Rapid detection of COVID-19 clusters in the United States using a prospective space-time scan statistic: An update

Alexander Hohl[1], Eric Delmelle[2], Michael Desjardins[3]
[1]Department of Geography, University of Utah, UT, USA
[2]Department of Geography and Earth Sciences & Center for Applied Geographic Information Science, University of North Carolina at Charlotte, NC, USA
[3]Department of Epidemiology & Spatial Science for Public Health Center, Johns Hopkins Bloomberg School of Public Health, MD, USA
alexander.hohl@geog.utah.edu, eric.delmelle@uncc.edu, mdesjar3@jhmi.edu

Abstract

Novel coronavirus disease 2019 (COVID-19), caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is a pandemic with 1,420,299 confirmed cases and 85,992 total deaths within the United States as of May 15th, 2020. As the number of cases continues to climb, detecting clusters of COVID-19 is critical to alleviate the strain on our public health system through improved resource allocation and decision-making. Here, we report on an analysis of daily case data at the county level using the prospective spatial-temporal scan statistic. In previous work, we performed the analysis for March 27th 2020 [1], and here we report updated results as of April 27th 2020, producing a new set of "active" and emerging clusters present. Our analysis resulted in sixteen significant space-time clusters of COVID-19 at the county level in the U.S. during the time span of March 22nd - April 27th. The space-time pattern of significant clusters mirrors active and emerging disease hot-spots at the end of our study period. The statistic can be rerun to support timely surveillance of COVID-19, as demonstrated here.

1 Introduction

Novel coronavirus disease 2019 (COVID-19) is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). As of May 15th, 2020, a total of 4,483,864 confirmed cases and 303,825 deaths from COVID-19 have been reported globally. With 1,420,299 cases and 85,992 deaths, the United States is currently most affected by the disease globally [2]. Patients may exhibit symptoms like fever, shortness of breath, and cough [3], but more severe cases can lead to pneumonia and death [4]. The elderly, as well as people with preexisting conditions, are especially at risk of a severe outcome of COVID-19 [5].

Disease surveillance refers to systematic collection and analysis of data on disease outbreaks for guidance of public health response, such as rapid testing, allocation of resources, or social distancing measures. It is thus an essential component of health emergency work and allows for timely recommendations for countermeasures [6]. The space-time scan statistic [7] is a widely used disease surveillance technique, as it identifies the geographic location, duration, strength and statistical significance of disease clusters. The statistic and its variants have been applied in many different settings: detecting unusual increases of hospital visits in NYC [8], analyzing the co-occurrence of Chikungunya and Dengue Fever in Colombia and Panama [9][10], identifying hot spots of crime activity [11], clustering geotagged tweets [12], and many more[1].

[1]For a comprehensive list of SatScan applications and methodological developments, see https://www.satscan.org/references.html

The prospective space-time scan statistic [8] is suitable for continuous surveillance and stands in contrast to its retrospective variant. It detects "active" or "emerging" clusters at the end of the study period while ignoring past clusters that may no longer constitute a public health threat. The prospective space-time scan statistic can be rerun as the phenomenon of interest unfolds, for tracking the development of existing clusters and for detecting new ones. Applications include syndromic surveillance in NYC [13], an early detection system for West Nile Virus [14], and lastly, COVID-19 in the United States as of March 27, 2020 [1].

Here, we report significant active clusters of COVID-19 in the contiguous United States during the time period of January 22^{nd} - April 27^{th}, 2020, thereby extending and updating an earlier study that focused on the time period of January 22^{nd} - March 27^{th} [1]. Our work contributes to surveillance efforts of COVID-19 by identifying areas of significantly elevated disease risk. It informs health authorities by providing guidance on allocating resources for contact tracing, rapid testing, and social distancing policy. This short article is structured as follows: Section 2 (Data and Methods) offers details on the COVID-19 case data, as well as theoretical background of the prospective space-time scan statistic. Section 3 (Results) contains an overview of the identified significant clusters. Section 4 (Discussion and Conclusions) summarizes strengths and weaknesses of our approach and offers an outlook on current and future work.

