



Irregular Shaped Small Nodule Detection Using a Robust Scan Statistic

Ali Abolhassani¹ · Marcos O. Prates²  · Safieh Mahmoodi³

Received: 22 June 2021 / Revised: 9 March 2022 / Accepted: 24 June 2022 /
Published online: 26 August 2022

© The Author(s) under exclusive licence to International Chinese Statistical Association 2022

Abstract

The spatial scan statistics based on the Poisson and binomial models are the most common methods to detect spatial clusters in disease surveillance. These models rely on Monte-Carlo simulation which are time consuming. Moreover, frequently, datasets present over-dispersion which cannot be handled by them. Thus, we have the following goals. First, we propose irregularly shaped spatial scan for the Bell, Poisson, and binomial. The Bell distribution has just one parameter but it is capable of handling over-dispersed datasets. Second, we apply these scan statistics to big maps. A fast version, without Monte-Carlo simulation, for the proposed Poisson and binomial scans is introduced. Intensive simulation studies are carried out to assess the quality of the proposals. In addition, we show the time improvement of the fast scan versions over their traditional ones. Finally, we end the paper with an application on the detection of irregular shape small nodules in a medical image.

Keywords Bell distribution · Linear time subset scan · Minimum spanning tree · Scan statistic · Validity Index

✉ Marcos O. Prates
marcosop@est.ufmg.br

Ali Abolhassani
ali.abolhassani@azaruniv.ac.ir

Safieh Mahmoodi
mahmoodi@cc.iut.ac.ir

¹ Department of Applied Mathematics, Azarbaijan Shahid Madani University, Tabriz, Iran

² Departamento de Estatística, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

³ Department of Mathematical Sciences, Isfahan University of Technology, Isfahan, Iran

1 Introduction

Are points randomly distributed in space or time? This is a common question in almost any field of sciences, for example in astronomy [1, 2], image analysis [3, 4], data mining [5], criminology [6, 7], ecology [8, 9], geography [10, 11], pattern recognition [12], biology [13], forestry [14, 15], epidemiology [16, 17], etc. The answer to this question depends on the position and the number of individuals with some characteristics concerning each other. The final goal is to verify whether “spatial clustering has happened”.

The spatial scan statistic presented by Kulldorff and Nagarwalla [18], is a tool that became popular for the detection and inference of spatial clusters, especially in epidemiology. The detection of disease outbreaks at the earliest possible time is essential for public health centers to develop appropriate public policies. Therefore, the spatial scan statistic is one of the most used statistics to help epidemiologists detect and evaluate spatial clusters in disease surveillance. In summary, this method scans a map with circular windows to determine the most likely zone to be a spatial cluster. Then, using Monte-Carlo simulation determine the significance of this zone.

With the new challenges presented by real data, many extensions over the traditional circular scan statistic [19] was proposed to accommodate large spatial data [4], zero-inflation [20], and over-dispersion and zero-inflation simultaneously [21, 22]. Another common challenge is the detection of the irregularly-shaped cluster. Many works have also been done in this direction [e.g., 23, 24, 25, 26, and others]. Assunção et al. [23] detect irregularly shaped clusters using the minimum spanning tree (MST) from graph theory. The construction of the MST can drastically reduce the number of candidate clusters making the method scalable for larger datasets. However, the method presents an overestimation of the cluster size (the so-called octopus effect). To control the overestimation effect, Costa et al. [24] proposed three spatial scan statistics to find irregularly shaped clusters. These three proposals were constructed upon an early stopping rule, a double connection requirement, and maximum linkage criteria. Another solution was proposed by Zhou et al. [25]. Their method is detailed in Subsect. 2.3.2. For an up-to-date review of scan statistics, we refer to Abolhassani and Prates [27] which includes irregularly shaped scan statistics and many other subjects in this area.

To the best of our knowledge, in epidemiology, it is common for researchers to scan small maps, with less than 1000 counties. This is done in practice because of the large number of candidate circular windows in the scan process and the need for Monte-Carlo hypothesis testing that makes its use of big data inappropriate. In this paper, we have considered a medical image as the study map. The medical image can be considered as a large map. Each pixel of the image can be thought of as a county of a map. The darkness of a pixel corresponds to the number of cases in that pixel. To detect spatial clusters on a medical image (i.e., a big map), we present an irregularly-shaped version of the Poisson, binomial, and Bell scan statistics. Moreover, a fast algorithm to handle big data problems is introduced for the Poisson and binomial scans.

We show by extensive simulation that the irregular Bell is a robust scan alternative to model miss-specification in comparison with the binomial and Poisson models and that the fast scan alternatives can provide about 50% speedup in the computational time in comparison with its traditional Monte-Carlo implementation. Despite the robustness of the irregular Bell scan, we were unable to provide a fast version of its algorithm. When working with event data, Castellares et al. [28] showed that the likelihood calculation time can be drastically reduced. However, this is not related to the goal of the current study.

The rest of this paper is as follows: Sect. 2 provides a broad review of the scan statistics, including in Sect. 2.4 the recent Bell scan statistic that has only one parameter and is suitable for over-dispersed data. An algorithm to detect irregularly-shaped clusters for the Bell distribution with traditional and fast algorithms for binomial and Poisson distributions is presented in Sect. 3. Section 4 presents an intensive simulation study. An application for real data is done in Sect. 5. Finally, we conclude in Sect. 6.

2 A Review on Scan Statistic

2.1 Circular Scan Statistic

Kulldorff [19] proposed a Likelihood Ratio Test (LRT) to detect spatial clusters. To do so, consider an inhomogeneous Poisson point process over k regions or locations in a study area. Let X_i be the number of cases in region i with corresponding at-risk population n_i under unit-specific relative risk ζ_i such that $X_i \sim \text{Poisson}(n_i \zeta_i)$. Furthermore, let Z be a subset of indices $1, 2, \dots, k$, describing a given zone, which represents a candidate cluster. Define \mathcal{Z} as a collection of all candidate clusters. Kulldorff [19] formulated a scan statistic that compares the total number of case-counts in zone Z , $X_Z = \sum_{i \in Z} X_i$, against the total number of case-counts in \bar{Z} (i.e. the zone defined by the areas of the map not in Z), $X_{\bar{Z}} = \sum_{i \in \bar{Z}} X_i$, controlled by the corresponding population counts, that are, $n_Z = \sum_{i \in Z} n_i$ and $n_{\bar{Z}} = \sum_{i \in \bar{Z}} n_i$ within and without zone Z , respectively. Let $n = n_Z + n_{\bar{Z}}$ and $X = X_Z + X_{\bar{Z}}$, and assume that $\zeta_i = \zeta_Z$ for every region $i \in Z$ and that $\zeta_i = \zeta_{\bar{Z}}$ for every region $i \in \bar{Z}$. The hypothesis of interest is given by

$$H_0 : \forall Z \in \mathcal{Z}, \quad \zeta_Z = \zeta_{\bar{Z}} = \zeta_0, \quad \text{vs} \quad H_1 : \exists Z \in \mathcal{Z} \text{ such that } \zeta_Z > \zeta_{\bar{Z}}, \quad (1)$$

where H_0 implies that there is a constant risk, while H_1 implies that there is at least one cluster defined by a zone $Z \in \mathcal{Z}$ such that $\zeta_Z > \zeta_{\bar{Z}}$. Thus, for every candidate cluster Z , the likelihood function $L(Z) = L(Z, \zeta_Z, \zeta_{\bar{Z}})$ is given by

$$L(Z) = \prod_{i \in Z} \frac{e^{-n_i \zeta_Z} (n_i \zeta_Z)^{x_i}}{(x_i)!} \prod_{j \notin Z} \frac{e^{-n_j \zeta_{\bar{Z}}} (n_j \zeta_{\bar{Z}})^{x_j}}{(x_j)!}.$$

Additionally, to present how to find the most likely cluster (MLC), Kulldorff [19] developed a LRT defined by:

$$\lambda(Z) = \frac{\sup_{\zeta_Z > \zeta_{\bar{Z}}} L(Z, \zeta_Z, \zeta_{\bar{Z}})}{\sup_{\zeta_Z = \zeta_Z = \zeta_0} L(Z, \zeta_0)} I(\zeta_Z > \zeta_{\bar{Z}}) \tag{2}$$

with $\lambda(Z) = 1$, if $\zeta_Z < \zeta_{\bar{Z}}$ that allows to decide whether or not the MLC is statistically significant, meaning whether the area(s) included in the most likely cluster really incorporate(s) an abnormally high number of cases.

