Statistics Surveys
Vol. 15 (2021) 111–153
ISSN: 1935-7516
https://doi.org/10.1214/21-SS132

An up-to-date review of scan statistics[*]

Ali Abolhassani

Department of Mathematics, Azarbaijan Shahid Madani University, Tabriz, Iran
e-mail: ali.abolhassani@azaruniv.ac.ir

and

Marcos O. Prates

Department of Statistics, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
e-mail: marcosop@est.ufmg.br

Abstract: Scan statistics have been a very important and active area of statistical research in the past three decades. Detecting areas with a significant concentration of points is an important task in understanding the underlying phenomena in many fields such as: epidemiology, politics, crime analysis, zoology, etc. This study reviews how scan statistics have developed in the last three decades, the main concerns of researchers in scan statistics, and how researchers have approached these concerns.

Keywords and phrases: Literature review, scan statistics, spatial statistics, spatio-temporal statistics.

Received January 2021.

1. Introduction

Do points distribute in a region (in time, space or space-time) completely at random? This question about the dispersion of points in time, space or space-time frequently appears in astronomy (Mo and White, 1996; Adelberger et al., 2005; Gladders and Yee, 2000; Kim et al., 2002), image analysis (Haralick and Dinstein, 1975; Zhang and Zhu, 2012), data mining (Han, Kamber and Tung), criminology (Murray, Grubesic and Wei, 2014; Sherman and Weisburd, 1995; Harries, 1999; Eck et al., 2005), ecology (Stohlgren et al., 1999; Myers et al., 2000), geography (Ord and Getis, 1995; Anselin, 1995; Murray and Estivill-Castro, 1998; Grubesic, 2006; Yamada and Rogerson, 2008), pattern recognition (Haralick and Kelly, 1969), biology (Gutteridge, Bartlett and Thornton, 2003), forestry (Culvenor et al., 1998; Bar-Hen, Emily and Picard, 2015; Lee et al., 2017), epidemiology (Kulldorff, 1997; Duczmal et al., 2011; Wieland et al., 2007; Xu and Gangnon, 2016; Gangnon and Clayton, 2004; Lee, Gangnon and Zhu, 2017; Lee et al., 2021; Gangnon and Clayton, 2000, 2003; Yan and Clayton, 2006; Gangnon and Clayton, 2007; Gangnon, 2010a), climate-health (Lee, Sun and Chang, 2020), and others. The answer to the question depends on the positions of the observed points. The presence of "extra" individuals or

[*]Marcos O. Prates acknowledges partial funding support from CNPq grants (436948/2018-4 and PQ-307457/2018-4), and FAPEMIG grant (PPM-00532-16).

objects in some part of space is defined as one or more “clusters”. Cluster detection is very important in daily life. For example, in epidemiological problems quick detection of spatial clusters of cases can alert the government to take actions in hot spot areas before occurring an outbreak related to a contagious disease.

Scan statistics are the most important tool for detecting clusters. A scan statistic has the objective of detecting and evaluating the statistical significance of clusters that cannot be explained by the assumption of randomness. This is done by moving a window over the study region and identifying a region, if there is one, with a higher concentration of points than should occur by chance. Detection of such regions is traditionally performed by maximizing a likelihood ratio as will be shown.

Researchers encounter at least nine types of datasets where they might wish to identify spatial clusters using scan statistics: 1) point data; 2) case-control data; 3) aggregated data; 4) spatio-temporal data; 5) spatial survival data; 6) event data; 7) multinomial data; 8) ordinal data; and 9) time series data.

When Choynowski (1959) studied brain tumors, he considered statistical assumptions for a map and explored whether there was any city on the map with a statistically significant cluster. Since the occurrence of brain tumors is rare, he used Poisson distributions to calculate the probability of the number of sick people in each city. However, at that time, he tested the cities separately, so multiple testing problems occurred. Moreover, in his method, the border of clusters must be the same as the border of the cities. To deal with these drawbacks, Openshaw et al. (1987) proposed a graphical method called Geographical Analysis Machine (GAM). In their method, they first considered a grid on the map and then put circles around the grid points such that their radii increased step by step. Minimum and maximum radii were pre-selected. By Monte Carlo methods, one could determine which circle contained significantly more sick people in comparison to those outside the circle. Any significant circle was plotted on the map. In the end, the algorithm provided a map with some circles on it. These areas were labelled as spatial clusters. The GAM method raises at least four concerns: 1) a large number of circles must be checked to find clusters (the size of the candidate class is too large); 2) it only detects circular clusters, but in practice, clusters can have irregular shapes; 3) Monte Carlo hypothesis testing is time-consuming; and 4) due to multiple testing problems, GAM indicates some areas as clusters even if the points are distributed completely at random. Hence, researchers should choose a very small α, for example 0.001.

The above-mentioned methods were presented in previous works to detect spatial clusters. In the next section, we present the well-known spatial scan statistics. Section 2.2 is devoted to regularly and irregularly shaped spatial clusters, while Section 2.3 contains an overview of Bayesian scan statistics. In Section 3, the method of spatial cluster detection without Monte Carlo hypothesis testing is explored. Section 4 covers spatial clustering for event data. In Section 5, spatial scan statistics for general graphs are presented. Section 6 describes spatial scan statistics for continuous data. In Section 7, scan statistics for zero-inflated count data are discussed. Section 8 discusses nonparametric methods of spatial cluster detection. Section 9 presents spatio-temporal cluster detection methods. Section 10 is devoted to regression and spatial clustering. Multiple spatial cluster detection is discussed in Section 11. Section 12 is devoted to recent developments of scan statistics. In Section 13 we summarize some programs and packages for detecting spatial clusters. Finally, a conclusion is presented in Section 14.

2. Spatial Scan Statistics

2.1. Regularly Shaped Clusters

Kulldorff and Nagarwalla (1995) introduced a likelihood ratio test (LRT) to find spatial clusters. Their method uses "variable window scan statistics" and is sometimes called "moving window analysis" in engineering. In the scan statistics method, the authors called sick people "points" or "cases". The others individuals are called "controls". Under the alternative hypothesis, existence of a spatial cluster, it is assumed that the probabilities for a point (individual) in sub-region z or z^c to be a case are respectively p and q with $p > q$ where z^c is the complementary sub-region of z in the map, whereas under the null hypothesis all individuals have the same probability p of being a case. Thus, clustering detection is performed by testing:

$$H_0 : p = q, \quad \forall z \in \mathcal{Z} \qquad \text{vs} \qquad H_1 : p > q, \quad \exists z \in \mathcal{Z}. \tag{2.1}$$

Kulldorff and Nagarwalla (1995) considered only circular sub-regions as candidates for z, i.e., the candidate class, $\mathcal{Z}$, is the set of all circles with predefined maximum radius. Under the binomial or Poisson model, they calculated the likelihood of the study region under H_0 and H_1. Therefore, they were able to use the LRT to perform the hypothesis testing in (2.1). In particular, they calculated $\lambda(z) = \frac{L(z)}{L_0}$, for each z in $\mathcal{Z}$, where $L(z)$ and L_0 are respectively the likelihood of the data under H_1 and H_0, and determined the sub-region z which maximizes $\lambda(z)$. This z is called the most likely cluster (MLC) and $\lambda_R = \max_z \lambda(z)$ was defined as the scan statistic. Since the distribution of $\lambda(z)$ is unknown, they used Monte Carlo methods to decide whether the MLC is a significant cluster or not. Thus, to calculate the p-value, one can use the following steps. 1) simulate the number of cases in each cell using multinomial distribution under the null hypothesis, 2) calculate the scan statistic for the simulated map, 3) repeat steps 1 and 2, 999 times, and 4) calculate the p-value as $\dfrac{[\sum_{i=1}^{999} I(\lambda_i > \lambda_R)] + 1}{1000}$ where $I()$ is the indicator function.

This scan statistic method has some drawbacks: 1) it detects just one circular cluster but neither multiple nor irregular ones, 2) if the true cluster is irregularly shaped, circular scan statistics will detect either bigger clusters (overestimation) or smaller clusters (underestimation), and 3) applying Monte Carlo is time-consuming. Later, Kulldorff et al. (2006) introduced an elliptical spatial scan statistic that allows the detection of non-circular clusters. Although more flexible, this scan statistic cannot detect arbitrary irregular clusters either.

It is important to mention that scan statistics can be extended to deal with different kinds of data, for example event data (Rosychuk, Huston and Prasad, 2006), ordinal data (Jung, Kulldorff and Klassen, 2007), continuous data (Kulldorff, Huang and Konty, 2009; Zhang, Zhang and Lin, 2012), multivariate data (Cucala et al., 2017), and others.

Zhang and Lin (2014) noticed that the likelihood ratio statistic is a special case in the family of power divergence (PD) goodness-of-fit statistics, so the classical spatial scan test can be extended to the family of PD spatial scan tests. Their method is convenient because it can be combined with generalized linear models (GLMs). Besides this approach, the Wald-based spatial statistic (Liu, Liu and Zhang, 2018) is another alternative that can be combined with GLMs to detect spatial clusters.

The main advantage of scan statistics is specificity. In this context, specificity means identifying a single cluster responsible for the rejection of the null hypothesis (Gangnon and Clayton, 2004). However, scan statistics can be biased towards finding clusters in areas with greater spatial resolution. In other words, they tend to cherry-pick clusters in areas with a large number of geographically small cells (Gangnon and Clayton, 2004).

The weighted average likelihood ratio (WALR) test (Gangnon and Clayton, 2001) is an alternative method to detect spatial clusters instead of the traditional maximum likelihood approach. The test has a natural interpretation as the marginal likelihood ratio of a one cluster model to the no clustering model and with choice of the correct weights, it identifies clusters with higher power and less bias. The test is defined as unbiased, if, under the null hypothesis, each cell in the study region has an equal chance of belonging to the detected cluster. Furthermore, Gangnon and Clayton (2004) proposed two other scan statistics: a scan statistic based on a penalized likelihood ratio and a localized version of the WALR test. In these methods, the authors took advantage of the specificity of the scan statistic and the unbiasedness of the WALR test.

Based on the studies of Gangnon and Clayton (2001) and Gangnon and Clayton (2004), the spatial scan statistics method has high power in areas with fine geographic resolution and low power in areas with coarse geographic resolution. According to Gangnon (2010a), in real applications, there are many overlapping clusters in urban areas, while rural areas have few. The spatial scan statistic does not account for these local variations in the multiplicity problem. Hence, Gangnon (2010a) proposed two new spatially varying multiplicity adjustments for spatial cluster detection, one based on a nested Bonferroni adjustment and one based on local averaging. In fact, they proposed the local average likelihood ratio scan (LALRS) statistic and an unweighted version of the WALRS statistic, which are applicable in any setting.

Up to now, there are very few methodologies that quantify the uncertainty of a detected cluster. Along these lines, Lee et al. (2017) develop a new method for the quantification and visualization of uncertainty associated with a detected cluster.

2.2. Irregularly Shaped Spatial Clusters

2.2.1. Minimum Spanning Trees in Spatial Cluster Detection

To find an irregularly shaped cluster, Assunção et al. (2006) proposed the dynamic minimum spanning tree (dMST) method. This method can quickly find irregularly shaped clusters. A spanning tree is a sub-graph of a connected graph. It is a tree that contains all vertices of the graph. The minimum spanning tree (MST) for a weighted graph is a spanning tree that has minimum weight. The dMST method not only detects irregularly shaped clusters but also decreases the cardinality of the candidate class i.e., $\mathcal{Z}$, from many circles to just I candidates where I is the number of cells in the study region.

In this method, centroids v_i and v_j of cells i and j are connected to each other using an edge if these two cells are neighbors. Hence, corresponding to a map, we have an undirected graph. Under the Poisson assumption, the Kullback-Leibler divergence is calculated and used as weight $w(i, j)$ that is allocated to edge (v_i, v_j). The weights reflect the dissimilarity in the density of the number of "points" between two cells. Heavy weight means large dissimilarity between densities of the cells. Using the algorithm of Prim (1957), the MST can be found. The elimination of an edge from the MST separates it into two subtrees. Assunção et al. (2006) considered the smaller sub-tree as a candidate and calculated the likelihood ratio for this candidate using the LRT of Kulldorff and Nagarwalla (1995). Then, they returned the eliminated edge to its place and removed another edge. Again, the MST is separated into two sub-trees. The smaller sub-tree is considered the second candidate and the LRT is calculated for this sub-tree. This procedure goes on for all sub-graphs to find the MLC. Then, using the Monte Carlo procedure, one decides about the significance of the MLC as a spatial cluster.

Although the dMST method improves the scan statistic in two aspects (the flexible shape and reduced number of total candidates), it still has some deficiencies: 1) it detects just one irregularly shaped cluster; 2) it commonly detects a cluster bigger than the true cluster (overestimation or octopus effect), and 3) the Monte Carlo procedure is time-consuming.

As mentioned, one of the disadvantages of the dMST is the overestimation or octopus effect of the detected cluster. To control the overestimation, Costa, Assunção and Kulldorff (2012) proposed three spatial scan statistics to find irregularly shaped clusters. Their methods impose ad-hoc constraints on the cluster shape using three criteria: edMST (early stopping dMST), double connection, and mlink (maximum linkage).

The main difference between the edMST and the unrestricted dMST methods is that the edMST approach stops when the neighbors of a candidate do not increase the likelihood function. In the double connected spatial scan statistic, to be considered as a valid area for inclusion in the cluster, the candidate neighbor must be connected to at least two units of the current candidate and is aggregated to it if it increases the LRT. There is an exception in the first step when the current candidate has only one unit. If no neighbor satisfies the connection condition, then the algorithm stops and starts again from another unit. The results obtained by this method are more compact clusters than with the edMST method. Finally, the mlink spatial scan statistic is an alternative that searches only among neighbors with maximum connections to the current candidate and evaluates whether or not it must be included. For all these methods, Monte Carlo simulation is applied to determine the significance.

Recently, Zhou, Shu and Su (2015) presented another alternative named Adaptive minimum spanning tree (AMST), which is detailed in the next subsection.

2.2.2. Adaptive Minimum Spanning Trees in Detection of Irregularly Shaped Spatial Clusters

The adaptive minimum spanning tree (AMST) method for cluster detection was introduced by Goura, Rao and Reddy (2011). In this method, a validity index is introduced. It helps researchers to find the best partition of a graph. This index is defined as:

$$\mathrm{val}_{\mathrm{index}} = \frac{\mathrm{Intra}_{\mathrm{dist}}}{\mathrm{Inter}_{\mathrm{dist}}}$$

The numerator measures the distance inside partitions of the graph and the denominator measures the distance between partitions. According to Jain and Dubes (1988), this criterion is a measure to reflect the graph separation. Formally,

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

$$\mathrm{Inter}_{\mathrm{dist}} = \max |\lambda_{C^i} - \lambda_{C^j}|^2.$$

Note that λ_{C^i} is the rate of the points or the cases in the sub-partition C^i (i.e., the ratio between the total number of cases in sub-partition C^i and its population) and $\lambda_{C^{ij}}$ is the point rate of cell j in the sub-partition C^i (i.e., the ratio among the number of cases in cell j and the population of sub-partition C^i). These parameters can be estimated using the maximum likelihood method where K is the number of sub-graphs (sub-partitions) in the best partition of the original graph. According to Zhou, Shu and Su (2015), the minimum value for the validity index corresponds to the best partition of the graph. Thus, Algorithm 1 describes how to use the AMST to find spatial clusters.

Algorithm 1 Adaptive Minimum Spanning Tree Algorithm

```
1: find a MST using the Prim algorithm,
2: order the edge weights of the MST,
3: the second lightest weight is considered as a threshold. Remove this edge and all edges
   with heavier weight (in this step only two centroids are connected, and they are those
   with the lightest edge of the graph),
4: compute the validity index for the obtained partition,
5: add the next lightest edge to the previous partition and go back to step 4 until getting
   the initial MST.
```

According to Zhou, Shu and Su (2015), a partition with minimum validity index corresponds to the best partition of the MST, because the smallest validity index corresponds to the partition in which sub-graphs are more compact and have larger separation (gap) between sub-graphs.

Algorithm 1 was designed by Zhou, Shu and Su (2015) to control the overestimation problem presented in Assunção et al. (2006). Also, using the linear time subset scan (LTSS) property (Neill, 2012), the cardinality of $\mathcal{Z}$ (Zhou, Shu and Su, 2015) is reduced drastically.

Neill (2012) proved that the scan statistic and some of its extensions satisfy the LTSS condition. This property makes it possible to find an exact and efficient optimization over the subsets. LTSS can be used in problems related to scan statistics. Instead of considering all scan windows to find the MLC, one just needs to consider ordered windows. Suppose K is the number of sub-partitions in the best partition of the MST. Also, suppose that $PF(i)$ is a priority function of sub-partition $i = 1, \ldots, K$ (for example, the incidence rate of the ith sub-partition). It is necessary to calculate $PF(\cdot)$ for all sub-partitions. Let $o_{(j)}$ be the index of a sub-partition with j-th greatest priority function $PF(\cdot)$ such that $o_{(j)} \in \{1, \ldots, K\}$. Consider subsets $o_i = \{o_{(1)}, \ldots, o_{(i)}\}$ for $i = 1, \ldots, K$. LTSS proves that, instead of checking all 2^K candidates to find the MLC, one can just consider K subsets i.e., $\{o_1, \ldots, o_K\}$.

Similar to the other spatial cluster detection methods, after finding the MLC between ordered windows, Monte Carlo helps researchers to do hypothesis testing in cluster detection. This method is faster than the MST method and tackles the overestimation problem. Therefore, Zhou, Shu and Su (2015) improved the scan method to find irregularly shaped clusters in two ways: 1) by using AMST they obtained a solution for the overestimation problem of Assunção et al. (2006); and 2) by the LTSS property, they drastically decreased the cardinality of candidate classes. Furthermore, Yin and Mu (2018), by combining the MST and restricted likelihood ratio method (Tango, 2008), presented a hybrid method to detect irregularly shaped clusters more quickly.

It is important to mention the adaptive procedure proposed by Zhang and Zhu (2012) for irregularly shaped clusters in space. Their method, named spatial multiresolution cluster detection (MCD), is more effective in the detection of irregularly shaped clusters without the requirement of heavy computation. Hence, it is suitable for cluster detection for large spatial datasets, for example, images and fMRI data. Another interesting work is the particle swarm optimization method to optimize the scanning window for detecting irregular spatial clusters (Izakian and Pedrycz, 2012). As a side note, for data that can be hierarchically represented in trees, Prates, Assunção and Costa (2012) proposed a flexible scan statistic test to detect clusters using minimum description length penalization, which helps to prevent the detection of oddly shaped clusters.

2.3. Bayesian Scan Statistics

Traditionally, the scan statistics method is based on hypothesis testing and does not produce useful estimates of disease rates or cluster risks. Gangnon and Clayton (2000) developed a Bayesian inferential procedure to fit specific models for spatial clustering using cell count data. Their strategy incorporates ideas from image analysis, Bayesian model averaging, and model selection. They proposed a model for clustering in which the study region is divided into several components: a large background area and a relatively small number of clusters. A common rate (or covariate-adjusted risk) within each component is assumed. As an advantage, with their method, it is possible to obtain estimates for the disease rates.

Later, Gangnon and Clayton (2003) proposed a hierarchical model for spatially clustered disease rates. They developed a Bayesian approach capable of estimating the parameters of the hierarchical spatial clustering model. To be able to perform inference in the model, they implemented and used a reversible jump Markov chain Monte Carlo (RJMCMC) algorithm (Green, 1995). Their method allows for multiple clusters and produces posterior estimates of cell-specific and cluster-specific relative risk as well as cell-specific probabilities of cluster membership. Also, posterior inference about the number of clusters in the data is possible.

Yan and Clayton (2006) extended the spatial cluster method of Gangnon and Clayton (2003) to accommodate spatio-temporal cluster data. Again, the inference is performed using the RJMCMC algorithm. Like Gangnon and Clayton (2003), their method produces important information about the cluster detection, such as the number of clusters, the probability of each cell belonging to a cluster, and cell-specific relative risks. The authors argued that the method is parsimonious relative to fitting several spatial cluster models over time and is sensitive in detecting spatio-temporal clusters. Also, they mentioned that the method is more appropriate for datasets having a cluster structure in comparison with some Gaussian Markov random field-based models (e.g., Waller et al., 1997). Yan and Clayton (2006) suggested extending their method to find irregularly shaped clusters in big maps.

Gangnon and Clayton (2007) revisited Gangnon and Clayton (2003), including both spatial clustering and non-spatial random effects. In the prior work, the number of clusters was treated as a parameter to be estimated, requiring the use of the RJMCMC algorithm for inference. As an alternative, they considered models with a fixed, but overly large, number of clusters. Using the fixed values for the number of clusters, they were able to estimate the disease risks. The Bayes factor (the ratio of the posterior odds to the prior odds) was used to identify the local clusters. Fixing the number of clusters and not estimating them provides computational advantages since Bayesian inference avoids the RJMCMC procedure. Moreover, by not using the RJMCMC, it is easier to monitor convergence of the Markov Chain.

Neill, Moore and Cooper (2005) introduced a natural Bayesian extension of scan statistics. Their method is based on a Poisson distribution with a conjugate Gamma-Poisson model. Assume that we have I cells in the map. For cell i, i.e. s_i, the number of cases is C_i and the baseline (for example, population at risk) is b_i. The goal is to find if there is any spatial region S (set of locations s_i) for which the counts of cases are significantly higher than expected, given the baselines.

In other words, assume $C_i \sim \text{Poisson}(qb_i)$, where q is the (unknown) underlying disease rate and the aim is to compare the null hypotheses H_0 of a uniform disease rate $q = q_{\text{all}}$ to the set of alternative hypothesis $H_1(S)$, which is $q = q_{\text{in}}$ for all $s_i \in S$, and $q = q_{\text{out}}$ for all $s_i \in G - S$ for some $q_{\text{in}} > q_{\text{out}}$. A hierarchical Bayesian model where the disease rates q_{in}, q_{out}, and q_{all} are themselves drawn from Gamma distributions is assumed. Thus, under the null hypothesis H_0, $q = q_{\text{all}}$ for all $s_i \in G$, where $q_{\text{all}} \sim Ga(\alpha_{\text{all}}, \beta_{\text{all}})$. Under the alternative hypothesis $H_1(S)$, $q = q_{\text{in}}$ for all $s_i \in S$ and $q = q_{\text{out}}$ for all $s_i \in G - S$, where q_{in} is drawn from $Ga(\alpha_{\text{in}}, \beta_{\text{in}})$ and $q_{\text{out}} \sim Ga(\alpha_{\text{out}}, \beta_{\text{out}})$. It is also necessary to choose the priors for α and β, which is the most challenging task in any Bayesian analysis. To set appropriate priors, the authors assume that they have access to a large historical data set, and set α and β by matching the moments of each Gamma distribution to these historical values; for more detail see Neill, Moore and Cooper (2005). On the other hand, computation of the posterior probabilities $P(H_1(S)|D)$ of an outbreak in each region S, and the probability $P(H_0|D)$ that no outbreak has occurred, given dataset D, can be achieved by:

$$P(H_0|D) = \frac{P(D|H_0)P(H_0)}{P(D)}$$

and

$$P(H_1(S)|D) = \frac{P(D|H_1(S))P(H_1(S))}{P(D)}$$

where $P(D) = P(D|H_0)P(H_0) + \sum_S P(D|H_1(S))P(H_1(S))$. The prior choice of $P(H_0)$ and $P(H_1(S))$ and the calculation of $P(D|H_1(S))$ and $P(D|H_1(S))$ are discussed in detail by Neill, Moore and Cooper (2005). Therefore, one can return all regions with their posterior probability of belonging to the cluster and the overall probability of an outbreak.

This Poisson-Gamma Bayesian scan was extended to a multivariate version, coined the multivariate Bayesian scan statistic (MBSS) by Neill and Cooper (2010). As shown by the authors, the MBSS presents advantages over previous event detection approaches, including improved accuracy of detection, easy interpretation and visualization of results, and the ability to model and accurately differentiate between multiple event types. Next, Neill (2011) extended the MBSS to detect irregularly shaped clusters in multivariate data. Cançado, Fernandes and da Silva (2017) introduced a Bayesian zero-inflated beta-binomial scan statistic to handle zero-inflated data (for more details see Section 7.3). Overall, Bayesian scan statistics have advantages compared to frequentest approaches: 1) they usually have higher detection power; and 2) are faster (since there is no need to perform randomization testing).

3. Spatial Cluster Detection without Monte Carlo Hypothesis Testing

3.1. Cluster Evaluation Permutation Procedure

Turnbull et al. (1989) introduced the cluster evaluation permutation procedure (CEPP). To detect spatial clusters, they assumed that the map has I cities (cells) with $N = \sum_{i=1}^{I} n_i$ and $X = \sum_{i=1}^{I} X_i$ representing the total population and the total cases, respectively. They built a two-dimensional window for each cell such that this window contains neighbors of that cell and the total population in that window is equal to a predefined value R. The procedure for constructing the window is as follows. For each cell $i = 1, 2, \ldots, I$, if $n_i < R$, then the closest cell (based on Euclidean distance) is added to cell i. Suppose the closest cell to cell i is cell j. If $n_i + n_j = R$, then it is said that the window is built. If $n_i + n_j < R$, then cell j is completely absorbed in the window. If $n_i + n_j > R$, just a fraction of cell j, i.e., $\frac{R-n_j}{n_j}$ will be added to build the window. This process continues until the final construction of the window. Therefore, the i-th window contains cell i and its nearest neighbors, and there are I windows such that each of them contains exactly R people. The null hypothesis, H_0, is the case that people are distributed completely at random on the map. With this assumption, Turnbull et al. (1989) used $M_R = \max(X_{1R}, \ldots, X_{IR})$ as the test statistic for hypothesis testing such that X_{iR}, $i = 1, 2, \ldots, I$ (the number of cases in i-th window) are identical but not independent random variables. With this definition they were able to find the distribution of M_R under H_0.

Therefore, the CEPP method has at least three advantages in comparison with the GAM method: 1) there is no need to use Monte Carlo methods to compute the p-value; 2) it is not necessary to test many circles (the candidate class is not big); and 3) it is not necessary to apply it to small α values. However, it depends on specifying the size of R for the window and this is highly controversial.

3.2. Method of Besag and Newell

The CEPP method (Section 3.1) has some challenges in its computational aspects. Thus, Besag and Newell (1991) presented a method that is computationally more efficient. Consider n_i, X_i and H_0 as before. Besag and Newell (1991) tried to find a statistic with a known distribution to detect spatial clusters. They constructed this statistic for hypothesis testing by answering the following question: "Starting from an arbitrary cell, what is the minimum number of cells necessary to add to this cell to achieve a predefined number of sick people?". A small value for this statistic means there is a concentration of sick people in the neighborhood of the starting cell. To formulate this idea mathematically, they selected a sick person and denoted the cell of this person as A_0. Other cells in the map are called $A_1, A_2, \ldots$ based on the Euclidean distance of their centroids to the center of A_0 (that is, the nearest cell to A_0 is A_1 and so on). They defined $D_i = \sum_{j=1}^{i} X_{j-1}$ and $U_i = \sum_{j=1}^{i} n_{j-1}$. The test statistic is defined as $M = \min\{i, D_i \geq k\}$. Since small values of M mean there is a cluster around A_0, the test's significance level is: $Pr(M \leq \text{observed value}|H_0)$. In turn, $M > m$ means there are fewer than k sick people among the population with size U_m. The probability of having fewer than k sick people among population with the size U_m is the summation of hypergeometric probabilities. To find the $Pr(M < m + 1)$, Besag and Newell (1991) used a Poisson approximation to the hypergeometric distribution. Using this probability, significant clusters were determined and plotted on the map.

In this method, no Monte Carlo procedure is needed. However, there are two criticisms: 1) Euclidean distance may not be a good criterion to order cells — other distances may be more appropriate to reflect the similarities between cells; and 2) distributions for the number of cases in a cell are not considered.

3.3. Method of Soltani and Aboukhamseen

Soltani and Aboukhamseen (2015) introduced an alternative way to find spatial clusters using scan statistics. Consider the hypothesis test in (2.1) and suppose that $G = Z_1 \bigcup Z_2 \bigcup \ldots \bigcup Z_I$ is the study region. Also assume that $X_{z,+}$ and $n_+(G)$ are respectively the number of points (cases) in zone z and the total number of points (cases) in the study region. For an individual in study region G, suppose A_z denotes that this individual is in zone z and B_+ denotes that a person is labeled as a case. Soltani and Aboukhamseen (2015) defined the count measure μ on $(G, \mathcal{F})$ such that $\mu(z)$ and $\mu(G)$ are respectively the number of individuals in z and G, and $\mathcal{F}$ is a sigma field of subsets of G. Both $\mu(z_i), i = 1, ..., I$ and $\mu(G)$ are known and the probability of the event A_z is given by $\nu(z) = \frac{\mu(z)}{\mu(G)}$. With these assumptions, they proved that the hypothesis test in (2.1) is equivalent to the hypothesis test

$$H_0 : P_{z|+} = \nu(z) \quad vs. \quad H_1 : P_{z|+} > \nu(z) \tag{3.1}$$

where $P_{z|+}$ is the probability of belonging to zone z given that the label of the individual is $+$. Also, they proved that the exact and asymptotic distribution of points in zone z under H_0 are:

$$X_{z,+} \sim Bin(n_+(G), \nu(z)),$$

and

$$X_{z,+} \xrightarrow{D} N(n_+(G)\nu(z), n_+(G)\nu(z)[1 - \nu(z)]). \tag{3.2}$$

where $\xrightarrow{D}$ means convergence in distribution. Hence, using (3.1) and (3.2), zone z is a significant spatial cluster of level α if

$$\frac{\frac{X_{z,+}}{n_+(G)} - \nu(z)}{\sqrt{\frac{\nu(z)[1-\nu(z)]}{n_+(G)}}} > z_\alpha. \tag{3.3}$$

The main advantage of this method is the elimination of the Monte Carlo procedure to detect significant spatial clusters. Like most other scanning methods, the method requires the maximum size of the scanning window. When the appropriate size of the cluster is unknown and many cluster areas are expected to happen, the scanning procedure is repeated by varying the window size. This practice induces a multiple testing problem that is not considered by the authors.

3.4. Method of Aboukhamseen

Aboukhamseen, Soltani and Najafi (2016) extended the spatial scan statistic to the situation where the total population of the study in region G is unknown. They supposed that $n_+(G)$ is a random variable with a Poisson$(\lambda(G))$ distribution where $\lambda(G)$ is unknown and that $X_{z,+}|n_+(G)$ ∼Bin$(n_+(G), \nu(z))$. They proved that $X_{z,+}, X_{+,z^c}$ are independent under null hypothesis with marginal distribution Poisson$(\lambda(G)\nu(z))$ and Poisson$(\lambda(G)(1-\nu(z)))$ respectively. Using these facts one can prove that the scan window $W = \{X_{z,+}, z \subset G\}$ follows a ν homogeneous Poisson point process with rate $\{\lambda(G)\nu(z), z \subset G\}$ on $\mathcal{Z} = \{z, z \subset G\}$. Since $n_+(G)$ has a Poisson distribution, it is possible to find a one-sided or two-sided confidence interval for $\lambda(G)$. Also, using the marginal distribution of $X_{z,+}$, it is possible to find confidence intervals for $\lambda(G)\nu(z)$ (Garwood, 1936). Dividing the above-mentioned confidence intervals, one can find the confidence interval for $\nu(z)$. By considering the hypothesis testing (3.1) and the confidence interval for $\nu(z)$, it is possible to determine the significance of the MLC.

4. Spatial Clustering for Event Data

In most classical problems of scan statistics, researchers consider case-control data. However, in some problems, the consideration of disease-related events helps to perform a more adequate analysis. For example, suppose researchers want to know which hospitals in the study region have a heavy burden. If the researchers just consider the number of cases, their analysis will be biased because it is likely there are some cases who visit the hospital once while other cases may visit more than once. Hence the second group of cases places a higher burden to the hospital than to the first group. With this in mind, Rosychuk, Huston and Prasad (2006) considered maps for event data to determine spatial clusters. They proposed a compound Poisson model to detect spatial clusters for event data. Their strategy is based on the method of Besag and Newell (1991) which does not need Monte Carlo simulation.

To find event spatial clusters, they assumed that the number of events in a study region is a random sum of random individual events and has a compound Poisson distribution. Suppose that X_{ia} is the number of individuals in cell i with exactly a events and x_{ia} is the observed value of this random variable. Let $X_i = \sum_a X_{ia}$ be the number of cases in cell i while V_i is the random variable corresponding to the total number of events in cell i. So, $X = \sum_{i=1}^I X_i$ and $V = \sum_{i=1}^I V_i$ are respectively the total number of cases and the total number of events in a study region with I cells. The test statistic is similar to the one presented by Besag and Newell (1991), i.e., to construct a scan window, they combine cells to include at least k^* events. The test statistic for cell i is the number of cells combined with cell i to construct the scan window:

$$L_i{}^* = \min\left\{q : k^* \leq \sum_{p=0}^{q} V_{i_p}\right\}.$$

Suppose that l_i is the observed value for ${L_i}^*$. A small value of l_i means that one finds more events in small windows and is a sign that those windows form a spatial cluster. Suppose that $n_{i:l}$ and $X_{i:l}$ are respectively the total population and the total number of cases for l_i nearest neighbors of cell i. Since the number of events in the window corresponding to cell i follows a Poisson distribution under the null hypothesis, the estimate of the parameter for this distribution is $\lambda_{i:l_i} = n_{i:l_i}X/n$ where n is the population size of the study region. Let Y_j be the number of events for individual j, $j = 1, 2, 3, \ldots, X_{i:l_i}$; hence; $V_{i:l_i}$, the total number of events in l_i nearest neighbor of cell i, is $V_{i:l_i} = \sum_{i=1}^{X_{j:l_i}} Y_j$. For all $a \geq 1$, let $= Q(a) = Pr(Y - j = a)$. Let $Pr_{i:l_i}(b) = Pr(V_{i:l_i} = b)$. Then, the significance level can be written as $Pr({L_i}^* < l_i) = 1 - \sum_{b=0}^{k^*-1} Pr_{i:l_i}(b)$. Using the following recursive relation (Ross, 2014)

$$\begin{aligned} Pr_{i:l_i}(0) &= \exp(-\lambda_{i:l_i}) \\ Pr_{i:l_i}(b) &= \frac{\lambda_{i:l_i}}{b} \sum_{a=1}^{b} aQ(a)Pr_{i:l_i}(b-a), \qquad b \geq 1, \end{aligned} \tag{4.1}$$

one can determine the p-value by (4.1). Notice that in practice $Q(a)$ is unknown but can be treated as $Q(a) = x_a/x$. Now it is possible to discuss the hypothesis testing to detect spatial clusters as before.

The method of Rosychuk, Huston and Prasad (2006) has some benefits: 1) it is suitable for event data, 2) it can apply to case-control data (assuming that each case only has one event), and 3) it does not need Monte Carlo simulation. However, there are serious drawbacks in the strategy of Rosychuk, Huston and Prasad (2006): 1) computing the given recursive relation is time-consuming even for a small value of b; 2) to apply this method, one needs a predefined cluster size which is not known in practice; and 3) the significance level depends on this size. Castellares, Prates and Abolhassani (2019) noticed that with the relationship between the compound Poisson distribution and the Neyman type A, there is no need to use the recursive relation (4.1) to achieve the LRT. This strategy makes the cluster detection process computationally feasible.

Furthermore, an approximation to the probability of the events using the negative binomial distribution to detect spatial clusters for events was proposed by Chang and Rosychuk (2015). Because of the combinatorial coefficients in the negative binomial probability mass function, calculation of the likelihood is time-consuming for large datasets. The proposal of Castellares, Prates and Abolhassani (2019) shows that the execution time for calculating the likelihood function is faster than the recursive formula (Rosychuk, Huston and Prasad, 2006) and the negative binomial method (Chang and Rosychuk, 2015) in spatial cluster detection.

5. Scan Statistics for General Graphs

One possible way to explain relationships between nodes in networks is to use graphs. Sometimes researchers are interested in finding which nodes have common features or play a similar role in networks. These kinds of nodes can be treated as clusters. One of the earliest studies about scan statistics in graphs is the work of Priebe et al. (2005) on Enron graphs. They applied scan statistics to detect anomalies in time series of Enron Email graphs.

Marchette (2012) extended the idea of scan statistics in graphs to detect anomalies in time series of graphs. An anomaly is defined as a small region of vertices with unusually high connectivity between themselves in comparison with other regions. Consider a graph $G = (V(G), E(G))$, where V and E are respectively vertices and edges. The idea of a graph invariant property is the main key in the method of Marchette (2012); this is a function $\psi : G \longrightarrow \mathbb{R}$ that does not depend on how the graph is presented.

Let the set of k-neighbors of vertex v, be $N_k(v)$; this is the set of all vertices u in the graph such that the minimum number of edges needed to reach u from v is less than $k+1$. Given a subset U of $V(G)$ let $\Omega(U) = (U, E(U))$ be the induced sub-graph of the original graph G. The locality statistic is defined as $\Psi_k(v) = \psi(\Omega(N_k(v)))$. The scan statistic at scale k is given by $M_k(G) = \max_{v \in V} \Psi_k(v)$.

A time aspect can also be added to a graph. Let $\{G_t\}$ be a collection of graphs with time index t. In this collection of graphs, only the edges change in time, so the vertices are fixed for all times. Adding a time index to graphs not only helps researchers find clusters by comparing sub-graphs but also makes it possible to find anomalies by comparing regions to their past history. As before the locality statistic is $\Psi_k^t(v)$ which is the cardinality of the sub-graph induced by $N_k(v)$.

Large values of

$$\Psi_k^t(v) = \frac{\psi_k^t(v) - \mu_k^t(v)}{\max\{1, \sigma_k^t(v)\}}$$

indicate that there is an anomaly. Parameters $\mu_k^t(v)$ and $\sigma_k^t(v)^2$ are respectively the mean and the variance of $\psi_k^{t-w}(v), \ldots, \psi_k^{t-1}(v)$. Hence, the maximum of $\Psi_k^t(v)$ can be used as scan statistic at time t.

To detect the clustering pattern, Wang and Phoa (2016) defined a scan statistic for three different features: 1) Structure (S); 2) Attribute (A); and 3) both Structure and Attribute (SA), in social networks. These features are defined in Zhou, Cheng and Yu (2009). To become familiar with them, consider the example of "coauthor network" of Zhou, Cheng and Yu (2009). In this network, each node represents an author and vertex connectivity shows the relationship between authors. In the structural based cluster concept, in the "coauthor network", nodes with close connectivity form a cluster (they could have different topics), but from the standpoint of attribute-based clustering, topics are considered. Thus, authors in a cluster work on the same topics.

For a network, consider a graph G as before; suppose that $k = \{k_1, \ldots, k_{|V|}\}$ are the degrees of the vertices. Let k_G be the sum of all degrees and $|E(G)|$ be the total number of edges. Based on Erdos and Rényi (1960), the expected number of edges between any two vertices v_i and v_j is $e_{ij} = k_i k_j / (2|E(G)|)$ for $i \neq j$ and $e_{ii} = k_i^2 / (4|E(G)|)$. Wang et al. (2008) introduced the following scan statistic for detecting spatial clusters based on the "structure" in graphs:

$$\lambda_S(Z) = \left(\frac{|E_z|}{\mu(z)}\right)^{|E_z|} \left(\frac{|E(G)| - |E_z|}{\mu(G) - \mu(z)}\right)^{|E(G)|-|E_z|}.$$

where $\mu(z) = \dfrac{{k_z}^2}{4|E(G)|}$ and $\mu(G) = \dfrac{{k_G}^2}{4|E(G)|}$.

Wang and Phoa (2016) considered the “attribute” network. To find spatial clusters of attributes they considered a graph G in which there is an attribute X_i associated with each vertex. Thus now $G = (V, E, X)$ with $(X = x_1, ..., x_{|V|})$. Wang and Phoa (2016) considered four possible distributions for each X_i: 1) binomial; 2) Poisson; 3) normal; or 4) multinomial.

1) binomial distribution:

$$\lambda_A(z) = n_z \ln\left(\frac{p_{11}}{p_0}\right) + (N_z - n_z)\ln\left(\frac{1-p_{11}}{1-p_0}\right) \tag{5.1}$$

$$+(n_G - n_z)\ln\left(\frac{p_{10}}{p_0}\right) + ((N_G - N_z) - (n_G - n_z))\ln\left(\frac{1-p_{10}}{1-p_0}\right)$$

where N_G and n_G are respectively population size inside G and number of population with the particular attribute under study. The ratio of people with a particular attribute inside sub-graph z (i.e., inside scanning window) is p_{10}. Also, p_{11} is the ratio of that attribute in sub-graph $\bar{z}$ (i.e., outside scanning window). Finally, p_0 is the ratio of that attribute in G.

2) Poisson distribution:

$$\lambda_A(z) = n_z \ln\left(\frac{p_{11}}{p_0}\right) + (n_G - n_z)\ln\left(\frac{p_{10}}{p_0}\right) \tag{5.2}$$

where the values for p_{11}, p_{10} and p_0 are similar to previous paragraph.

3) normal distribution:

$$\lambda_A(z) = n\ln(\sqrt{\hat{\sigma}^2}) - n\ln(\sqrt{(2\hat{\sigma}_z^2)}) \tag{5.3}$$

where n is the total number of nodes, $\hat{\sigma}^2$ is the variance of all the x_i’s and $\hat{\sigma}_z^2 = (\sum_{i\in z}(x_i - \bar{x}_z)^2 - \sum_{j\in z^c}(x_j - \bar{x}_{z^c})^2)/n$.

4) multinomial distribution:

$$\lambda_A(z) = \sum_k \left(n_{zk}\ln\left(\frac{n_{zk}}{n_z}\right) + (n_k - n_{zk})\ln\left(\frac{n_k - n_{zk}}{n - n_z}\right) - n_k\ln\left(\frac{n_k}{n}\right)\right) \tag{5.4}$$

where k is the number of categories, n is the total number of the nodes, n_z is the number of nodes in sub-graph z, n_{zk} is the total number of nodes in z whose attribute is of category k and finally, n_k is the total number of nodes in the whole graph in category k.

If “attribute” and “structure” are independent, then the scan statistic for both “structure and attribute” is $\lambda_{SA}(Z) = \lambda_S(Z) + \lambda_A(Z)$. The sub-graph z which maximizes the scan statistic is the MLC. To test the significance of the MLC from the attribute point of view, a randomized permutation procedure is applied as follows: for situations in which the underlying distribution is normal, first, assign observed values to nodes randomly and calculate the scan statistic. The process is repeated many times (e.g., 999) and the scan statistic is calculated for each repetition. To calculate the p-value, one needs to sort these 999 simulated values with the observed value of the scan statistic from the real dataset. The p-value is given by $(M+1)/1000$, where M is the number of simulated values greater than observed value of the scan statistic. In the case of the binomial and Poisson distributions, the process is similar. Instead of assigning the observed values to nodes, the values for nodes are generated under the null hypothesis using a multinomial distribution.

In the case of detecting a spatial cluster of the graph structure, randomly assigning degrees to vertices is impossible, because if the degree is randomly assigned to nodes, it is likely that the nodes and edges will not construct a valid graph (Sierksma and Hoogeveen, 1991). Hence, to obtain the significance of MLC, it is suggested to apply a probabilistic method and use the expected degree based on the random graph model (Erdos and Rényi, 1960). Let (k_i, k_j) be the degree of nodes i and j respectively. Then the expected number of edges is $e_{ij} = \frac{k_i k_j}{2|E(G)|}$ and all random graphs are generated with the same expected degrees. Therefore, it is possible to perform a Monte Carlo hypothesis testing procedure. Fortunato (2010); Woodall et al. (2017), studied cluster detection in networks in more detail.

6. Spatial Scan Statistics for Continuous Data

The spatial scan statistics method is commonly applied to count datasets. However, some researchers are interested in applying methods to find spatial clusters in the case of continuous spatial datasets where the measurements might have distributions such as, for example, normal, exponential, Weibull, etc. This section discusses how to construct spatial scan statistics for continuous spatial datasets.

6.1. Normal and Multivariate Gaussian Scan Statistics

To find spatial clusters of low weight infants in New York City, Kulldorff, Huang and Konty (2009) proposed a normal scan statistic. As before, suppose there are I cells in a map and the total population in the study region G is N individuals i.e., $N = \sum_{k=1}^{m} N_k$ where N_k is the population of cell k. Suppose each cell has one or more individuals with spatial location $i, i = 1, \ldots, I$ and $x_s = \sum_{i \in s} x_i$, where x_s is the summation of the weights in sub-region s.

Detecting spatial clusters is equivalent to performing the following hypothesis test: H_0: all observations come from the same normal distribution vs. H_1: there is at least one sub-region where the mean of observations is more (less) than outside of this sub-region. To perform this hypothesis test, Kulldorff, Huang and Konty (2009) constructed a scan statistic based on the likelihood ratio test, i.e., $\max_z(\ln L_z)/(\ln L_0)$, where L_z and L_0 are respectively the likelihoods under H_1 and H_0. The likelihood ratio depends on sub-region z via the mean of the sum of deviations from the mean inside and outside sub-region z, i.e., the likelihood ratio will be a maximum when σ_z is maximized where σ_z is the above-mentioned mean of deviations. Hence, the MLC is a sub-region z that maximizes σ_z. After finding the MLC the significance of the MLC of the spatial cluster can be obtained using randomization.

Sometimes it is interesting to consider the correlation between variables in spatial clustering problems. Cucala et al. (2017) applied a multivariate Gaussian scan statistic to find spatial clusters. This method is more powerful than its independent version.

6.2. Weighted Normal Spatial Scan Statistics

Huang et al. (2009) introduced a weighted normal scan statistic to detect spatial clusters for continuous measures. The weight variable δ_z is considered to reflect the uncertainty of the regional measures in zone z. Let δ_z be a known measure proportional to the inverse of the uncertainty in zone z. It is possible to measure different values for studied data from different sources. For example, it is possible to record different values for pollution data from different locations in a cell and report their average value as w_z. The weight δ_z can be considered as the inverse of the variance of the w_zs. In some situations it is not possible to calculate the variance of w_z so one can use population size or the number of cases as a proxy for the inverse variance.

To construct a spatial scan statistic Huang et al. (2009) first assume that for $z \in Z$, $w_z|\delta_z \sim N\left(\mu_z, \dfrac{{\sigma_G}^2}{\delta_z}\right)$ and for $z \in Z^c$, $w_z|\delta_z \sim N\left(\mu_z^c, \dfrac{{\sigma_G}^2}{\delta_{z^c}}\right)$. Suppose that μ_z is the mean of measurements inside zone z, and $\dfrac{{\sigma_G}^2}{\delta_z} = \sigma_{w_z}^2$ is the variance of w_z. It is assumed that given δ_z the w_zs are independent normal with the same mean and different variances. In view of these assumptions,

$$L(\mu_z, \mu_{z^c}, {\sigma_G}^2) = \textstyle\prod_{z\in Z} f_z(\mu_z, \sigma_{w_z}) \prod_{z\in Z^c} f_{z^c}(\mu_{z^c}, \sigma_{w_{z^c}}).$$

Hence, the MLEs for the parameters can be found, say $\hat{\mu}_z$, $\hat{\mu}_{z^c}$ and $\hat{\sigma}_G^2$. Using these, the estimate $\hat{\sigma}_{w_z}^2$ is obtained. To construct the spatial scan statistic, it is necessary to find the MLC by maximizing the likelihood ratio. Therefore, it is necessary to know for which sub-region z the log likelihood function is maximized under H_1, i.e., by maximizing $\ln L(\mu_z, \mu_{z^c}, {\sigma_G}^2)$, the MLC is determined. Huang et al. (2009) proved that maximizing the log likelihood function is equivalent to maximizing $\dfrac{(\sum_{z\in Z}\delta_z w_z)^2}{\sum_{z\in Z}\delta_z} + \dfrac{(\sum_{z\in Z^c}\delta_z w_z)^2}{\sum_{z\in Z^c}\delta_z}$ with respect to z in the candidate class. They used Monte Carlo hypothesis testing to determine the significance of the MLC.

6.3. Spatial Scan Statistics for Survival Data

To deal with continuous spatial data in spatial clustering problems Huang, Kulldorff and Gregorio (2007) presented the exponential spatial scan statistic for censored and uncensored continuous survival data. This model is suitable for finding spatial clusters of lifetimes. The assumptions are as follows: T_i is the exponentially distributed survival time of individual i, θ_{in} is the mean of T_i inside zone z, and θ_{out} is the mean of T_i outside zone z. The goal is to test if there is at least one sub-region in study region G (with total population N) such that $\theta_{\text{in}} > \theta_{\text{out}}$ or not. They assumed that for a fixed censoring time L_i, T_i is observed if and only if $T_i \leq L_i$, otherwise right censoring is present. Hence the observed time for individual i is $t_i = \min(T_i, L_i)$. To determine which lifetime observation is censored, they used an N dimensional vector γ such that i-th element is 1 when censoring happens and 0 otherwise. With this vector, one can count the number of non-censored individuals in any zone z, i.e., $r_{\text{in}} = \sum_{i \in z} \gamma_i$. Using this notation the likelihood of a zone z is:

$$L(z, \theta_{\text{in}}, \theta_{\text{out}}) = (\theta_{\text{in}})^{-r_{\text{in}}} e^{-\sum_{i\in z} t_i/\theta_{\text{in}}} (\theta_{\text{out}})^{-r_{\text{out}}} e^{-\sum_{i\in z^c} t_i/\theta_{\text{out}}}.$$

To detect a spatial cluster, as in the other methods mentioned above, the likelihood ratio test is used to find the most likely cluster. Any sub-region z which maximizes

$$\lambda(z) = \frac{\max_{z,\theta_{\text{in}} \neq \theta_{\text{out}}} L(z, \theta_{\text{in}}, \theta_{\text{out}})}{\max_{z,\theta_{\text{in}} = \theta_{\text{out}}} L(z, \theta_{\text{in}}, \theta_{\text{out}})},$$

is the MLC. In the presence of censored data, after estimating parameters, the likelihood ratio is:

$$\lambda(z) = \frac{\max_z \left(\frac{r_{\text{in}}}{\sum_{i\in z} t_i}\right)^{r_{\text{in}}} \left(\frac{r_{\text{out}}}{\sum_{i\in z^c} t_i}\right)^{r_{\text{out}}}}{\left(\frac{R}{\sum_{t\in G} t_i}\right)^{R}} I\left(\frac{\sum_{i\in z} t_i}{r_{\text{in}}} > \frac{\sum_{i\in z^c} t_i}{r_{\text{out}}}\right).$$

Since the denominator is independent of z, to find the MLC it is only necessary to determine for which candidate z the numerator is maximized. After determining the MLC, it is necessary to generate simulated data under H_0 to determine the significance of the MLC. Since the distribution of survival times is unknown, it is impossible to generate datasets under the null hypothesis. One way to handle this limitation is permutation of observed pairs $\{(t_i, \gamma_i), i = 1, \ldots, n\}$ between individuals. The position of individuals is fixed in the map since they are in the real dataset. To find the exact distribution of λ, one needs to calculate λ for $N!$ permutations, which is time-consuming even for small N. Therefore, Huang, Kulldorff and Gregorio (2007) proposed to use random selection of 9999 permutations instead of all $N!$ permutations. Finally, the MLC is significant at level α if the p-value is less than α (i.e., $RK/(1 + 9999) < \alpha$, RK is the rank of the scan statistic for the real dataset among the original and 9999 permuted datasets). Similarly, for non-censored data, the scan statistic is constructed to find spatial clusters.

Because the exponential distribution has just one parameter, it is sensitive to modest variations (Bhatt and Tiwari, 2014). Therefore, Bhatt and Tiwari (2014) proposed a more robust alternative for the exponential distribution using the Weibull distribution with two parameters. The method constructing the spatial scan statistic for this distribution is similar to the previous one.

7. Scan Statistics for Zero-inflated Count Data

7.1. Zero-inflated Poisson Scan Statistics

Excess of zeros can be observed in a given dataset. In spatial cluster detection, the extra zeros can lead to biased inferences (Gómez-Rubio and López-Quílez, 2010; Cançado, da Silva and da Silva, 2014). Cançado, da Silva and da Silva (2014) considered the zero-inflated Poisson distribution to detect spatial clusters for these types of datasets. Suppose that the number of cases in sub-region z has a zero-inflated Poisson distribution, i.e., $X_z \sim ZIP(p, n_z\theta_z)$ such that n_z and p are respectively the population size in zone z and the probability of being structurally zero for a cell. The hypothesis test

$$H_0 : \forall z \in \mathcal{Z} \quad \theta_z = \theta_0 \quad vs. \quad H_1 : \exists z \in \mathcal{Z} \quad \theta_z > \theta_0$$

is used to detect spatial clusters. Following Kulldorff (1997), the likelihood ratio test is used to perform hypothesis testing. Hence,

$$L(p, z, \theta_z, \theta_0) = \prod_{i \in z} f_i(x_i) \prod_{i \in z^c} f_j(x_j)$$

such that f_i's and f_j's are respectively the probability mass functions (pmf) of ZIP$(p, n_i\theta_z)$ and ZIP$(p, n_j\theta_0)$. Using the pmf of the zero-inflated Poisson distribution, it is impossible to find the MLE for the parameters. By applying the method of Lambert (1992), they constructed a new likelihood, based on the knowledge of the different kinds of zeros, i.e., structural or sampling zeros. Their method is as follows.

Suppose that $\delta = (\delta_1, \ldots, \delta_m)$ is a vector where δ_i indicates if the zero in cell i is structural (δ_i=1). Hence, δ_i is a binary variable, $\delta_i \sim Ber(p)$. To find the MLE of the parameters, Cançado, da Silva and da Silva (2014) worked with bivariate data (x_i, δ_i), $i = 1, \ldots, m$ and found the likelihood using the pmf of (x_i, δ_i), namely,

$$\begin{aligned} L_1(\theta_z, \theta_0, p, \delta) &= p^{\sum_{i=1}^m \delta_i}(1-p)^{m-\sum_{i=1}^m \delta_i} e^{-\theta_z \sum_{i \in z} n_i(1-\delta_i)} \theta_z^{\sum_{i \in z} x_i(1-\delta_i)} \\ &\times \; e^{-\theta_0 \sum_{i \in z^c} n_i(1-\delta_i)} \theta_0^{\sum_{i \in z^c} x_i(1-\delta_i)}. \end{aligned}$$

Therefore, one can find a closed form for the MLE of the parameters. However, the vector δ is commonly unknown and needs to be estimated. An EM algorithm is proposed to estimate δ, so the MLE of the parameters can be found. Now, one can find the sub-region z which maximizes the likelihood ratio $\lambda(z)$. This sub-region is the MLC. Since the distribution of $\max_z \lambda(z)$ is unknown, Monte Carlo simulation is used for hypothesis testing to find spatial clusters.

7.2. Zero-inflated Binomial Scan Statistics

Cançado, Fernandes and da Silva (2017) introduced a zero-inflated binomial scan statistic. Suppose that m, n_i, x_i, N, C are respectively, the total number of cells, the population and cases in cell i, the total population of the study region and the total cases indicated on the map. The aim is to find sub-region z for which the probability of being a case in z, i.e., θ_z, is greater than the probability of being a case outside z, say θ_{z^c}, when the dataset has extra zeros. Let X_z be the number of cases in zone z such that $X_z \sim Bin(n_z, \theta_z)$ where n_z is the population in zone z. To work with datasets with extra zeros, it is necessary to use a mixture model, i.e., a degenerate distribution at zero and a binomial model for nonzero counts. But using this mixture, the MLEs do not have closed-form expressions. Therefore, constructing a scan statistic using the mixed distribution is not trivial. Then, Cançado, Fernandes and da Silva (2017) proposed to use the method in Lambert (1992) to construct their scan statistic. They defined the vector $\delta = (\delta_1, \ldots, \delta_m)$ similar to the previous subsection. Hence, $\delta_i \sim Ber(p)$. To obtain the scan statistic, one needs to maximize the likelihood ratio function (Kulldorff and Nagarwalla, 1995). So, Cançado, Fernandes and da Silva (2017) calculated $L_0(\delta, \mathbf{x})$ and $L_1(\delta, \mathbf{x})$ to compute zero-inflated binomial scan statistics. i.e., $\lambda(z)$. The zero-inflated binomial scan statistic $\lambda(z)$ is a function of $\theta_z, \theta_{\bar{z}}, p, \theta_0$ and vector δ. To find the MLC one needs to calculate the MLE for these unknown parameters and δ. The MLE for the parameters is obtained easily, but to estimate δ one must rely on the EM algorithm. After estimating the unknown parameters and vector δ, the MLC is obtained by maximizing $\lambda(z)$. As in previous cases, the distribution of $\lambda(z)$ is unknown; therefore these authors also used Monte Carlo for hypothesis testing.

Additionally, de Lima et al. (2015) suggested applying a zero-inflated double Poisson model to detect spatial clusters for zero-inflated and overdispersed spatial data. In turn, Zhang et al. (2017) theoretically discussed the asymptotic properties of spatial scan statistics and considered overdispersion of lung cancer in Texas.

7.3. Bayesian Beta-binomial Scan Statistics

Cançado, Fernandes and da Silva (2017) proposed the Bayesian zero-inflated binomial scan statistic to detect spatial clusters. To construct their beta-binomial scan statistic, they assumed $(x_i|\delta_i) \sim Bin(n_i, \theta_i)$ and considered $\text{Beta}(\alpha_0, \beta_0)$ as a prior for θ_0. The posterior is then $\text{Beta}(C + \alpha_0, N - C + \beta_0)$ where C and N are total cases and total population respectively. To find the Bayesian spatial scan statistic it is necessary to find the marginal likelihood under H_0 and H_1. After selecting adequate priors $P(H_z)$ and $P(H_0)$, which are respectively the zone prior probability and the prior probability of having no cluster, the Bayesian beta binomial spatial scan is presented in Algorithm 2.

Algorithm 2 Bayesian Beta Binomial Algorithm

```
1: For each candidate sub-region $z$, calculate $S_z = P(x|H_z)P(H_z)$ and $\sum_z S_z$.
2: Compute $S_0 = P(x|H_0)P(H_0)$ and obtain $P(x)$ using $\sum_z S_z$ and $S_0$.
3: Compute $P(H_z|x)$ for each $z$ in the candidate class.
```

Any sub-region z which maximizes $P(H_z|x)$ is a MLC. To determine $P(H_z)$, α_z, β_z, $P(H_0)$, α_0, and β_0, one can use historical information or non-informative priors.

The advantage of the Bayesian method in spatial cluster detection problems is the freedom of this method from Monte Carlo simulation in determining significance, but computing statistical power makes no sense for this method. Hence, the authors suggested the use of the Bayes factor (BF) as an alternative criterion for power which is defined as:

$$BF = \frac{P(x|H_z)}{P(x|H_0)}.$$

A $BF > 1$ indicates that H_z is more strongly supported by the observed data x. Thus, after finding the MLC, the BF can be used to decide about the significance of the MLC. $BF > 1$ indicates that the detected MLC is significant. There is at least one drawback of this method: for large values of N and/or C, the method leads to numerical instability. To mitigate this problem, the logarithm of the probabilities in the above-mentioned computations can be used.

7.4. Bayesian Zero-inflated Binomial Scan Statistics

In the same paper, Cançado, Fernandes and da Silva (2017) proposed an extension of the beta-binomial scan statistic presented in the previous subsection. The Bayesian zero-inflated binomial method is proposed as follows: suppose $(X_i|\theta_i, p) \sim ZIB(n_i, \theta_i, p)$, $\theta_i \sim$Beta(α_i, β_i) and $p \sim$Beta(α_p, β_p). The null hypothesis is $H_0 : \theta_i = \theta_0$, $\alpha_i = \alpha_0$ and $\beta_i = \beta_0$. As in the previous subsection, the MLC is a sub-region z which maximizes $P(H_z|x)$. Like Algorithm 2, Algorithm 3 below has three steps.

Algorithm 3 Bayesian Beta Binomial Algorithm

```
1: Compute $S_z = P(x|H_z, \delta)P(H_z, \delta)$, for each $z$ in the candidate class.
2: Compute $P(x|H_0, \delta)P(H_0, \delta)$ and also compute $P(x)$ by adding the values in steps 1 and 2.
3: Obtain $P(H_z|x, \delta) = \frac{P(x|H_z, \delta)P(H_z, \delta)}{P(x)}$.
```

Any sub region z which maximizes $P(H_z|x)$, is the MLC area. As before the BF is the criterion chosen to ascertain the significance of the MLC. But since in practice the vector δ is unknown, it must be estimated. A Gibbs sampler is used to estimate the δ parameter vector.

8. Nonparametric Methods in Spatial Cluster Detection

8.1. Distribution-free Scan Statistics Based on the Index of Concentration

Cucala (2014) extends scan statistics to point processes using a distribution free scan statistic. He proves that the distribution-free scan statistic is completely equivalent to the Gaussian-based scan statistic presented by Kulldorff, Huang and Konty (2009).

The distribution-free scan statistic considered by Cucala (2014) has homoscedastic and heteroscedastic versions. It is developed based on the assumption that $\{(x_i, s_i), i = 1, \ldots, m\}$ are realizations such that s_i and x_i are respectively the coordinate of the centroid and the associated measure for cell i.

To construct the distribution-free scan statistic in the homoscedastic version, it is assumed that $X_1, \ldots, X_m$ are independent and identically distributed (i.i.d.) random variables related to the measure in cell i with $E(X_i) = \nu, Var(X_i) = \sigma^2$, for all i, and $Cov(X_i, X_j) = 0$ if $i \neq j$. Consider that the mean and variance are unknown. As before any connected sub-region z is a candidate for being a spatial cluster. If the mean of measures in a candidate region z, $\bar{\mu}(z) = \frac{\sum_{i=1}^{m} X_i I(s_i \in z)}{\sum_{i=1}^{m} I(s_i \in z)}$, is significantly higher (lower) than the mean outside z, z will be considered a spatial cluster. Assume $D(z) = \bar{\mu}(z) - \bar{\mu}(z^c)$ is the difference of means inside and outside z. Under the null hypothesis $E(D(z)) = 0$ and $Var(D(z)) = \sigma^2(\frac{1}{n(z)} + \frac{1}{n(z^c)})$, where $n(z)$ is the total number of cells in sub-region z.

Moreover, Cucala (2014) introduced an index of concentration defined by

$$I(z) = \frac{\mu(z) - \mu(z^c)}{\sqrt{Var(D(z))/\sigma^2}}.$$

Since $E(I(z)) = 0$, $Var(I(z)) = \sigma^2$ and these values do not depend on $n(z)$, $I(z)$ can be used to find potential clusters having different population sizes. The next step is to determine the MLC, i.e., a connected sub-region which maximizes (minimizes) the index of concentration to detect spatial clusters of hot spots (cold spots). Cucala (2014) introduced three scan statistics: 1) the positive scan statistic, $\lambda_P = \max_{z \in \mathcal{Z}} I(z)$ which finds hot spots; 2) the negative scan statistic, $\lambda_N = \min_{z \in \mathcal{Z}} I(z)$, which finds cool spots; and 3) the global scan statistic, $\lambda = \max_{z \in \mathcal{Z}} |I(z)|$. After determining the MLC, because the distribution of the scan statistic is unknown, one needs to use Monte Carlo to determine the significance level of the spatial cluster. It is impossible to generate datasets under H_0, because it is assumed that this method is distribution-free. Hence, the author used random labeling to evaluate significance.

To construct the distribution-free heteroscedastic version of the scan statistic, it is supposed that the variances of measures are not equal and they depend on the weight related to cell i, i.e., $Var(X_i) = \frac{\sigma^2}{\delta_i}$ for all i. As before, σ^2 is unknown. Suppose that x_i is the mean of measures in cell i and this mean is the mean of δ_i cases. Under these assumptions the population in zone z is $n(Z) =$

$\sum_{i=1}^{m} \delta_i I(s_i \in z)$ and the mean of measures in z is $\bar{\mu}(z) = \frac{\sum_{i=1}^{m} \delta_i X_i I(s_i \in z)}{n(z)}$. Note that $I(z), \lambda_p, \lambda_z, \lambda$ are the same as the homoscedastic versions. Again, using random labeling one can ascertain the significance of the MLC.

8.2. Distribution-free Scan Statistics Based on the Wilcoxon Test

Jung and Cho (2015) presented a nonparametric method to find a spatial cluster for continuous datasets using the Wilcoxon rank-sum test. This method not only is an alternative to the normal scan statistic (Kulldorff, Huang and Konty, 2009), but also can be applied to heavy-tailed and skewed distributions.

Suppose F_{in} and F_{out} are CDFs of measures inside and outside a sub-region $z \in \mathcal{Z}$ and N is the population size of the study region G. To find a spatial cluster consider the following hypothesis test:

$$H_0 : F_{\text{in}} = F_{\text{out}} \quad \forall z \in \mathcal{Z} \quad vs. \quad H_1 : F_{\text{in}}(x) = F_{\text{out}}(x - \triangle), \quad \exists z \in \mathcal{Z}$$

where $\triangle$ is the location shift of the CDF of the outside relative to the inside candidate. If $\triangle > 0$, ($\triangle < 0$), the measurements tend to be higher (lower) inside candidate z compared to outside it. Since the distribution of measurements is unknown, it is impossible to define a scan statistic, so the authors suggested using the Wilcoxon rank-sum test. To apply this test, suppose that the rank of the measurement i, i.e., the rank of the x_i, is o_i. The Wilcoxon rank-test for candidate z is $W_z = \sum_{i \in z} o_i$. Using the normal approximation for W_z, i.e., $\frac{W - E(W_z)}{\sqrt{Var(W_z)}}$ is approximately normal$(0, 1)$, a p-value can be calculated. A sub-region z which has minimum p-value is considered the MLC. The main benefit of this method is its flexibility in the application for heavy-tailed and skewed distributions. Another clear advantage is that this method does not require Monte Carlo simulation and computing the p-value is simple.

9. Spatio-temporal Clusters

Most of the proposed cluster detection methods are based only on the spatial aspect, without the time dimension. However, the time dimension can be an important factor in cluster detection problems. Knox and Bartlett (1964) pioneered the study of space-time clustering by proposing a test. Their method is based on counting the number of pairs of events which are close in space and time simultaneously. A large number of these events indicates that events form spatio-temporal clusters. The Knox and Bartlett (1964) test was generalized by Mantel (1967). These tests are appropriate when the interest is to know the existence of a space-time cluster and sound an alarm without identifying its location and duration. In other words, these tests are suitable to declare that, for example, there is a disease outbreak.

To find the location and time period of an occurring cluster, the "space-time scan statistic" is a common extension of the purely spatial scan. It is a powerful statistical framework for the analysis of point processes. In this type of scan, the goal is to detect regions of space-time where the counts are significantly higher than expected. Kulldorff et al. (1998) were the first to discuss the time dimension in cluster detection using scan statistics. In space-time scan statistics, instead of a circular window, a cylindrical window scans the map such that its base is for scanning geographical area and its height corresponds to the time dimension. This window moves in space and time to find the MLC. Afterward, the significance of the MLC is determined by Monte Carlo simulation. Kulldorff et al. (1998) tried to find spatio-temporal clusters of brain cancer in New Mexico.

Later, Kulldorff (2001) applied the space-time scan statistic to prospective disease surveillance. The method was illustrated for thyroid cancer among men in New Mexico in 1973-1992. Also, Elias et al. (2006) tried to find spatio-temporal clusters of Meningococcal disease in Germany in a study in which specimens were obtained during 42 months.

The space-time scan statistic was also applied by Tonini, Tuia and Ratle (2009) to detect spatio-temporal clusters in fire sequences to find active fires in the state of Florida (US) during 2003-2006. According to their work, statistically significant clusters were detected in time and space. Clusters of forest fires are more frequent in hot seasons (spring and summer). This information helps authorities to take preventive measures at the correct time and space. Another application of space-time scan statistics was presented by Carneiro et al. (2007) involving American visceral leishmaniasis in the state of Bahia, Brazil, covering the 11-year period from 1994 to 2004.

Although there are many applications of the “space-time scan statistic”, Tango, Takahashi and Kohriyama (2011); Correa, Assunção and Costa (2015); Gangnon (2010b); Tango (2016) criticized the use of the prospective space-time scan statistic. To improve some of its problems, Assunção et al. (2007) proposed a score-based space-time scan statistic which is discussed in the next subsection. Also, Prates, Kulldorff and Assuncao (2014) presented a simulation study showing that the relative risk estimates for the space-time scan statistic must be defined with care and presented bias in its estimation, while the relative risk estimator is well defined for the purely spatial scan situations being not biased as the true relative risk of the cluster increases.

9.1. A Score-based Space-time Scan Statistic

To find spatio-temporal clusters, Assunção et al. (2007) proposed a new scan statistic that detects clusters in time and space in a point process by scanning three dimensions (two dimensions for space and one dimension for time). As before, a hypothesis testing method is applied such that the null hypothesis is that the underlying point process is a homogeneous Poisson point process with separable space-time intensity versus the alternative hypothesis of the existence of at least one space-time cluster.

To create the space-time scan statistic, Assunção et al. (2007) assumed a Poisson point process in a space-time region $\mathcal{A} = A \times [0, \mathcal{T}]$, such that A is a two dimensional area and $\mathcal{T}$ stands for the study time. They denote the space-time intensity by $\lambda(x, y, t)$. Under the null hypothesis, the intensity is separable, i.e.,

$\lambda(x, y, t) = \lambda_S(x, y)\lambda_T(t)$. Hence, under this hypothesis, the likelihood is a separable function of time and space, which are (functional) nuisance parameters. By introducing an alternative to the null model, they obtain a closed form score test statistic. They choose $C = C_s \times C_T$ as a fixed and arbitrary cylinder, such that C_S denotes space and C_T is the time period. The alternative hypothesis is:

$$H_{C,\epsilon} : \lambda(x, y, t) = \lambda_{\mathcal{S}}(x, y)\lambda_{\mathcal{T}}(t)\big(1 + \epsilon I_C(x, y, t)\big)$$

where $\epsilon > 0$ and I is an indicator function. This hypothesis expresses the deviation of the point process from the separability hypothesis. Considering this alternative hypothesis, the likelihood depends on ϵ, $\lambda_S(x, y)$ and $\lambda_{\mathcal{T}(t)}$. But $\lambda_S(x, y)$ and $\lambda_{\mathcal{T}(t)}$ are unknown, so applying the LRT method of Kulldorff (1997) is impossible. Therefore, Assunção et al. (2007) suggested a locally most powerful test based on the score statistic:

$$U_C = \frac{N(C) - N(C_S \times [0, T])N(A \times C_T)/N(A \times [0, T])}{\sqrt{N(C_S \times [0, T])N(A \times C_T)/N(A \times [0, T])}}$$

where $N(C) \sim \text{Poisson}(E_0(N(C)))$, $N(C_S \times [0, T])$ is the number of events inside the cylinder in area S with height T, $N(A \times [0, T])$ is the total number of events and $N(A \times C_T)$ is the number of events in a cylinder with base A and height T.

The most likely spatio-temporal cluster is a cylinder C which maximizes U_C i.e.,

$$U = \sup_C \{U_C\}. \tag{9.1}$$

Since the sampling distribution of (9.1) is unknown, Monte Carlo hypothesis testing is suggested (Dwass, 1957). For this test, the spatial locations of the events are fixed and by permutation of time $t_i, i = 1, 2, \ldots, n$ datasets are generated based on the null hypothesis (n is the total number of events). If the rank of U obtained by a real dataset is in the k-th largest in comparison to the values of U in the $m - 1$ generated datasets under null hypothesis, then $p = k/m$ is the one-sided significance level. This Monte Carlo hypothesis test to detect spatio-temporal clusters is naive and time-consuming. Hence, Assunção et al. (2007) created a more workable test process. Suppose that cylinder C_1 is contained in cylinder C_2 i.e.,

$$C_1 = C_{S_1} \times C_{T_1} \subset C_2 = C_{S_2} \times C_{T_2}.$$

Thus, $N(C_{S_1} \times [0, T])N(A \times C_{T_1}) \le N(C_{S_2} \times [0, T])N(A \times C_{T_2})$. On the other hand, $U = \dfrac{N(C) - \mu}{\sqrt{\mu}}$ is a decreasing function with respect to μ, so $U_{C_1} \le U_{C_2}$. Therefore, they showed that it is enough to scan all distinct subsets of events and their associated enveloping cylinders. Based on these facts, an improved version of the naive Monte Carlo test is presented.

9.2. A Surveillance Method

Assunção and Correa (2009) proposed a method to detect clusters in space-time based on the Shiryaev-Roberts statistic and its martingale property (Kenett and

Pollak, 1996). Suppose that $\lambda(x,y,t)$ is the unknown rate of a Poisson point process in $\mathbb{R}^3$ such that this process is observed in $\mathcal{A} \times (0,\mathcal{T}]$, where $\mathcal{A}$ represents the space component and $(0,\mathcal{T}]$ the time component. Assume that the event (x_i, y_i, t_i) is observed at time t_i. Hence, at time t_n, the number of observed events is exactly n. Imagine a cylinder $C_{k,n}$ with circular base $B(k,\rho)$ and the length of the interval $(t_k, t_n]$ as its height, i.e., $C_{k,n} = B(k,\rho) \times (t_k, t_n]$, such that ρ is the radius of its base and $t_n > t_k$. Considering $N(C_{k,n})$ to be the number of cases inside the cylinder following a Poisson$(\mu(C_{k,n}))$, where $\mu(C_{k,n}) = \int_{C_{k,n}} \lambda(x,y,t)dxdydt$, and assuming $\lambda_{\mathcal{A}}(x,y)$ and $\lambda_{\mathcal{T}}(t)$ are the marginal spatial and temporal densities, Assunção and Correa (2009) discussed the surveillance method in space-time. Their assumption is that the intensity function is separable, i.e., the intensity is proportional to the product of the time and space intensity components. They defined a coefficient of proportionality $\mu = \mu(\mathcal{A} \times (0,\mathcal{T}])$ where μ is the expected number of events in the study space. Therefore, the intensity in the presence of a cylindrical cluster is as follows:

$$\lambda(x,y,t) = \mu\lambda_{\mathcal{A}}(x,y) \times \lambda_{\mathcal{T}}(t)\big(1 + \epsilon I_{C_{k,n}}(x,y,t)\big) \tag{9.2}$$

such that I is an indicator function which shows whether or not an event belongs to the cylinder, and the constant $\epsilon > 0$ is the relative change in event's intensity within the cylinder.

Assuming there is no emerging cluster, the likelihood of the space-time Poisson process for n observed events is (Streit, 2010):

$$L_\infty = \left(\prod_{i=1}^{n} \lambda(x_i, y_i, z_i)\right) \times \exp\left(-\int_{\mathbb{R}^3} \lambda(x,y,t)dxdydt\right). \tag{9.3}$$

The emerging cluster at time $t_k < t_n$ is calculated by (9.3) using the intensities in (9.2). Let L_k be the likelihood of the space-time Poisson processes when n events have been observed. The test statistic is given by $R_n^{\text{STCD}} = \sum_{k=1}^{n} \dfrac{L_k}{L_\infty} = \sum_{k=1}^{n} \Lambda_{k,n}$, where $\Lambda_{k,n} = (1+\epsilon)^{N(C_{n,k})} \exp\left(-\epsilon\mu(C_{n,k})\right)$. To perform the hypothesis test, a value "A" must be defined as threshold. The null hypothesis (there is no spatio-temporal cluster) is rejected if the test statistic exceeds "A". The determination of "A" and ϵ are further discussed by Veloso et al. (2017). Since $\mu(C_{k,n})$ is unknown it is necessary to estimate it. This estimate is given by:

$$\frac{(\text{events in a cylinder with height } t_n) \times (\text{ events in time interval}(t_k, t_n] \text{ in } \mathcal{A})}{\text{total events}}. \tag{9.4}$$

This estimation function was proposed by Assunção and Correa (2009). However, Veloso et al. (2017) identified an error in this estimator and proposed a modified version. In the Assunção and Correa (2009) version, it was assumed that if the actual time is t_n, then the most recent event is included in the total number of events in the time interval $(t_k, t_n]$ and it may be included in the number of events in the disk $B(k,\rho)$. In the Veloso et al. (2017) version, to preserve the martingale property, the parameter estimation at time t_n should not depend on the observations at the current time t_n. Therefore, Veloso et al. (2017) modified the estimation of $\mu(C_{k,n})$ by excluding the actual time t_n. Their modification changes the denominator of (9.4) to the number of total events minus 1 and the numerator by excluding t_n from the interval. Besides this modification, Veloso et al. (2017) proposed an algorithm for automatic detection of multiple emerging space-time clusters. They also proposed how to automatically estimate the variable ϵ instead of fixing it.

10. Regression and Spatial Clustering

10.1. Transformation Method to Detect Spatial Clusters in Case-control Data

Aggregated data are used for spatial clusters on maps, where detailed information is lost in comparison with case-control data. In case-control data, the coordinate of each case is known on the map, while for aggregated data researchers only know the total number of cases and population in each cell. Therefore, when case-control data are available, it is recommended to use all the information available instead of aggregating it.

Demattei, Molinari and Daurès (2007) presented a method based on regression models and data transformation to detect multiple irregularly shaped spatial clusters. This method is an extension of the method of Molinari, Bonaldi and Daurés (2001). Their method selects the best model based on a double maximum test of H_0: uniform distribution of cases vs. H_1: there is at least one spatial cluster. Suppose there are n cases in study region G with total population N and $X_1, \ldots, X_n$ are i.i.d. random variables with density $h(x)$ which denote the place of cases in G. Using the coordinates of the cases, the authors introduced two variables, "distance" and "order". The distance variable is the distance between a point and its nearest neighbor, and the order variable is the order of selection of the cases. The order variable is denoted by t. If there is a cluster in G, cases in this cluster will have consecutive selection order and the interior distances will be less than the distances outside the cluster.

To find spatial clusters using this method, suppose that x_k, $k = 1, \ldots, n$ is an observation of X_k and $x_{(k)}$ is the k-th selection when x_1 is chosen arbitrarily from all observed values of cases and given $\{x_{(1)}, \ldots, x_{(k)}\}$, $x_{(k+1)}$ is the nearest case from $x_{(k)}$ among the $n - k$ remaining cases which are not selected yet. Suppose that D_k is the distance between $X_{(k)}$ and $X_{(k+1)}$ and d_k is its observed value. The cumulative distribution and density of D_k are defined as G_k and g_k respectively. Let d_k^w be the ratio between d_k and the expected value of D_k given that the values of $x_1, \ldots, x_k$ are known. Under the null hypothesis, if $d_k^w > 1$ then the observed distance is greater than its expected value, while for $d_k^w < 1$ the observed distance is less than its expected distance. Therefore, the null hypothesis will not be rejected if d_k^w is statistically close to one. The way to calculate $E_{H_0}(D_k | X_1 = x_1, \ldots, X_k = x_k)$ is given in Demattei, Molinari and Daurès (2007). Using the ordered weighted distances, it is possible to find the location of the spatial clusters.

Consider ordered pairs (k, d_k^w) with $k = 1, 2, \dots, n-1$. To detect the spatial cluster bounds, Demattei, Molinari and Daurès (2007) applied a regression of weighted distance on the order of selection. Under the no-cluster hypothesis, the appropriate model will be constant i.e.,

$$f(t) = \bar{d} = \frac{1}{T}\sum_{k=1}^{T} d_k^w, \qquad \forall t = 1, 2, \dots, T,$$

thus, free of t. If there is one cluster in the dataset, there will be two breaks (jump and fall) in the model, at points T_1 and T_2. In other words, from point 1 to point T_1 the value of the model is $\bar{d}_{[1:T_1]}$, from point T_1+1 to T_2, the value of the model is $\bar{d}_{[T_1+1:T_2]}$ and from point T_2+1 to point T, the value of the model is $\bar{d}_{[T_2+1:T]}$. T_1 and T_2 are called breaks (cluster bounds). At these points, i.e., T_1 and T_2, the value of the model rises or falls. $\bar{d}$ is the mean weighted distances in the corresponding interval i.e., $\bar{d}_{[a:b]}$ which is the mean distance based on ${d^w}_a, {d^w}_{a+1}, \dots, {d^w}_b$. The spatial cluster areas are the points for which the mean distance is low. To determine breaks, i.e., T_i's, the following strategy is used: consider ϵ as the minimum ratio of points which are between two breaks (minimum size of potential cluster), for example, 0.1, and $\Delta_\epsilon = \{(T_1, \dots, T_m), \forall i = 1, 2, \dots, m+1, card([T_{i-1}+1, T_i]) \geq |T\epsilon|\}$. The breaks are the points that minimize the squared error between d_t^w and $f(t)$. These points can be found by using the computer program developed by Bai and Perron (2003). After finding the breaks, to visualize the cluster area, Demattei, Molinari and Daurès (2007) suggested a disc-based method and a Voronoi tessellation method (Allard and Fraley, 1997), which are complementary to each other. A researcher may find many models with different breaks, so to select the best model it is suggested to perform hypothesis testing of no breaks vs. k breaks using the statistic proposed by Bai and Perron (1998) and to determine the significant models. To select the best model among the significant models, they suggested using the double maximum test (Bai and Perron, 1998). After selecting the best model, it is necessary to determine the p-value associated with each cluster by Monte Carlo simulation.

Indeed, in this method, when using the ordering of cases the spatial data are transformed into a one-dimensional point process. This method has at least two advantages: 1) it can detect irregularly shaped clusters; and 2) it has low computational demands. However, there are some drawbacks: 1) it is necessary to determine ϵ; and 2) it is necessary to select the upper bound for the number of breaks. It is important to mention that even under the null hypothesis, the distances may not be distributed identically when their dependency structure is complex. Hence, Cucala (2009) proposed a method similar to the method of Demattei, Molinari and Daurès (2007), based on a specific distribution property and introduced a flexible spatial scan test for case-control data. The aim of this method is to detect a sub-region in which the number of cases is abnormally high.

Consider an ordered sample with size n from uniform $(0,1)$. Suppose $\{S_1 \dots, S_{n+1}\}$ are the length of the intervals constructed from this ordered sample. Let

A_1 be a sub-region of study region G such that the distance of points inside this sub-region and the border of G is less than D_1 (distance between the first selected point and the border of G). Cucala (2009) assumed that A'_1 is proportional to the population in sub-region A_1, and proved that under H_0, the distribution of A'_1 is the same as the distribution of S_1. Similarly considering A_2 as a sub-region external to A_1, where the distance between points inside it and the first chosen point (i.e., X_1) is less than D_2, the distribution of A'_2, is the same as the distribution of S_2 under the null hypothesis. Likewise, A'_2 is proportional to the population in A_2. Under H_0, the area spacing $\{A'_1, \dots, A'_{n+1}\}$ has the same distribution as the uniform spacing i.e., $\{S_1, \dots, S_{n+1}\}$. Considering $T_i = \sum_{j=1}^{i} A'_j, 1 \leq i \leq n$, under H_0, $\{T_1, \dots, T_n\}$ are distributed as n ordered statistics from a uniform $(0, 1)$ (because of the distributional property), so the cluster detection in $\{T_1, \dots, T_n\}$ corresponds to cluster detection in the study region. Therefore, Cucala (2009) transformed the spatial cluster detection problem into a one-dimensional cluster detection problem. This kind of cluster detection can be done by applying the concentration index method of Cucala (2008).

Zhang and Lin (2009) presented a model-based approach that is equivalent to the spatial scan statistics method. Their method can be applied to overdispersed data. Furthermore, it is interesting for practitioners to identify clusters of spatial units with distinct patterns in a regression coefficient. Lee, Gangnon and Zhu (2017) proposed a formal statistical methodology by focusing on spatially varying coefficient regression methods such as geographically weighted regression models. They developed this new method for spatial cluster detection with a covariate. Detection of a single circular cluster and multiple clusters are possible with this new method. A limitation is that it allows for only one covariate. Relaxation of this restriction is desirable in order to explore the study region and allow for the detection of irregularly shaped clusters.

10.2. Spatio-temporal Cluster Detection

Demattei and Cucala (2010) extended the transformation method and spacing method mentioned above to find spatio-temporal clusters. They introduced a spatio-temporal distance and based on the ordering technique, tried to find spatio-temporal clusters. Not only did this add time to the spatial cluster detection method but also their method can be applied to find multiple clusters in case-control data. Multiple cluster detection is discussed in Section 11. In essence, Demattei and Cucala (2010) presented a method to find multiple clusters in time and space.

It should be mentioned that Kulldorff et al. (1998) pioneered the discussion of the detection of clusters in time and space. Section 9 presents many alternatives for spatio-temporal cluster detection.

11. Multiple Spatial Cluster Detection in Study Regions

As discussed before, to find a spatial cluster in the study region, researchers perform hypothesis testing – H_0: there is no spatial cluster vs. H_1: there is at least one sub-region z classified as a spatial cluster in the study region. In all of the above-mentioned studies, after finding the MLC, authors discuss the statistical significance and explore weaker clusters i.e., the second most likely cluster (SLC) in the study region (which does not overlap with the MLC). To find the significance of the SLC, it is necessary to apply a sequential method where cases and populations inside the MLC are removed from the dataset, and then, the standard spatial scan statistic is applied.

To find multiple spatial clusters, Li et al. (2011) proposed an alternative method. Under two conditions, the combination of sub-region z_i and z_j, i.e., (z_i, z_j), is considered as a two-cluster candidate in the case of multiple clusters:

1. z_i and z_j do not overlap, i.e., in centroids, cases, and controls.
2. The population of z_i and z_j is less than, for example, 50 percent of the total population.

In this respect, the study region is divided into three disjoint areas, z_i, z_j, and $G \cap (z_j, z_i)^c$. Suppose that the elements of an ordered triple (x_1, m_1, p_1) are respectively the number of cases, the population size, and the incidence probability in z_i. The triples (x_2, m_2, p_2) and (x_3, m_3, p_3) are corresponding triples for sub-region z_j and $G \cap (z_j, z_i)^c$. The goal is to perform the following hypothesis test

$$H_0 : p_1 = p_2 = p_3 \quad vs. \quad H_1 : p_1 > p_3, p_2 > p_3 \quad \text{in at least for one sub-region z}$$

For any two-cluster candidates, x_1, x_2 and x_3, are observations of an independent Poisson distribution. Li et al. (2011) found the likelihood ratio to construct the scan statistic. Any two-cluster sub-regions which maximize the likelihood ration are called the Most Likely Two-Cluster (MLTC). After finding the MLTC, it is necessary to check whether or not the two-cluster area is significant as a spatial cluster. If it is significant, one should investigate which sub-region is the first suspected cluster and which one is the second suspected one. After detecting the second suspected sub-region, using a sequential method, its significance should be tested. To find the second MLC, it is not necessary to search the study region to detect a sub-region which has the second greatest likelihood ratio given no overlap with the MLC because the second most likely cluster is exactly the second suspected sub-region in the two-cluster method.

Others have also investigated this topic, e.g., Zhang, Assunção and Kulldorff (2010), who also discussed how to determine multiple spatial clusters. Jung, Kulldorff and Richard (2010) constructed a spatial scan statistic for multinomial data. Wu and Glaz (2015) suggested a new adaptive procedure for a multiple window scan statistic. Wan et al. (2012) used an ant colony optimization to detect multiple spatial clusters. This method was proposed only for regional spatial count data. The ant colony optimization multiple cluster detection method was compared with three other methods: genetic algorithm-based spatial cluster detection (Duczmal et al., 2007), circular scan (Kulldorff, 1997), and flexible shape spatial scan statistic (Tango and Takahashi, 2005). The ant colony method outperformed the competitors in simulations.

The above-mentioned methods, i.e., Zhang, Assunção and Kulldorff (2010) and Li et al. (2011), do not allow overlapping clusters. The forward stepwise and forward stagewise methods proposed by Xu and Gangnon (2016) are effective in detection of multiple overlapping spatial clusters.

12. Recent Developments of Scan Statistics

Recently, Li et al. (2019) considered two types of scales: the aggregation level of the input data and the population threshold used in the cluster detection. They stated that these scales can affect the results of cluster detection using scan statistics. This effect is called detection inconsistency. To show this, Li et al. (2019) considered different scale settings and used two measures (i.e., the distance between cluster centers and the Jaccard index (Jaccard, 1901)). They measured the constancy of the detected clusters and studied three levels of aggregation (county, town, and a 900m grid) and three population thresholds (10%, 25%, and 50%). Four main results were obtained:

1. For a strong cluster and in a place with the high population density, the method is not highly sensitive to the data aggregation level. For weak true clusters and/or for less populous areas, the detection results from the different scale settings can be inconsistent.
2. The method's sensitivity to the population threshold is determined by the actual size of the true cluster.
3. The results show the superiority of a regular grid with fine resolution over the subjectively defined areal units.
4. When the population threshold is not smaller than 50%, a county-level analysis may have good quality when the disease has a strong clustering pattern in a place with high population density. However, county-level data should not be used to detect small clusters and with small population thresholds.

Jung (2019) extended scan statistics to matched case-control data. Since the Bernoulli-based scan statistic (2.1) is for independent observations, and since the case and control measures within a matched pair are not independent, the Bernoulli-based scan statistic is not suitable for matched case-control data. Hence, Jung (2019) designed hypothesis tests based on the odds ratio and used McNemar's test statistic and the Wald-type test statistic to detect spatial clusters. Their simulation study showed that the proposed methods had higher power and higher accuracy to detect spatial clusters for matched case-control data than the Bernoulli-based spatial scan statistic.

Ishioka et al. (2019) used the zero-suppressed binary decision diagram (Minato, 1993) to handle the large cardinality of the candidate class. This method can be compared with the method of AMST (Zhou, Shu and Su, 2015) in detecting irregularly shaped spatial clusters. Also, Desjardins, Hohl and Delmelle (2020) used a prospective space-time scan statistic to detect clusters of Covid-19 in the United States.

Since the Poisson-based spatial scan statistic detects larger clusters by absorbing insignificant neighbors with non-elevated risks, Lee and Jung (2019) suspected that the spatial scan statistic for ordinal data may also have similar undesirable outcomes. Hence, they applied a restricted likelihood ratio to the spatial scan statistic for ordinal outcome data to circumvent this problem. Through a simulation study, they demonstrated not only that original spatial scan statistic suffer from overdetection but also that their proposed methods have reasonable or better performance compared with the original methods.

For spatio-temporal count data with an excess of zeros, Allévius and Höhle (2019) proposed an unconditional space-time scan statistic. Moreover, a functional-model-adjusted spatial scan statistic was presented by Ahmed and Genin (2020). This new spatial scan statistic is designed to adjust cluster detection for longitudinal confounding factors indexed in space. This scan statistic was developed using generalized functional linear models in which the longitudinal confounding factors were considered to be functional covariates.

In many environmental applications, the response variables are spatially correlated. As an extension of the method proposed by Lee, Gangnon and Zhu (2017), Lee, Sun and Chang (2020) proposed a mixed effect model for spatial cluster detection to take the spatial correlation into account. The method developed can identify multiple potentially overlapping clusters. Recently, Lee et al. (2021) introduced a varying coefficient regression method to detect spatio-temporal clusters. They extended the spatial-only varying coefficient regression model to the spatio-temporal setting including flexible temporal patterns. The method relies on the detection of a potential cylindrical cluster of the regression coefficients, and is based on testing whether the regression coefficient is the same or not over the entire spatial domain for each time point. Additionally, it can detect multiple clusters.

13. Software and Packages

Up to this point in the paper we have provided a broad overview of a variety of scan statistics. We believe it is also important to give a guide to users about the available implementation tools. Therefore, a short review of software and `R` packages that work with scan statistics is now presented.

The package `SaTScan` (`https://www.satscan.org`) is one of the most complete programs in scan statistics and implements many methods (Kulldorff and Nagarwalla, 1995; Kulldorff, 1997, 2001; Kulldorff et al., 2006; Huang, Kulldorff and Gregorio, 2007; Huang et al., 2009, and others). The `rsatscan` package (Kleinman, 2015) uses `R` to create the files needed to execute the `SaTScan` software. `FleXScan` (`http://www.niph.go.jp/soshiki/gijutsu/index_e.html`) is another free software package for detecting spatial clusters (Tango and Takahashi, 2005; Tango, 2008). The package `rflexscan` (Otani and Takahashi, 2020) is a wrapper for the `FleXScan` software. The software package `TreeScan` (`https://www.treescan.org/`) implements the scan statistic proposed by Kulldorff, Fang and Walsh (2003). Finally, `ClusterSeer` is a program developed at **BioMedware**. This software handles many cluster detection methods (Turnbull et al., 1989; Besag and Newell, 1991; Kulldorff, 1997). It is important to mention that this stand alone program is available for Windows users.

A diversity of R packages also provide independent implementation of scan methods or introduce new scan statistics as alternatives. DCluster (Gómez-Rubio, Ferrándiz-Ferragud and Lopez-Quílez, 2005) implements the traditional scan (Kulldorff, 1997) with bootstrap and Gumbel alternatives to calculate the cluster significance. The AMOEBA package (Valles, 2014) includes a function to detect spatial clusters based on the Getis-Ord statistic. A Bayesian Dirichlet process for spatial clusters is available at PReMiuM (Liverani et al., 2015). A version of the scan statistic to detect clusters in social networks is available at SNscan (Wang, Hsu and Phoa, 2016). The package graphscan (Loche et al., 2016) implements the distribution free methods of Cucala (2008, 2009). The surveillance package (Meyer et al., 2017) implements many space-time scan methods. ClustGeo (Chavent et al., 2017) implements hierarchical clustering with spatial constraints to create spatial partitions of a map. The scanstatistics package (Allévius, 2018) (https://github.com/BenjaK/scanstatistics) implements a number of spatial (Poisson, negative binomial, zero-inflated Poisson), space-time and the negative binomial Bayesian scan statistics. SpatialEpi (Kim and Wakefield, 2018) has an implementation of the Bayesian cluster method from Wakefield and Kim (2013) and other traditional scan statistics. SpatialEpiApp (Moraga, 2017) is a package which provides a Shiny application for spatial and space-time scan statistics. Recently, the DClusterm package (Gómez-Rubio et al., 2019) implements a model-based approach using dummy variables in GLMs. A large variety of independent implementations of scan methods (e.g., Turnbull et al., 1989; Besag and Newell, 1991; Tango and Takahashi, 2005; Assunção et al., 2006; Kulldorff et al., 2006; Costa, Assunção and Kulldorff, 2012; Neill, 2012) are available in the smerc package (French, 2020a). Finally, for case-control data, the package smacpod (French, 2020b) has some spatial cluster methods.

14. Conclusion

In this paper, a detailed review of the development of scan statistics over the past three decades is presented. Scan statistics were initially proposed in the sixties and gained greater attention in the nineties with the advance of computational power. Our main goal here is to provide an up-to-date overview of scan statistics methods, their diversity of applications and some available software for practitioners.

The scan statistics method is based on the likelihood ratio test. The work of Kulldorff and Nagarwalla (1995) is the starting point for using scan statistics to find circular clusters. Subsequently, detecting non-circular clusters gained importance. The circular scan method was extended to detect non-circular clusters using MST (Assunção et al., 2006). This extension not only helps researchers in the detection of non-circular clusters but also increases the speed of cluster detection. AMST (Zhou, Shu and Su, 2015) is an alternative to the MST method of Assunção et al. (2006). This alternative made it possible to detect multiple non-circular clusters.

Monte Carlo hypothesis testing methods play a crucial role in scan statistics, but are time-consuming. With this in mind, Turnbull et al. (1989); Besag and Newell (1991); Soltani and Aboukhamseen (2015); Aboukhamseen, Soltani and Najafi (2016) presented some solutions to eliminate the Monte Carlo procedure in cluster detection.

The Poisson and binomial models are the most traditional models in the context of scan statistics. But these models are not appropriate for continuous data. Thus, the normal, multivariate normal and weighted normal scan statistics were introduced by researchers. Besides these models, the exponential scan statistic (Huang, Kulldorff and Gregorio, 2007) is constructed to deal with censored and uncensored survival data.

The inflation of zeros in a given dataset is another challenge in cluster detection. The need for zero-inflated models led to the construction of zero-inflated Poisson, zero-inflated binomial, and Bayesian beta-binomial scan statistics. The variety of these models tempted researchers to introduce nonparametric methods to detect spatial clusters. Cucala (2014) showed that the distribution-free scan statistic is equivalent to the Gaussian-based scan method presented by Kulldorff, Huang and Konty (2009). Another nonparametric method, based on the Wilcoxon test, was presented by Jung and Cho (2015).

Naturally, spatial scan statistics have evolved to handle spatio-temporal clusters by including the time component in the analysis. Pioneers in this topic were Knox and Bartlett (1964); their method evolved in many directions, as described by Mantel (1967); Kulldorff et al. (1998); Kulldorff (2001); Assunção et al. (2007); Assunção and Correa (2009); Veloso et al. (2017), among others. Applications of these methods are still relevant, e.g., to the detection of Covid-19 clusters in the United States (Desjardins, Hohl and Delmelle, 2020).

Available software and R packages that implement different types of scan statistics are presented (Section 13). This section provides a shortcut for practitioners who want to apply the methods discussed as well as researchers who want to compare new proposals with existing ones.

The supply of networks, graphs, and maps grows daily. Therefore, beyond the epidemiological applications, the topic of scan statistics is still very active in networks, trajectories and text streams. Recently, the issue of inference for networks, trajectories and text streams has gained attention. For example, detecting the busiest nodes in a network and using mobility information to detect sources of diseases or extract core knowledge from texts are tasks that can use scan statistics to provide adequate answers (Assunção, Souza and Prates, 2020).

With the increasing collection of data, researchers need to scan ever larger maps to detect spatial clusters. Although some incipient methods for scanning large maps are starting to appear (Assunção, Souza and Prates, 2020), we still see the need to extend scan methods to efficiently find spatial clusters in big maps. It seems that the use of parallel computation, Bayesian methods, and graph theory can help researchers to tackle this challenge. On the other hand, to the best of our knowledge, there is little work on detecting spatial clusters for proportional data (de Lima et al., 2016) and no parametric scan method to detect spatial clusters in the presence of heavy tailed and/or asymmetric spatial data. We believe that working with stable spatial models can be a way to improve scan statistics methods.

References

Aboukhamseen, S., Soltani, A. and Najafi, M. (2016). Modelling cluster detection in spatial scan statistics: Formation of a spatial Poisson scanning window and an ADHD case study. *Statistics & Probability Letters* **111** 26–31. MR3474778

Adelberger, K. L., Steidel, C. C., Pettini, M., Shapley, A. E., Reddy, N. A. and Erb, D. K. (2005). The Spatial Clustering of Star-forming Galaxies at Redshifts $1.4 \lesssim z \lesssim 3.5$. *The Astrophysical Journal* **619** 697.

Ahmed, M.-S. and Genin, M. (2020). A functional-model-adjusted spatial scan statistic. *Statistics in Medicine* **39** 1025–1040.

Allard, D. and Fraley, C. (1997). Nonparametric maximum likelihood estimation of features in spatial point processes using Voronoi tessellation. *Journal of the American Statistical Association* **92** 1485–1493.

Allévius, B. and Höhle, M. (2019). An unconditional space-time scan statistic for ZIP-distributed data. *Scandinavian Journal of Statistics* **46** 142–159. MR3915270

Allévius, B. (2018). scanstatistics: space-time anomaly detection using scan statistics. *Journal of Open Source Software* **3** 515.

Anselin, L. (1995). Local indicators of spatial association-LISA. *Geographical Analysis* **27** 93–115.

Assunção, R. M., Souza, R. C. S. N. P. and Prates, M. O. (2020). New Frontiers for Scan Statistics: Network, Trajectory, and Text Data. In *Handbook of Scan Statistics* (J. Glaz and M. Koutras, eds.) 1–24. Springer New York.

Assunção, R. and Correa, T. (2009). Surveillance to detect emerging space–time clusters. *Computational Statistics & Data Analysis* **53** 2817–2830. MR2667592

Assunção, R., Costa, M., Tavares, A. and Ferreira, S. (2006). Fast detection of arbitrarily shaped disease clusters. *Statistics in Medicine* **25** 723–742. MR2225159

Assunção, R., Tavares, A., Correa, T. and Kulldorff, M. (2007). Space-time cluster identification in point processes. *Canadian Journal of Statistics* **35** 9–25. MR2345372

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica* **66** 47–78. MR1616121

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics* **18** 1–22.

Bar-Hen, A., Emily, M. and Picard, N. (2015). Spatial cluster detection using nearest neighbor distance. *Spatial Statistics* **14** 400–411. MR3431048

Besag, J. and Newell, J. (1991). The detection of clusters in rare diseases.

Journal of the Royal Statistical Society. Series A (Statistics in Society) **154** 143–155.

Bhatt, V. and Tiwari, N. (2014). A spatial scan statistic for survival data based on Weibull distribution. *Statistics in Medicine* **33** 1867–1876. MR3256908

Cançado, A. L., da Silva, C. Q. and da Silva, M. F. (2014). A spatial scan statistic for zero-inflated Poisson process. *Environmental and Ecological Statistics* **21** 627–650. MR3279582

Cançado, A. L., Fernandes, L. B. and da Silva, C. Q. (2017). A Bayesian spatial scan statistic for zero-inflated count data. *Spatial Statistics* **20** 57–75. MR3654003

Carneiro, D. D., Bavia, M. E., Rocha, W. J., Tavares, A. C., Cardim, L. L. and Alemayehu, B. (2007). Application of spatio-temporal scan statistics for the detection of areas with increased risk for American visceral leishmaniasis in the state of Bahia, Brazil. *Geospatial Health* **2** 113–126.

Castellares, F., Prates, M. O. and Abolhassani, A. (2019). Comments on "A spatial scan statistic for compound Poisson data". *Statistics in Medicine* **38** 1297–1299.

Chang, H.-M. and Rosychuk, R. J. (2015). A spatial scan statistic for compound Poisson data, using the negative binomial distribution and accounting for population stratification. *Statistica Sinica* **25** 313–327. MR3328817

Chavent, M., Kuentz, V., Labenne, A. and Saracco, J. (2017). ClustGeo: Hierarchical Clustering with Spatial Constraints R package version 2.0. MR3856335

Choynowski, M. (1959). Maps based on probabilities. *Journal of the American Statistical Association* **54** 385–388.

Correa, T. R., Assunção, R. M. and Costa, M. A. (2015). A critical look at prospective surveillance using a scan statistic. *Statistics in Medicine* **34** 1081–1093. MR3322735

Costa, M. A., Assunção, R. M. and Kulldorff, M. (2012). Constrained spanning tree algorithms for irregularly-shaped spatial clustering. *Computational Statistics & Data Analysis* **56** 1771–1783. MR2892376

Cucala, L. (2008). A hypothesis-free multiple scan statistic with variable window. *Biometrical Journal: Journal of Mathematical Methods in Biosciences* **50** 299–310. MR2420271

Cucala, L. (2009). A flexible spatial scan test for case event data. *Computational Statistics & Data Analysis* **53** 2843–2850. MR2667594

Cucala, L. (2014). A distribution-free spatial scan statistic for marked point processes. *Spatial Statistics* **10** 117–125. MR3280094

Cucala, L., Genin, M., Lanier, C. and Occelli, F. (2017). A multivariate Gaussian scan statistic for spatial data. *Spatial Statistics* **21** 66–74. MR3692177

Culvenor, D. S., Coops, N., Preston, R. and Tolhurst, K. G. (1998). A spatial clustering approach to automated tree crown delineation. In *Proc. of the International Forum on Automated Interpretation of High Spatial Resolution Digital Imagery for Forestry* 67–80.

de Lima, M. S., Duczmal, L. H., Neto, J. C. and Pinto, L. P. (2015). Spatial scan statistics for models with overdispersion and inflated zeros. *Statistica Sinica* **25** 225–241. MR3328812

de Lima, M. S., dos Santos, V. S., Duczmal, L. H. and da Silva Souza, D. (2016). A spatial scan statistic for beta regression. *Spatial Statistics* **18** 444–454. MR3575501

Demattei, C. and Cucala, L. (2010). Multiple spatio-temporal cluster detection for case event data: an ordering-based approach. *Communications in Statistics-Theory and Methods* **40** 358–372. MR2765861

Demattei, C., Molinari, N. and Daurès, J.-P. (2007). Arbitrarily shaped multiple spatial cluster detection for case event data. *Computational Statistics & Data Analysis* **51** 3931–3945. MR2364501

Desjardins, M. R., Hohl, A. and 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. *Applied Geography* **118** 102202.

Duczmal, L., Cançado, A. L., Takahashi, R. H. and Bessegato, L. F. (2007). A genetic algorithm for irregularly shaped spatial scan statistics. *Computational Statistics & Data Analysis* **52** 43–52. MR2409963

Duczmal, L. H., Moreira, G. J., Burgarelli, D., Takahashi, R. H., Magalhães, F. C. and Bodevan, E. C. (2011). Voronoi distance based prospective space-time scans for point data sets: a dengue fever cluster analysis in a southeast Brazilian town. *International Journal of Health Geographics* **10** 29.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics* **28** 181–187. MR0087280

Eck, J., Chainey, S., Cameron, J. and Wilson, R. (2005). Mapping Crime: Understanding Hotspots.

Elias, J., Harmsen, D., Claus, H., Hellenbrand, W., Frosch, M. and Vogel, U. (2006). Spatiotemporal analysis of invasive meningococcal disease, Germany. *Emerging Infectious Diseases* **12** 1689.

Erdos, P. and Rényi, A. (1960). On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* **5** 17–60. MR0125031

Fortunato, S. (2010). Community detection in graphs. *Physics Reports* **486** 75–174. MR2580414

French, J. (2020a). smerc: Statistical Methods for Regional Counts R package version 1.3.3.

French, J. (2020b). smacpod: Statistical Methods for the Analysis of Case-Control Point Data R package version 2.1.

Gangnon, R. E. (2010a). Local multiplicity adjustments for spatial cluster detection. *Environmental and Ecological Statistics* **17** 55–71. MR2594935

Gangnon, R. E. (2010b). A model for space–time cluster detection using spatial clusters with flexible temporal risk patterns. *Statistics in Medicine* **29** 2325–2337. MR2759949

Gangnon, R. E. and Clayton, M. K. (2000). Bayesian detection and modeling of spatial disease clustering. *Biometrics* **56** 922–935.

Gangnon, R. E. and Clayton, M. K. (2001). A weighted average likelihood ratio test for spatial clustering of disease. *Statistics in Medicine* **20** 2977–2987.

Gangnon, R. E. and Clayton, M. K. (2003). A hierarchical model for spatially clustered disease rates. *Statistics in Medicine* **22** 3213–3228. MR2015035

Gangnon, R. E. and Clayton, M. K. (2004). Likelihood-based tests for localized spatial clustering of disease. *Environmetrics: The Official Journal of the International Environmetrics Society* **15** 797–810.

Gangnon, R. and Clayton, M. K. (2007). Cluster detection using Bayes factors from overparameterized cluster models. *Environmental and Ecological Statistics* **14** 69–82. MR2345667

Garwood, F. (1936). Fiducial limits for the Poisson distribution. *Biometrika* **28** 437–442.

Gladders, M. D. and Yee, H. (2000). A new method for galaxy cluster detection. I. The algorithm. *The Astronomical Journal* **120** 2148.

Gómez-Rubio, V. and López-Quílez, A. (2010). Statistical methods for the geographical analysis of rare diseases. *Advances in Experimental Medicine and Biology* **686** 151–171.

Gómez-Rubio, V., Moraga, P., Molitor, J. and Rowlingson, B. (2019). DClusterm: model-based detection of disease clusters R package version 0.2-3.

Goura, V., Rao, N. M. and Reddy, M. R. (2011). A dynamic clustering technique using minimumspanning tree. In *Proceedings of the 2nd International Conference on Biotechnology and Food Science (IPCBEE'11), IACSIT Press, Singapore* 66–70.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. MR1380810

Grubesic, T. H. (2006). On the application of fuzzy clustering for crime hot spot detection. *Journal of Quantitative Criminology* **22** 77.

Gutteridge, A., Bartlett, G. J. and Thornton, J. M. (2003). Using a neural network and spatial clustering to predict the location of active sites in enzymes. *Journal of Molecular Biology* **330** 719–734.

Gómez-Rubio, V., Ferrándiz-Ferragud, J. and Lopez-Quílez, A. (2005). Detecting clusters of disease with R. *Journal of Geographical Systems* **7** 189–206.

Han, J., Kamber, M. and Tung, A. K. H. Spatial Clustering Methods in Data Mining. In *Geographic Data Mining and Knowledge Discovery* 188–217. Taylor & Francis.

Haralick, R. and Dinstein, I. (1975). A spatial clustering procedure for multi-image data. *IEEE Transactions on Circuits and Systems* **22** 440–450.

Haralick, R. and Kelly, G. (1969). Pattern recognition with measurement space and spatial clustering for multiple images. *Proceedings of the IEEE* **57** 654–665.

Harries, K. D. (1999). Mapping Crime: Principle and Practice Technical Report, US Department of Justice, Office of Justice Programs, National Institute of Justice, Crime Mapping Research Center.

Huang, L., Kulldorff, M. and Gregorio, D. (2007). A spatial scan statistic for survival data. *Biometrics* **63** 109–118. MR2345580

Huang, L., Tiwari, R. C., