are computed within each cylinder, which are derived from the total number of centroids captured in each cylinder.

We selected the discrete Poisson data model, where we assume that the COVID-19 cases follow a Poisson distribution according to the population of the geographic region. The null hypothesis H_0 states that the model reflects a constant risk with an intensity μ, which is proportional to the at-risk population. The alternative hypothesis H_A states that the number of observed COVID-19 cases exceeds the number of expected cases derived from the null model (elevated risk within a cylinder). The expected number of COVID-19 cases (μ) under the null hypothesis H_0 is derived as follows in Equation 1:

$$\mu = p * \frac{C}{P} \tag{1}$$

with p the population in i; C the total COVID-19 cases in the U.S.; and P the total estimated population in the U.S. Note that the model assumes that the population is static for each location at each time period.

A maximum likelihood ratio test is used to identify scanning windows with an elevated risk for COVID-19, which is defined in Equation 2:

$$\frac{L(Z)}{L_0} = \frac{\left(\frac{n_Z}{\mu(Z)}\right)^{n_Z} \left(\frac{N - n_Z}{N - \mu_{(Z)}}\right)^{N-n_Z}}{\left(\frac{N}{\mu(T)}\right)^{N}} \tag{2}$$

with $L(Z)$ the likelihood function for cylinder Z, and L_0 the likelihood function for H_0; n_Z the number of COVID-19 cases in a cylinder; $\mu(Z)$ the number of expected cases in cylinder Z; N the total number of observed cases for the entire U.S. across all time periods; and $\mu(T)$ the total number of expected cases in the study area across all time periods. The cylinder has an elevated risk when the likelihood ratio is greater than 1, that is $\frac{n_Z}{\mu(Z)} > \frac{N-n_Z}{N-\mu_{(Z)}}$. Furthermore, the space-time scan statistic uses different cylinder sizes, and the cylinder with the highest likelihood ratio (maximum) is the most likely cluster. Monte Carlo testing is utilized (999 simulations) to assess the statistical significance of space-time clusters. Each simulation is conditioned to the same number of cases, and the likelihood is computed, so we obtain 999 likelihood ratios for each candidate cluster representing the distribution of the likelihood ratio under H_0. Secondary clusters are also reported if they are statistically significant at the $p < 0.05$ level.

To circumvent the assumption that the relative risk of COVID-19 is homogenous throughout a significant space-time cluster, we also report and visualize the relative risk for each U.S. county that belongs to a cluster. The relative risk (RR) for each location belonging to a cluster is derived from Equation 3:

$$RR = \frac{c/e}{(C-c)/(C-e)} \tag{3}$$

Where c is the total number of COVID-19 cases in a county, e is the total number of expected cases in a county, and C is the total number of observed cases in the U.S. RR is the estimated risk within a location divided by the risk outside of the location (i.e. everywhere else). For example, if a county has a RR of 2.5, then the population within that county are 2.5 times more likely to be exposed to COVID-19. The reported clusters also have a relative risk, which is derived the same way as Equation 3; but the clusters RR is estimated risk (observed/expected) divided by the risk outside of the cluster.

The incubation period of COVID-19 can be up to 2 weeks, so we detected active clusters that spanned $\leq$ 42 days, which is approximately three incubation periods of onset of the most current COVID-19 case in the dataset. The results identify statistically significant emerging clusters of COVID-19 in the U.S. at the county level between January 22nd-March 9th, 2020 in section 3.1 and between January 22nd-March 27th, 2020 in section 3.2. As the pandemic continues, new data can be added the prospective space-time scan statistic to monitor active clusters and identify areas that no longer are experiencing excess incidence based on available confirmed cases (i.e. areas that no longer have an excess public health risk).