2 Data and Methods

We collected case data from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University[2]. The repository is updated daily and contains case counts for the United States at the county level. We gathered daily case counts for the contiguous United States during the time period since the first confirmed case in the U.S. to the date of performing the analysis (January 22^{nd} - to April 27^{th}, 2020). In addition, we obtained 2018 ACS 5-year estimates of the total population of each county, which we joined to the corresponding county polygons obtained from the U.S. Census. Therefore, we ignored cases assigned to cruise ships "Diamond Princess" and "Grand Princess", as well as cases that were assigned to states rather than counties. As the data set contains cumulative case counts for each day of our study period (Figure 1), we subtracted the previous day's count to obtain the number of new cases per day for each county.

We applied the prospective Poisson space-time scan statistic [8] to detect active clusters of COVID-19. The space-time scan statistic identifies the most likely clusters from a set of cylindrical candidate clusters of varying size and geographic location (county centroids are candidate locations). The base of the cylinder corresponds to the circular spatial scanning window, whereas the height corresponds to the temporal window. We set the maximum spatial and temporal scanning window size to include no more than 10% of the population at-risk and 50% of the study period, respectively. We chose the minimum number of cases per cluster to 5 and the minimum duration to 2 days. These parameter settings are consistent with our earlier efforts of identifying clusters of COVID-19 [1]. The expected and observed numbers of cases are computed for each cylinder and a maximum likelihood test is performed. The null hypothesis H_0 states that risk within the cylinder is not different from the outside, whereas the alternative hypothesis H_a states that risk inside the cylinder is elevated. The number of expected cases μ is computed using Equation 1:

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

where p is the population inside the cylinder, C the total number of cases in the U.S., and P the total population in the U.S.

[2]https://github.com/CSSEGISandData/COVID-19

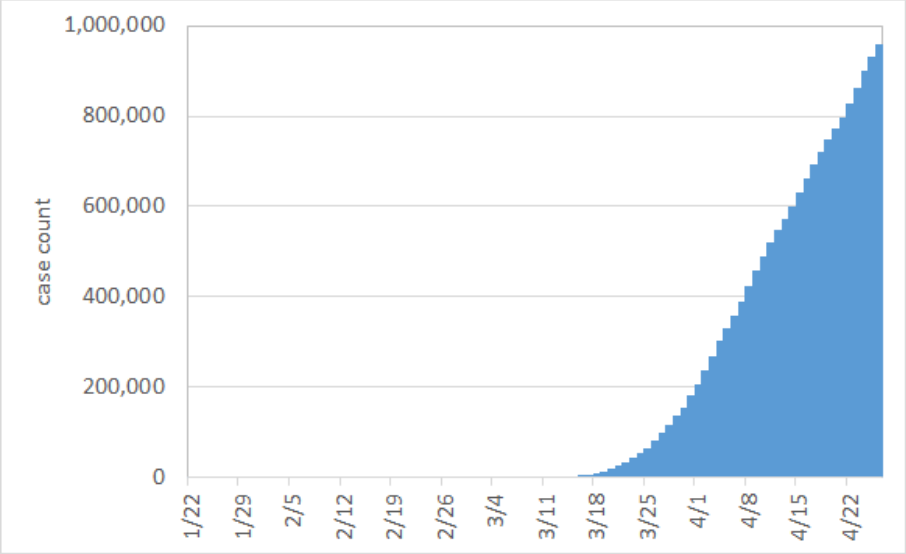

Figure 1: Cumulative number of COVID-19 cases in the contiguous United States between January 22nd and April 27th, 2020.

A likelihood ratio test is used to identify clusters of elevated disease risk using Equation 2:

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

where $L(Z)$ is the likelihood function for candidate cylinder Z, L_0 the likelihood function for H_0; n_z the number of COVID-19 cases inside a cylinder; $\mu(Z)$ the expected number of cases in cylinder Z; N the number of observed cases for the entire U.S. during the entire study period; and $\mu(T)$ the total number of expected cases in the study area across all time periods. If a cylinder exhibits a likelihood ratio greater than 1, that is $\frac{n_Z}{\mu(Z)} > \frac{N-n_Z}{N-\mu_{(Z)}}$, its risk is elevated. Out of all the candidate cylinders, the most likely cluster is denoted as the cylinder that has the highest likelihood ratio. We assess statistical significance using 999 Monte Carlo simulations, to obtain 999 likelihood ratios for each candidate cluster that form a distribution under H_0. In addition, we report secondary clusters if statistically significant at the $p < 0.001$ level.