The clustering method introduced by Kulldorff [19] is called scan statistic and has some limitations. First, it cannot detect non-circular clusters. Second, the distribution of the test statistic λ is unknown. To solve the first limitation, Kulldorff et al. [29] proposed an elliptic spatial scan statistic, but this method does not detect irregularly shaped clusters well. To find irregularly-shaped clusters, the Minimum Spanning Tree (MST) was first introduced [23] and later the Adaptive MST (AMST) method [25]. They not only enabled researchers to find irregularly shaped clusters but also decreased the scanning time of the map in two ways: (1) by decreasing the cardinality of the candidate class, and (2) by applying Linear Time Subset Scan (LTSS) property [30]. More details about these methods are discussed in Sect. 2.3.

The second limitation of the scan statistic was solved by Soltani and Aboukhamseen [31, 32]. They found the exact distribution for λ (Sect. 2.2) which allowed us to modify the Poisson and binomial scan introduced by Zhou et al. [25] to a faster alternative. Hence, these modified versions are more adequate to big maps.

2.2 Spatial Clustering Without Monte-Carlo

The advantage of Soltani and Aboukhamseen [32] and Aboukhamseen et al. [31] methods were the elimination of Monte-Carlo in the scan statistic procedure. Consider the hypothesis testing (1), and suppose that $G = Z_1 \cup Z_2 \cup \dots \cup Z_k$ is the studying region, also $X_{+,Z}$ and $n_+(G)$ are the number of points (cases) in Z and G , respectively. Let A_Z be the event that an individual is in a zone Z and B_+ stands for the event that an individual in G is a case that has characteristic + (for example an infected person). Consider $P_{+|Z} = P(B_+|A_Z)$ and let μ be a counting measure on (G, \mathcal{F}) such that $\mu(Z)$ is the number of people in Z and \mathcal{F} is a sigma-field on G containing Z_1, \dots, Z_k . Both $\mu(Z_i), i = 1, \dots, k$ and $\mu(G)$ are known. The probability of A_Z is defined as $\nu(Z) = \frac{\mu(Z)}{\mu(G)}$. According to the above notations, Soltani and Aboukhamseen [32] proved that (1) is equivalent to

$$H_0 : P_{Z|+} = \nu(Z) \quad \text{vs} \quad H_1 : P_{Z|+} > \nu(Z). \tag{3}$$

In addition, they found the exact and asymptotic distribution of points in zone Z under null hypothesis H_0 , as follows:

$$\begin{aligned} X_{+,Z} &\sim \text{Bin}(n_+(G), \nu(Z)), \\ X_{+,Z} &\implies N(n_+(G)\nu(Z), n_+(G)\nu(Z)[1 - \nu(Z)]), \end{aligned}$$

where (\implies) denotes convergence in distribution.

According to the above facts, zone Z is a cluster in level α if

$$z = \frac{\frac{X_{+,Z}}{n_+(G)} - v(Z)}{\sqrt{\frac{v(Z)[1-v(Z)]}{n_+(G)}}} > z_\alpha. \tag{4}$$

In this method of cluster detection, one does not need Monte-Carlo hypothesis testing, hence it is quicker than the method of Kulldorff and Nagarwalla [18].

Aboukhamseen et al. [31] developed a spatial scan statistic for a situation in which $n_+(G)$ is a random variable from a Poisson distribution with unknown parameter $\lambda(G)$. Also they consider $X_{+,Z}|n_+(G) \sim Bin(n_+(G), v(Z))$. Since $n_+(G)$ has Poisson distribution with parameter $\lambda(G)$ and using the marginal distribution of $X_{+,Z}$, they gave a confidence interval for $v(Z)$. Based on the value of $v(Z)$, one can decide that if the null hypothesis in (3) will be rejected or not. The large discrepancy between the lower limit of the confidence interval and computed statistic $v(Z)$ means smaller p-value and hence greater significance of z .

Aboukhamseen et al. [31] proposed

$$\pm 0.25z_{\alpha/2} + \sqrt{z_{\alpha/2}^2 + 4n_+(G)}$$

as a confidence interval for $\lambda(G)$. However, we notice that this formula is not correct and suggest the use of

$$n_+(G) + \frac{1}{2}z_{\alpha/2}^2 \left(1 \pm \sqrt{1 + \frac{4n_+(G)}{z_{\alpha/2}^2}} \right)$$

to find correct confidence interval for $\lambda(G)$. In this work we use the latter confidence interval.

2.3 Irregular-Shape Clusters

2.3.1 Minimum Spanning Tree

As mentioned before, in the scan statistic method [18], the class of candidates (i.e., circular zones) is too big and it is necessary to compute the likelihood ratio too many times. Further, the scan statistic is capable of detecting only circular clusters. Hence, Assunção et al. [23] solved these two undesired problems by using an MST. In their work, they proposed the MST technique to detect irregularly-shaped clusters. The method is as follows:

Consider a map and mark the center of each county by its centroid. Connect any two counties if they are neighbors. Construct an in-directed graph corresponding to the map. Give a weight $W(i, j)$ to each the edge (v_i, v_j) , where $W(i, j)$ is computed by the Kullback–Liebler (KL) divergence. Weights reflect dissimilarity of density

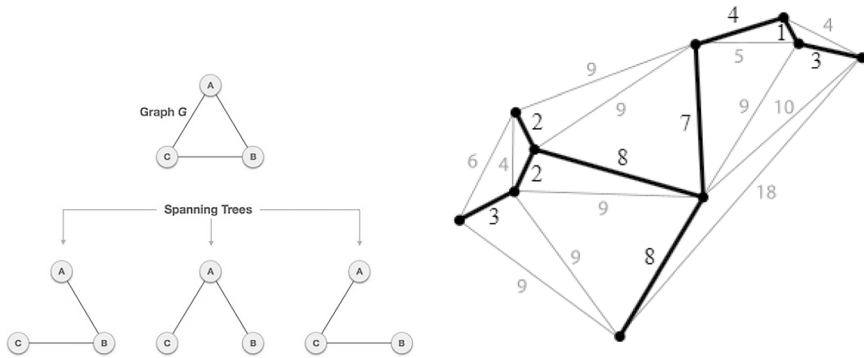


Fig. 1 Left: A Graph and its spanning trees. Right: A graph with weights on edges is in gray color. Black lines show MST

corresponding to counts of cases between two counties. High $W(i, j)$ means high dissimilarity between the density of county i and j .

A spanning tree τ of a graph G is a sub-graph of G which is a tree and contains all nodes of G (Fig. 1 left). The MST is a spanning tree which has the minimum weight. Figure 1 right, shows the representation of an MST. Prim [33] proposed one of the simplest algorithms to build an MST and this method was used by Assunção et al. [23].

After building the MST, Assunção et al. [23] reduced the class of the candidates from many circular zones to n candidates (n is the number of nodes) by removing one edge at a time from the original MST. In other words, by removing one edge from the MST, two sub-graphs appear. Assunção et al. [23] considered the smallest one as a candidate. Then they return the eliminated edge to its place and remove another edge. Again the smallest sub-graph is considered as the second candidate. This procedure continues until getting the n -th candidate. After obtaining the class of all candidates (with the cardinality of n), they compute λ for each element of this class to determine the MLC. In the next step, Monte-Carlo hypothesis testing is used to decide whether the MLC is significant as a cluster or not.