3. Results

3.1 County-level results – January 22nd-March 9th, 2020

Table 1 provides the characteristics of the statistically significant emerging space-time clusters of COVID-19 at the county level from January 22nd and March 9th, 2020. Cluster 1 is found in the northwestern U.S. and includes 23 counties with a RR > 1 (i.e. more observed than expected cases). King County in Washington has a RR of 135.4 with 82 observed cases at the time of this study, and Santa Clara County in California contained 36 observed cases and a RR of 62. Cluster 2 only contains one county (Westchester) in New York with a RR of 639 and 97 observed cases. Cluster 3 contains counties in the mid-Atlantic region of the U.S. with nine counties exhibiting a RR > 1. Nassau County in Long Island, New York contained the highest

RR of 80.4 with 17 observed cases. Cluster 4 is in eastern Texas and contains two counties with a RR > 1 (Fort Bend – RR = 47.9; Harris – RR = 8). Cluster 5 is located in northern Georgia with 4 counties with an elevated relative risk: Polk (RR = 104.7), Fulton (RR = 21.3), Cobb (RR = 17.7), and Cherokee (RR = 17.5). Cluster 6 is located in the Midwest, where Summit County, Colorado (RR = 250.9) and Johnson County, Iowa (RR = 154.9) exhibits the highest relative risk. Cluster 7 contains two counties in southern California: Los Angeles (RR = 6.8) and Orange (RR = 4.9). Finally, Cluster 8 is located in southern Florida and contains 4 counties with an elevated risk: Charlotte (RR = 56), Manatee (RR = 52.6), Lee (RR = 27.5), and Broward (RR = 16).

Table 1: Emerging space-time clusters of COVID-19 from January 22nd-March 9th, 2020 at the county-level (RR = relative risk)

Cluster	Duration (days)	*p*	Observed	Expected	RR	# of counties	# of counties with RR >1
1	Feb 29th - Mar 9th	<0.001	207	7.9	43.2	107	23
2	Mar 4th - Mar 9th	<0.001	97	1.5	639	1	1
3	Mar 5th - Mar 9th	<0.001	53	5.1	11.3	66	9
4	Mar 5th - Mar 9th	<0.001	12	0.9	13.1	12	2
5	Mar 3rd - Mar 9th	<0.001	10	0.6	16.3	12	4
6	Mar 6th - Mar 9th	0.001	17	2.8	6.3	552	10
7	Mar 4th - Mar 9th	0.002	16	2.5	6.4	2	2
8	Mar 7th - Mar 9th	0.017	8	0.5	14.4	13	4