We map the distribution of risk inside significant cylinders as relative risk (RR), which is the risk within a location divided by the risk outside. Therefore, for each county that belongs to a cluster, we compute and report Equation (3):

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

where c is the total number of cases for a given county, e the number of expected cases in a county, and C the number of observed cases in the U.S. Similarly, we report RR for clusters.

3 Results

Table 1 shows significant clusters of COVID-19 in the U.S. at the county level, as detected by prospective Poisson space-time scan statistic for January 22^{nd} - April 27^{th}, 2020. As expected, New York City and its neighboring counties form the strongest cluster of observed cases (Cluster 1). However, in terms of relative risk, Cluster 10 (Bledsoe County, TN) ranks higher, as it has an expected number of cases of 1.8 due to low population. The spatial distribution of clusters shows that hot spots of COVID-19 can be found in every major region in the U.S., suggesting a widespread presence of the virus.

As compared to significant clusters resulting from our previous analysis [1] of the January 22^{nd} - March 27^{th}, 2020 period, we observe less clusters (16 vs. 26), but their size generally increased (average number of counties per cluster: 79 vs. 14), and exhibit higher relative risk (average relative risk: 24.5 vs 5.7). These observations mirror the spread and increasing number of confirmed cases as the virus permeates our society, as well as spatial diffusion as exemplified by person-to-person disease transmission.

It is worth noting the clusters that disappeared (became insignificant) since March 27^{th}, 2020: the Salt Lake City (UT) region, Kershaw County (SC), southeastern Indiana, the San Francisco Bay area (CA), the Denver (CO) area. The reason for vanishing clusters may be increased awareness, successful measures to flatten the curve (i.e. social distancing, wearing of masks and gloves), or they might simply be an indication of a locally maturing epidemic. Also, the number of cases in those areas may simply be too low compared to surrounding areas. In order words, the magnitude of these clusters was not enough to hold the 'cluster' denomination again in the current analysis.

Cluster	Duration (days)	p-value	Observed	Expected	RR	# of counties	# of counties with RR > 1
1	Mar 22^{nd} - Apr 27^{th}	< 0.001	447,076	36,714.3	21.9	59	52
2	Mar 28^{th} - Apr 27^{th}	< 0.001	89,980	24,296.0	3.9	129	68
3	Apr 2^{nd} - Apr 27^{th}	< 0.001	83,984	23,838.7	3.7	221	122
4	Mar 26^{th} - Apr 27^{th}	< 0.001	20,761	2,886.3	7.3	29	29
5	Mar 27^{th} - Apr 27^{th}	< 0.001	15,325	4,637.1	3.3	2	2
6	Apr 1^{st} - Apr 27^{th}	< 0.001	18,559	7,254.5	2.5	123	106
7	Apr 11^{th} - Apr 27^{th}	< 0.001	8,055	1,729.0	4.6	127	38
8	Apr 11^{th} - Apr 27^{th}	< 0.001	2,878	285.7	10.0	15	9
9	Mar 31^{st} - Apr 27^{th}	< 0.001	17,738	8,700.2	2.0	1	1
10	Apr 24^{th} - Apr 27^{th}	< 0.001	577	1.8	318.1	1	1
11	Apr 3^{rd} - Apr 27^{th}	< 0.001	2,324	424.9	5.4	7	6
12	Apr 1^{st} - Apr 27^{th}	< 0.001	26,594	17,862.0	1.5	424	154
13	Apr 26^{th} - Apr 27^{th}	< 0.001	8,463	4,044.4	2.1	14	9
14	Apr 17^{th} - Apr 27^{th}	< 0.001	982	454.4	2.1	63	18
15	Apr 5^{th} - Apr 27^{th}	< 0.001	4,797	3,510.6	1.3	47	13
16	Apr 24^{th} - Apr 27^{th}	< 0.001	614	256.4	2.3	1	1

Table 1: Active space-time clusters of COVID-19 from January 22^{nd}-April 27^{th}, 2020 at the county level (RR = relative risk) .

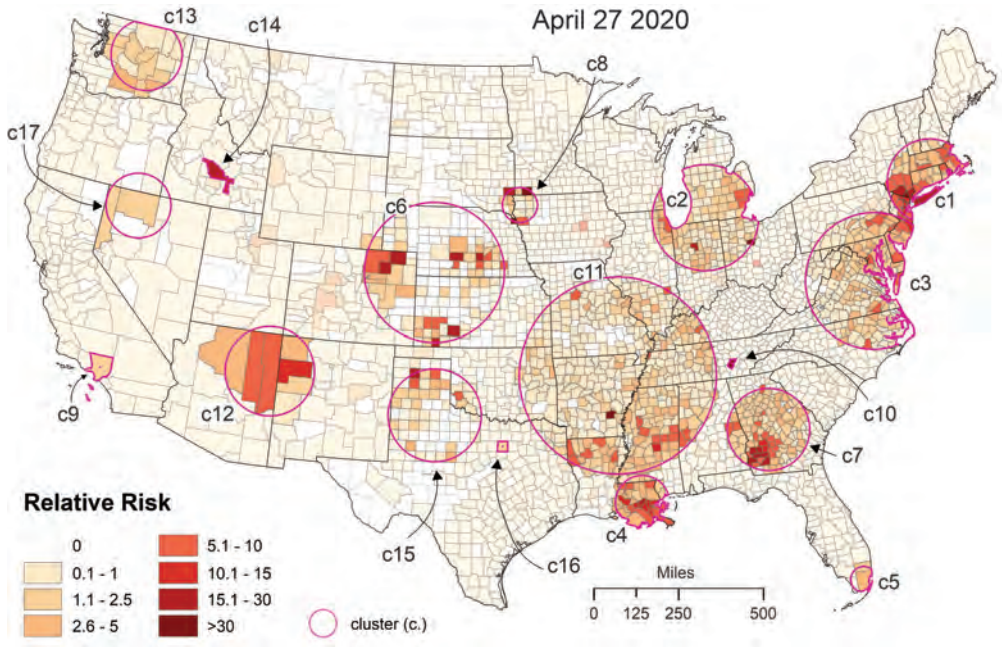

Figure 2: Cumulative number of COVID-19 cases in the contiguous United States between January 22^{nd} and April 27^{th}, 2020.

4 Discussion and Conclusions

In this paper, we report active and emerging clusters of COVID-19, as identified by the prospective Poisson space-time statistic. This analysis is a continuation of previous work where we applied the same methodology [1]. However, as the pandemic of COVID-19 progresses, we used an updated dataset which includes one more month of records than the previous analysis (January 22^{nd} - April 27^{th} vs. January 22^{nd} - March 27^{th}). The prospective space-time scan statistic [8] is a practical tool for disease outbreak monitoring, which allows health authorities to make informed decisions about allocating resources for maximizing the impact their response. Its appeal lies in the ability to detect active or emerging clusters, and it can be rerun during the course of an epidemic, as exemplified here, for updated reporting. Timely disease surveillance is especially important due to the current efforts of relaxing social distancing measures.

Despite its many appeals, the prospective space-time scan statistic used in our study has limitations: First, the statistic in its basic form only allows for clusters of circular shape. This property has been discussed extensively in the scientific community, which resulted in many efforts to relax the circular dictate and allow for clusters of arbitrary shape [15, 16, 17, 18]. Second, we only used confirmed cases, which may not reflect the true magnitude and spatial distribution of the virus, even though there are reports of asymptomatic carriers [19]. This limitation points towards current testing practices and could be addressed by increasing the availability and usage of test sets. Third, further analysis is needed to explore dynamics of COVID-19 within clusters. Some of the significant clusters were very large and exhibited considerable variation of relative risk within. Therefore, analyzing the spread of the virus locally (i.e. within the affected regions) is needed for a more effective response.

