Analysis of distribution characteristics of COVID-19 in America based on space-time scan statistic

Yuexu Zhao and Qiwei Liu*

College of Economics, Hangzhou Dianzi University, Hangzhou, China

Based on the epidemic data of COVID-19 in 50 states of the United States (the US) from December 2021 to January 2022, the spatial and temporal clustering characteristics of COVID-19 in the US are explored and analyzed. First, the retrospective spatiotemporal analysis is performed by using SaTScan 9.5, and 17 incidence areas are obtained. Second, the reliability of the results is tested by the circular distribution method in the time latitude and the clustering method in the spatial latitude, and it is confirmed that the retrospective spatiotemporal analysis accurately measures in time and reasonably divides regions according to the characteristics in space. Empirical results show that the first-level clustering area of the epidemic has six states with an average relative risk of 1.28 and the second-level clustering area includes 18 states with an average relative risk of 0.86. At present, the epidemic situation in the US continues to expand. It is necessary to do constructive work in epidemic prevention, reduce the impact of epidemic, and effectively control the spread of the epidemic.

KEYWORDS
the spatiotemporal analysis, COVID-19, scan statistic, spatial aggregation, Omicron

Introduction

Coronavirus disease 2019 (COVID-19) refers to the new coronavirus infection in 2019 caused by acute respiratory infectious diseases in the majority of patients; some of them will develop as severe cases and even results in death. Since the large-scale outbreak of new corona pneumonia in 21 January 2020, the economy in the US has been gradually affected. Besides, the epidemic is a public emergency that all countries in the world have to face. As one of the most serious epidemic countries, the US has accumulated 74,741,586 new coronavirus cases as on 31 January 2022, and there is a certain aggregation tendency in time and space. Scan statistic is a method to test whether there is an aggregation of diseases, and detect whether the abnormal increase of diseases in time and space is caused by random variation. It has been widely used in infectious diseases, cardiovascular diseases, and other fields as a spatial statistical method in epidemic statistics.
In 1965, Joseph (1) proposed the concept of scan statistic. Kulldorff et al. (2–5) proposed the spatial scan statistic, and applied scan statistic to analyze the breast cancer mortality in the US. For example, they utilized the dynamic variable scanning window to detect the leukemia data in northern New York, and used the log-likelihood ratio to determine the cluster with the highest degree of aggregation. They also proposed many statistical models of spatiotemporal scanning, such as retrospective space-time scan statistic in Bernoulli model or Poisson model, prospective space-time scan statistic, space-time rearrangement scan statistic, and elliptical spatial scan statistic in periodic geographic disease monitoring. As the research goes further, Jung et al. (6, 7) proposed ordinal model scan statistic in 2007, which had excellent performance compared with Bernoulli scan statistic for binary classification of prostate cancer data. Huang et al. (8, 9) proposed the spatial scan statistic based on the exponential model for the male survival data with prostate cancer in the US in 2007; this method could be applied to the survival data and pure spatial data. In order to study the spatial heterogeneity continuously measured in the population data, the weighted normal spatial scan statistic was proposed and applied to the two-stage lung cancer survival research in 2009. Barbara (10) found that Cutl’s method was more effective than Kulldorff’s scan statistic for irregular shape spatiotemporal clusters, and for cylindrical spatiotemporal clusters; these two methods had similar results. Li et al. (11) analyzed the fund sustainability. Yin (12) carried out the research on application in early warning of infectious diseases, and graded the data of provinces and cities. Ma et al. (13) selected the optimal spatial scale through the number of signals in the monitoring of infectious diseases. So far, scan statistic has been widely used in disease prevention, including tuberculosis, schistosomiasis, and hand, foot, and mouth disease.

The majority of the abovementioned studies explore the spatial aggregation of various infectious diseases. COVID-19 is a highly contagious disease, which has seriously affected people's lives since its outbreak, and has a great threat to people's health. Hohl et al. (14) used the daily new coronavirus case data provided by the John Hopkins University at the county level, and applied SaTScan to conduct a prospective space-time analysis, and detected the active clusters in various provinces and cities in the US. To avoid using prospective space-time scan statistic to identify emergence of COVID-19 disease
TABLE 1  Retrospective spatiotemporal analysis.