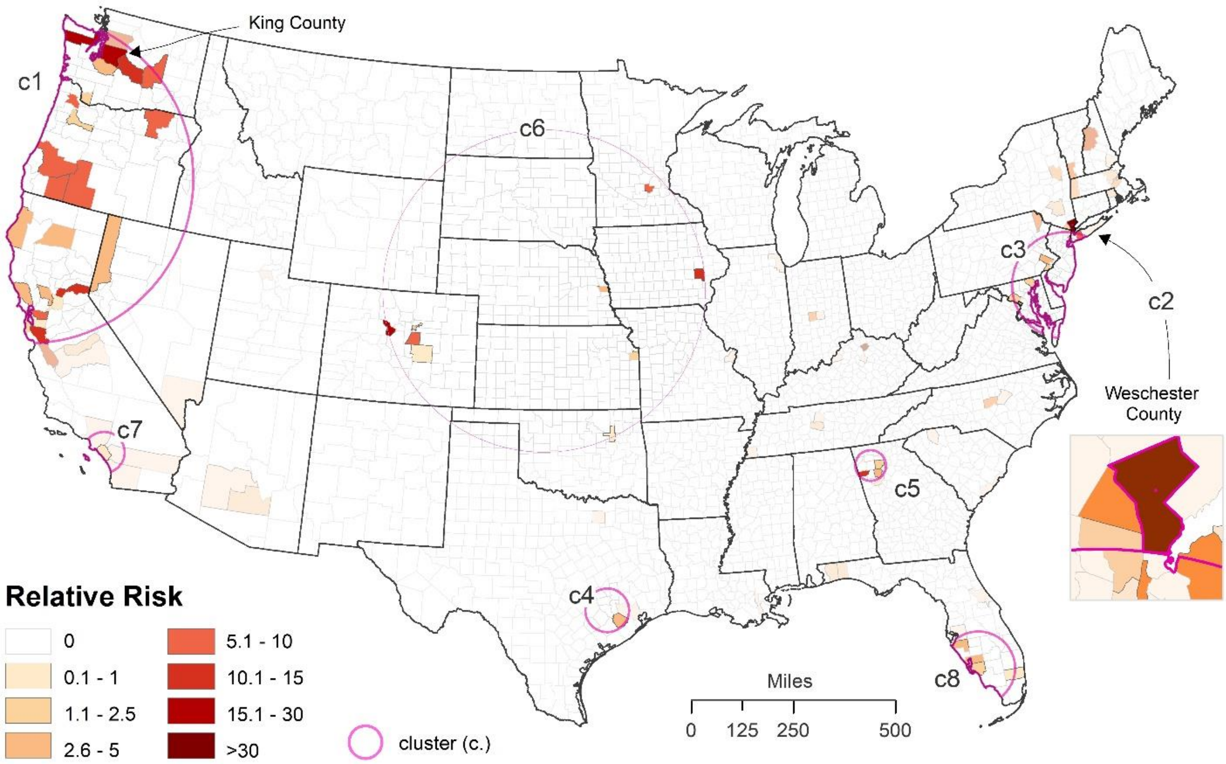

Figure 2: Spatial distribution of emerging space-time clusters of COVID-19 at the county-level from January 22nd-March 9th, 2020

Figure 2 illustrates the extent of eight emerging space-time clusters of COVID-19 at the county-level from January 22nd to March 9th, 2020. We highlight both King (Washington) and Westchester (New York) counties, which are known as the first major hotspots of the outbreaks in the US. King County is known to have the first U.S. case of COVID-19, which was introduced by recent travelers in China; leading to deadly outbreaks in nursing homes and the surrounding area (Bryson-Cahn et al. 2020). Westchester County includes the city of New Rochelle, which was the location of New York's initial outbreak and was subject to a containment zone spanning a one-mile radius (Wallis 2020). The Bay area in California (especially San Francisco) has also been as a major hotspot of COVID-19, which was one of the first areas in the U.S. to implement a "shelter-in-place" order (Fracassa 2020). Counties with a relative risk of 0 are more transparent to focus solely on the counties with an elevated risk that "contribute" to the emerging clusters. Figure 1 indicates that many densely populated counties were within an emerging cluster across the U.S., while we continue to monitor the outbreaks and detect new clusters in the following section using eighteen more days of case data.

3.2 County-level results – January 22nd-March 27th, 2020

Table 2 summarizes the characteristics of the twenty-six statistically significant emerging space-time clusters of COVID-19 at the county level between January 22nd and March 27th, 2020. Cluster 1 (the most likely cluster) contains 14 counties in New York (NY), Connecticut, and New Jersey, and Manhattan, NY exhibits the highest RR of 96.8; which was also the highest RR in the U.S. at the time of the analysis. Cluster 2 contains 3 counties in Michigan, and Wayne County exhibiting the highest RR of 4.9. Cluster 3 contains two parishes in the southeastern part of Louisiana and included the New Orleans consolidated city-parish exhibiting a RR of 9.0.

Clusters 4, 9, 10, 12-16, 23, and 26 contains one county each: Cook, Illinois (RR = 3.1), Blaine, Idaho (RR = 19.1) which includes the town of Sun Valley and is considered the Idaho COVID-19 hotspot at the time of this publication, Marion, Indiana (RR = 3.7), Summit, Utah (RR = 8.2), Cleburne, Arkansas (RR = 12.8), Caddo, Louisiana (RR = 4.5), Bartow, Georgia (RR = 3.7), Kershaw, South Carolina (RR = 4.5), Clark, Arkansas (RR = 5.3) and Wasatch, Utah (RR = 4.2), respectively. Cluster 5 contains 4 counties in northern Washington State, with Snohomish County exhibiting the highest RR of 2.6. Cluster 6 contains 5 counties in Georgia, and Dougherty County exhibits the highest RR of 8.6. Cluster 7 contains 3 counties in Colorado with a RR > 1, and Gunnison County exhibiting the highest RR of 9.8. Cluster 8 is the largest cluster that contains 43 counties throughout New York State, Ohio, Pennsylvania, West Virginia,

Virginia, North Carolina, Maryland, and New Jersey with a RR > 1, with Monmouth County, New Jersey exhibiting the highest RR of 8.4.

Cluster 11 contains 3 counties in Florida, and Broward County exhibits the highest RR of 2.2. Cluster 17 contains 11 counties in Georgia with a RR > 1, and Carroll County exhibits the highest RR of 3.9. Cluster 18 contains 4 counties in Indiana with a RR > 1, and Decatur County exhibits the highest RR of 11.5. Cluster 19 contains 2 counties in Missouri, and St. Louis exhibits the highest RR of 1.9. Cluster 20 contains two counties in California, and San Francisco exhibits the highest RR of 1.9. Cluster 21 contains 3 counties in Tennessee with a RR > 1, and Davidson County exhibits the highest RR of 1.7. Cluster 22 contains 3 counties in Colorado, and Denver exhibits the highest RR of 1.7. Cluster 24 contains 3 counties in Alabama, and Walker County exhibits the highest RR of 3.0. Finally, Cluster 25 contains 5 counties in Mississippi with a RR > 1, and Quitman County exhibits the highest RR of 6.9.

Cluster 11 contains 3 counties in Florida, and Broward County exhibits the highest RR of 2.2. Cluster 17 contains 11 counties in Georgia with a RR > 1, and Carroll County exhibits the highest RR of 3.9. Cluster 18 contains 4 counties in Indiana with a RR > 1, and Decatur County exhibits the highest RR of 11.5. Cluster 19 contains 2 counties in Missouri, and St. Louis exhibits the highest RR of 1.9. Cluster 20 contains two counties in California, and San Francisco exhibits the highest RR of 1.9. Cluster 21 contains 3 counties in Tennessee with a RR > 1, and Davidson County exhibits the highest RR of 1.7. Cluster 22 contains 3 counties in Colorado, and Denver exhibits the highest RR of 1.7. Cluster 24 contains 3 counties in Alabama, and Walker

County exhibits the highest RR of 3.0. Finally, Cluster 25 contains 5 counties in Mississippi with a RR > 1, and Quitman County exhibits the highest RR of 6.9.

Figure 3 shows the locations and spatial patterns of the twenty-six emerging space-time clusters of COVID-19 at the county level in the U.S. between January 22nd and March 27th, 2020. Adding updated COVID-19 case data produced eighteen more emerging clusters than our analysis in section 3.1. The resulting space-time clusters are smaller in size and more "intense" when running the prospective statistic between January 22nd and March 27th. Notably, the relative risk decreased in Washington State's counties, especially King County where the COVID-19 outbreak was first introduced in the U.S. It is important to highlight that the relative risk throughout the U.S. increased using case data until March 27th; compared to the first analysis in section 3.1 that ended on March 9th. Furthermore, the northeastern U.S. is clearly the epicenter of COVID-19 in the country as shown in Figure 2. Figure 2 also shows that some clusters in Figure 1 have "disappeared" (e.g. southern California and Texas), likely due to increases in testing and vast increases of confirmed cases in many locations after March 9th. Overall, the reported space-time clusters in Table 2 and Figure 2 tell a story of the rapid COVID-19 dispersal and transmission across the U.S.

Table 2: Emerging space-time clusters of COVID-19 from January 22nd-March 27th, 2020 at the county level (RR = relative risk)

Cluster	Duration (days)	*p*	Observed	Expected	RR	# of counties	# of counties with RR >1
1	Mar 19th - Mar 27th	<0.001	56,189	3,343.8	33.1	14	14
2	Mar 21st - Mar 27th	<0.001	3,036	835.8	3.7	3	3
3	Mar 19th - Mar 27th	<0.001	1,477	228.0	6.5	2	2
4	Mar 24th - Mar 27th	<0.001	1,953	636.4	3.1	1	1
5	Mar 17th - Mar 27th	<0.001	1,929	1,032.9	1.9	4	4
6	Mar 20th - Mar 27th	<0.001	251	35.3	7.1	5	5
7	Mar 11th - Mar 27th	<0.001	218	30.5	7.2	4	3
8	Mar 13th - Mar 27th	<0.001	3,214	2,173.1	1.5	273	43
9	Mar 8th - Mar 27th	<0.001	93	4.8	19.1	1	1
10	Mar 25th - Mar 27th	<0.001	323	87.9	3.7	1	1
11	Mar 26th - Mar 27th	<0.001	630	294.0	2.1	3	3
12	Mar 19th - Mar 27th	<0.001	95	11.6	8.2	1	1
13	Mar 23rd - Mar 27th	<0.001	49	3.8	12.8	1	1
14	Mar 25th - Mar 27th	<0.001	100	22.2	4.5	1	1
15	Mar 20th - Mar 27th	<0.001	98	26.1	3.7	1	1
16	Mar 21st - Mar 27th	<0.001	63	14.1	4.5	1	1
17	Mar 26th - Mar 27th	<0.001	294	189.7	1.5	14	11
18	Mar 26th - Mar 27th	<0.001	44	12.5	3.5	8	4
19	Mar 26th - Mar 27th	<0.001	146	79.8	1.8	2	2
20	Mar 26th - Mar 27th	<0.001	175	101.5	1.7	2	2
21	Mar 24th - Mar 27th	<0.001	205	127.2	1.6	4	3
22	Mar 25th - Mar 27th	<0.001	198	125.8	1.5	3	3
23	Mar 23rd - Mar 27th	0.003	18	3.4	5.3	1	1
24	Mar 25th - Mar 27th	0.003	143	86.4	1.6	3	3
25	Mar 26th - Mar 27th	0.004	48	19.1	2.5	8	5
26	Mar 23rd - Mar 27th	0.019	21	5.1	4.1	1	1