Although the MST method improved the scan statistic in two aspects (i.e., allows flexible shape for the candidate cluster and reduces the class of candidates from a large number of circles to n candidates), it has some deficiencies. First, it just detects one cluster on the map. Second, it tends to detect clusters that are bigger than their actual size. Third, it still requires Monte-Carlo hypothesis testing. To solve the first problem, Zhou et al. [25] introduce the AMST method that is discussed in the following subsection.

2.3.2 Adaptive Minimum Spanning Tree

As mentioned, the MST method detected only one cluster on the map. Although researchers can remove some heavy edges to find two or more clusters, determining

the number of elimination is not trivial and is a drawback. To overcome this problem, Zhou et al. [25] proposed the AMST method. In this method, one does not need to have prior knowledge about parameters such as the number of clusters and the initial cluster center.

For the AMST method, the concept of validity index is important and it is defined as:

$$\text{val}_{\text{index}} = \text{Intra}_{\text{dist}} / \text{Inter}_{\text{dist}}, \tag{5}$$

such that $\text{Intra}_{\text{dist}}$ measures compactness of sub-partitions of a graph while $\text{Inter}_{\text{dist}}$ measures isolation of them, i.e., separation between the sub-partitions.

The Intra cluster distance and Inter cluster distance are defined as follows:

$$\text{Intra}_{\text{dist}} = \sum_{i=1}^K \sum_{j \in C} |\lambda_{C^{ij}} - \lambda_{C^i}|^2 / K.$$

This is the average of the sum square of the difference of all the incidence rates within a sub-partition, from the rate of that sub-partition. In addition,

$$\text{Inter}_{\text{dist}} = \max_{ij} |\lambda_{C^i} - \lambda_{C^j}|^2.$$

This is the maximum diversification of the point rates of any two sub-partitions with rates λ_{C^i} and λ_{C^j} . Note that λ_{C^i} is the expected point rate of sub-partition C^i and $\lambda_{C^{ij}}$ is the point rate of county j in candidate sub-partition C^i . We call $|\lambda_{C^i} - \lambda_{C^j}|$ as the distance between sub-partitions i and j . K is the total number of sub-partitions after removing some edges in the minimum spanning tree. One can estimate $\lambda_{C^{ij}}$ and λ_{C^i} by using a maximum likelihood approach. By minimizing $\text{val}_{\text{index}}$, the best partition of the MST can be obtained. Then Zhou et al. [25] applied linear time subset scan property (LTSS) [30], on the best partition of the MST to find clusters.

The steps of the algorithm based on the AMST can be found in Zhou et al. [25]. Other proposals for validity indexes are presented in the Supplementary Material Section SM-1 and results were found to be similar to the former one.

2.4 The Bell Scan Statistic

Two common classical models to detect spatial clusters are the Poisson and binomial models. Although the Poisson distribution has just one parameter, it has a restriction of having the variance equal to the mean. Hence, this model is not suitable for over-dispersed data sets. In the case of the binomial model, it has two parameters and the index of dispersion (the ratio of variance to the expected value) is less than 1. Considering these facts, the Poisson and binomial distributions may not be appropriate to handle over-dispersed data.

The Bell distribution was introduced by Bell [34, 35], which has just one parameter and it can be applied to count data with over-dispersion. Random variable V has Bell distribution with parameter θ if its probability mass function (p.m.f.) follows:

$$P(V = v) = \frac{\theta^v e^{-\theta} + 1 B_v}{v!}, \quad \theta > 0, \quad v = 0, 1, 2, \dots, \quad (6)$$

such that B_v 's are Bell numbers, which equals to the v -th moment of Poisson distribution with rate 1.

In the Supplementary Material Section SM-2, we mention some useful properties which are important in simulation of data from Bell distribution.

Abolhassani et al. [21] presented the circular Bell and the zero-inflated circular Bell scan statistics. To construct the Bell scan statistic, they supposed that each cell i in a map, has an observed count of cases v_i , such that it is a realization of the Bell distribution with parameter $\theta_i = W_0(E_i \zeta_i)$, i.e., $V_i \sim \text{Bell}(\theta_i)$ with the expected count $E_i \zeta_i$, where E_i is a known value that one would like to control for (offset) and ζ_i is the relative risk. As before, any connected sub-region can be considered as spatial cluster candidate, and \mathcal{Z} is the class of all candidates. They are interested to perform test (1).

The likelihood function under H_1 is written as follows:

$$L_1(Z, \zeta_i, \zeta_Z) = \prod_{i \in Z} \frac{(W_0(E_i \zeta_Z))^{v_i} e^{-W_0(E_i \zeta_Z)} + 1 B_{v_i}}{v_i!} \times \prod_{i \in \bar{Z}} \frac{(W_0(E_i \zeta_{\bar{Z}}))^{v_i} e^{-W_0(E_i \zeta_{\bar{Z}})} + 1 B_{v_i}}{v_i!},$$

and the likelihood under H_0 as:

$$L(Z, \zeta_0) = \prod_{i=1}^k \left(\frac{B_{v_i}}{v_i!} (W_0(E_i \zeta_0))^{v_i} e^{(1 - W_0(E_i \zeta_0))} \right).$$

The derivative of the Lambert function is given by $W'_0(x) = \frac{W_0(x)}{x(1 + W_0(x))}$. Thus, to find the MLE of ζ_0 under H_0 , the $\ln L(Z, \zeta_0)$ is calculated by:

$$L_0 = \ln L(Z, \zeta_0) = \sum_{i=1}^k \ln \left(\frac{B_{v_i}}{v_i!} \right) + \sum_{i=1}^k v_i \ln(W_0(E_i \zeta_0)) + k - \sum_{i=1}^k e^{W_0(E_i \zeta_0)}$$

Hence

$$\frac{dL_0}{d\zeta_0} = \sum_{i=1}^k \frac{v_i}{\zeta_0(1 + W_0(E_i \zeta_0))} - \sum_{i=1}^k \frac{E_i}{(1 + W_0(E_i \zeta_0))} = 0,$$

which can be solved numerically. Similarly for H_1 , the parameters ζ_Z and $\zeta_{\bar{Z}}$ can be obtained. Likewise Kulldorff [19], to find a spatial cluster, they calculated (2). Let $\lambda = \max_Z \lambda(Z)$ be the Bell spatial scan statistic. Since the denominator is not dependent on Z , it is sufficient to maximize the numerator of $\lambda(Z)$. Any Z which maximizes $\lambda(Z)$ is the MLC. After determining the MLC, Monte-Carlo simulation can be employed to check its significance. Clearly, as in the Poisson and binomial scan statistics, the Bell distribution is able to control for any important factor such as, population size when $E_i = n_i$ and perform the analysis over the relative risk ζ_i .

However, in real life, we can find maps for which the population size of cells is the same. For example, consider a medical image. Each pixel can be considered as a cell with the same population at risk. The darkness of each pixel corresponds to the number of cases in that pixel.