| Cluster | Start date | The number of states | The actual value | Value of expectation | ˜R | LLR | p  |
|---------|------------|----------------------|-----------------|---------------------|----|-----|----|
| 1       | Dec 1      | 6                    | 4,157,904       | 3,363,765.04        | 1.28 | 101,152.5 | 0.0000 |
| 2       | Dec 1      | 18                   | 5,817,663       | 6,526,216.68        | 0.86 | 52,598.68 | 0.0000 |
| 3       | Dec 1      | 1                    | 1,891,720       | 2,301,321.62        | 0.81 | 42,332.33 | 0.0000 |
| 4       | Dec 1      | 5                    | 1,114,238       | 1,418,944.71        | 0.78 | 37,219.81 | 0.0000 |
| 5       | Dec 1      | 4                    | 1,965,177       | 2,274,140.59        | 0.85 | 24,001.90 | 0.0000 |
| 6       | Dec 1      | 2                    | 945,695         | 1,169,284.65        | 0.80 | 23,886.30 | 0.0000 |
| 7       | Dec 1      | 1                    | 672,175         | 845,809.54          | 0.79 | 19,780.36 | 0.0000 |
| 8       | Dec 1      | 6                    | 882,793         | 1,034,449.98        | 0.85 | 12,161.49 | 0.0000 |
| 9       | Dec 1      | 2                    | 3,455,831       | 3,236,835.98        | 1.08 | 8,298.06  | 0.0000 |
| 10      | Dec 1      | 1                    | 1,844,383       | 1,700,649.87        | 1.09 | 6,333.58  | 0.0000 |
| 11      | Dec 1      | 2                    | 1,351,620       | 1,228,440.97        | 1.11 | 6,284.42  | 0.0000 |
| 12      | Dec 1      | 1                    | 535,773         | 465,366.48          | 1.15 | 5,172.76  | 0.0000 |
| 13      | Dec 1      | 1                    | 1,110,569       | 1,011,672.39        | 1.10 | 4,878.70  | 0.0000 |
| 14      | Dec 1      | 1                    | 423,426         | 485,990.37          | 0.87 | 4,288.31  | 0.0000 |
| 15      | Dec 1      | 3                    | 1,056,440       | 990,205.76          | 1.07 | 2,254.76  | 0.0000 |
| 16      | Dec 1      | 3                    | 2,193,108       | 2,263,168.50        | 0.97 | 1,198.58  | 0.0000 |
| 17      | Dec 1      | 1                    | 381,903         | 355,779.67          | 1.07 | 949.66    | 0.0000 |

The table presents the results of retrospective spatiotemporal analysis. Each row represents a cluster, with columns for the cluster number, start date, number of states, actual value, value of expectation, relative risk (˜R), log-likelihood ratio (LLR), and the significance level (p).

Methodology

Retrospective spatiotemporal analysis needs to build a scanning window to judge the number of diseases inside and outside the window. Since the scanning statistics involve time and space, the scanning window is in the form of a cylinder; the height of the cylinder represents the time, and the bottom area of the cylinder represents the area. The location and size of the scan window are dynamic, as it is unknown when and where the COVID-19 outbreak will occur.

In the analysis process, a position is randomly selected as the scanning center, and then, the cylindrical scanning window changes continuously. The cluster of geographic size of the scanning window ranges between zero and a predefined upper limit. There are several ways to determine the value of upper bound, for example, one can take the percentage of number of people at risk of disease or radius value of circle as the upper bound. In this article, we use the former method. The time length of the scan window specifies the maximum time frame according to the percentage of the entire study cycle or the specific number of days.

To determine the possibility of aggregation, the actual number of patients and the number of regional populations are calculated to obtain the theoretical number of patients, and the log-likelihood ratio (LLR) is constructed by using the actual and theoretical number of patients inside and outside the window; the relative risk (˜R) is calculated to evaluate the strength of aggregation. Since the scanning window undergoes a dynamic change, numerous scanning windows will be generated during the scanning process. For controlling the false-positive rate at a certain level, the window with the largest LLR is
TABLE 2 Numbers of cases and morbidities in the top 20 states.

| State          | Population   | Case   | Morbidity | Rank |
|----------------|--------------|--------|-----------|------|
| Rhode Island   | 1,097,379    | 152,016| 0.138526   | 1    |
| New York       | 20,201,249   | 2,060,920| 0.102019  | 2    |
| Massachusetts  | 7,029,917    | 692,497| 0.098507   | 3    |
| Delaware       | 989,948      | 94,921 | 0.095885   | 4    |
| New Jersey     | 9,288,994    | 859,151| 0.092491   | 5    |
| South Carolina | 5,118,425    | 466,698| 0.091180   | 6    |
| Wisconsin      | 5,893,718    | 535,773| 0.090906   | 7    |
| Kansas         | 2,937,880    | 264,696| 0.090098   | 8    |
| Alaska         | 733,391      | 65,639 | 0.089501   | 9    |
| Hawaii         | 1,455,271    | 129,489| 0.088566   | 10   |
| Utah           | 3,271,616    | 289,753| 0.087318   | 11   |
| Illinois       | 12,812,508   | 1,110,569| 0.086679  | 12   |
| Louisiana      | 4,657,757    | 399,318| 0.085732   | 13   |
| Florida        | 21,538,187   | 1,844,383| 0.084768  | 14   |
| North Carolina | 10,439,388   | 884,922| 0.084129   | 15   |
| Kentucky       | 4,505,836    | 381,903| 0.084019   | 16   |
| West Virginia  | 1,793,716    | 151,977| 0.084019   | 17   |
| Vermont        | 643,077      | 54,458 | 0.084019   | 18   |
| California     | 39,538,223   | 3,326,342| 0.084019  | 19   |
| Arizona        | 7,151,502    | 600,864| 0.084019   | 20   |

TABLE 3 Results of circular distribution analysis.

| Cluster | $r$   | $r_0$ | $p$   |
|---------|-------|-------|-------|
| 1       | 0.492 | 0.013 | 0.001 |
| 2       | 0.476 | 0.011 | 0.000 |
| 3       | 0.539 | 0.019 | 0.001 |

$E(A) = \frac{n_A}{m_A} \times m_A$ (1)

$E(T) = \sum E(A)$ (2)

The probability density function of specific points observed at region $x$ is as follows:

$$f(x) = \begin{cases} 
    \frac{pE(x)}{pE(A)+q[E(T)−E(A)]}, & x \in A \\
    \frac{qE(x)}{pE(A)+q[E(T)−E(A)]}, & x \notin A 
\end{cases}$$ (3)

where $p$ is the ratio of actual incidence to expected incidence in window $A$, $q$ is the ratio of actual incidence to expected incidence outside window $A$, and the probability of any specific point is independent of all other points, one can also refer to Tang et al. (18) and Yang (19).

If $p > q$, the likelihood function $LR(A, p, q)$ is denoted by:

$$LR(A, p, q) = \frac{e^{-n_T}}{n_T!} \left( \frac{n_A}{E(A)} \right) ^ {n_A} \left( \frac{n_T - n_A}{E(T) - E(A)} \right) ^ {n_T - n_A} \prod_{x_i \in A} E(x_i)$$ (4)
Otherwise, the likelihood function $LR_0$ (based on invalid hypothesis) is

$$LR_0 = \frac{e^{-\lambda T}}{n_T!} \left( \frac{n_T}{E(T)} \right)^{n_A} \prod_{x_i \in A} E(x_i) \quad (5)$$

Test statistic for spatiotemporal scan $\lambda$ is defined as follows:

$$\lambda = \frac{\text{Sup}_{A,p > q} LR(A, p, q)}{\text{Sup}_{A,p=q} LR(A, p, q)} \quad (6)$$

According to Equations (4) and (5), we have

$$\lambda = \sup_A \left( \frac{n_A}{E(A)} \right)^{n_A} \left( \frac{n_T - n_A}{E(T) - E(A)} \right)^{n_T - n_A} \left( \frac{n_T}{E(T)} \right)^{-n_T} \text{I}(B) \quad (7)$$

$$B = \left\{ \frac{n_A}{E(A)} > \frac{n_T - n_A}{E(T) - E(A)} \right\} \quad (8)$$

In formula (7), $\text{I}(\cdot)$ is a characteristic function. The ratio of the actual incidence to the expected incidence in window $A$ is greater than the ratio of the actual incidence to the expected incidence outside window $A$. This is a measure of how risk within a cylinder differs from risk outside.

Next, we use Monte Carlo random method to simulate the $p$ value of LLR to determine whether the aggregation is statistically significant. First, we simulate $c$ random datasets, calculate the maximum LLR for each dataset, and rank it with the real LLR from big to small. If the real value rank is $R$, then we have

$$p = R(c + 1)^{-1} \quad (9)$$

If $p < 0.05$, we reject the original assumption. The relative risk of each aggregation is as follows:

$$\tilde{R} = \left( \frac{n_A}{E(A)} \right) \left( \frac{n_T - n_A}{E(T) - E(A)} \right)^{-1} \quad (10)$$

**Empirical analysis**

**Source of the data**

This paper selects 50 states from the US for research. The data of the COVID-19 mainly came from the data of the New York Times, including the date of diagnosis, the current area, and the source of infection of patients with COVID-19. Demographic data mainly came from the US 2020 census data, the basic geographic information data of each state were derived from Google satellite map data, and the latitude and longitude coordinates mainly chose the state capital as the center position.

The variant data are derived from the Centers for Disease Control and Prevention (CDC) study that tracks the proportion of variants estimated from weekly random sampling in the Department of Health and Human Services region followed by gene sequencing tests across the region, which we use to estimate the number of variant infections in the state over a week with new cases per day.

**Parameter setting**

We use the retrospective spatiotemporal analysis method, and choose the discrete Poisson model. The scanning time is set to be from 1 December 2021 to 31 January 2022, and the time interval is 1 day. As COVID-19 is highly contagious, the population with a ceiling of 50% in the space window is at risk, and the maximum circle size file is set at 30% of the population, rather than 30% of the regular population, and the regional overlap is set at zero. Referring to a large number of relevant literatures, combined with the actual situation, it is known that the outbreak of COVID-19 is fast, and the incubation period is short. Besides, the inaction of the US government to manage the outbreak makes the cycle longer. The daily pattern of COVID-19 changes rapidly, so the minimum time cluster is set to 1 day. In the test window, the number of Monte Carlo random simulation is set to 999.

**Description analysis**

In the population distribution, the US COVID-19 has nothing to do with gender, and included patients mainly in the age group of 44 to 59 years. In terms of time distribution, the US had the largest number of new cases on 10th January, with 1,420,374 cases. On 3rd, 18th, and 24th January, more than 1 million new cases were added daily with 1,003,751, 1,173,885, and 1,025,999 cases, respectively. In terms of the overall trend, the outbreak in the early stages of each state is relatively serious, and the number of confirmed cases has experienced a short lag and rapid growth. From Figure 1, we can see that the overall epidemic situation has not been effectively controlled, so the number of confirmed cases has increased cumulatively, having a certain increasing tendency. In the regional distribution, cases were mainly concentrated in the east and west of the US, and California has the largest number of confirmed cases, followed by New York.

**Space-time analysis**

A retrospective spatiotemporal analysis is carried out in the US. After SatScan 9.5 is run, 17 clustering areas are obtained and arranged from large to small according to the log-likelihood
ratio, and we obtain $p < 0.01$. The clustering areas are tested by the aboriginality test. The specific data are summarized in Table 1.

The cluster areas are mainly concentrated in Connecticut, Rhode Island, New York, Massachusetts, New Hampshire, and New Jersey from 1 December 2021 to 31 January 2022. The log-likelihood ratio is 101,152.56, and the aggregation is the highest, with a relative risk of 1.28. It also shows that the aggregation of COVID-19 in the six places during this period is strong. From 1 December 2021 to 31 January 2022, 18 states, such as Colorado, become the second agglomeration, with a log-likelihood ratio of 52,598.68 and a relative risk of 0.86. Texas from 1 December 2021 to 31 January 2022 is one of the three types of gathering areas, with a log-likelihood ratio of 42,332.33 and a relative risk of 0.81.

Combined with the daily incidence of each state, it can be observed that the starting time of the gathering area is just the time for the sudden increase of the confirmed cases of COVID-19 in the region, and the end time is the time for the growth rate of the confirmed cases to begin to decline. Combined with Table 2, the incidence of the four states involved, Rhode Island, New York, Massachusetts, and New Jersey, accounts for the top five regions of the incidence of COVID-19 in the US, and New York is the city with the second largest number of confirmed cases. Although Massachusetts and New Hampshire have fewer confirmed cases than New York, they are geographically close to New York, where the epidemic is relatively serious.