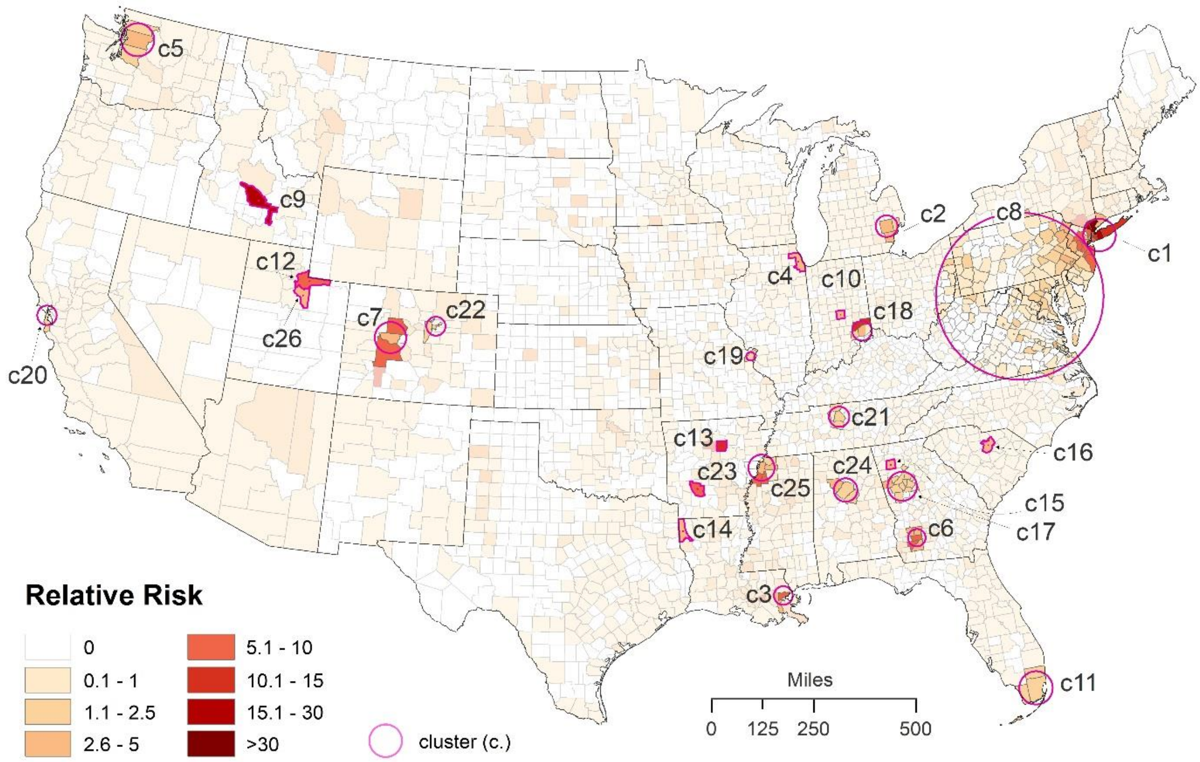

Figure 2: Spatial distribution of emerging space-time clusters of COVID-19 at the county level from January 22nd-March 9th, 2020

4. Discussion

In this paper, we utilized a prospective space-time scan statistic to detect emerging clusters of COVID-19 in the United States at the county level, providing results at two distinct time periods. To our knowledge, this study is the first one that utilizes space-time scan statistics to detect emerging clusters of COVID-19 in the United States. The prospective scanning statistic is a valuable surveillance tool to monitor disease outbreaks as they unfold (Kulldorff and Kleinman 2015). We suggest that prospective scanning statistics should be utilized in the suite of tools available to public health departments and researchers. It is important to conduct rapid statistical analysis to supplement basic case and disease rate maps available to better understand the highest risk areas of COVID-19; and how risk will progress throughout the duration of this pandemic. Since March 18th, 2020, each of the 50 U.S. states and Washington D.C. reported a confirmed case of COVID-19 (Dong et al. 2020). The prospective approach utilized in this study can be useful for state and local health departments to monitor the outbreaks in a timely fashion.

The main strength of the prospective approach is the ability to add updated COVID-19 counts and rerun the statistic to identify new emerging clusters; while also tracking the previously detected clusters to determine if they are growing or shrinking in magnitude. Doing so can help determine if current mitigation and isolation techniques are effective at curbing the spread of COVID-19. We demonstrate the notion of the prospective approach by presenting results between January 22nd – March 9th, 2020, and January 22nd – March 27th, 2020. The updated results in section 3.2 showcase the evolution of the COVID-19 outbreaks in the U.S., while 18 more clusters were detected using the updated daily case data. Notably, Manhattan became the epicenter of COVID-19 in the U.S., with a staggering 25% of the confirmed cases across the country. Furthermore, New Orleans and the Fort Lauderdale/Miami areas became hotspots in the southern U.S. Wayne County, Michigan contains Detroit, which also was detected as one of the major hotspots in the U.S when adding the updated daily cases to the prospective scan statistic.

One way to further evaluate the evolution of the detected clusters is to relax the statistical significance required (i.e. $p < 0.05$) and rerun the analysis at numerous spatial and temporal scales. As a result, we can identify locations that may become significant in a few days or a week's time but is beyond the scope of this exploratory paper. Furthermore, the incidence rates are not uniform across the U.S. Population density, age groups, and state and local mitigation measures will influence COVID-19 transmission and the magnitude of current and newly detected space-time clusters.

Healthcare facilities and resources will continue to be tested as more cases are suspected and confirmed with increases in testing (Heymann and Shindo 2020; Yee et al. 2020). Isolation measures and intensive contact tracing can successfully control COVID-19 outbreaks and reduce the burden facing hospitals and healthcare providers (Hellewell et al. 2020). Enhanced hygiene and stricter social distancing measures are required to reduce SARS-CoV-2 circulation, especially when community transmission is detected (Dalton et al. 2020). Availability of public datasets are also critical to increase surveillance efforts across the globe and corresponding areas facing substantial increases in transmission (Sun et al. 2020). Confirmed case counts are not enough to understand the true magnitude of the COVID-19 pandemic. Compiling datasets that include suspected, probable, and negative test counts can substantially improve surveillance efforts and our understanding of COVID-19 transmission dynamics (Lipsitch et al. 2020).