Given the limitations of our approach, future work should focus on exploring patterns of COVID-19 induced mortality and its relationships with socioeconomic characteristics of affected regions. Especially the distribution of at-risk groups should be considered and analyzed in relation to mortality patterns, in order to better inform authorities on their mitigation efforts. In addition, there is an urgent need to study the relationship between reactions of humans (e.g., fear, panic, hate) to the current pandemic of COVID-19 and the characteristics of their residential neighborhood. Federal, state, and local countermeasures, as well as resourcing and funding initiatives, must consider demographic and socioeconomic factors and their association with people's response to disease outbreaks.

References

[1] M.R. Desjardins, A. Hohl, and E.M. Delmelle. Rapid surveillance of covid-19 in the united states using a prospective space-time scan statistic: Detecting and evaluating emerging clusters. *Applied Geography*, 118:102202, 2020.

[2] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*.

[3] Qiurong Ruan, Kun Yang, Wenxia Wang, Lingyu Jiang, and Jianxin Song. Clinical predictors of mortality due to covid-19 based on an analysis of data of 150 patients from wuhan, china. *Intensive care medicine*, pages 1–3, 2020.

[4] Elisabeth Mahase. Coronavirus: covid-19 has killed more people than sars and mers combined, despite lower case fatality rate, 2020.

[5] Zunyou Wu and Jennifer M McGoogan. 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*, 323(13):1239–1242, 2020.

[6] Weizhong Yang. *Early Warning for Infectious Disease Outbreak: theory and practice*. Academic Press, 2017.

[7] Martin Kulldorff. A spatial scan statistic. *Communications in Statistics-Theory and methods*, 26(6):1481–1496, 1997.

[8] Martin Kulldorff. 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, 2001.

[9] MR Desjardins, A Whiteman, I Casas, and E Delmelle. Space-time clusters and co-occurrence of chikungunya and dengue fever in colombia from 2015 to 2016. *Acta tropica*, 185:77–85, 2018.

[10] Ari Whiteman, MR Desjardins, GA Eskildsen, and JR Loaiza. Detecting space-time clusters of dengue fever in panama after adjusting for vector surveillance data. *PLoS neglected tropical diseases*, 13(9):e0007266, 2018.

[11] Mingxuan Han, Michael Matheny, and Jeff M Phillips. The kernel spatial scan statistic. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 349–358, 2019.

[12] Roberto CSNP Souza, Renato M Assunção, Daniel B Neill, and Wagner Meira Jr. Detecting spatial clusters of disease infection risk using sparsely sampled social media mobility patterns. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 359–368, 2019.

[13] Richard Heffernan, Farzad Mostashari, Debjani Das, Adam Karpati, Martin Kulldorff, and Don Weiss. Syndromic surveillance in public health practice, new york city. 2004.

[14] Farzad Mostashari, Martin Kulldorff, Jessica J Hartman, James R Miller, and Varuni Kulasekera. Dead bird clusters as an early warning system for west nile virus activity. *Emerging infectious diseases*, 9(6):641, 2003.

[15] Toshiro Tango and Kunihiko Takahashi. A flexibly shaped spatial scan statistic for detecting clusters. *International journal of health geographics*, 4(1):11, 2005.

[16] Martin Kulldorff, Lan Huang, Linda Pickle, and Luiz Duczmal. An elliptic spatial scan statistic. *Statistics in medicine*, 25(22):3929–3943, 2006.

[17] Luiz Duczmal, André LF Cançado, Ricardo HC Takahashi, and Lupercio F Bessegato. A genetic algorithm for irregularly shaped spatial scan statistics. *Computational Statistics & Data Analysis*, 52(1):43–52, 2007.

[18] Daniel P De Oliveira, Daniel B Neill, James H Garrett Jr, and Lucio Soibelman. Detection of patterns in water distribution pipe breakage using spatial scan statistics for point events in a physical network. *Journal of Computing in Civil Engineering*, 25(1):21–30, 2011.

[19] Yan Bai, Lingsheng Yao, Tao Wei, Fei Tian, Dong-Yan Jin, Lijuan Chen, and Meiyun Wang. Presumed asymptomatic carrier transmission of covid-19. *Jama*, 323(14):1406–1407, 2020.