### Circular distribution analysis

Since the research time is 62 days, we divide 360° evenly over each day, then 1 day is equivalent to 5.81°, and 1 h is equivalent to 0.21°. To avoid the infinite calculation, the calculation time of

| Cluster | $\bar{x}$ | $s$   | Peak incidence | Epidemic peak period | Peak period span (Day) |
|---------|--------|------|-----------------|----------------------|------------------------|
| 1       | 32.7887| 73.5005 | Dec 6          | Dec 6–Dec 19         | 14                     |
| 2       | 87.1093| 69.7240 | Dec 16         | Dec 4–Dec 28         | 25                     |
| 3       | 77.1640| 63.6210 | Dec 14         | Dec 3–Dec 24         | 12                     |

**FIGURE 2**

Coordinated hierarchical cluster analysis.
TABLE 5 Correspondence tables of states.

| Code | State         | Code  | State            |
|------|---------------|-------|------------------|
| US-01| Alabama       | US-30| Montana          |
| US-02| Alaska        | US-31| Nebraska         |
| US-04| Arizona       | US-32| Nevada           |
| US-05| Arkansas      | US-33| New Hampshire    |
| US-06| California    | US-34| New Jersey       |
| US-08| Colorado      | US-35| New Mexico       |
| US-09| Connecticut   | US-36| New York         |
| US-10| Delaware      | US-37| North Carolina   |
| US-12| Florida       | US-38| North Dakota     |
| US-13| Georgia       | US-39| Ohio             |
| US-15| Hawaii        | US-40| Oklahoma         |
| US-16| Idaho         | US-41| Oregon           |
| US-17| Illinois      | US-42| Pennsylvania     |
| US-18| Indiana       | US-44| Rhode Island     |
| US-19| Iowa          | US-45| South Carolina   |
| US-20| Kansas        | US-46| South Dakota     |
| US-21| Kentucky      | US-47| Tennessee        |
| US-22| Louisiana     | US-48| Texas            |
| US-23| Maine         | US-49| Utah             |
| US-24| Maryland      | US-50| Vermont          |
| US-25| Massachusetts | US-51| Virginia         |
| US-26| Michigan      | US-53| Washington       |
| US-27| Minnesota     | US-54| West Virginia    |
| US-28| Mississippi   | US-55| Wisconsin        |
| US-29| Missouri      | US-56| Wyoming          |

The peak day and peaks of each cluster are summarized in Table 4.

Results comparison

The peak periods calculated by spatial-temporal scanning analysis are compared with those calculated by circular distribution method, as shown in Table 6. Combined with the actual situation, the peak period of the disease obtained by the circular distribution method is similar to the epidemic situation of COVID-19 in the region. The peak period of the disease obtained by the spatial-temporal scanning method is longer than that of the circular distribution method, and the time is generally advanced. Spatiotemporal scanning analysis can send early warning signals for COVID-19, which is a kind of fulminant and fast infectious disease, and has a higher practical value for disease prevention and control.

The clustering areas obtained by spatiotemporal scanning analysis are compared with the classification results obtained by the system clustering method, as shown in Table 7. In the Spatiotemporal Scanning Analysis (STSA), Hierarchical Cluster Analysis (HCA), and Coordinated Hierarchical Cluster Analysis (CHCA), US-44 (Rhode Island), US-25 (Massachusetts), US-09 (Connecticut), US-34 (New Jersey), and US-33 (New Hampshire) are classified into the first category, while US-36 (New York) is classified into the first category by spatiotemporal scanning. Combined with the actual situation, it can be seen that the results have a great relationship with the cumulative confirmed cases. After adding the coordinate index, the results are highly correlated with the geographical location, and the case information is weakened. The spatiotemporal scanning method makes good use of the information of regional population, case information, geographical location, and other information to give a reasonable clustering area. In terms of disease prevention and control, spatiotemporal scanning method can better provide theoretical basis for its adaptation to local conditions.

Through the abovementioned comparative analysis, it can be seen that the circular distribution method and the space-time scan method have a certain overlap interval in the peak
TABLE 6  Comparison of STSA and CDA.

| Cluster | Fastigium (STSA) | Span (Day) | Fastigium (CDA) | Span (Day) |
|---------|------------------|------------|-----------------|------------|
| 1       | Dec 1–Jan 31     | 62         | Dec 6–Dec 19    | 14         |
| 2       | Dec 1–Jan 31     | 62         | Dec 4–Dec 28    | 25         |
| 3       | Dec 1–Jan 31     | 62         | Dec 3–Dec 24    | 12         |

TABLE 7  Comparison of STSA and (C)HCA.

| Cluster | State (STSA) | State (HCA) | State (CHCA) |
|---------|--------------|-------------|--------------|
| 1       | US-09 US-44 US-36 US-25 US-33 US-34 | US-15 US-02 US-10 US-54 US-33 US-50 | US-09 US-10 US-13 US-17 US-18 US-20 |
|         |              | US-38 US-35 US-28 US-40 US-09 US-05 | US-21 US-24 US-25 US-26 US-29 US-33 |
|         |              | US-21 US-22 US-55 US-20 US-45 US-49 | US-34 US-37 US-39 US-42 US-44 US-47 |
|         |              | US-04 US-34 US-25 US-44 | US-50 US-51 US-54 US-55 |
| 2       | US-08 US-56 US-35 US-49 US-46 US-31 | US-08 US-01 US-27 US-29 US-47 US-18 | US-01 US-04 US-05 US-08 US-15 US-22 |
|         | US-20 US-40 US-38 US-04 US-30 US-19 | US-53 US-51 US-13 US-37 US-26 US-17 | US-28 US-35 US-40 US-45 US-49 |
|         | US-16 US-29 US-27 US-48 US-05 US-32 | US-42 US-39 | |
| 3       | US-48        | US-36 US-12 US-48 US-06 | US-12 US-36 US-06 US-48 |
| 4       | US-53 US-41 US-16 US-30 US-32 | US-32 US-19 US-41 US-24 US-46 US-31 | US-02 US-16 US-19 US-23 US-27 US-30 |

FIGURE 3  Proportion of COVID-19 variants.
Distribution of Omicron cases from Jan 25 to Jan 31, 2022.

**Omicron variation**

In the study of infectious diseases, we cannot ignore the situation of some variants. Based on the time node selected in this paper, the first Omicron case was reported in the US on 1 December, so we are paying attention to Omicron at this stage. Next, we need to know more about Omicron. In fact, Omicron has a significant growth advantage over Delta, leading to rapid spread in the community with higher levels of incidence than previously seen in this pandemic. With the sharp increase of cases and the scarcity of medical resources, we should also give importance to its dissemination.

From Figure 3, we can see that the B.1.167.2 (Delta) accounted for 99.25% on 4 December, while B.1.1.529 (Omicron) accounted for a low proportion. After 2 weeks, the proportion of Omicron increased rapidly, reaching 40.64%, while the corresponding Delta decreased to 58.08%. After another week, the proportion of Omicron exceeded Delta, becoming the largest
variant of infection. After the following 5 weeks, the proportion reached 95.31%. Within 2 months, Omicron became the mutant with the largest proportion of infection, and its propagation speed was very fast.

Based on the CDC’s tracking data of variants and the prediction of the proportion of variants, we calculate the number of variants per week in different regions according to the new cases per day and the proportion of variants per week in the corresponding region. The following figure clearly shows the cumulative number of Omicron cases in the last week. In order to show the map integrally, US-02 and US-15 have changed the actual location in the map. According to the number clustering, the map is divided into four categories. We can find the features in Figure 4.

The numbers of items in clusters 1–4 are 18, 29, 2, 1, respectively, and the cluster centers are 89,703, 24,703, 217,010, 507,330 respectively, in Figure 4. The number of Omicron cases increases rapidly in 2 months, with the largest number of cases in one state, US-06, accumulating to 507,330 per week, and also with the largest number in neighboring states.

To summarize, it can be clearly seen that the rate of infection of Omicron increased rapidly, and there is a trend of diffusion from the middle to the surrounding. Population flow is one of the reasons for the rapid spread of virus. The reality of the spread from densely populated cities to other cities can also be observed from the distribution map of Omicron cases.

Conclusion

In this paper, a retrospective spatiotemporal analysis of confirmed cases of COVID-19 in 50 states of the US is carried out. The first cluster is Connecticut, Rhode Island, New York, Massachusetts, New Hampshire, and New Jersey. The second cluster comprises 18 states, and the three types of gathering area is Texas. Through observation, it can be seen that the geographical location of the capital belonging to the same type of gathering area is relatively close. There is minimal difference between the gathering time and the peak time of newly confirmed cases daily, and the incidence of prominent gathering areas is higher. The reliability test of space-time scan results show that space-time scan has the advantages of accurate measurement in time and reasonable division of regions according to characteristics in space. On the basis of making full use of the existing time and spatial information, a spatiotemporal scanning analysis accurately locates the clustering area, timing and quantifying the corresponding clustering area, and evaluating the risk degree of the region, as we know that a high level of economic development and perfect medical conditions have played a positive role in the recovery of patients. From the analysis of this paper, spatiotemporal scanning analysis has greatly improved the timeliness and effectiveness of early warning of diseases, and can provide scientific basis for early prevention and control of diseases.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

YZ conceived and designed the study. QL analyzed the data. YZ and QL contributed to the writing of the manuscript. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Naus J. The distribution of the size of the maximum cluster of points on a line. J Am Stat Assoc. (1965) 60:532–8. doi: 10.2307/2282688
2. Kulldorff M. Prospective time periodic geographical disease surveillance using a scan statistic. J Royal Statist Soc. (2001) 164:61–72. doi: 10.2307/2680534
3. Kulldorff M, Heffernan R, Hartman J, et al. A space-time permutation scan statistic for the early detection of disease outbreaks. PLoS Med. (2005) 2:e59. doi: 10.1371/journal.pmed.0020059
4. Kulldorff M, Huang L, Pickle L, Duczmal L. An elliptic spatial scan statistic. Stat Med. (2006) 25:3929–43. doi: 10.1002/sim.2490
5. Kulldorff M, Mostashari F, Duczmal L, Katherine Yih W, Kleinman K, Platt R. Multivariate scan statistic for disease surveillance. Stat Med. (2007) 26:1824–33. doi: 10.1002/sim.2818

6. Jung I, Kulldorff M, Klassen A. A spatial scan statistic for ordinal data. Stat Med. (2007) 26:1594–607. doi: 10.1002/sim.2607

7. Jung I, Kulldorff M, Richard OJ. A spatial scan statistic for multinomial data. Stat Med. (2010) 29:1910–8. doi: 10.1002/sim.3951

8. Huang L, Kulldorff M, Gregorio D. A spatial scan statistic for survival data. Biometrics. (2007) 63:109–18. doi: 10.1111/j.1541-0420.2006.00661.x

9. Huang L, Tiwari RC, Zou Z, Kulldorff M, Feuer EJ. Weighted normal spatial scan statistic for heterogeneous population data. J Am Stat Assoc. (2009) 104:886–98. doi: 10.1198/jasa.2009.ap07613

10. Wieckowska B, Górna I, Trojanowski M, Pruciak A, Stawinska-Witoszyńska B. Searching for space-time clusters: the CutL method compared to Kulldorff’s scan statistic. Geospatial Health. (2019) 14:314–20. doi: 10.4081/gh.2019.791

11. Li D, Fang Z, Yu Y. Scan statistic: a new method to detect the persistence of fund performance. Operat Res Manag. (2006) 15:82–7. doi: 10.3966/j.issn.1007-3221.2006.01.018

12. Yin F. Application of Spatio-Temporal Scan Statistic in Early Warning of Infectious Diseases. Chengdu: Sichuan University (2007).

13. Má Y, Li X, Zhang Y. Spatial scale selection of scan statistic in infectious disease surveillance. Modern Prevent Med. (2011) 38:1601–4.

14. Höhl A, Delmelle EM, Desjardins MR, Lan Y. Daily surveillance of COVID-19 using the prospective space-time scan statistic in the United States. Spat Spatiotemporal Epidemiol. (2020) 34:100354. doi: 10.1016/j.sste.2020.100354

15. Xu FY, Beard K. A comparison of prospective space-time scan statistic and spatiotemporal event sequence based clustering for COVID-19 surveillance. PLoS ONE. (2021) 16:e0252990. doi: 10.1371/journal.pone.0252990

16. Höhl A, Delmelle E, Desjardins M. Rapid detection of COVID-19 clusters in the United States using a prospective space-time scan statistic: an update. SIGSPATIAL Special. (2020) 12:27–33. doi: 10.1145/3404111.3404116

17. Pei T, Wang X, Song C, Liu Y, Huang Q, Shu H, et al. Research progress on spatiotemporal analysis and modeling of COVID-19 epidemic. J Geo-Inform Sci. (2021) 23:188–210. doi: 10.12082/dipxxs.2021.200434

18. Tang X, Zhou H. Scanning statistics and its application in epidemiology. China Health Stat (2011) 28:332–7. doi: 10.3969/j.issn.1002-3674.2011.03.042

19. Yang L. The Distribution Characteristics of COVID-19 in Wenzhou City Based on Scan Statistic. Hangzhou: Hangzhou Dianzi University (2020).