Despite the strengths of our study, there are limitations worth mentioning. First, there are many counties that were included in the clusters that did not contain any reported cases of COVID-19; however, this is due to the scanning process (an artifact of the statistic) and is circumvented by reporting the relative risk for the locations that belong to each cluster. Second, the case data only include confirmed cases and it is important to highlight that suspected and probable cases are not considered due to unavailability and uncertainty. Therefore, the true magnitude of the COVID-19 pandemic and will not be known for some time. Third, more local-level surveillance and studies are required to understand the transmission dynamics of the current and future emerging clusters; as SaTScan is an exploratory statistic. Fourth, COVID-19 is more severe for the elderly and those with preexisting medical conditions. Future studies can implement case/control cluster techniques with death and case counts (e.g. space-time Bernoulli models), while simultaneously adjusting for age and other relevant covariates. It is also possible to adjust for younger age groups to examine if mitigation guidelines have been successful in any way. Finally, this study utilized COVID-19 case data up until March 27th, 2020. Therefore, the magnitude and number of emerging clusters in our county-level analysis is likely much higher as cases continue to increase across the U.S.

5. Conclusion

We utilized publicly available case data from Johns Hopkins University's Center for Systems Science and Engineering to detect emerging space-time clusters of COVID-19 at the county level in the United States for two separate time periods. We suggest that the counties belonging to emerging clusters should be prioritized when allocating resources and implementing various quarantine and isolation measures to slow viral transmission. COVID-19 and general infectious disease surveillance can benefit from our prospective approach by monitoring outbreaks as they happen as new data becomes available. We emphasize the importance of focusing surveillance on emerging and active clusters during epidemics, essentially dismissing previous clusters that do not threaten public health that would appear in a retrospective analysis. Furthermore, data sharing and availability is crucial and allows a variety of researchers to contribute to our knowledge of COVID-19 and epidemiology, in general. Geographers can play a vital role in mitigating disease transmission, and this study is one example of the plethora of methods that can be implemented in a limited timeframe to effectively inform public health officials and decision-makers about spatial and space-time transmission dynamics.

References

Anastassopoulou, C., Russo, L., Tsakris, A., & Siettos, C. (2020). Data-Based Analysis, Modelling and Forecasting of the novel Coronavirus (2019-nCoV) outbreak. *medRxiv*.

Bryson-Cahn, C., Duchin, J., Makarewicz, V. A., Kay, M., Rietberg, K., Napolitano, N., ... & Lynch, J. B. (2020). A Novel Approach for a Novel Pathogen: using a home assessment team to evaluate patients for 2019 novel coronavirus (SARS-CoV-2). *Clinical Infectious Diseases*.

Dalton, C., Corbett, S., & Katelaris, A. (2020). Pre-emptive low cost social distancing and enhanced hygiene implemented before local COVID-19 transmission could decrease the number and severity of cases. *Available at SSRN 3549276*.

Desjardins, M.R., Hohl, A., Delmelle, E., & Casas, I. (2020). Identifying and visualizing space-time clusters of vector-borne diseases. In *Geospatial Technology for Human Well-Being and Health*. Springer. Forthcoming.

Desjardins, M. R., Whiteman, A., Casas, I., & Delmelle, E. (2018a). Space-time clusters and co-occurrence of chikungunya and dengue fever in Colombia from 2015 to 2016. *Acta tropica*, *185*, 77-85.

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*.

Fracassa, D. (2020). Bay Area coronavirus shutdown: How life will change with shelter-in-place order. *San Francisco Chronicle*. Retrieved from: https://www.sfchronicle.com/bayarea/article/Bay-Area-to-shelter-in-place-What-you-need-15135087.php [last accessed 22 March 2020].

Hellewell, J., Abbott, S., Gimma, A., Bosse, N. I., Jarvis, C. I., Russell, T. W., ... & Flasche, S. (2020). Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. *The Lancet Global Health*.

Heymann, D. L., & Shindo, N. (2020). COVID-19: what is next for public health? *The Lancet*.

Hong, N., He, J., Ma, Y., Jiang, H., Han, L., Su, L., ... & Long, Y. (2020). Evaluating the secondary transmission pattern and epidemic prediction of the COVID-19 in metropolitan areas of China. *medRxiv*.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., ... & Cheng, Z. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, *395*(10223), 497-506.

Jones, R. C., Liberatore, M., Fernandez, J. R., & Gerber, S. I. (2006). Use of a prospective space-time scan statistic to prioritize shigellosis case investigations in an urban jurisdiction. *Public Health Reports*, *121*(2), 133-139.

Kulldorff, M., & Kleinman, K. (2015). Comments on 'A critical look at prospective surveillance using a scan statistic'by 'T. Correa, M. Costa, and R. Assunção. *Statistics in medicine*, *34*(7), 1094.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics-Theory and methods*, *26*(6), 1481-1496.

Kulldorff, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *164*(1), 61-72.

Kulldorff, M. (2018). SaTScan™ user guide for version 9.6. [https://www.satscan.org/].

Kulldorff, M., Athas, W. F., Feurer, E. J., Miller, B. A., & Key, C. R. (1998). Evaluating cluster alarms: a space-time scan statistic and brain cancer in Los Alamos, New Mexico. *American journal of public health*, *88*(9), 1377-1380.

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., ... & Xing, X. (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*.

Lipsitch, M., Swerdlow, D. L., & Finelli, L. (2020). Defining the epidemiology of Covid-19—studies needed. *New England Journal of Medicine*.

Liu, Y., Gayle, A. A., Wilder-Smith, A., & Rocklöv, J. (2020). The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of travel medicine*.

Mahase, E. (2020). Coronavirus: covid-19 has killed more people than SARS and MERS combined, despite lower case fatality rate. *BMJ, 368*.

Owusu, C., Desjardins, M. R., Baker, K. M., & Delmelle, E. (2019). Residential mobility impacts relative risk estimates of space-time clusters of chlamydia in Kalamazoo County, Michigan. *Geospatial health*, *14*(2).

Roser, M., & Ritchie, H. (2020). Coronavirus Disease (COVID-19). *Our World in Data*.

Ruan, Q., Yang, K., Wang, W., Jiang, L., & Song, J. (2020). Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China. *Intensive Care Medicine*, 1-3.

Sanche, S., Lin, Y. T., Xu, C., Romero-Severson, E., Hengartner, N. W., & Ke, R. (2020). The Novel Coronavirus, 2019-nCoV, is Highly Contagious and More Infectious Than Initially Estimated. *arXiv preprint arXiv:2002.03268*.

Sun, K., Chen, J., & Viboud, C. (2020). Early epidemiological analysis of the coronavirus disease 2019 outbreak based on crowdsourced data: a population-level observational study. *The Lancet Digital Health*.

Wallis, C. (2020). Life in the Containment Zone. What it’s like in New Rochelle, N.Y., the site of the state’s biggest coronavirus outbreak. *Scientific American*. Retrieved from: https://blogs.scientificamerican.com/observations/life-in-the-containment-zone/ [last accessed 22 March 2020].

Whiteman, A., Desjardins, M.R., Eskildsen, G.A., & Loaiza, J.R. (2019) Integrating vector surveillance data to improve space-time risk estimation of dengue fever in Panama. *PLOS Neglected Tropical Diseases*, *13*(9): e0007266.

Wu, Z., & McGoogan, J. M. (2020). Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention. *Jama*.

Yee, J., Unger, L., Zadravecz, F., Cariello, P., Seibert, A., Johnson, M. A., & Fuller, M. J. (2020). Novel coronavirus 2019 (COVID-19): Emergence and implications for emergency care. *Journal of the American College of Emergency Physicians Open*.

Yih, W. K., Deshpande, S., Fuller, C., Heisey-Grove, D., Hsu, J., Kruskal, B. A., ... & Puga, E. (2010). Evaluating real-time syndromic surveillance signals from ambulatory care data in four states. *Public Health Reports*, *125*(1), 111-120.

Yin, F., Li, X., Ma, J., & Feng, Z. (2007). The early warning system based on the prospective space-time permutation statistic. *Wei sheng yan jiu= Journal of hygiene research*, *36*(4), 455-458.

Yu, K. D. S., & Aviso, K. B. (2020). Modelling the Economic Impact and Ripple Effects of Disease Outbreaks. *Process Integration and Optimization for Sustainability*, 1-4.