Thus, when $E_i = E, \forall i$, the hypothesis testing in (1) is equivalent to the hypothesis testing presented in (7):

$$H_0 : \theta_Z = \theta_{\bar{Z}} = \theta_0, \quad \forall Z \in \mathcal{Z} \quad \text{vs} \quad H_1 : \exists Z \in \mathcal{Z} \text{ such that } \theta_Z > \theta_{\bar{Z}}. \quad (7)$$

The likelihood under H_1 can be simplified as follows:

$$\begin{aligned} L_1(Z, \theta_Z, \theta_{\bar{Z}}) &= \prod_{i \in Z} \frac{\theta_Z^{v_i} e^{-\theta_Z+1} B_{v_i}}{v_i!} \times \prod_{i \in \bar{Z}} \frac{\theta_{\bar{Z}}^{v_i} e^{-\theta_{\bar{Z}}+1} B_{v_i}}{v_i!} \\ &= \left(\prod_{i=1}^k \frac{B_{v_i}}{v_i!} \right) \theta_Z^{\sum_{i \in Z} v_i} e^{-(\theta_Z+1) \sum_{i=1}^k I_i(Z)} \theta_{\bar{Z}}^{\sum_{i \in \bar{Z}} v_i} e^{-(\theta_{\bar{Z}}+1) \sum_{i=1}^k I_i(\bar{Z})}, \end{aligned}$$

where $I(\cdot)$ is the indicator function and k is the number of areas (cells) on the map. The MLEs for the parameters under H_1 can be obtained by:

$$\frac{\partial \ln L_1}{\partial \theta_{\bar{Z}}} = 0 \implies \frac{\sum_{i \in \bar{Z}} v_i}{\theta_{\bar{Z}}} - e^{\theta_{\bar{Z}}} \sum_{i=1}^k I_i(\bar{Z}) = 0.$$

Therefore, $\hat{\theta}_{\bar{Z}} = W_0(\bar{v}_{\bar{Z}})$, where $\bar{v}_{\bar{Z}} = \sum_{i \in \bar{Z}} v_i / \sum_{i=1}^k I_i(\bar{Z})$. The likelihood under H_0 is given by:

$$L(Z, \theta_0) = \prod_{i=1}^k \frac{B_{v_i}}{v_i!} \theta_0^{\sum_{i=1}^k v_i} e^{(1-e^{\theta_0})k},$$

and the MLE for θ_0 is of the form $\hat{\theta}_0 = W_0(\bar{v})$. Therefore, under this restriction, the Bell scan statistic has close form and can be directly obtained by θ in (2).

In this paper, we extend the circular Bell scan statistic proposed by Abolhasani et al. [21] to the irregular Bell scan. The algorithm of this scan is presented in Subsect. 3.2.

3 Fast Irregular Shape Cluster

In this section, we present three algorithms to find irregularly shaped spatial clusters. Two of them (i.e., Poisson and binomial) do not need Monte-Carlo simulation. The third algorithm (i.e., Bell) is a robust scan method, it is suitable for over-dispersed data sets but requires Monte-Carlo simulation. All of these algorithms are suitable for big maps.

3.1 Fast Algorithm for Binomial and Poisson Models

As mentioned, Zhou et al. [25] proposed the AMST method for detecting irregular shape clusters fast. However, this method needs a large number of simulated data sets (for example 20, 000) to obtain the high percentiles of the test statistic. In this section, we propose an algorithm that increases the speed of the method of Zhou et al. [25] eliminating the need of Monte-Carlo simulation [31, 32]. The new algorithm for Poisson model (Algorithm 1) is as follows:

Algorithm 1 Fast Poisson scan statistics algorithm.

- 1: Find a MST based on the KL divergence [23].
 - 2: Find the best partition of the MST based on a val_{index} .
 - 3: Find the candidate class, \mathcal{Z} , by the LTSS property [30].
 - 4: For a $Z \in \mathcal{Z}$, obtain $\nu(Z)$ and its confidence interval (Section 2.2).
 - 5: Based step 4, decide if the null hypothesis is rejected or not.
 - 6: Repeat steps 4 and 5 for all Z 's in the candidate class one by one.
 - 7: Show the significant clusters on the map
-

In the case of the binomial model we propose Algorithm 2:

Algorithm 2 Fast binomial scan statistics algorithm.

- Steps 1-2 are the same as in the previous algorithm.
 - 3: Count the individuals and cases in G , i.e. $\mu(G)$ and $n_+(G)$.
 - 4: Obtain $\mu(Z)$ and $X_{+,Z}$ where Z is in the candidate class.
 - 5: Compute z in (4).
 - 6: The null hypothesis is rejected if $z > z_\alpha$ where z_α is obtained by the normal table.
 - 7: Repeat steps 5 to 7 for all Z 's in the candidate class one by one.
 - 8: Show the significant clusters on the map.
-

3.2 Bell Model

To find irregular spatial clusters based on the Bell model, we need to calculate the KL divergence for this distribution. Let under H_1 in (1) $V_i \sim Bell(W_0(E_i \zeta_i))$ and $V_j \sim Bell(W_0(E_j \zeta_j))$, such that under H_0 , we have $\zeta_i = \zeta_j = \zeta$. The KL divergence is:

$$W(i, j) = E_i \zeta_i \left(\ln \frac{W_0(E_i \zeta_i)}{W_0(E_i \zeta)} \right) + E_j \zeta_j \left(\ln \frac{W_0(E_j \zeta_j)}{W_0(E_j \zeta)} \right) + e^{W_0(E_i \zeta)} - e^{W_0(E_i \zeta_i)} + e^{W_0(E_j \zeta)} - e^{W_0(E_j \zeta_j)}. \quad (8)$$

Under the constraints of (7) the divergence is given by:

$$W(i, j) = \hat{\theta}_i \exp \hat{\theta}_i \log \frac{\hat{\theta}_i}{\hat{\theta}} + \hat{\theta}_j \exp \hat{\theta}_j \log \frac{\hat{\theta}_j}{\hat{\theta}} - (\exp \hat{\theta}_i - \exp \hat{\theta}) - (\exp \hat{\theta}_j - \exp \hat{\theta}) \tag{9}$$

where

$$\hat{\theta}_i = W_0(v_i), \quad \hat{\theta}_j = W_0(v_j), \quad \hat{\theta} = W_0((v_i + v_j)/2).$$

After the KL divergence determination, we propose Algorithm 3 for the irregular Bell scan:

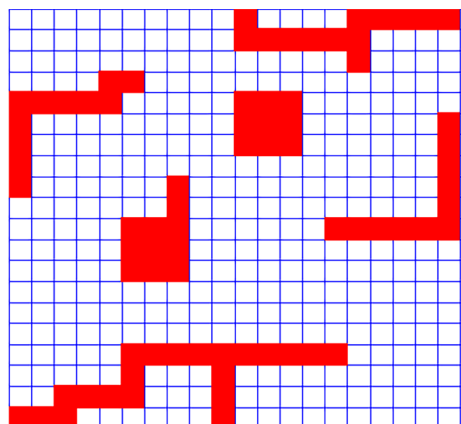
Algorithm 3 Irregular Bell scan statistics algorithm.

- 1: Find the MST based on the KL divergence using (8) (or (9), when E_i 's are equal).
 - 2: Find the best partition of the MST based on a val_{index} .
 - 3: Find the class of the candidates by the LTSS property [30].
 - 4: For the first candidate in the candidate class, using Monte-Carlo decide that if it is significant or not as a cluster.
 - 5: Repeat step 4 for all Z 's in the candidate class one by one.
 - 6: Show the significant clusters on the map.
-

4 Simulation

In this section, following the type of maps from our application, we study maps similar to the one presented in Fig. 2. Our simulation is based on three main scenarios to detect irregular shape spatial clusters and each scenario has 3 steps where the relative risk of the cluster areas is increased in each scenario. Three spatial scans (Bell, binomial, Poisson) are compared based on these scenarios.

Fig. 2 Study region with cluster area in red color. The shape of clusters is not circular (Color figure online)



4.1 Poisson Maps

In the first scenario, we generate the map with 20×20 cells using Poisson distribution. The population of each cell is constant and set as 1000. We consider irregular shape clusters with different shapes in the map: (1) L shape, (2) circular, (3) circular with tail, (4) snake and (5) snake with two heads. These shapes are shown with red color in Fig. 2. Inside the red areas, we generate the number of cases using Poisson(12) and outside of those areas, using Poisson(10), which means a higher relative risk inside the cluster of 20%. Then, we apply three different algorithms, i.e., Ir-Poisson (Algorithm 1), Ir-binomial (Algorithm 2), and Ir-Bell (Algorithm 3) to detect clusters. We repeat this process 200 times. Using four criteria we compare the three algorithms. These criteria are biasness, recall, precision and harmonic mean of precision and recall ($F1$), which are as follows.

First, Prates et al. [36] discussed the relative risk and biasness in spatial scan statistics. The bias is defined as the true ratio of the parameters inside and outside the cluster to the ratio of their estimated value. Bias values near 1 mean that the selected clusters are better to estimate the relative risk between the cases inside and outside the clusters than detected clusters with a bigger or smaller value for biasness. The precision and recall are two famous criteria in clustering problems which are defined as:

$$\text{Recall} = \frac{|\text{Detected cluster} \cap \text{True cluster}|}{|\text{True cluster}|},$$

$$\text{Precision} = \frac{|\text{Detected cluster} \cap \text{True cluster}|}{|\text{Detected cluster}|},$$

and also

$$F1 = \frac{2(\text{recall} \times \text{precision})}{\text{recall} + \text{precision}}$$

such that $|A|$ is the cardinality of set A .

The results for this simulation are shown in Fig. 3. According to this figure, the recall for Ir-binomial is higher, but its precision is lower than the other scans. This means this model leads to over-estimation in cluster detection. The Ir-Poisson and Ir-Bell have very similar behavior in precision and recall. The bias values are almost the same for the three models. In the case of $F1$, Ir-Poisson and Ir-Bell are very similar and some times the $F1$ for them reaches to above 0.5, where Ir-binomial scan cannot reach that.

In the next step of the simulation, we change the parameter inside the cluster to 20 and consider the parameter outside cluster 10 providing a relative risk of 100%. The results of cluster detection are shown in Fig. 4. The recall (precision) for Ir-Poisson is higher (lower) than other scans. Since the recall for Ir-Poisson is near 1 and its precision is high, it means the true clusters are detected with few non-cluster areas also included as clusters. Ir-binomial has more bias, and the other scans are very similar to each other in this case. $F1$ for Ir-Bell and Ir-binomial are a little higher than Ir-Poisson.

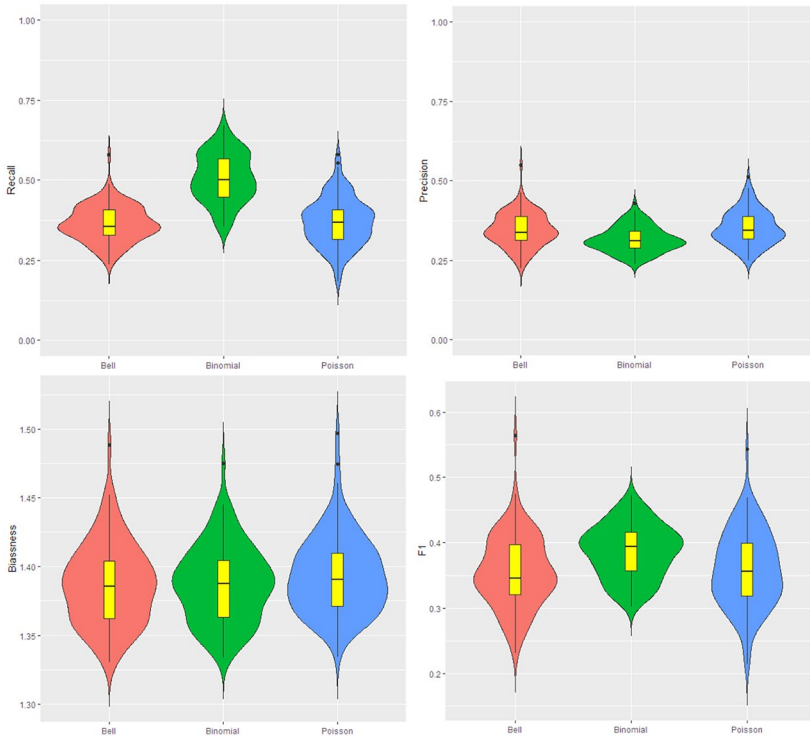


Fig. 3 Data for the map are generated from the Poisson with parameter 12 and 10 respectively inside and outside cluster. From the top left to the bottom right: the violin plot for the recall, precision, bias, and $F1$. The number of iteration is 200. The red, green and blue colors are respectively for the Ir-Bell, Ir-binomial and Ir-Poisson scans (Color figure online)

To have a better vision about the performance of three scans, we select the first 50 iterations of simulations and plot precision and recall point-wise in Fig. 5. Based on this figure, in the case of Ir-Poisson and Ir-Bell, precision is always under recall which is not true for Ir-binomial. Correlation of recall and precision for Ir-Bell, Ir-binomial, and Ir-Poisson are 1, -0.17 , -0.03 . This means we have more over-estimation and under-estimation in applying Ir-binomial. Considering these facts and the graph of bias value, we believe that Ir-Poisson and Ir-Bell detect clusters better in this scenario comparing to the Ir-binomial.

The increasing of the parameter inside the cluster from 20 to 40, causes recall, precision, bias, and $F1$ to become very close to 1, as expected because in this step the distinction between the cluster areas in comparison to the non-cluster areas is very large.

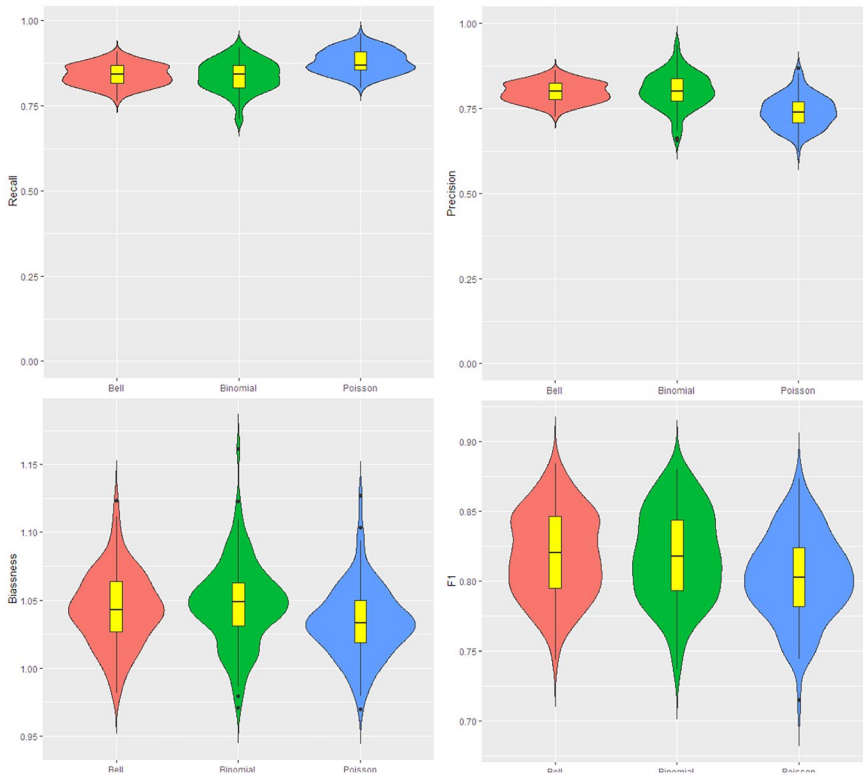


Fig. 4 Data for the map are generated from the Poisson with parameter 20 and 10, respectively, inside and outside cluster. From the top left to the bottom right: the violin plot for the recall, precision, bias, and $F1$. The number of iteration is 200. The red, green and blue colors are respectively for the Ir-Bell, Ir-binomial and Ir-Poisson scans (Color figure online)

4.2 Binomial Maps

The results of cluster detection for binomial maps for different scenarios are presented and discussed in details in the Supplementary Material Section SM-3. Briefly, the Ir-Bell and Ir-binomial perform better in irregular shape cluster detection comparing to the Ir-Poisson scan.

4.3 Bell Maps

In the first step of this scenario, the cluster areas in the map are generated from a $Bell(W_0(12))$. A $Bell(W_0(10))$ is used to generate cases outside the cluster areas. Therefore, we guarantee a relative risk of 20% inside the cluster. We apply three different scans (Ir-Bell, Ir-Poisson, Ir-binomial) to detect clusters. The results of

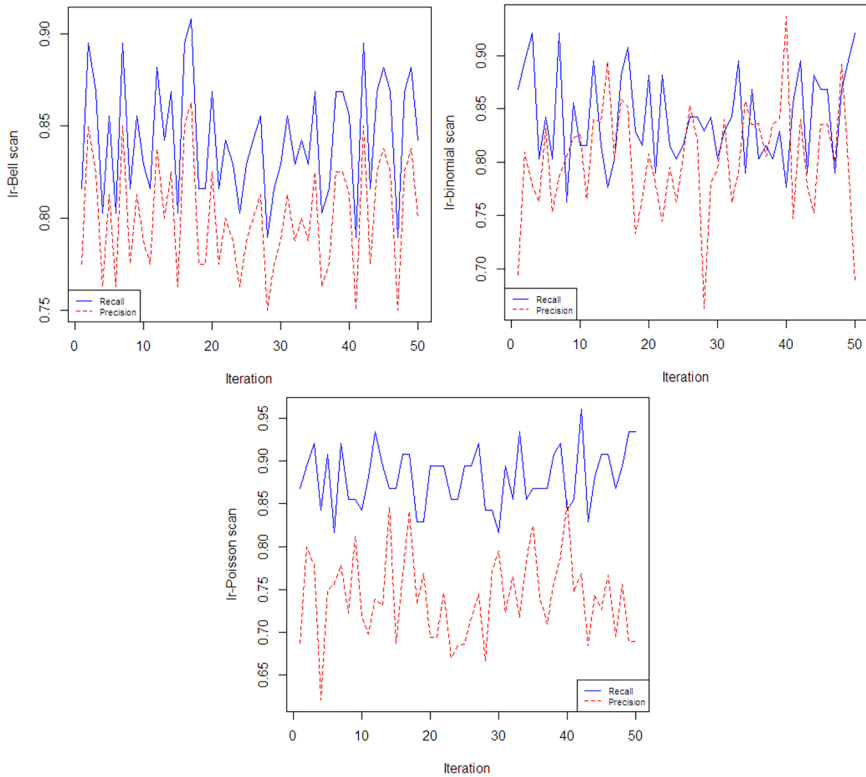


Fig. 5 Variation of the recall and precision in the first 50 iteration of the simulation study for the scenario Poisson(10)-Poisson(20). The Ir-Bell, Ir-binomial and Ir-Poisson are presented

comparison are presented in Fig. 6. We can see that the Ir-Bell scan outperforms the other methods under this scenario. Notice that $F1$ is much higher than the other two with a smaller bias.

Increasing $W_0(12)$ to $W_0(20)$ leads us to declare that the Ir-Bell scan has better performance in cluster detection. Because bias values for this model are smaller than the two other models, and its $F1$ is higher. The results of the cluster detection are shown in Fig. 7. The three models have almost the same recall but the precision for Ir-Bell is considerably higher than the other two. This leads to high $F1$ and better bias value for the Ir-Bell scan. Finally, we increase the parameter inside the cluster to $W_0(40)$. In this case, the three scans have perfect performance and all criteria are close to 1.

Overall, we can conclude that the Ir-Bell scan is robust to other generation schemes (model misspecification) and outperform the other scans when is the true distribution. It is a strong candidate to consider when analysing real data.

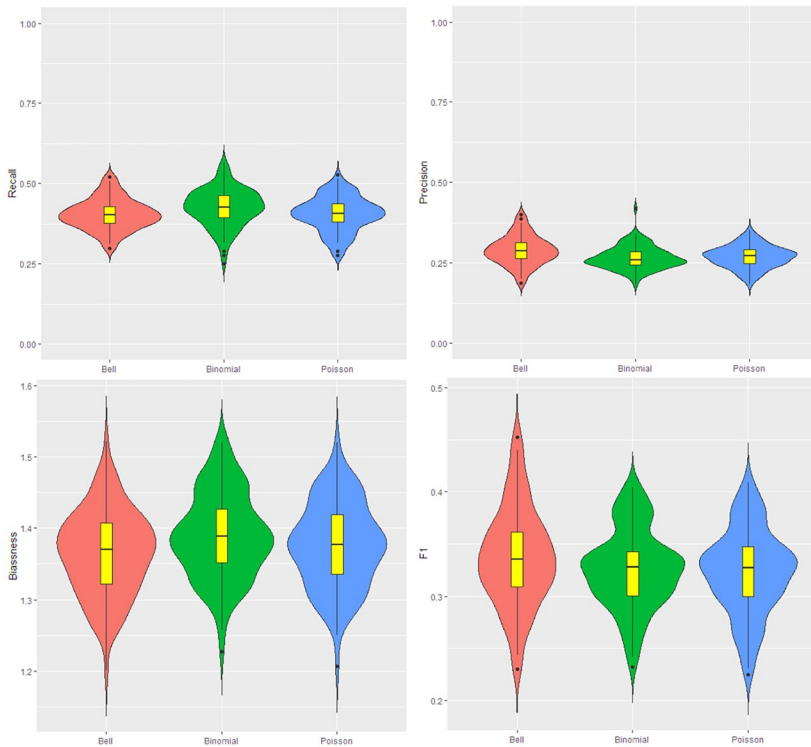


Fig. 6 Data for the map are generated from the Bell with parameter $W_0(12)$ and $W_0(10)$ respectively inside and outside cluster. From the top left to the bottom right: the violin plot for the recall, precision, bias, and $F1$ in 200 iteration. The red, green and blue colors are for the Ir-Bell, Ir-binomial and Ir-Poisson scans (Color figure online)

5 Application

5.1 Irregularly Shaped Spatial Clusters in a Medical Image

In this section, a real data set is studied. Since we concluded in Sect. 4 that the Ir-Bell is a more robust scan statistics, we will proceed with our analysis using the Ir-Bell scan.

Any image can be considered as a map with many cells. Therefore, the scan statistic method was applied by Popescu and Lewitt [37] to detect circular small nodules on a medical image. We use the same image to detect small nodules with more details (irregular shape nodules) and compare the performance of our algorithm with their results. This image has $205 \times 205 = 42,025$ pixels.

The location and the size of the true clusters are not known in practice. Hence, choosing the size of the scanning window in spatial clustering problems is a

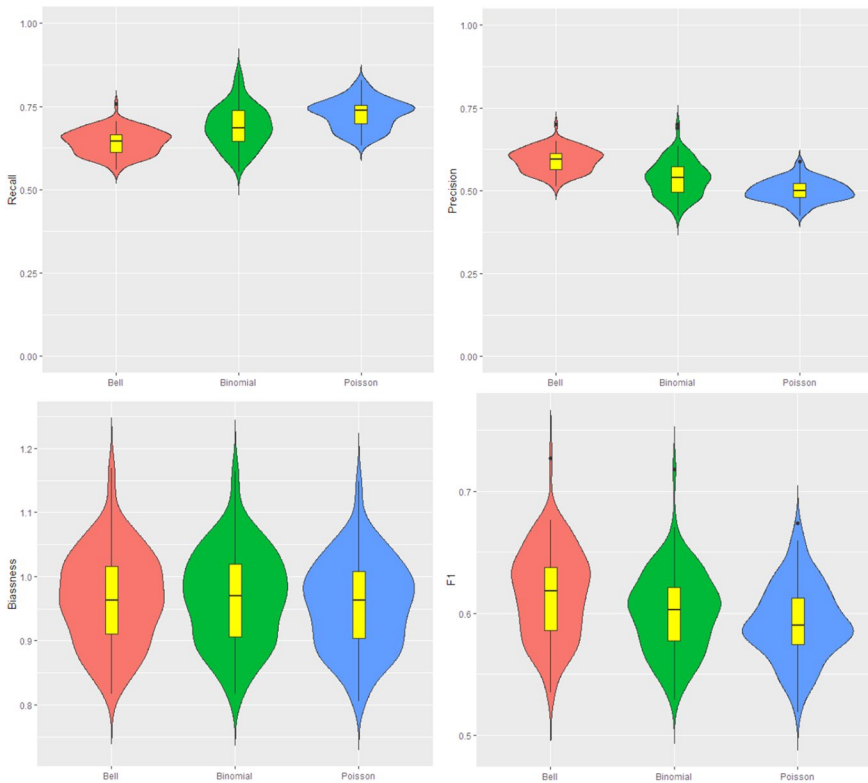


Fig. 7 Data for the map are generated from the Bell with parameter $W_0(20)$ and $W_0(10)$ respectively inside and outside cluster. From the top left to the bottom right: the violin plot for the recall, precision, bias, and F1. The number of iteration is 200. The red, green and blue colors are respectively for the Ir-Bell, Ir-binomial and Ir-Poisson scans (Color figure online)

challenge for researchers. According to Kim and Jung [38], little research has been done on the maximum scan window size or maximum reported cluster size. Wang et al. [39] stated that the maximum window size effects on the size of detected clusters. According to their paper, historical information and information about the real

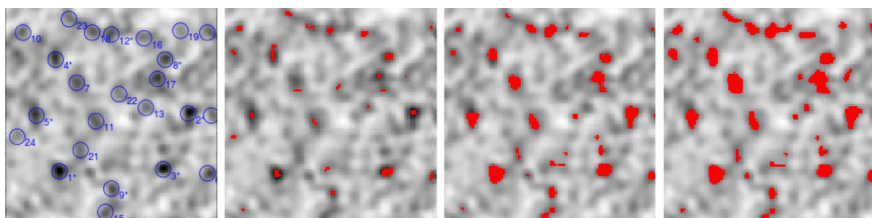


Fig. 8 From the left to the right: (1) detection of circular small nodules in a medical image by Popescu and Lewitt [37]. (2), (3) and (4): Detection of small irregular shape nodules in a medical image by the Ir-Bell scan statistic using, respectively, 1%, 5% and 10% of total image as window size

cluster can help to determine the maximum size of the scanning window. For example, a researcher may be interested in finding small clusters (such as small cancerous glands), in which case it is recommended that the maximum size of the window be considered small. But, sometimes the goal of research is to find larger clusters, such as the clusters in the Covid-19 disease, which may even involve half of the population. In such cases, the researcher can consider the size of the scanning window to be large. Therefore, a suitable window size can be determined by the researcher's prior information and experience.

Currently, the Gini coefficient and the maximum clustering set-proportion statistic (MCS-P) are the most common choices to select appropriate window size without any prior information [39]. Nevertheless, in our application we have prior information about the clusters [37], and since our goal is to provide algorithms for spatial cluster detection, we do not focus on the Gini coefficient to determine window size. Popescu and Lewitt [37] considered very small scan window size (about 0.5% of the image) such that the total detected area is about 10% of the total image because the detection of small nodules is the study objective.

The left side of Fig. 8 shows the circular scan window and the detected clusters. Our goal is to scan this image to find irregularly shaped clusters and compare the results with the result of Popescu and Lewitt [37]. We use the proposed Ir-Bell, Ir-Binomial, and Ir-Poisson scan algorithms and consider the maximum size window varying from 1%, 5%, and 10% of the image. As previously mentioned, the location and size of the correct clusters are unknown in real data. On the other hand, the maximum scanning window size affects clustering results. Based on the results of Popescu and Lewitt [37], we choose equal window sizes. These choices have the following advantages: scanning by 1% determines the center of the cluster, in other words, where a nodule starts to grow. Scanning by 5% and 10% helps us to see whether increasing the window size has a significant effect on cluster detection or not. The significant difference can be examined through eye comparisons. As we can see, there is no significant difference between the latter two results. It is worth noting that the method of Popescu and Lewitt [37] has at least two disadvantages: first, it detects clusters in a circular shape, and second, as can be seen in Fig. 8 (left side),

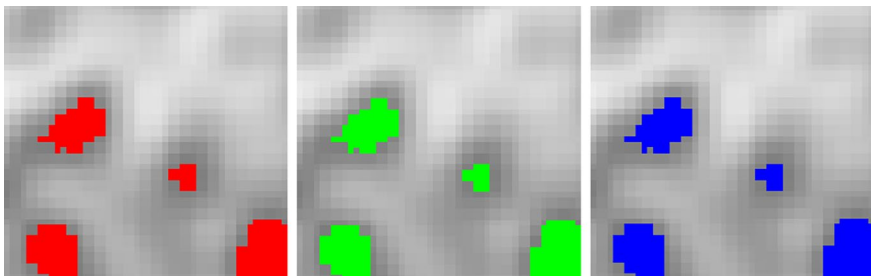


Fig. 9 From the left to the right: (1) Irregular shape clusters detected by the Ir-Bell model based on Monte-Carlo. (2) Irregular shape clusters detected by the Ir-Binomial model based on Monte-Carlo and without it. (3) Irregular shape clusters detected by the Ir-Poisson model with and without Monte-Carlo. The explored area is 50×50 pixels on the top left of medical image

the radius of blue circles are equal. It is expected that the size of cancerous glands in the body does not have such restriction, as can be seen in the other windows of Fig. 8.

The results of cluster detection in the medical image are in Fig. 8 for Ir-Bell scan. The regions of detected clusters by the Ir-Bell scan algorithm are very similar to the locations of the clusters detected by Popescu and Lewitt [37]. However, our algorithm is capable of providing more details in the shape of the clusters, avoiding over or under detection of the clusters areas. It should be noted that by applying Ir-Binomial and Ir-Poisson scans, similar results are obtained. In Fig. 9, a part of the image is selected and magnified to see the same performance of Ir-Bell, Ir-Poisson, and Ir-Binomial more clearly. They perform equally for different scanning methods.

5.2 Execution Time for New Algorithms

According to Zhou et al. [25], 20, 000 random data sets are needed to find irregular shape clusters based on Monte-Carlo hypothesis testing. The scan process for this number of data sets is troublesome when the map is big. Hence the elimination of Monte-Carlo from irregular shape cluster detection can decrease detection time making the methodology better prepared for the real-life challenges of nowadays.

Algorithms 1 and 2 (Ir-Poisson and Ir-binomial respectively) in our paper are independent of the Monte-Carlo method. Therefore, we can compare the execution time to detect irregular shape clusters with and without Monte-Carlo hypothesis testing.

To this aim, we select just 50×50 pixels in the top left of the medical image. This partial area is selected because the traditional algorithms require 20, 000 iterations of Monte-Carlo and this study would be time-consuming for the whole image with 42, 025 pixels. The detected clusters by the different methods are shown in Fig. 9. As expected, all models return the same clusters for the fast and slow versions. Unlike Popescu and Lewitt [37], we do not have access to a cluster of high computational performance. Instead, we used our R [40] coding in a desktop computer core i5 with 4Gb of RAM and Windows 7. Under such a configuration, the Ir-Bell method takes about 8 h to scan the entire image.

The execution times and p values are in Table 1. First, it is important to emphasize that the p values returned by Monte-Carlo and theoretical are the same. Also, this table reveals the advantage of the elimination of the Monte-Carlo procedure in decreasing detection time which is decreased by an order of 50%. All of our codes

Table 1 Execution time in minutes for detecting spatial clusters with and without the Monte-Carlo procedure

	Bell	Binomial	Poisson
With Monte-Carlo	30.41	43.28	36.62
Without Monte-Carlo	–	17.44	16.01
p -value	< 0.0002	< 0.0002	< 0.0002

are performed in R and more improvement can be done in the execution time if a better implementation of the methods is explored.

6 Conclusion

In this paper, we introduce new approaches to handle big maps in spatial clustering problems. To do this, three scan statistics are presented: Ir-Poisson scan, Ir-binomial scan (Sect. 3.1), and Ir-Bell scan (Sect. 3.2).

By our simulation studies, we show that the Ir-Bell scan statistic outperforms the traditional Poisson and binomial scan statistics in cluster detection when it is the true distribution.

We apply our methods to a medical image. The results verify the results of Popescu and Lewitt [37], however, provide more insights and richness in terms of interpretation, since the shape of the detect cluster are more precise. Moreover, using our naive R implementation, we show that, with the same results, the fast scan versions of the Ir-Poisson and Ir-binomial (Algorithms 1 and 2) perform at least two times faster than the traditional ones that rely on Monte-Carlo simulation.

Finally, as future work, we are interested in studying and extending these irregular shape cluster detection to their zero-inflated fast versions. Nowadays it is common to have data sets that are zero-inflated. Thus, zero-inflated methods have become relevant to provide more realistic, precise, and adequate analysis for the data.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s12561-022-09353-7>.

Acknowledgements Marcos O. Prates would like to acknowledge partial financial support from CNPq Grants 436948/2018-4, PQ-307457/2018-4 and PQ-309186/2021-8.

References

1. Adelberger KL, Steidel CC, Pettini M, Shapley AE, Reddy NA, Erb DK (2005) The spatial clustering of star-forming galaxies at redshifts $1.4 \lesssim z \lesssim 3.5$. *Astrophys J* 619(2):697
2. Mo H, White SD (1996) An analytic model for the spatial clustering of dark matter haloes. *Mon Not R Astron Soc* 282(2):347–361
3. Haralick R, Dinstein I (1975) A spatial clustering procedure for multi-image data. *IEEE Trans Circuits Syst* 22(5):440–450
4. Zhang L, Zhu Z (2012) Spatial multiresolution cluster detection method. [arXiv:1205.2106](https://arxiv.org/abs/1205.2106)
5. Han J (2001) Spatial clustering methods in data mining: a survey. *Geographic data mining and knowledge discovery*, pp 188–217
6. Harries KD et al (1999) Mapping crime: principle and practice. Technical report, US Department of Justice, Office of Justice Programs, National Institute of Justice
7. Murray AT, Grubestic TH, Wei R (2014) Spatially significant cluster detection. *Spat Stat* 10:103–116
8. Myers N, Mittermeier RA, Mittermeier CG, Da Fonseca GA, Kent J (2000) Biodiversity hotspots for conservation priorities. *Nature* 403(6772):853

9. Stohlgren TJ, Binkley D, Chong GW, Kalkhan MA, Schell LD, Bull KA, Otsuki Y, Newman G, Bashkin M, Son Y (1999) Exotic plant species invade hot spots of native plant diversity. *Ecol Monogr* 69(1):25–46
10. Grubestic TH (2006) On the application of fuzzy clustering for crime hot spot detection. *J Quant Criminol* 22(1):77
11. Yamada I, Rogerson P (2008) Statistical detection and surveillance of geographic clusters. Chapman and Hall/CRC, Boca Raton
12. Haralick R, Kelly G (1969) Pattern recognition with measurement space and spatial clustering for multiple images. *Proc IEEE* 57(4):654–665
13. Gutteridge A, Bartlett GJ, Thornton JM (2003) Using a neural network and spatial clustering to predict the location of active sites in enzymes. *J Mol Biol* 330(4):719–734
14. Bar-Hen A, Emily M, Picard N (2015) Spatial cluster detection using nearest neighbor distance. *Spat stat* 14:400–411
15. Culvenor DS, Coops N, Preston R, Tolhurst KG (1998) A spatial clustering approach to automated tree crown delineation. In: Proceedings of the international forum on automated interpretation of high spatial resolution digital imagery for forestry, pp 67–80
16. Duczmal LH, Moreira GJ, Burgarelli D, Takahashi RH, Magalhães FC, Bodevan EC (2011) Voronoi distance based prospective space-time scans for point data sets: a dengue fever cluster analysis in a southeast Brazilian town. *Int J Health Geogr* 10(1):29
17. Wieland SC, Brownstein JS, Berger B, Mandl KD (2007) Density-equalizing euclidean minimum spanning trees for the detection of all disease cluster shapes. *Proc Natl Acad Sci* 104(22):9404–9409
18. Kulldorff M, Nagarwalla N (1995) Spatial disease clusters: detection and inference. *Stat Med* 14(8):799–810
19. Kulldorff M (1997) A spatial scan statistic. *Commun Stat Theory Methods* 26(6):1481–1496
20. Caçado AL, da Silva CQ, da Silva MF (2014) A spatial scan statistic for zero-inflated poisson process. *Environ Ecol Stat* 21(4):627–650
21. Abolhassani A, Prates MO, Castellares F, Mahmoodi S (2020) Zero-inflated bell scan: a more flexible spatial scan statistic. *Spat Stat* 36:100433
22. de Lima MS, Duczmal LH, Neto JC, Pinto LP (2015) Spatial scan statistics for models with overdispersion and inflated zeros. *Stat Sin* 25:225–241
23. Assunção R, Costa M, Tavares A, Ferreira S (2006) Fast detection of arbitrarily shaped disease clusters. *Stat Med* 25(5):723–742
24. Costa MA, Assunção RM, Kulldorff M (2012) Constrained spanning tree algorithms for irregularly-shaped spatial clustering. *Comput Stat Data Anal* 56(6):1771–1783
25. Zhou R, Shu L, Su Y (2015) An adaptive minimum spanning tree test for detecting irregularly-shaped spatial clusters. *Comput Stat Data Anal* 89:134–146
26. Yin P, Mu L (2018) A hybrid method for fast detection of spatial disease clusters in irregular shapes. *GeoJournal* 83(4):693–705
27. Abolhassani A, Prates MO (2021) An up-to-date review of scan statistics. *Stat Surv* 15:111–153
28. Castellares F, Prates MO, Abolhassani A (2019) Comments on “a spatial scan statistic for compound poisson data”. *Stat Med* 38(7):1297–1299
29. Kulldorff M, Huang L, Pickle L, Duczmal L (2006) An elliptic spatial scan statistic. *Stat Med* 25(22):3929–3943
30. Neill DB (2012) Fast subset scan for spatial pattern detection. *J R Stat Soc Ser B* 74(2):337–360
31. Aboukhamseen S, Soltani A, Najafi M (2016) Modelling cluster detection in spatial scan statistics: formation of a spatial poisson scanning window and an adhd case study. *Stat Probab Lett* 111:26–31
32. Soltani A, Aboukhamseen S (2015) An alternative cluster detection test in spatial scan statistics. *Commun Stat Theory Methods* 44(8):1592–1601
33. Prim RC (1957) Shortest connection networks and some generalizations. *Bell Syst Tech J* 36(6):1389–1401
34. Bell ET (1934) Exponential numbers. *Am Math Mon* 41(7):411–419
35. Bell ET (1934) Exponential polynomials. *Ann Math* 35(2):258–277
36. Prates MO, Kulldorff M, Assunção RM (2014) Relative risk estimates from spatial and space-time scan statistics: are they biased? *Stat Med* 33(15):2634–2644
37. Popescu LM, Lewitt RM (2006) Small nodule detectability evaluation using a generalized scan-statistic model. *Phys Med Biol* 51(23):6225
38. Kim S, Jung I (2017) Optimizing the maximum reported cluster size in the spatial scan statistic for ordinal data. *PLoS ONE* 12(7):e0182234

39. Wang W, Zhang T, Yin F, Xiao X, Chen S, Zhang X, Li X, Ma Y (2020) Using the maximum clustering heterogeneous set-proportion to select the maximum window size for the spatial scan statistic. *Sci Rep* 10(1):1–14
40. R Core Team (2019) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.