AN IRREGULAR SPATIAL CLUSTER DETECTION COMBINING THE GENETIC ALGORITHM

Tao Wang[1], Yitong Zhao[2], Yonglin Lei[3] and Mei Yang[4], Shan Mei[5]

[1,3,4,5] College of System Engineering, National University of Defence Technology, Changsha, China
[2]66136 Troop of PLA, Beijing, China

ABSTRACT

Spatial cluster detection is widely used for disease surveillance, prevention and containment. However, the commonly used clustering methods cannot resolve the conflicts between the accuracy and efficiency of the detection. This paper proposes an improved method for flexibly-shaped spatial scanning, which can identify irregular spatial clusters more accurately and efficiently. By using a genetic algorithm, we also accelerate the detection process. We convert geographic information to a network structure, in which nodes represent the regions and edges represent the adjacency relationship between regions. According to Kulldorff's spatial scan statistics, we set the objective function. A constraint condition based on the spectral graph theory is employed to avoid disconnectedness or excessive irregularity of clusters. The algorithm is tested by analysing the simulation data of H1N1 influenza in Beijing. The results show that compared with the previous spatial scan statistic algorithms, our algorithm performs better with shorter time and higher accuracy.

KEYWORDS

Spatial cluster detection; flexibly-shaped spatial scanning; H1N1 influenza in Beijing

1. INTRODUCTION

In the area of public health security, the space disease surveillance study was conducted to identify and predict the spatial distribution of disease transmission at the beginning of the disease outbreak, which can help disease containment.

Commonly, spatial cluster detection algorithms search the epidemic regions by setting the scan window to a specific shape, such as circle[1, 2] (Circle Scan), rectangle[3], and oval[4]. These algorithms are fast, but sometimes the outbreak regions can be irregularly shaped, many regions without decease outbreak are often included in detection, which can result in loss of detection performance[5].

In order to solve the problem of the above algorithms, Tango et al. proposed a spatial cluster algorithm that can detect arbitrary shape outbreak regions, called " flexibly shaped spatial scan statistic"[5] (Flex Scan). Compared with the previous algorithms, the Flex Scan algorithm detects

Natarajan Meghanathan (Eds) : CCSIT, NCWMC, DaKM - 2019
pp. 09-20, 2019. © CS & IT-CSCP 2019 DOI: 10.5121/csit.2019.90802

the epidemic regions more accurate. However, since the Flex Scan algorithm needs to traverse the entire region and its nearest K-1 regions, As the number of epidemic regions increases, the computational efficiency of the algorithm will decrease[5].

Some scholars hope to use the heuristic algorithm to accelerate the process of detection. Duczmal et al. used simulated annealing algorithms to check criminal clusters in large cities[6]. Neill et al. proposed a fast space scanning algorithm (LTSS) by building scoring functions to accelerate the calculation process[7]. Although the heuristic algorithms have fast computing speed, most of the heuristic algorithms can only detect regions which are much larger than the actual outbreak regions or result in multiple non-connected outbreak regions. Plus, the accuracy is still lower than common algorithms.

The main contribution of this paper is to balance the accuracy of detection and the efficiency of computation. We explore the improved flexibly shaped spatial scan statistic (I Flex Scan) algorithm by combining the Circle Scan algorithm and the Flex Scan algorithm. Compared with the FLEXSCAN algorithm, which determines the epidemic area by traversal, we use the genetic algorithm to accelerate the process. By using this algorithm, we determine the constraints according to the spectral graph theory, so that the range of the epidemic area is not too scattered.

2. PROBLEM DESCRIPTION AND MATHEMATICAL MODEL

2.1. TRANSFORM THE MAP INTO A NETWORK STRUCTURE

To illustrate this scenario, Figure 1depictsthe disease outbreak problem, which shows the cholera outbreak along a winding river floodplain. In the left part, each region represents a county. We convert the left part to a connected graph $G = (V, E)$ where nodes $v \in V$ represent the regions with features values $x_v$representing the patient number, which is shown in the right part. Our goal is to find the most possible outbreak regions over all connected sub-graphs.

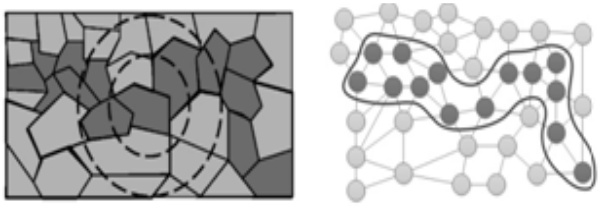

Figure 1. Use spatial cluster detection to find the outbreak region

In a map divided by blocks, the boundaries of each block can be seen as a connection of some coordinate points, usually represented by the latitude and longitude. So we think that all geographical blocks are irregular polygons made up of the connection points of coordinates. We can use Arc GIS software to calculate the geometric centre of each geographic polygon and get the adjacency list of each block. The latitude and longitude of the geometric centre is used as the coordinates of the nodes in the network, and the adjacency list is used to express the adjacency relationship between the nodes, then the map is transformed into a network structure.

Figure 2 illustrates the process of transforming a map into a network structure, where Figure 2(a)gives the example of a map divided by blocks, Figure 2(b) shows the result of calculating the

geometric center of each geographic polygon, Figure 2(c) shows the result of calculating the adjacency relationship, and Figure 2(d) is the result of transforming the map into network structure.

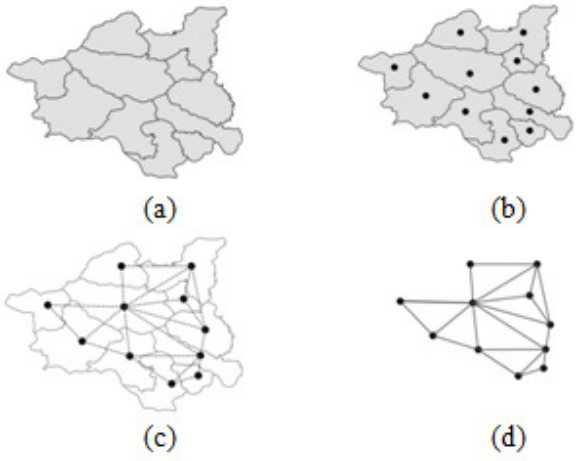

Figure 2. The example of transform map into network structure

2.2. OBJECTIVE FUNCTION

Given a network into M nodes, with total population N and C total cases, let the component Z be any set of cluster nodes. Under the null hypothesis (there are no clusters in the network), the number of cases in each node follows a Poisson distribution. Define $L(Z)$ as the likelihood under the alternative hypothesis that there is a cluster in the component Z, and$L_0$the likelihood under the null-hypothesis. The component Z with the maximum likelihood is defined as the most likely cluster. If $\mu(Z)$ is the expected number of cases inside the component Z under the null hypothesis, $c(Z)$ is the number ofcases inside Z. Under the Poisson distribution, the test statistic, which was constructed with the likelihood ratio test[1,2], is given by:

$$LR(Z)=\left[\frac{c(Z)}{\mu(Z)}\right]^{c(Z)}\left[\frac{C-c(Z)}{C-\mu(Z)}\right]^{C-c(Z)} I(Z)$$

Where$I(Z)$ represents an indicator function, when$\frac{c(Z)}{\mu(Z)} > \frac{C-c(Z)}{C-\mu(Z)}$,$I(Z)=1$,and 0 otherwise. According to the experiments of Tango et al. [5], for irregular clustering, there are some misjudged nodes in the components identified by the Circle Scan algorithm.

In order to make up for the shortcomings of previous algorithms for irregular clustering, we combine Circle Scan and Flex Scan to propose an improved flexible spatial scan statistics (I Flex Scan) algorithm. I Flex Scan is based on Circle Scan to locate the scope of the epidemic. On this basis, the spatial cluster is determined by removing the normal nodes, which is misjudged by Circle Scan.

Figure 3 shows the process of I Flex Scan algorithm, in which the blue area is the approximate epidemic range determined by Circle Scan algorithm. At the same time, the grey node is the

normal node, the deep grey node is the epidemic node, and the orange node is the misjudgment node that needs to be removed.

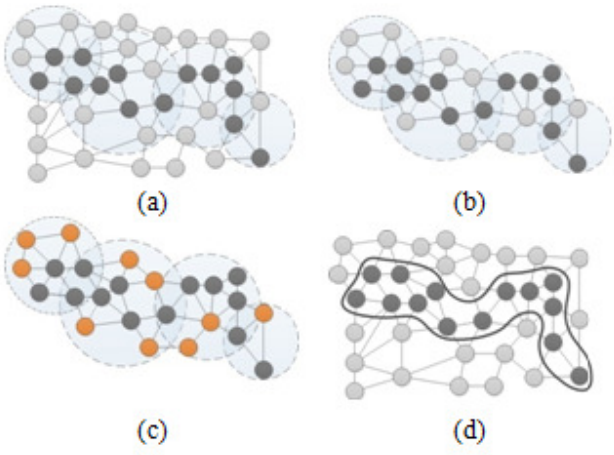

Figure 3. The key steps of the I Flex Scan algorithm

In order to find misjudged nodes, we need to change our targets into finding clusters with fewer cases. Let vector$\boldsymbol{x} = \{x_1, x_2, \dots, x_m\}^T$ represent the difference between the number of cases in each node and the global maximum number of cases in all nodes. It can be show that: $x_i = c_{\max} - c_i$ Where$i$denotes the ID of nodes,c_{max}denotes the maximum number of cases in all nodes, c_i denotes the number of case in node i.

In order to make it easier to optimize, let vector$\boldsymbol{F} = \{f_1, f_2, \dots, f_m\}$represents the state of the nodes, where $f_i = 1$ means node i is belong to the cluster, and 0 otherwise. Then $LR(Z)$can be expressed in the following form:

$$LR(Z) = \left(\frac{F\boldsymbol{x}}{1_m F^T \bar{X}}\right)^{Fx} \left(\frac{(1_m - F)\boldsymbol{x}}{\left(m - 1_m F^T\right)\bar{X}}\right)^{(1_m - F)x} I(Z)$$

Where $\bar{X}$is the mean value of x_1 to x_m,1_m is all one vector, and $I(Z)$ represents an indicator function, when$\frac{F\boldsymbol{x}}{\mathbf{1_m}F^T\bar{X}} > \frac{(\mathbf{1_m}-F)\boldsymbol{x}}{(m-\mathbf{1_m}F^T)\bar{X}}$,$I(Z) = 1$, and 0 otherwise.

2.3. CONNECTIVITY CONSTRAINTS

Assuming that the epidemic is transmitted through direct contact between patients, the outbreak should be geographically distributed.

In a network structure, the one can be expressed as the connectivity between nodes. Influenced by modern traffic, cluster of infectious diseases may appear in several network components that are not connected to each other, but the internal nodes of each component are still connected. Our algorithm can return each component separately as a possible outbreak result, so it is necessary to consider the connectivity constraints between nodes.

The network connectivity constraints based on spectral graph theory[8], the connectivity constraints can be expressed by algebraic form, based on the following theorem:

Theorem 1: let G be an undirected un weighted graph whose adjacency matrix is W and the Laplace matrix is L. Then the algebraic multiplicity corresponding to the zero eigen value of L is equal to the number of connected components in G.

Given a graph G with un weighted adjacency matrix A, the next formula is used as the connectivity constraint of the network nodes selected by the F:

$$Q(M,\gamma) = diag\left((A \circ M - \gamma M)1_n\right) - A \circ M + \gamma M \geq 0$$

Where$M = FF^T$, γ is a positive parameter,$A \circ M$denotes elementwise matrix multiplication:$(A \circ M)_{ij} = A_{ij}M_{ij}$.

If$Q\ (M,\gamma)$ is positive semidefinite matrix, then internal nodes are communicated with each other.γ is used to regulate the connectivity between network nodes.

3. THE GENETIC ALGORITHM APPROACH

3.1. ENCODING AND INITIALIZATION

Encode is the primary and the key step to apply the genetic algorithm. And the encoding methods affect the crossover operator and the mutation operator, which determines the efficiency of genetic evolution. According to the characteristic of this problem, we use the binary code method.

Let$t_1$andt_2 are the beginning and ending days of the epidemic data respectively, where $t_1$andt_2 are integers, vector F represents the selected state of the nodes. We have two different decision variables in our problem, so the chromosomes of the genetic algorithm are divided into two parts, shown as in Fig. 4.

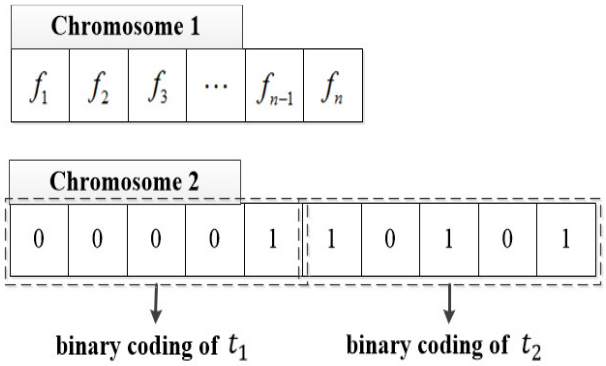

Figure 4. Structure of chromosome

Chromosome 1 represents the vector F, where $f_i = 1$ means node i belongs to the cluster, and vice versa. Chromosome 2 includes$t_1$andt_2, assuming that the outbreak time range is not more

than a month, t_1, t_2 can be represented by 5 bit binary numbers.As shown in Figure 4, the encoding of chromosome 2 indicates that the data range of our algorithm is within aninterval[1].

3.2. CROSSOVER AND MUTATION OPERATORS

The method we use to select two individuals from the parent generation is binary tournament. Then, we generate a binary string which has the same length as the parent individual randomly, in which 0 represents non-exchange and 1 represents exchange. The number of 1s is generated randomly from 1 to 3. The specific operation steps are as follows:

- Determine the number of individuals to choose.
- Select individuals randomly from the population, according to $LR(Z)$ value, choose the best individual as the offspring.
- Repeat the previous step until a new generation is formed.

As illustrated in Figure 5, according to the template of the binary string, we can cross two parent individuals and obtain new offspring.

Parent1	0	1	0	1	0	1
Parent2	1	1	0	0	1	1

Template: 1 0 0 1 0 0

Offspring1	1	1	0	0	0	1
Offspring2	0	1	0	1	1	1

Figure 5.Crossover operation of chromosome

The creation of offspring is given as follows:

- Select parent individuals $(G_i^{1,l}, G_i^{2,l})$ by binary tournament;
- Generate a random number $r \in [0,1]$and set the crossover possibilityP_c. If $r < P_c$, go to the next step, otherwise, this is non-exchange;
- Generate a binary string randomly and create offspring $(G_i^{1,l+1}, G_i^{2,l+1})$ using the crossover operator.

The crossover operator can yield good individual coding structure from a global perspective and is near the optimal solution.

Using mutation operator can improve the local search ability of genetic algorithm, maintain the population diversity and prevent premature convergence. The mutation operator was set to randomly change the value of one locus in chromosome 1 and 2, respectively.

3.3. I FLEX SCAN ALGORITHM

Based on the content discussed beyond, the basic process of the algorithm in this paper is described below and shown in Figure 6:

- Using Circle Scan algorithm to narrow the search range of epidemic area;
- Set the parameters such as population number, maximum generation number, and crossover and mutation probabilities;
- Generate an initial populationP_0;
- Adjust the initial solution set to satisfy the constraint conditions. Evaluating objective functions. Obtain the first generation P_1. Set thegeneration count$N_{gen} = 1$;
- Perform crossover and mutation for populationP_1, obtain the offspring generationQ_1, adjust the solution set to satisfy the constraint conditions, and evaluate objective functions for individual inQ_1;
- Merge the parent population P_1 and offspring populationQ_1,And according to the objective function value, sort the objective functions in descending order;
- Select individuals located at the front of the set, obtain the new parent population P_2;
- Record P_1, let $P_1 = P_2$, and increment N_{gen};
- Repeat Steps 5 to 8 until the count reaches the maximum generation number.

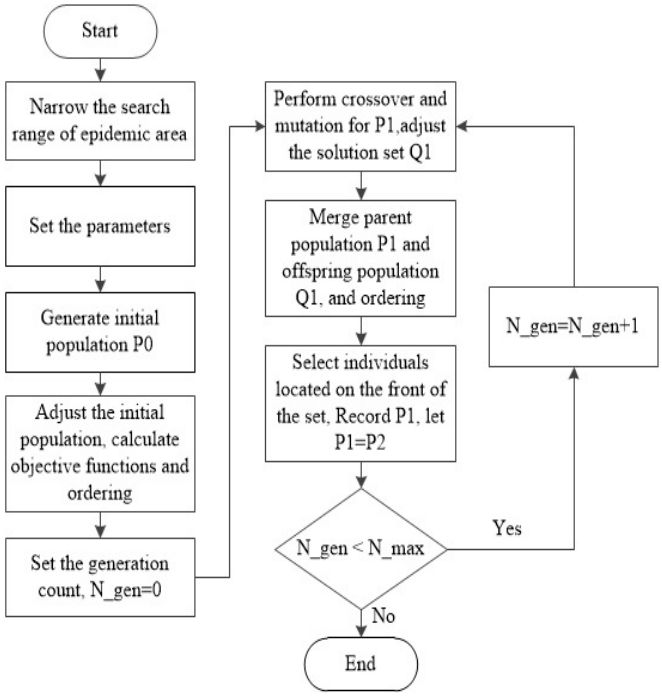

Figure 6.Flowchart for I Flex Scan algorithm

4. EXPERIMENTAL RESULTS AND ANALYSES

In this section, we compared the performance of our algorithm with the Circle Scan and Flex Scan using a simple network and simulated H1N1 flu data[9].

4.1. A SIMPLE NETWORK EXPERIMENT

The ground truth on a 30-node grid is shown in Figure 7(a). The input x is noiseless: let$x_i = 1$ for red nodes and 0 elsewhere. In each algorithm, suppose the upper limit of the abnormal node is 15, Figure 7(b) to 7(e) show the result of each algorithm when the LR value is maximum, where the red node is the abnormal node judged by the algorithm, and the blue node is the normal node judged by the algorithm.

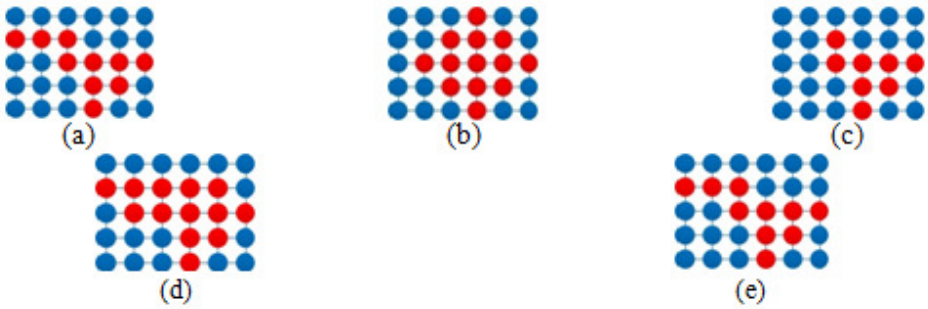

Figure 7.Comparison of three algorithms in a simple network. (a) shows ground truth. (b) shows result of Circle Scan. (c) shows result of Flex Scan. (d) and (e) show result of I Flex Scan when γ=0.01 and γ=0.005.

By observing Figure 7 (b), we can see that the result of Circle Scan algorithm is approximately circular. Even in the ideal case without noise data, the result is greatly different from the ground truth and unable to reflect the true distribution range of abnormal nodes.

Figure 7 (c) shows the results of the Flex Scan algorithm and found that the results of Flex Scan are similar to the ground truth. Notice that the result contains only 10 red nodes. This shows that the Flex Scan algorithm requires a larger node limit to find all the abnormal nodes, thus requiring longer running time.

Figure 7 (d) and (e) show the results of I Flex Scan algorithm under different parameterγ. The largerγ value in Figure 7 (d) results in a thicker connection between the red nodes. Al though the result contains all the red nodes in the ground truth, there is still a mis judgment. While the value of γ in Figure 7 (e) is small, so that IFlexScan allows the connections between nodes to be less tight and the red nodes correspond to the real situation.

The experimental results show that, compared with the previous algorithm, I Flex Scan algorithm can find abnormal nodes well, but for different problems, we need to dynamically adjustγ to optimize the result.

4.2. AN EXPERIMENT BASED ON BEIJING H1N1 FLU SIMULATION DATA

The genetic algorithm is applied for the study of clusters of high incidence of H1N1 simulation data. We use the same simulation data set with 4 simulated irregularly shaped clusters, which has been used in our previous research[9].

The performance of the algorithm is measured by recall, precision and F-Measure. In order to explain the meaning of each index, set the positive sample is the outbreak area and the negative sample is the normal area, the following concepts are used:

- TP (True Positive): The number of positive samples correctly predicted by the algorithm.
- TN (True Negative): The number of negative samples correctly predicted by the algorithm.
- FP (False Positive): The number of negative samples that are incorrectly predicted by the algorithm as a positive sample.
- FN (False Negative): The number of positive samples that are incorrectly predicted by the algorithm as negative samples.

The expression of recall and precision is shown in the following formula:

$$\mathrm{Re}call = \frac{TP}{TP + FN}$$

$$\mathrm{Pr}ecision = \frac{TP}{TP + FP}$$

Combine these two indicators to generate F-Measure, as shown in the following formula:

$$F_1 = \frac{2 \times \mathrm{Re}call \times \mathrm{Pr}ecision}{\mathrm{Re}call + \mathrm{Pr}ecision}$$

Figures 8 to 11 show the best results of the three algorithms for the four sets of simulation data respectively, where (a) is the result of Circle Scan; (b) is the result of Flex Scan; (c) is the result of I Flex Scan. The area of the outbreaks that are correctly predicted (TP corresponding area) is depicted in blue. The normal area that is correctly predicted (TN corresponding area) is depicted in white. The normal area that is predicted to be the outbreak area (FP corresponding region) is depicted in red. The outbreak area that is predicted to be the normal area (FN corresponding region) is depicted in yellow.

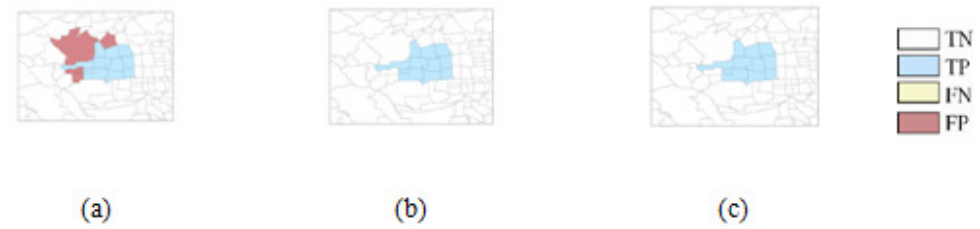

Figure 8. In Experiment 1, the best results for each algorithm.

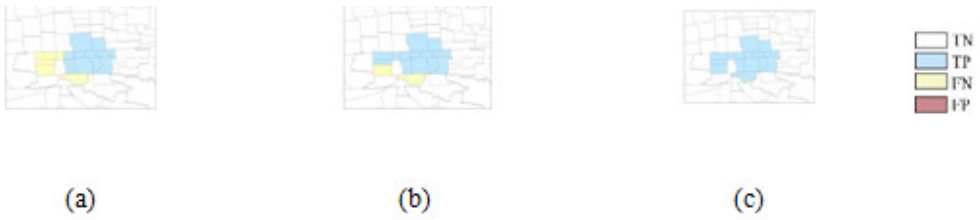

Figure 9. In Experiment 2, the best results for each algorithm.

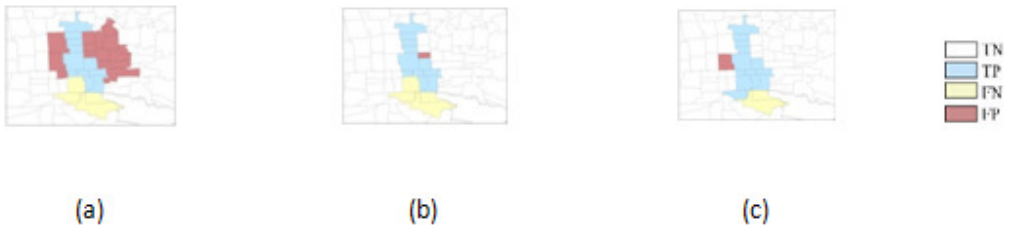

Figure 10. In Experiment 3, the best results for each algorithm.

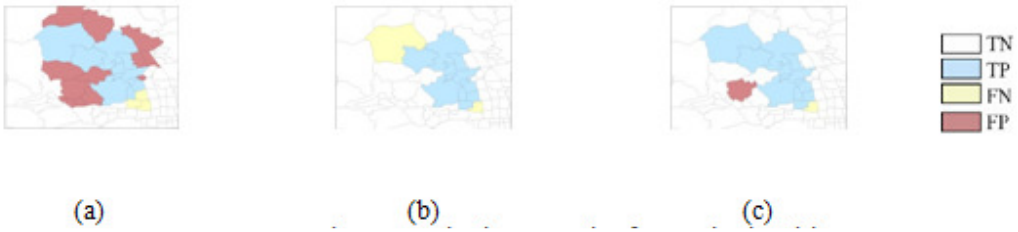

Figure 11. In Experiment 4, the best results for each algorithm.

The Circle Scan algorithm works well for regularly shaped outburst areas, but it still has miss judgment compared to Flex Scan and I Flex Scan algorithms. For irregularly shaped outburst are as, the results of Circle Scan are approximately circular with a large number of FP blocks, that is,the Circle Scan algorithm misjudges the normal area as an outbreak. Flex Scan and I Flex Scan results more in line with the actual situation.

Through the above evaluation methods, Table 1 shows the detailed results comparison.

Table 1 . The algorithms performance comparison.

Experiment	Algorithms	*TP*	*TN*	*FP*	*FN*	□□□□□□	□□□□□□□□□	$□_1$
	Circle Scan	11	294	4	0	1	0.73	0.84
1	Flex Scan	11	298	0	0	1	1	1
	I Flex Scan	11	298	0	0	1	1	1
	Circle Scan	10	294	0	5	0.67	0.91	0.77
2	Flex Scan	13	294	0	2	0.87	1	0.93
	I Flex Scan	15	294	0	0	1	1	1
	Circle Scan	11	280	14	4	0.73	0.44	0.55
3	Flex Scan	11	293	1	4	0.73	0.92	0.81
	I Flex Scan	13	293	1	2	0.87	0.93	0.90
	Circle Scan	9	285	12	3	0.75	0.43	0.55
4	Flex Scan	10	297	0	2	0.83	1	0.91
	I Flex Scan	12	296	1	0	1	0.92	0.96

Table 2 shows the four clusters of Figure 8(c) to Figure 11(c) and the runtimes required by the I Flex Scan to achieve the best results, where LLR = log(LR) is used instead of LR.

Table 2. The four clusters and runtimes of I Flex Scan.

Experiment	Case	Population	Incidence	LLR	Runtimes(second)
1	375	111182	0.0034	177.8	2
2	251	54503	0.0046	107.4	4
3	397	74942	0.0053	189.5	8
4	357	82038	0.0044	192.6	6

The results show that, for irregular outbreaks, I Flex Scan can find the clusters with high LLR value and the runtimes is shorter than 10 second.

5. CONCLUSIONS

We described a method for the detection of irregular spatial clusters, called I Flex Scan, which uses the spatial scan statistic in maps divided into finite numbers of regions. We use the genetic algorithm to improve operation efficiency, and set the constraints based on spectral graph theory, which can guarantee the connectivity of zones.

The algorithm is tested by analysing the simulation data of H1N1 influenza in Beijing. The results show that I Flex Scan can find irregularly-shaped connected clusters. And compared with the previous spatial scan statistic algorithms, our algorithm performs better with shorter time and higher accuracy. We believe that our study encourages further investigations for the use of genetic algorithms for epidemiological studies.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the support of National Science Foundation of China under grant 71373282.

REFERENCES

[1] Kulldorff M 1997 A spatial scan statistic J. Communications in statistics-theory and methods 26(6). 1481-96.

[2] Huang L, Kulldorff M and Gregorio D 2007 A spatial scan statistic for survival data J. Biometrics 63(1). 109 -18.

[3] Neill D B, Moore A W 2004 A Fast Multi-Resolution Method for Detection of Significant Spatial Disease Clusters J. Advances in Neural Information Processing Systems 13(4). 651 - 658.

[4] Kulldorff M, Huang L and Pickle L 2006 An elliptic spatial scan statistic J. Statistics in medicine 25(22). 3929-43.

[5] Tango T, Takahashi K 2005 A flexibly shaped spatial scan statistic for detecting clusters J. International journal of health geographic 4(1). 11.

[6] Duczmal L, Kulldorff M and Huang L 2006 Evaluation of spatial scan statistics for irregularly shaped clusters J. Journal of Computational and Graphical Statistics15(2). 428-42.

[7] Neill D B 2012 Fast subset scan for spatial pattern detection J. Journal of the Royal Statistical Society: Series B (Statistical Methodology)74(2). 337-60.

[8] Chung F R 2012 Spectral graph theory C.Regional Conference Series in Mathematics92. 212.

[9] Zhao Y, Mei S and Zhang W 2017 Irregular spatial cluster detection based on h1n1 flu simulation in Beijing C. the Asian Simulation Conference. 2017. 169-79.

AUTHORS

Shan Mei is working with National University of Defence Technology. He is interested in Complex Systems and Unmanned Aerial Vehicles.

Yonglin Lei is working with National University of Defence Technology. He is interested in Model-driven Architechure.

Mei Yang is working with National University of Defence Technology. She is interested in Modelling and Simulation.

Shan Mei is working with National University of Defence Technology. She is interested in Complex Systems and Complex Networks.