Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Rapid surveillance of COVID-19 by timely detection of geographically robust, alive and emerging hotspots using Particle Swarm Optimizer

Ankita Wadhwa *, Manish Kumar Thakur

Department of Computer Science Engineering and IT, Jaypee Institute of Information Technology, A-10 Sector 62, Noida, UP, 201309, India

ARTICLE INFO
Keywords:
Coronavirus
COVID-19
Hotspot detection
Spatio-temporal analysis
Particle swarm optimizer

ABSTRACT
A novel virus, called severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has rapidly become a pandemic called Coronavirus disease 2019 (COVID-19). According to the World Health Organization, COVID-19 was first detected in Wuhan city in December 2019 and has affected 216 countries with 9473214 confirmed cases and 484249 deaths globally as on June 26th, 2020. Also, this outbreak continues to grow in many countries like the United States of America (U.S.), Brazil, India, and Russia. To ensure rapid surveillance and better decision-making by government authorities in different countries, it is vital to identify alive and emerging hotspots within a country promptly. State-of-the-art methods based on space-time scan statistics (like SaTScan) are not geographically robust. Also, due to the enumeration of many Spatio-temporal cylinders, the computation cost of Spatio-temporal SaTScan (ST-SaTScan) is very high. In the applications like COVID-19 where we need to detect the emerging hotspots daily as soon as the new count of cases gets updated, ST-SaTScan seems inefficient. Therefore, this paper proposes a Particle Swarm Optimizer-based scheme to timely detect geographically robust, alive, and emerging COVID-19 hotspots in a country. Timely detection can help government officials design better control strategies like increasing testing in hotspots, imposing stricter containment rules, or setting up temporary hospital beds. Performance of ST-SaTScan and proposed scheme have been analyzed for four worst-hit U.S. states for the incubation period of 14 days between June 11th, 2020, and June 24th, 2020. Results indicate that the proposed scheme detects hotspots of a higher likelihood ratio (a measure to indicate the significance of hotspot) than ST-SaTScan in significantly less time. We also applied the proposed scheme to detect the emerging COVID-19 hotspots in all states of the U.S. During the study period, the proposed scheme has detected 104 emerging COVID-19 hotspots.

1. Introduction

Coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), was first identified in December 2019 in Mainland China, with its epicenter at Wuhan City (C. Wang, Horby, Hayden, & Gao, 2020). The most common symptoms of COVID-19 include fever, cough, shortness of breath, and fatigue; however, severe cases may suffer from organ failures, heart attacks, pneumonia, or even death (Sanche et al., 2020). The novelty of COVID-19, its highly contagious nature, and lack of appropriate medicinal knowledge resulted in its rapid outbreak worldwide. Consequently, COVID-19 was declared a pandemic by the World Health Organization on March 11th, 2020 (Cucinotta & Vanelli, 2020). On June 27th, 2020, there were 9473214 confirmed cases and 484249 deaths globally. However, the death rate of COVID-19 lies between 1% and 5%, and 80% of the total cases are mild ones (Ruan, Yang, Wang, Jiang, & Song, 2020).

The lack of proper medication and vaccines for COVID-19 forced the governments of different countries to impose a complete lockdown on the entire country. Depending on the outbreak’s severity, many countries strictly implemented the lockdown; however, governments in some countries have taken various other preventive measures. These measures include restrictions on international and domestic flights, restricted inter-territorial movements for non-essential commodities, etc. The complete lockdown and other preventive measures had severe and apparent effects on the economies, resulting in lifting the lockdowns even with a surge of confirmed COVID-19 cases. However, in fighting the outbreak of a highly infectious disease like COVID-19 and also to ensure little adversity in economies, it is essential to target the prioritized locations (known as hotspots) for stricter containment measures, increased testing, and ramping up of medical facilities (Desjardins, Hohl,
For many years, spatial hotspot detection techniques have been applied to detect significant spatial hotspots for various applications. These techniques can also be used to identify COVID-19 hotspots and thus prioritize them for immediate actions and preventive measures.

Formally, a spatial hotspot is defined as an area (or zone) where cases inside the area are significantly greater than the number of cases outside that area (Eftelioglu, Tang, & Shekhar, 2015). In addition to spatial dimensions, time is used as a third dimension in Spatio-temporal analysis, which helps identify active, emerging, diminishing, or persistent hotspots. In monitoring outbreaks like COVID-19, decision-makers are often interested in those Spatio-temporal hotspots which are active at the end of the study period, called alive or emerging hotspots.

Researchers have made significant contributions to detect spatial and spatio-temporal hotspots of different shapes useful in various application domains in the past few years. Some of the contemporary contributions made in this regard are discussed next.

1.1. Related work

The contemporary methods for spatial and spatio-temporal hotspot detection can be broadly classified into two categories: clustering-based approaches and Scan-statistics-based approaches. Clustering-based techniques assume hotspots to be clusters of spatial locations, where, location points within a cluster are closer, and the points in different clusters are farther from each other. Considering this interpretation, numerous approaches have been proposed to detect the hotspots which utilize the basic principles of clustering. These works identify hotspots using conventional clustering methods. Recently, many authors (Kumar, n.d.; Lawson, 2010; Shi & Pun-Cheng, 2019) consolidated and reported the details of some of the clustering based (partition based, density based etc.) hotspot detection methods.

In one contribution, authors (Tayal et al., 2014) developed a partition based clustering approach for crime hotspot detection and criminal identification on dataset of India. They used k-means clustering to identify circular hotspots. Certain limitations of k-means clustering in context of hotspot detection were observed. First, the value of k (number of clusters) needs to be specified, which may not be known beforehand. Second, k-means simply partitions all the points into k clusters, where each cluster contains approximately same number of points. Thus, this type of distribution may not represent significant hotspot regions. Further, in another contribution, authors (F Di Martino, Loia, & Sessa, 2008) extended k-means by incorporating fuzzy logic to determine the value of k automatically. Furthermore, another work (Ferdinando Di Martino & Sessa, 2013) introduced a Fuzzy PSO based clustering approach to identify spatial hotspots. The authors claimed that Fuzzy PSO approach for partition clustering is efficient in determining the optimal value of k in k-means, is robust to outliers and takes less computational cost than Fuzzy k-means approach. Thus, they proved their scheme to be more efficient for large datasets.

Apart from partition-based clustering, numerous conventional density-based clustering methods like DBSCAN (Ester, Kriegel, Xu, & Miinch, 1996), ST-DBSCAN (Birant & Kut, 2007), CLIQUE (Agrawal, Gehrke, Gunopulos, & Raghavan, 1998), STING (W. Wang, Yang, & Muntz, 1997), etc. have been used for identifying spatial and spatio-temporal hotspots. Out of these approaches DBSCAN (Ester et al., 1996) is the most widely used method which groups nearby dense points in the given spatial space. The grouping contains at least minpts points and then also includes other points within ε distance (called ε neighbourhood) of these points. In addition, ST-DBSCAN (Birant & Kut, 2007) is the extension of DBSCAN proposed by authors to deal with spatio-temporal data. Choosing optimal values of user specific parameters viz. minpts and ε in DBSCAN and ST-DBSCAN requires domain expertise. To select the optimal values of DBSCAN related parameters, researchers have proposed different solutions. For instance, in one work, authors (Scitovski, 2018) proposed a modified DBSCAN called rough DBSCAN which automatically identifies the DBSCAN related parameters. They also applied their proposed scheme to identify earthquake prone zones. In another similar study for earthquake zoning, authors of (Ankita & Thakur, 2018) presented a meta-heuristic algorithm to identify DBSCAN related parameters and showed effectiveness of density based clustering schemes for irregular shaped spatial hotspot detection. From the related work discussed in clustering based hotspot detection category, it can be observed that clustering-based techniques offers low computational complexity; however, they may result in false hotspots due to lack of statistical significance tests (Eftelioglu, Shekhar, Kang, & Farah, 2016). Further, clustering-based approaches need some preliminary information related to the method, which may not be available beforehand.

On the other hand, scan statistics-based approaches use a scan window of the desired shape (circular, ring, ellipse, cylindrical etc.) and a statistical test to identify significant hotspots. Kulldorff has made the initial contribution in scan statistics-based methods for detecting spatial hotspots of circular and elliptical hotspots (Martin Kulldorff, 1997; Martin Kullendorff, Huang, Pickle, & Duczmal, 2006). The work was extended on spatio-temporal domain by Kullendorf (Martin Kullendorff, 2001) and is popularly known as ST-SaTScan. Kullendorf’s space–time scan statistic has been used in literature to detect emerging hotspots of various diseases like influenza (Li, Fu, Lin, & Jiang, 2019), brain cancer (Martin Kullendorff, Athas, & Miller, 1998), malnutrition (Verdiana & Widyaniingsih, 2017), legionnaires disease (Edens et al., 2019) and most recently COVID-19 (Desjardins et al., 2020). ST-SaTScan uses a cylindrical window of three dimensions where the base represents the space and height represents the time. This cylindrical window moves in the spatial and temporal domain in search space to enumerate candidate cylindrical hotspots. Spatially, many cylinder bases with different geographical centers and radius are generated by considering each county (or equivalent administrative units) as center and distance to every other county (or equivalent administrative units) as radii (Martin Kullendorff, 2001). Thus, spatially SaTScan allows only the location of counties to be the center of the cylinder. Due to the selection criteria of center and radii, SaTScan may miss detecting some significant hotspots where the center of hotspot does not lie exactly on the location of a county and hence is not geographically robust (not sensitive to geographical barriers). Temporally, for each cylinder base, consideration of different starting dates results in different cylinder heights. Hence, many candidate cylindrical hotspots with varying bases and heights are generated. A statistical measure called Log-Likelihood Ratio (Martin Kullendorf, 1997), indicating the strength of a hotspot, is then calculated for each cylinder. The significance of a hotspot is confirmed by relying on randomization tests to determine whether the observed pattern is significant or has occurred by chance. Due to the generation of a large number of cylinders, followed by a randomization test, SaTScan is a high computational cost method.

In subsequent years various researchers have tried to improve Kullendorf’s approach in terms of shape and efficiency by applying pruning, filtering, and speedups. In one contribution, authors (Eftelioglu et al., 2015) improved the SaTScan algorithm for circular hotspot detection by using a cubic grid circle listing algorithm. They applied an upper bound on each grid’s likelihood ratio to detect circular hotspots. Through their
approach, authors identify non-contiguity crime hotspots separated by rivers, mountains, etc. However, their approach is limited to the spatial domain. In another contribution, the authors (Tang, Eftheligouli, & Shekhar, 2015) proposed a fast elliptical hotspot detection method and detected crime hotspots in Denver city. Using a lookup table and upper bound pruning technique, the authors achieved lower computational complexity than Kulldorff’s elliptical hotspot detection method.

Another study (Eftheligouli, Shekhar, et al., 2016) proposed a ring-shaped hotspot detection method based on dual grid pruning and found ring-shaped crime hotspots in San Diego city. Also, a technique for linear hotspot detection was proposed by (Tang, Eftheligouli, Oliver, & Shekhar, 2017) and used to detect pedestrian fatality hotspots in Florida. The author proposed an efficient method by applying the shortest path tree pruning and speeding up the Monte Carlo simulations.

Also, irregularly shaped hotspots have been detected in (Duczmal, Cançado, Takahashi, & Bessegato, 2007) by using a Genetic Algorithm based scan statistics approach for optimizing the shape of the scan window. Another approach for irregularly shaped hotspot detection based on a multi-objective Genetic algorithm was proposed by (Cançado et al., 2010). It optimizes two objectives: likelihood ratio and the penalty function for restricting excessive freedom in shape. Finally, in a study, authors (Izakian & Pedrycz, 2012) used Particle Swarm Optimization (PSO) to optimize the shape of scanning window hotspot detection.

In general, the discussed contributions fall under two broad categories (Clustering-based and Scan statistics based) and suggests following: (a) clustering-based hotspot detection techniques may detect false hotspots and also needs prior information like the number of clusters (Eftheligouli, Shekhar, et al., 2016) whereas, (b) scan statistics based approaches rules out the false hotspots and detects such hotspots which are statistically significant but at a high computation cost (c) the modifications to improve SaTScan/ST-SaTScan available in the literature are limited to optimizing the shape of hotspots and not in terms of improving its efficiency or geographical robustness.

2. Highlights and outline of the paper

Considering the limitations of discussed approaches, we have proposed a novel Particle Swarm Optimizer based scheme for detecting geographically robust Spatio-temporal hotspots in this paper. The proposed approach facilitates the timely identification of geographically robust Spatio-temporal, alive, and emerging COVID-19 hotspots and contributes to the ongoing COVID-19 surveillance. We have also presented the comparison between the proposed approach and the traditional SaTScan method using county-level COVID-19 data of a few worst-hit states of the United States of America (U.S.), viz. New York, Texas, Florida, and California. The comparison is made using time taken to detect the hotspots and the quality of detected hotspots (in terms of Log-Likelihood Ratio). Motivated by the reduced computational cost taken by the proposed scheme to identify hotspots in worst-hit U.S. states, the PSO-based scheme is further applied to identify all U.S. states’ alive and emerging hotspots. In all these analyses (either comparison with SaTScan over worst-hit U.S. states or finding emerging hotspots of all the U.S. states), the COVID 19 data of the U.S. from June 11th, 2020 to June 24th, 2020 is considered.

The rest of the paper is organized as follows: Section 2 reviews the Kulldorff’s SaTScan method adopted for COVID-19 hotspots detection. It also forms a baseline for our proposed approach. Section 3 presents the proposed Particle Swarm Optimizer based approach for detecting alive and emerging COVID hotspots and is referred to as ST-PSO-GRHD (Spatio Temporal PSO-based Geographically Robust Hotspot Detection). Section 4 details the experimental studies and results. It includes brief details of the datasets, followed by a comparative analysis of both approaches on four worst-hit states of the U.S. using day-wise confirmed cases count. In this section, the performance of the proposed PSO-based approach is also presented to detect alive and emerging hotspots of the entire U.S. (all the 3026 counties). Finally, conclusions and future scope of the work are discussed in Section 5.

3. Spatio-temporal SaTScan: A review in the context of COVID-19 hotspots

SaTScan is one of the often-used approaches (Martin Kulldorff, 1999) or software (M Kulldorff, 2018) to detect spatial, temporal, and Spatio-temporal clusters (or hotspots). The basis of SaTScan is scan statistics. In search of active and emerging COVID-19 hotspots, SaTScan has been explained here to detect cylindrical hotspots, i.e., Spatio-temporal (S.T.) hotspots. Depending on data availability, SaTScan can be used either for point data (each point represents the location of one event occurrence) or aggregated data (e.g., county or equivalent administrative block level). The considered data for COVID-19 is available at the aggregate level, i.e., county-wise data of U.S. states (The New York Times, 2020); therefore, SaTScan has been explained here in the context of aggregate data. In this study, the considered administrative unit is U.S. county.

For Spatio-temporal analysis, Spatio-temporal SaTScan (ST-SaTScan) moves a cylindrical scan window of varying radii and heights in the search space S to detect areas of unusually high concentration of an event. Each cylindrical zone Z(xc, yc, r, t) is defined using the center of the base (xc, yc) representing latitude and longitude of a county, radius of the base r, and height of the cylinder t representing last t days of the study period. To detect active and emerging hotspots, the duration of cylindrical window is considered between any one day of study period as the starting date and the last day of the study period as the end date. All the candidate hotspots are enumerated. The detailed procedure of enumerating cylinders is presented in Step 1 of Algorithm 1.

Further, a hypothesis test is required, and for each cylinder, the following question needs to be addressed: “Is there any difference in the concentration of events inside vs. outside the cylinder?”. In ST-SaTScan, the hypothesis test is constructed with two outcomes, null hypothesis H0 and alternate hypothesis H1. The null hypothesis states that “the number of occurrences of an event follows complete spatial randomness (CSR) in the study area S”, whereas the alternate hypothesis states that “the frequency of occurrence of an event is higher within the cylinder as compared to outside”.

In the hypothesis test, the strength of a cylindrical window is calculated to check its candidature for a hotspot. Log-likelihood ratio (LLR) (Xie, 2019) & Shi & Pun-Cheng, 2019) is one of the often-used measures in ST-SaTScan to compute the strength and accordingly decide the candidature of the obtained cylindrical window as a hotspot. Let A be the total number of cases in all counties. Let c(Z) be the actual observed number of cases within all counties of cylinder Z. As per the null hypothesis H0, the expected number of cases within Z denoted by E(Z) can be defined using (1) by the unitary method.

\[
C_{x}(Z) = \frac{A \cdot \text{pop}(Z) \cdot t}{\text{pop}(S) \cdot T} \tag{1}
\]

where \(\text{pop}(Z)\) is the population of all counties within the circular base of cylinder Z, \(\text{pop}(S)\) is the total population of study area S, T is the number of days under consideration for analysis. Next, the probability that a particular case lies within the hotspot Z if H0 were true is given by \(C_{x}(Z)\) and for the case to be outside Z is \(1 - C_{x}(Z)\). If all the cases are uniformly distributed and H0 is true, the total probability that exactly \(C_{x}(Z)\) points are present within Z is given by \(P_{0}\) as shown in (2).
Further, the alternate hypothesis $H_1$ considers that $H_0$ is not true. In this case, the probability that any case lies within the hotspot $Z$ is given by $\frac{c_a(Z)}{A}$ and that the case lies outside $Z$ is $\frac{c_e(Z)}{A}$. The total probability that exactly $c_a(Z)$ points are present within $Z$ if the actual distribution is true is given by $P_1$, as shown in (3).

$$P_1 = \left( \frac{c_a(Z)}{A} \right) \times \left( \frac{A - c_a(Z)}{A} \right)^{A - c_a(Z)}$$  \hspace{1cm} (3)

The likelihood ratio is defined as $P_1$. Taking logarithm on both sides, the expression for LLR is shown in (4).

$$LLR(Z) = \log \left( \frac{c_a(Z)}{c_e(Z)} \right) + \left( \frac{A - c_a(Z)}{A} \right)^{A - c_a(Z)} \times I(c_a(Z), c_e(Z))$$  \hspace{1cm} (4)

where $I(c_a(Z), c_e(Z))$ is a binary indication function. $I()$ has a value of 1 when a zone $Z$ has more activity points than expected ($c_a(Z) > c_e(Z)$); otherwise, it is set to 0 to prevent the detection of low cases areas.

Along with LLR, its distribution under the null hypothesis is also required to perform the hypothesis test. This distribution indicates what to expect when there are no hotspots in the study area. A randomization test is carried out to obtain this distribution as follows: Monte Carlo Simulations (MCS) is used to perform the randomization test and get the distribution of LLR under $H_0$. In total, $m$ random datasets in the study area are generated using the Poisson point process (CSR) (assuming that COVID-19 cases follow a Poisson distribution according to the population of the geographic area). All possible cylinders are enumerated for each of the $m$ random datasets. As the higher value of $\frac{P_1}{P_0}$ (likelihood ratio) indicates that more cases have been reported than expected, the highest LLR value out of all cylinders is stored. This is done for all $m$ datasets, and the values are stored in a list. The list is named as maxLLRs. Next, to obtain the distribution, which suggests what type of LLR value to expect in the same study area when there are no hotspots, the maxLLRs list is sorted in descending order.

After enumerating all the candidate cylinders and their corresponding LLR values, the cylinder with maximum LLR is kept among overlapping cylinders, and the rest are removed. The statistical significance of each cylinder is then calculated using its LLR value and distribution of LLR under the null hypothesis. A cylindrical zone $Z$ is referred as a statistically significant hotspot based on its p-value (referred as significance level). Relative position $p$ of LLR(Z) in maxLLRs list is determined first to calculate the p-value of the cylindrical zone $Z$. The p-value is then calculated by taking the ratio between $p$ and $m + 1$, where $m$ is the number of Monte Carlo Simulations. If the p-value of a cylindrical zone $Z$ is less than the desired level of significance $\alpha$, then $Z$ is considered as a statistically significant alive or emerging hotspot; otherwise, the alternate hypothesis $H_1$ is rejected for that $Z$.

Obviously, the number of enumerated cylinders is very large, so each cylinder is expanded till the user-inputted maximum spatial bound (radius in km) and maximum temporal bound (duration in days) is reached. As detailed in Section 4, experimental analysis is carried out with user-inputted spatial bound, and temporal bound is taken as 50% of the study period. However, decision-makers of different agencies can decide these bounds so that the government can adequately look and cater to each hotspot. The ST-SaTScan scheme for detecting alive and emerging COVID-19 hotspots is presented in Algorithm 1.

**Table 1**

| Input Description | Description |
|-------------------|-------------|
| Cases [C][T]      | C: number of counties | T: number of days in the study period. Cases(i,j): represents the number of confirmed COVID-19 cases reported in ith county on jth day. |
| Population[C]     | The population of C counties |
| min_rad, max_rad  | The lower and upper bound of desired hotspots |

**Table 2**

| Decision Variable | Description |
|-------------------|-------------|
| $(x_1, y_1)$       | Center of the base of cylinder (a latitude-longitude pair representing a location in search space S) |
| $r$               | The radius of the cylinder |
| $t$               | Height of cylinder from the end of the study period (representing time window (T-t+1)$^p$ to T$^p$ day). This allows only alive and emerging hotspots. |

The computational complexity of ST-SaTScan scheme presented in Algorithm 1 is computed as follows: Step 1 of the algorithm enumerates all cylinders by generating O($C^2$) spatial circles, and for each spatial base, O(T) temporal windows are generated. For each of the O($C^2T$) cylindrical windows, LLR computation takes O(C) time. Hence, Step 1 requires O($C^2T$) time. Step 2 of the ST-SaTScan scheme, repeats Step 1 for ‘m’ number of times to find the cylinder with maximum LLR in each Monte Carlo simulation. Thus, the computation time required for Step 2 is O($m \times C^2 \times T$) time. Step 3 checks each cylinder for its statistical significance and takes O($C^2T$) time. Hence, the total time complexity of ST-SaTScan is O($m \times C^2 \times T$).

As seen from Algorithm 1, SaTScan enumerates only those cylinders which are centered at the centroid of some county. Therefore, it may miss detecting some of the significant hotspots whose center lies at some location other than a county centroid (Eftelioglu, Tang, & Shekhar, 2016). Thus, the hotspots detected by ST-SaTScan may or may not be geographically robust and may report reduced significance. Geographical robustness can be added by allowing any random location (apart from the centroid of a county) to be selected as the center of the spatial base. But it will enlist an infinite number of cylinders. Thus, adding geographical robustness to Spatio-temporal hotspot detection makes it an NP-class problem. Considering the significance of geographical robustness in detecting COVID 19 alive and emerging hotspots, we propose a Particle Swarm Optimizer (PSO) based scheme for the same in the next section.

**Algorithm 1.** ST-SaTScan for detecting COVID-19 Hotspots
As discussed in Section 2, finding geographically robust hotspots using ST-SaTScan is an NP-class problem. Therefore, this section proposes an efficient scheme for detecting geographically robust alive and emerging COVID-19 hotspots using Particle Swarm Optimizer (PSO). The motivation to use PSO is to reduce the time taken to detect Spatio-temporal COVID-19 hotspots.

This section first gives a brief detailing of the Particle Swarm Optimizer. Secondly, we propose and model the Spatio-temporal hotspot detection problem as an optimization problem. Finally, a complete scheme for detecting significant COVID-19 Spatio-temporal hotspots along with its complexity analysis is presented. The proposed PSO-based scheme for Spatio-Temporal Hotspots Detection is now onwards referred to as ST-PSO-GRHD.

4. Proposed Scheme-Spatio Temporal Hotspot detection using PSO

As discussed in Section 2, finding geographically robust hotspots using ST-SaTScan is an NP-class problem. Therefore, this section proposes an efficient scheme for detecting geographically robust alive and emerging COVID-19 hotspots using Particle Swarm Optimizer (PSO). The motivation to use PSO is to reduce the time taken to detect Spatio-temporal COVID-19 hotspots.

This section first gives a brief detailing of the Particle Swarm Optimizer. Secondly, we propose and model the Spatio-temporal hotspot detection problem as an optimization problem. Finally, a complete scheme for detecting significant COVID-19 Spatio-temporal hotspots along with its complexity analysis is presented. The proposed PSO-based scheme for Spatio-Temporal Hotspots Detection is now onwards referred to as ST-PSO-GRHD.

Algorithm 1: ST-SaTScan for detecting COVID-19 Hotspots

Input:
  i) A matrix of cases named ‘cases’ where, cases(i, j) represents the number of COVID-19 cases in ith county (total C counties in the study area) on jth day (total T days in study duration)
  ii) Loc[C] and Pop[C] containing (latitude, longitude) pairs and population of each county respectively.
  iii) Max_rad representing upper bound on radius of the hotspot.
  iv) The number of Monte Carlo simulations ‘m’ and p-value threshold αp.

Output: Non-overlapping alive and emerging COVID-19 hotspots with p-value <= αp.

Algorithm:
Step 1: Enumerate cylinders and calculate their LLR
  for each county i do
    for each county j except i do
      rad=distance( Loc(i), Loc(j))
      if rad=max_rad
        break;
      end
    end
  end

Step 2: Generate distribution of LLR under H0
  maxLLRs=[];
  for each Monte Carlo simulation i= 1 to m
    cases’ = PoissonDistribution();
    Compute maximum LLR’ cylinder using step 1
    append (maxLLRs, LLR’);
  end
  maxLLRs=sort(maxLLRs, desc);

Step 3: Test cylinders generated in step 1 for statistical significance
  for each cylinder H in cylinders
    pos= the position of LLR of cylinder in the maxLLRs
    p-value_H = pos/m+1;
    if (p-value_H) < αp
      append (Hotspots_significant, H);
  end

4.1. Particle Swarm Optimizer

Particle Swarm Optimizer (PSO) was proposed in 1995 by Kennedy and Eberhart (Eberhart & Kennedy, 1995) to copy the optimization behavior seen in a group of organisms in nature like fishes and birds (called as particles) and use it to solve real-world optimization problems. In PSO, a group of particles coordinates to reach an optimal solution to a problem. Each particle i possesses a d-dimensional position xi (representing a solution) and a d-dimensional velocity, vi where d is the dimension of the search space. Each candidate solution is evaluated using a fitness function f(). The first step of the algorithm is random initialization of each particle’s position xi and zero initialization of its velocity vi. Each particle then iteratively moves to better positions in the search space using its own experience and the social experience of other particles. The movement of particles over the iterations is modeled using two equations, viz. velocity update (5) and position update (6). Each
particle’s movement is guided by its own local best position (called pbest) as well as the best position obtained by the entire group of particles (called gbest). The velocity of ith particle at t + 1th iteration and the new position of the particle are obtained using (5) and (6) respectively (Clerc & Kennedy, 2002).

\[
LLR(Z) = \log \left( \frac{(c_i(Z))^{c_i(Z)} \times ((A - c_i(Z))^c_i(Z))^{-1} + I(c_i(Z), c_i(Z))}{A_i} \right), c_i(Z) = \frac{A \times \text{pop}(Z) \times t}{\text{pop}(S) \times T}
\]  

\[x_i(t + 1) = x_i(t) + v_i(t + 1)
\]

where \(\omega\) represents the inertia weight of the particle, \(\text{rand}_1\) and \(\text{rand}_2\) are random numbers in the range \([0, 1]\), and \(\alpha\) and \(\beta\) represent the weightage given to cognitive component and social component of the swarm, respectively. The particles keep updating their positions and velocities until some convergence criteria are met. Convergence criteria can be either a maximum number of iterations or the same fitness if obtained for a fixed number of iterations. \(\chi\) in (5) is a constriction coefficient and is defined in (7).

\[
\chi = \frac{2}{\sqrt{2 - \psi} - \sqrt{\psi^2 - 4 \times \psi}}
\]

where \(\psi = \alpha + \beta\) should be greater than 4 to ensure stability. With the inclusion of constriction factor, PSO can obtain better quality solutions by maintaining a balance between exploration and exploitation (Clerc & Kennedy, 2002; Lim, Montakhab, & Nouri, 2009).

4.2. An optimisation model for geographically robust spatio-temporal hotspot detection

In this section, we model the Spatio-temporal hotspot detection problem as an optimization problem. The presented optimization model is aimed to detect alive and emerging COVID-19 hotspots. For the COVID-19 dataset having day-wise cases in C counties of U.S. for T number for days, the optimization model aims to find such an alive or emerging cylinder in search space with the highest likelihood ratio. The obtained cylinder with the highest likelihood ratio represents the most significant hotspot. The model is reused by removing counties obtained in the previous cylindrical hotspot to find the subsequent hotspots. The model is used repeatedly until a non-significant hotspot (identified using a hypothesis test) is obtained. The inputs, decision variables, fitness function, and constraints of the optimization model are presented in Table 1, Table 2, Equation (8), and Equations (9)–(12), respectively.

Maximize \(LLR(Z|x_c, y_c, r, t)\) where,

\[v_i(t + 1) = \chi \times [\omega \times v_i(t) + \alpha \times \text{rand}_1 \times (pbest - x_i(t)) + \beta \times \text{rand}_2 \times (gbest - x_i(t))]\]

Subject to

\[S_{min} \leq x_i \leq S_{max}\]

\[S_{min} \leq y_i \leq S_{max}\]

where \(S_{min}\) and \(S_{max}\) are decision variables for the model. The model aims to find the cylinder with the maximum likelihood ratio. To calculate the expected count of cases in a cylinder, we assume that COVID-19 cases follow a Poisson distribution of LLR under null hypothesis \(H_0\), (2) finding the cylindrical hotspot with maximum likelihood ratio in the search space \(S\), and (3) statistical significance using hypothesis test.

The first step of the proposed scheme involves generating the distribution of LLR under null hypothesis \(H_0\), which states that there is no significant hotspot in the search area. This distribution is called maxLLRs and is obtained using Monte Carlo simulations in a similar manner, as discussed in Section 2.

In the second step, the proposed scheme uses PSO to find the cylindrical hotspot with maximum likelihood ratio in the search space \(S\). The decision variables, fitness function, and constraints used by PSO are presented in Section 3.2. Initially, each of the \(K\) particles is randomly placed in the search space where each particle has 4 dimensions representing the center \((x_c, y_c)\), radius \(r\), and height \(t\) of the cylindrical window. To detect alive and emerging hotspots only, the height \(t\) of the cylindrical window is considered from the end date of the study period. The fitness value (LLR) of each particle is calculated using (8) to identify the global best (gbest) and personal best (pbest) position of each particle. While calculating LLR for a cylindrical zone, all those counties whose centroid location lies within the circular spatial base of the cylinder are considered inside the zone \(Z\). The sum of the day-wise count of cases from \((T-1)th\) to \(Tth\) day in counties inside \(Z\) is used as \(c_i(Z)\). Particles then update their velocities and positions based on pbest and gbest using (5) and (6). All the particles keep wandering in the search space to reach the optimum value of LLR until a convergence criterion is met. The ST-PSO-GRHD scheme converges if the best solution in the swarm does not improve continuously for a fixed number of iterations or a maximum number of iterations has been reached. After convergence, the global
best value gbest represents the cylindrical hotspot C.H. with the maximum LLR value.

The third step of the proposed scheme checks the statistical significance of hotspot C.H. using a hypothesis test. The p-value of C.H. is obtained in a similar way as discussed in Section 2. If the p-value of C.H. is less than $\alpha_p$, then we store C.H. in a list of significant cylindrical hotspots. Searching for the next non-overlapping hotspot begins by removing the counties within C.H. from county set C and repeating steps 2 and 3 with the remaining counties. If the p-value of any cylindrical hotspot obtained in step 2 is greater than $\alpha_p$, we say that C.H. is not a significant hotspot and terminate the algorithm. Finally, the list of significant hotspots is outputted. The complete ST-PSO-GRHD scheme is presented in algorithm 2.

Algorithm 2. ST-PSO-GRHD for detecting emerging and alive COVID-19 hotspots

Algorithm:
Significant_Cyl_Hotspots=[];

Step 1: Distribution of LLR under $H_0$
maxLLRs=[];
for each Monte Carlo simulation i= 1 to m
    cases’ = PointDistribution();
    (LLR’,H) = Cylinder_maxLLR(cases’, Loc, Pop, S, min_rad, max_rad);
    append(maxLLRs, LLR’);
end
maxLLRs=sort(maxLLRs, desc);

Step 2: Find cylinder with maximum LLR in S
(LLR’,CH)= Cylinder_maxLLR(cases, Loc, Pop, S, min_rad, max_rad);

Step 3: Test CH for statistical significance
pos= the position of LLR’ in the maxLLRs
p-value_CH = pos/m+1;
if (p-value_CH) < $\alpha_p$,
    append(significant_cyl_Hotspots, CH);
    C=C- {counties with centroid location inside CH};
goto step 2;
else
    print Significant_cyl_Hotspots;
end

function cylinder_maxLLR(cases[][], Loc[], Pop[], S,R, min_rad, max_rad)
Create and initialize swarm with K particles of dimension 4 within bounds
repeat
    for each particle i = 1 to K do
        fitness($X_i$) = LLR($X_i$, cases, loc, pop) using Equation 6
        if (fitness($X_i$) > fitness($pbest_i$)) then
            $pbest_i = X_i$
        end
        if (fitness($pbest_i$) > fitness($gbest$)) then
            $gbest = pbest_i$
        end
    end
    for each particle i = 1 to K do
        modify velocity using Equation 3
        modify position using Equation 4
    end
    until convergence criteria is met
return [fitness($gbest$), $gbest$]
4.4. Complexity analysis of ST-PSO-GRHD

The function, cylinder_maxLLR(), which uses PSO to detect cylinder with maximum LLR value, is used in two of the three major steps of the proposed ST-PSO-GRHD scheme. The function, cylinder_maxLLR() requires \( K \times 4 \) computational steps to initialize \( K \) particles each of 4 dimensions in the search space followed by calculation of their fitness; hence time complexity is \( O(K) \). Each particle’s fitness value calculation (LLR) requires \( O(C) \) computations. Based on the particle’s personal best and global best positions, each particle’s position and velocity are updated until convergence or for the maximum number of iterations ‘iter’. Thus, the time complexity of the function cylinder_maxLLR() is \( O((iter \times K \times C)) \). Step 1 of the algorithm runs ‘m’ Monte Carlo simulations and hence calls cylinder_maxLLR() function ‘m’ times, i.e. time complexity for Step 1 is \( O(m \times iter \times K \times C) \). In Step 2, cylinder_maxLLR () is called once; hence, the time complexity for Step 2 is \( O(iter \times K \times C) \). Finally, Step 3 calculates the p-value in constant time. However, Step 2 and Step 3 are repeated ‘n’ times, where \( n \) is the number of significant hotspots in the search space. Thus, the combined time complexity for Step 2 and Step 3 is \( O(n \times iter \times K \times C) \). Therefore, as \( m > n \), the time complexity of ST-PSO-GRHD is \( O(m \times iter \times K \times C) \).

5. Experimental studies and results

This section presents the experimental details and results obtained for ST-SaTScan and ST-PSO-GRHD schemes on the COVID-19 dataset of the United States of America (U.S.). First, we briefly describe the COVID-19 dataset we used for our experiments. Subsequently, we investigate and compare the performance of ST-SaTScan and ST-PSO-GRHD schemes in terms of quality of hotspot detected (effect of geographical robustness) and time taken for the detection. For analysis, county-wise data of the 4 worst-hit states (as on June 25th, 2020) in the U.S. is used. Finally, we present the alive, and emerging hotspots of the entire U.S. country detected using the proposed PSO-based scheme. The experiments are conducted on an Intel CORE i7 processor with 8 GB RAM using MATLAB R2018a.

5.1. Dataset

The COVID-19 case count data for the United States is obtained from the “COVID-19-data” repository published by the New York Times on their GitHub page. It is an ongoing repository maintained and published by The New York Times by compiling time series data from state governments and health departments. These data are freely available on their GitHub page (https://github.com/nytimes/covid-19-data) (The New York Times, 2020) at different levels of aggregation viz. National-level data, State-level data, and county-level data. To identify hotspots at the lowest level of detail, we have used county-level case count data for our study.

As the incubation period of coronavirus is 14 days, we have taken 14 days of data from June 11th, 2020 to June 24th, 2020. This data is available for 3026 counties of the U.S. The dataset contains a daily cumulative case count of different counties. Hence, to obtain a day-wise count for spatio-temporal analysis, the case count for each day in the study period is obtained by subtracting the previous day-case count from the current day case count. For some counties, the obtained day-wise count was negative for a few days. This indicates fewer cumulative cases on a particular day than the previous day, which is practically impossible. Therefore, we have cleaned the data by replacing these negative values with zero (assuming that the count of reported cases on that day is zero). Further, longitude and latitude pairs of the centroid of each U.S. county have been obtained from the Data.Healthcare.Gov web page, a federal government website managed by the Centers for Medicare and Medicaid Services in Baltimore, US (https://data.healthcare.gov/dataset/Geocodes-USA-with-Counties/52wv-g36k)(Geocodes USA with Counties, 2013). Also, the population of each U.S. County is required to generate the distribution of test statistics under the null hypothesis of no hotspots. The estimated 2019 population of each county is obtained from the United States Census Bureau website(http://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html) (County Population Totals: 2010–2019, n.d.).

5.2. Comparative analysis and discussions

This subsection presents the comparative analysis between ST-SaTScan and ST-PSO-GRHD schemes to detect COVID-19 hotspots. The comparison is made using the Log-likelihood ratio, a statistical indicator of the strength of the detected hotspots, and the time taken by both the schemes to detect COVID-19 hotspots. For comparison, we have used COVID-19 data of four worst-hit U.S. states (as on June 24th, 2020) viz. New York, California, Texas, and Florida. Execution time reported in both the schemes does not include the time for Monte Carlo simulations. For both the schemes, \( m \) (number of Monte Carlo Simulations) and \( \sigma_p \) (p-value threshold) are set to 99 and 0.01, respectively (\( m \) is computed experimentally as detailed in Appendix 1). In general, an upper bound (i.e., a maximum allowed value) for the cylinder base (spatial upper bound) is used to avoid detection of extremely large hotspots. This value is either chosen by experts or it is chosen proportionately to the overall study area under consideration. Therefore, the value of this upper bound (or maximum allowed radius) is not same for all the states and chosen as 300 km, 300 km, 200 km, and 400 km for California, New York, Florida and Texas respectively in our experiments. Similarly, to detect the emerging hotspots, an upper bound on height of the cylinder (temporal upper bound) is used. The temporal upper bound is set as 50% of the study period (June 11th, 2020 and June 24th, 2020) so that only emerging hotspots can be detected. The values of other PSO-specific parameters in ST-PSO-GRHD are presented in Table 3. These are the typical values of PSO parameters used in the literature (Lim et al., 2009) (Sengupta, Basak, & Peters, 2018).

Table 3 presents the characteristics of COVID-19 hotspots detected in four worst-hit U.S. states by ST-SaTScan and ST-PSO-GRHD schemes. These states include California State (58 counties), New York State (58 counties), Florida State (67 counties), and Texas State (252 counties) with a maximum radius of the circular base of the cylindrical hotspot (spatial upper bound) as 300 km (km), 300 km, 200 km, and 400 km respectively. As presented in Table 4, both the schemes detected 4

| Parameter                          | Value |
|------------------------------------|-------|
| Population Size                    | 50    |
| \( \alpha \) (cognitive component) | 2.05  |
| \( \beta \) (social component)     | 2.05  |
| \( \Omega \) (inertia weight)      | Linearly decreased from 0.9 to 0.2 |
| Maximum Number of iterations       | 500   |
| Number of iterations for convergence | 100   |

Further, the characteristics of the statistically significant emerging COVID-19 hotspots which are reported for both the schemes include the duration of the hotspot, count of counties in the hotspot, the observed and expected number of cases in the hotspot, LLR value of the hotspot, and time taken to detect them.
Table 4
COVID 19 emerging hotspots detected by ST-SaTScan and ST-PSO-GRHD schemes for four worst-hit U.S. states. The first column lists the name of the four U.S. states along with the count of counties in that state, and spatial upper bound in kilometers (km). The remaining fields represent the Characteristics (count of counties in the hotspot, observed and expected number of cases in the hotspot and respective LLR, followed by time taken in seconds to detect the hotspots) of the detected emerging hotspots.

| U.S. State, Count of counties in the state, Spatial upper bound while detecting the hotspots (in km) | Hotspot # | ST-SaTScan | ST-PSO-GRHD |
|---|---|---|---|
| | | Duration | Count of Counties in the Hotspot | Observed | Expected | LLR | Required Time (in seconds) | Duration | Count of Counties in the Hotspot | Observed | Expected | LLR | Required Time (in seconds) |
| California State, 58 counties, 300 km | 1 | 18th to 24th June | 1 | 10 | 17331 | 10545.4 | 2361.3 | 165.59 | 21st to 24th June | 5 | 12178 | 6003.5 | 2838.9 | 36.76 |
| | 2 | 22nd to 24th June | 2 | 2 | 2716 | 1756.68 | 232.71 | 22nd to 24th June | 12 | 2143 | 1076.9 | 419.04 | 22nd to 24th June | 2 | 449 | 227.77 | 83.95 |
| | 3 | 20th to 24th June | 2 | 3 | 1171 | 661.68 | 161.48 | 20th to 24th June | 2 | 509 | 390.87 | 16.41 |
| | 4 | 20th to 24th June | 3 | 3 | 727 | 449.04 | 73.02 | 20th to 24th June | 2 | 449 | 227.77 | 83.95 |
| New York State, 58 counties, 300 km | 1 | 18th to 24th June | 4 | 4 | 2694 | 1123.84 | 942.77 | 170.38 | 18th to 24th June | 1 | 2403 | 606.03 | 1708.78 | 41.58 |
| | 2 | 22nd to 24th June | 2 | - | - | - | - | - | 18th to 24th June | 1 | 138 | 85.09 | 13.97 |
| Florida State, 67 counties, 200 km | 1 | 18th to 24th June | 25 | 25 | 21774 | 14894.2 | 2359.98 | 429.06 | 18th to 24th June | 25 | 21774 | 14894.2 | 2359.98 | 65.34 |
| | 2 | 20th to 24th June | 2 | 2 | 170 | 40.7704 | 113.70 | 20th to 24th June | 3 | 215 | 53.58 | 137.62 |
| | 3 | 19th to 24th June | 2 | 1141 | 869.97 | 39.31 | 19th to 24th June | 1 | 1225 | 929 | 43.89 |
| | 4 | 23rd to 24th June | 2 | 84 | 12.49 | 88.61 | 23rd to 24th June | 4 | 127 | 58.54 | 29.96 |
| Texas State, 252 counties, 400 km | 1 | 19th to 24th June | 47 | 47 | 17489 | 9613.5 | 3418.44 | 67419.1 | 19th to 24th June | 47 | 17350 | 9442.19 | 3479.7 | 133.73 |
| | 2 | 22nd to 24th June | 21 | 21 | 809 | 328.30 | 251.26 | 22nd to 24th June | 21 | 832 | 321.64 | 283.03 |
| | 3 | 23rd to 24th June | 5 | 5 | 1908 | 1266.7 | 144.56 | 23rd to 24th June | 4 | 3298 | 2581.25 | 96.87 |
| | 4 | 18th to 24th June | 2 | 368 | 211.44 | 47.61 | 18th to 24th June | 2 | 371 | 212.62 | 48.41 |
| | 5 | 23rd to 24th June | 4 | 4 | 104 | 46.43 | 26.33 | 18th to 24th June | 1 | 43 | 7.91 | 37.76 |
| | 6 | - | - | - | - | - | - | - | 18th to 24th June | 1 | 37 | 6.26 | 35.03 |
emerging hotspots, each in California and Florida; however, counties’
count in detected hotspots is different. Further, ST-SaTScan detected 1
and 5 emerging hotspots in New York and Texas, respectively, whereas
ST-PSO-GRHD detected 2 and 6 emerging hotspots in New York and
Texas states, respectively.

The spatial distribution of emerging Spatio-temporal COVID-19
hotspots detected by both schemes in four worst-hit U.S. states,
California, New York, Florida, and Texas, have been presented in Fig. 1,
Fig. 2, Fig. 3, and Fig. 4, respectively. In these figures, the most sig-
ficant hotspot (hotspot with highest LLR) in a state is labeled as Hotspot
1. The second most significant hotspot in the state is Hotspot 2 and so on.
While comparing the performance (in terms of LLR of the detected
hotspots) of both the schemes, it is observed that the LLR values of most
likely hotspots detected by the proposed scheme for all the four states of

![Map of California with detected hotspots](image1)

![Map of New York with detected hotspots](image2)
the U.S. are higher than LLR of most likely hotspots detected by ST-SaTScan. For example, the LLR of hotspot #1 (i.e., most significant hotspot) detected by ST-SaTScan in California State is 2361.3. In contrast, ST-PSO-GRHD detected the most significant hotspot (hotspot #1) with improved LLR as 2838.9. Compared to ST-SaTScan, similar improvements in LLR have been observed in the detection of subsequent hotspots (hotspot #2, i.e., second most significant hotspot, etc.) in California State by ST-PSO-GRHD. Obviously, the detection of the hotspots with improved LLR is the impact of geographical robustness addressed in ST-PSO-GRHD.

Besides LLR, the performance of both schemes (ST-SaTScan, ST-PSO-GRHD) has also been compared using the time required to detect the hotspot(s). Although geographical robustness is not considered in ST-SaTScan (due to enumeration of an infinite number of circular bases), it has been observed that the time taken by ST-SaTScan is significantly higher than ST-PSO-GRHD. For example, the required time by ST-SaTScan to detect all hotspots in California State is 165.59 s, whereas it is 36.76 for ST-PSO-GRHD.

In another observation, there are around 4 times more counties in Texas than in California, but the required time by ST-SaTScan to detect all hotspots in Texas significantly increased by 1834 times of the required time for California. Compared to ST-SaTScan, more scalable performance has been shown by ST-PSO-GRHD where the time required to detect all hotspots in Texas is increased by 4 times the required time for California. In general, as the number of counties in the state under study increases, time taken by ST-SaTScan increases significantly. In contrast, the time taken by ST-PSO-GRHD increases almost linearly with an increase in the number of counties. The apparent reason behind the observed performance is the dependency of these schemes on the count of counties (C). Recall that the required time in ST-SaTScan is proportional to $C^3$ (Section 2) and may not be time-efficient in the detection of hotspots with bigger values of C. In contrast, the time required by ST-

---

Fig. 3. Spatial distribution of emerging Spatio-temporal COVID-19 hotspots detected by ST-SaTScan (left) and ST-PSO-GRHD (right) for Florida State. Hotspot with the highest LLR is labeled as Hotspot 1, the second-highest LLR is labeled as Hotspot 2, and so on.

Fig. 4. Spatial distribution of emerging Spatio-temporal COVID-19 hotspots detected by ST-SaTScan (left) and ST-PSO-GRHD (right) for Texas State. Hotspot with the highest LLR is labeled as Hotspot 1, the second-highest LLR is labeled as Hotspot 2, and so on.
PSO-GRHD is proportional to C (Section 3.4) and may be effective in the detection of hotspots with bigger values of C.

5.3. Analysis over two different timespans

In this section, we present and analyze the results of ST-PSO-GRHD over two incubation periods viz. May 28th, 2020 to June 10th, 2020, and June 11th, 2020 to June 24th, 2020. ST-PSO-GRHD is applied to the COVID-19 data of California, New York, Florida and Texas over two timespans. The spatial distribution of emerging Spatio-temporal COVID-19 hotspots detected over both the timespans in California, New York, Florida, and Texas have been presented in Fig. 5, Fig. 6, Fig. 7, and Fig. 8, respectively.

As observed from Fig. 5, three hotspots covering 8 counties were identified in California from May 28th, 2020 to June 10th, 2020. However, in the next 14 days, these hotspots expanded and reported 21 counties as the hotspots. These 21 counties include 7 out of 8 counties reported as emerging hotspots in the last 14 days. Similar observations were made for New York state. As evident from Fig. 6, New York reported only one hotspot consisting of New York city counties in the timespan May 28th, 2020 to June 10th, 2020. However, in the next 14 days, along with New York City, Oneida county also emerged as a hotspot.

Further, it is evident from Fig. 7 that the 18 counties which were part of hotspots identified in Florida from May 28th, 2020 to June 10th, 2020 are also part of 33 hotspot counties identified in the next 14 days i.e.,...
June 11th, 2020 to June 24th, 2020. Finally, the observations over two consecutive timespans for Texas state indicate that only 23 counties were emerging hotspots as on June 10th, 2020. However, in the next 14 days timespan, the count increased to 76 counties. It is evident from Fig. 8 that the majority of counties that were emerging hotspots in the first time span are also a part of emerging hotspots in the second time span.

Motivated with the performance (quality of the detected hotspots and time requirement) of the proposed ST-PSO-GRHD scheme over ST-SaTScan, we applied the proposed ST-PSO-GRHD scheme to detect the emerging hotspots in all the U.S. states (totaling 3026 counties). In the subsequent sections, we present the observed results.

5.4. Case study: emerging hotspots in the United States

This section presents the alive and emerging COVID-19 hotspots detected for all the states in the U.S. Considering the need of timely initiation of the preventive measures for COVID-19 outbreak and inefficiency of ST-SaTScan in handling large datasets (recall from Section 4.1, there are 3026 counties in the U.S. for which day-wise confirmed COVID-19 cases are available), ST-PSO-GRHD is applied to detect the hotspots for all the states in the U.S. For uniformity, the spatial upper bound (maximum radius of the circular base of the cylindrical hotspot) is kept as 100 km (km), and the temporal upper bound (height of the cylinder) is kept as 50% of the study period (between June 11th, 2020 and June 24th, 2020).
During the study period of 14 days, ST-PSO-GRHD detected 104 emerging COVID-19 hotspots. Out of 104, the characteristics of the 30 most significant emerging COVID-19 hotspots detected in the U.S. are presented in Table 5. These characteristics include the duration of the hotspot, its LLR value, the observed (or actual) number of cases, expected number of cases (calculated using (1)) and the number of counties forming the hotspot. The characteristics of the remaining 74 hotspots are listed in Appendix 2.

Out of 104, the most significant hotspot referred to as Hotspot #1 is found in Arizona State with an LLR value of 11840.71 and includes 2 counties viz. Yuma and Maricopa. A total of 14215 cases are observed in these two counties for the last 7 days of the study period, which is significantly greater than the expected count of 2802.95 cases. The second most significant hotspot, referred to as Hotspot 2 with 10087 observed cases, is detected in Florida and consists of 7 counties viz. Broward, Collier, Glades, Hendry, Martin, Miami Dade, and Palm Beach. The expected number of cases in these 7 counties is 4037.07, which is much less than the observed cases, and hence it is a significant hotspot with an LLR value of 3235.54. Further, the next three significant hotspots, i.e., Hotspot #3 (includes 4 counties with LLR as 3213.82), Hotspot #4 (includes 9 counties with LLR as 3213.79), and Hotspot #5 (includes 1 county with LLR as 2771.25), are detected in Texas, Florida, and California respectively. The observed duration for the first four most significant hotspots is between June 18th, 2020 and June 24th, 2020 whereas the duration for the fifth-most significant hotspot is between June 19th, 2020 and June 24th, 2020.

All the 104 alive or emerging hotspots in the U.S. detected by ST-PSO-GRHD are shown in Fig. 9. It can be observed that most of the hotspots are detected in California, Texas, Florida, New York, New Jersey, Arizona, Alabama, Mississippi, Utah, and North Carolina states of the U.S. Other states containing a few hotspots include Nevada, Colorado, Kansas, Idaho, South Dakota, Iowa, Minnesota, Tennessee, and Virginia. Out of 104, the maximum number of significant emerging hotspots is detected in Texas State whereas Michigan, Colorado, Ohio, and the Wisconsin States have the least number of hotspots.

6. Conclusion and future scope

This paper proposed a PSO-based geographically robust and efficient scheme for detecting Spatio-temporal alive and emerging hotspots. The proposed scheme, ST-PSO-GRHD, does not rely on considering a county centroid location as the center of the hotspot and thus detects more significant hotspots (evident with improved LLR) than ST-SaTScan. Also, the ability of ST-PSO-GRHD to promptly detect significant hotspots makes it a valuable surveillance scheme that can be used to detect emerging hotspots (or outbreaks) on a daily basis and as soon as they unfold. This study utilized the COVID-19 day-wise count of confirmed cases in the United States for one incubation period of 14 days between June 11th, 2020, and June 24th, 2020 to detect emerging hotspots. Performance of the ST-SaTScan and ST-PSO-GRHD schemes to detect emerging COVID-19 hotspots have been analyzed on 4 worst-hit states of the U.S. viz. New York, California, Texas, and Florida. It has been observed that ST-PSO-GRHD detects more significant hotspots (higher LLR) than ST-SaTScan quickly. Hence, the ST-PSO-GRHD method was applied to detect emerging hotspots in all states of the U.S. The counties belonging to 104 significant emerging hotspots detected by ST-PSO-GRHD should be given attention and priority in testing and allocating resources to slow down the virus transmission. Health officials and government agencies can use the rapid surveillance and statistical analysis provided by the proposed scheme to identify high-risk areas and take timely measures to contain the COVID-19 outbreak.

Besides the strengths and applicability of this study, aggregation of data at the lowest level can help officials further drill down to affected zones that may further improve the precision in handling the COVID-19 outbreak. Also, the addition of some relevant information in the context of COVID-19, viz. count of older people, count of people with preexisting health conditions, etc., in future studies may further strengthen the conclusions.
Funding

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

CRediT authorship contribution statement

Ankita Wadhwa: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. Manish Kumar Thakur: Conceptualization, Methodology, Validation, Formal analysis, Resources, Writing – review & editing, Supervision.

Declaration of competing interest

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.apgeog.2022.102719.

Appendix 1

Choosing the number of Monte Carlo Simulations.

Monte Carlo simulations are cost expensive. As each Monte Carlo simulation in ST-SaTScan takes O(|A|³) time, therefore the count of Monte Carlo simulations cannot be randomly selected. To decide an approximate number of Monte Carlo simulations, a simple procedure is to run a fixed number of simulations and then evaluate the mean and standard deviation of the test statistic (Lerche et al. https://doi.org/10.1260/014459805776986876). Then increase the number of simulations, calculate mean and standard deviation again, and compare the results with the previous set of simulations. Repeat until we find that the change in observed values with an increase in the number of simulations is very less. This experimentation was done with different simulations (9, 19, 29, 49, 99, 199, 299, 399, 499, and 999), and the corresponding results were observed. It was seen that for ‘m < 99’ the average of LLR was changing from one set of simulations to other. For ‘m > 99’, the average and standard deviation were almost the same. In comparison, the computation time increased with the increasing number of simulations. The results of a sample dataset are shown in the table below.

| m   | average   | Stdev  | Time(in sec) |
|-----|-----------|--------|--------------|
| 9   | 7.107938  | 2.879943 | 12.304297    |
| 19  | 5.803806  | 2.249776 | 25.678133    |
| 49  | 6.620155  | 2.869332 | 65.58123     |
| 99  | 6.209283  | 2.651164 | 125.32       |
| 199 | 6.488089  | 2.516975 | 246.85       |
| 299 | 6.316002  | 2.425302 | 384.927146   |
| 399 | 6.347084  | 2.561424 | 521.640005   |

(continued on next page)
Thus, considering the performance of Monte Carlo simulations and their computation time, m is chosen as 99.

Appendix 2

During the study period of 14 days, ST-PSO-GRHD detected 104 emerging COVID-19 hotspots in the U.S. Out of 104, the characteristics of 30 most significant alive or emerging COVID-19 hotspots have been detailed in Table 5. The characteristics of the remaining 74 hotspots (numbered 31 to 104) are listed below. These characteristics include the duration of Spatio-temporal hotspot, LLR value of the hotspot, the observed (or actual) number of cases in that hotspot, expected number of cases in that hotspot (calculated using (1)), and the number of counties forming the hotspot.

| Hotspot number | Duration       | LLR     | Observed | Expected | Number of Counties |
|----------------|----------------|---------|----------|----------|-------------------|
| 31             | 18th to 24th June | 289.9761 | 524      | 146.7059 | 2                 |
| 32             | 18th to 24th June | 281.2931 | 1225     | 571.2766 | 1                 |
| 33             | 18th to 24th June | 269.7471 | 1396     | 698.5231 | 3                 |
| 34             | 18th to 24th June | 268.1499 | 1085     | 489.8453 | 4                 |
| 35             | 18th to 24th June | 264.895  | 748      | 281.4323 | 1                 |
| 36             | 18th to 24th June | 263.6708 | 444      | 117.5777 | 8                 |
| 37             | 18th to 24th June | 240.8672 | 369      | 90.28718 | 5                 |
| 38             | 18th to 24th June | 237.2377 | 608      | 216.0992 | 1                 |
| 39             | 19th to 24th June | 234.3723 | 2111     | 783.0567 | 2                 |
| 40             | 18th to 24th June | 212.9769 | 1427     | 571.2766 | 13                |
| 41             | 18th to 24th June | 212.4057 | 1202     | 627.5904 | 20                |
| 42             | 19th to 24th June | 201.4234 | 502      | 175.3917 | 2                 |
| 43             | 19th to 24th June | 194.532  | 1088     | 560.4645 | 10                |
| 44             | 18th to 24th June | 192.3892 | 2668     | 1780.853 | 2                 |
| 45             | 19th to 24th June | 153.4877 | 419      | 154.5761 | 8                 |
| 46             | 19th to 24th June | 149.2828 | 104      | 10.02947 | 1                 |
| 47             | 18th to 24th June | 145.1174 | 295      | 90.06648 | 3                 |
| 48             | 18th to 24th June | 121.1934 | 941      | 540.8403 | 2                 |
| 49             | 18th to 24th June | 119.026  | 451      | 197.4701 | 3                 |
| 50             | 18th to 24th June | 115.6853 | 119      | 19.51784 | 1                 |
| 51             | 18th to 24th June | 115.1042 | 634      | 324.6583 | 12                |
| 52             | 19th to 24th June | 113.2196 | 2348     | 1693.732 | 7                 |
| 53             | 18th to 24th June | 112.6252 | 789      | 439.1247 | 7                 |
| 54             | 18th to 24th June | 107.8076 | 609      | 314.771 | 10                |
| 55             | 19th to 24th June | 105.7793 | 286      | 104.9098 | 5                 |
| 56             | 21st to 24th June | 103.5696 | 284      | 103.8541 | 9                 |
| 57             | 18th to 24th June | 102.9754 | 249      | 85.81601 | 1                 |
| 58             | 18th to 24th June | 91.39835 | 815      | 487.6019 | 1                 |
| 59             | 19th to 24th June | 88.51591 | 518      | 271.1683 | 3                 |
| 60             | 18th to 24th June | 86.3387  | 134      | 33.14848 | 5                 |
| 61             | 18th to 24th June | 85.3836  | 1647     | 1172.723 | 11                |
| 62             | 18th to 24th June | 84.1971  | 91       | 15.7869  | 1                 |
| 63             | 18th to 24th June | 80.44444 | 757      | 459.6297 | 1                 |
| 64             | 18th to 24th June | 77.53729 | 77       | 12.11201 | 1                 |
| 65             | 19th to 24th June | 77.53255 | 661      | 390.4847 | 5                 |
| 66             | 19th to 24th June | 69.65858 | 209      | 118.8152 | 2                 |
| 67             | 18th to 24th June | 66.60587 | 309      | 147.9007 | 1                 |
| 68             | 18th to 24th June | 66.01095 | 190      | 72.22951 | 5                 |
| 69             | 18th to 24th June | 62.83659 | 931      | 629.7371 | 12                |
| 70             | 18th to 24th June | 60.01844 | 95       | 23.89597 | 1                 |
| 71             | 18th to 24th June | 56.30142 | 124      | 40.00079 | 5                 |
| 72             | 18th to 24th June | 56.22117 | 201      | 85.5798  | 1                 |
| 73             | 18th to 24th June | 51.00192 | 43       | 5.489356 | 1                 |
| 74             | 18th to 24th June | 47.76887 | 102      | 32.22037 | 1                 |
| 75             | 18th to 24th June | 46.58505 | 37       | 4.346511 | 1                 |
| 76             | 19th to 24th June | 46.19377 | 596      | 391.1809 | 6                 |
| 77             | 21st to 24th June | 45.39449 | 655      | 440.5409 | 2                 |
| 78             | 18th to 24th June | 45.1104  | 164      | 70.39356 | 1                 |
| 79             | 18th to 24th June | 44.54471 | 212      | 102.534  | 3                 |
| 80             | 21st to 24th June | 42.63214 | 448      | 279.944  | 6                 |
| 81             | 18th to 24th June | 41.64245 | 104      | 36.36945 | 3                 |
| 82             | 20th to 24th June | 41.63051 | 189      | 89.64546 | 2                 |
| 83             | 21st to 24th June | 39.04052 | 420      | 263.9994 | 3                 |
| 84             | 18th to 24th June | 37.49152 | 89       | 30.1506  | 4                 |
| 85             | 18th to 24th June | 37.43705 | 184      | 90.15118 | 1                 |
| 86             | 18th to 24th June | 36.93593 | 47       | 9.682573 | 3                 |

(continued on next page)
References

Agrawal, R., Gehrke, J., Gunopulos, D., & Raghavan, P. (1998). Automatic subspace clustering of high dimensional data for data mining applications. *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, 94–105. https://doi.org/10.1145/276394.276418

Ankita, & Thakur, M. K. (2018). Modified DBCSCAN using particle swarm optimization for spatial hotspot identification. *Eleventh International Conference on Computing, Communication, Networking & Information Technology (ICCCN)*, 1–3. https://doi.org/10.1109/ICCCN.2018.8530538

Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data & Knowledge Engineering*, 60(1), 208–221. https://doi.org/10.1016/j.datak.2006.01.013

Cançado, A. L. F. F., Duarte, A. R., Duczmal, L. H., Ferreira, S. J., Fonseca, C. M., & Goiotti, E. C. D. M. (2010). Penalized likelihood and multi-objective spatial scans for the detection and inference of irregular clusters. *International Journal of Health Geographies*, 9(1), 55. https://doi.org/10.1186/1476-072X-9-55

Clerc, M., & Kennedy, J. (2002). The particle swarm - explosion, stability, and convergence in a multidimensional complex space. *IEEE Transactions on Evolutionary Computation*, 6(1), 58–73.

2019. (n.d.) County Population Totals. (2010). *United States Census Bureau*. Retrieved June 25, 2020, from https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html.

Cucinotta, D., & Vanelli, M. (2020). WHO declares COVID-19 a pandemic. *Acta BioMedica*, 91(1), 157–160. https://doi.org/10.23750/abm.v91i1.9397

Dejardin, M., R., Hohl, A., & Delmelle, E. M. (2020). Rapid surveillance of COVID-19 in the United States using a prospective space-time scan statistic: Detecting and evaluating emerging clusters. *March Applied Geography*, 118, Article 102202. https://doi.org/10.1016/j.mapplgeo.2020.102202.

Di Martino, F., & Sessa, S. (2013). A fuzzy particle swarm optimization algorithm and its application to hotspot events in spatial analysis. *Journal of Ambient Intelligence and Humanised Computing*, 4(1), 85–97. https://doi.org/10.1007/s12652-011-0096-5

Duczmal, L., Cançado, A. L. F. F., Takahashi, R. H. C., & Bessegato, L. F. (2007). ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data & Knowledge Engineering*, 60(1), 208–221. https://doi.org/10.1016/j.datak.2006.01.013

Eftelioglu, E., Shethkar, S., Kang, J. M., & Farah, C. C. (2016). Ring-shaped hotspot detection. *IEEE Transactions on Knowledge and Data Engineering*, 28(12), 3367–3381. https://doi.org/10.1109/TKDE.2016.2607200

Eftelioglu, E., Tang, X., & Shethkar, S. (2015). Geographically robust hotspot detection: A summary of results. *Proceedings - 15th IEEE International Conference on Data Mining Workshops*, ICDMW, 1447–1456. https://doi.org/10.1109/ICDMW.2015.159

Eftelioglu, E., Tang, X., & Shethkar, S. (2016). Geographically robust hotspot detection: A summary of results. *Proceedings - 15th IEEE International Conference on Data Mining Workshops*, ICDMW, 1447–1456. https://doi.org/10.1109/ICDMW.2015.159

Esteban, M., Kriel, H., Xu, X., & Münch, D., (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Geocodes USA with Counties*, 2010: 29–36.

Geocodes USA with Counties. (2013). *DataHealthCare.Gov*. https://data.healthcare.gov/dataset/Geocodes-USA-with-Counties/52wy-g3t86.

Izakian, H., & Pedrycz, W. (2012). A new PSO-optimized geometry of spatial and spatio-temporal scan statistics for disease outbreak detection. *Swarm and Evolutionary Computation*, 4, 1–11. https://doi.org/10.1016/j.swevo.2012.02.001

Kulldorff, M. (1997a). Spatial scan Statistic.Pdf (pp. 1481–1496).

Kulldorff, M. (1997b). A spatial scan statistic. *Communications in Statistics - Theory and Methods*, 26(6), 1481–1496. https://doi.org/10.1080/03610929708831995

Kulldorff, M. (1999). Spatial scan statistics: Models, calculations, and applications. *In* recent advances in scan statistics and applications (pp. 303–322). https://doi.org/10.1007/978-1-598-12733-4

Kulldorff, M. (2001). Prospective time period geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society - Series A: Statistics in Society*, 164(1), 61–72. https://doi.org/10.1002/sim.2089

Kulldorff, M. (2018). SatScan user guide 2006 (Vol. 8). *www.satscan.org*.

Kulldorff, M., Athas, W. A., & Miller, B. A. (1998). Evaluating cluster Alarms: A space-time scan statistic and brain cancer in Los Alamos, New Mexico. *American Journal of Public Health*, 88(9), 1377–1386.

Kulldorff, M., Huang, L., Pickle, L., & Duczmal, L. (2006). An elliptic spatial scan statistic. *Statistics in Medicine*, 25(22), 3929–3943. https://doi.org/10.1010/sim.2490

Kumar, M. V. (n.d.). Crime hotpot detection using spatial clustering. A literature review of related work.

Lawson, A. B. (2010). Hotspot detection and clustering: Ways and means. *Environmental and Ecological Statistics*, 17(2), 231–245. https://doi.org/10.1007/s10651-010-0142-8

Li, Z., Fu, J., Lin, G., & Jiang, D. (2019). Spatiotemporal variation and hotspot detection of the avian influenza A( H7N9) virus in China, 2013–2017. *International Journal of Environmental Research and Public Health*, 16(4), 1–13. https://doi.org/10.3390/ijerph16040648

Lim, S. Y., Montakhah, M., & Nouri, H. (2009). A construction factor based particle swarm optimization for economic dispatch. *ESTM 2009 - 2009 European Simulation and Modelling Conference. Modelling and Simulation, 305–311, 2009.

Martino, F. D., Loia, V., & Sessa, S. (2008). Extended fuzzy C-means clustering algorithm for hotspot events in spatial analysis. *International Journal of Hybrid Intelligent Systems*, 5(1), 31–44. http://iospress.metapress.com/content/g07h64x34328g2s82/

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*, 46(5), 846–848. https://doi.org/10.1007/s00134-020-05991-x

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. http://arslive.org/2020/03/2669.

Scitovszky, S. (2018). A density-based clustering algorithm for earthquake zoning. *Computers & Geosciences*, 110, 90–95. https://doi.org/10.1016/j.cageo.2017.08.014

Sengupta, S., Basak, S., & Peters, R. (2016). Particle swarm optimization: A survey of historical and recent developments with hybridization perspectives. *Machine Learning and Knowledge Extraction*, 1(1), 157–191. https://doi.org/10.3390/make10100010

Shi, Z., & Pan-Cheng, L. S. C. (2019). Spatiotemporal data clustering: A survey of methods. *ISPRS International Journal of Geo-Information*, 8(3).https://doi.org/10.3390/ijgi80300112

Tang, X., Eftelioglu, E., Oliver, D., & Shethkar, S. (2017). Significant linear hotspot discovery. *IEEE Transactions on Big Data*, 3(2), 140–153. https://doi.org/10.1109/TBDATA.2016.2631518

Taylor, D. K., Jain, A., Arora, S., Agarwal, S., Gupta, T., Tyagi, N., et al. (2014). Crime detection and criminal identification in India using data mining techniques. *AI & Society*, 30(1), 117–127. https://doi.org/10.1007/s10458-014-9539-6

The New York Times. (2020). COVID-19 data. https://github.com/nytimes/covid-19-data

Vediana, M., & Widyaningsih, Y. (2017). Hotspot detection using space-time scan statistics on children under five years of age in Depok. *AIP Conference Proceedings*, https://doi.org/10.1063/1.4979454, 1827.
Wang, C., Horby, P. W., Hayden, F. G., & Gao, G. F. (2020). A novel coronavirus outbreak of global health concern. *The Lancet*, 395(10223), 470–473. https://doi.org/10.1016/S0140-6736(20)30185-9

Wang, W., Yang, J., & Muntz, R. (1997). In *SING: A statistical information grid approach to spatial data mining* 1 introduction 2 related work (pp. 1–18).

Xie, Y. (2019). Significant DBSCAN towards statistically robust clustering. *International Symposium on Spatial and temporal databases*. ACM.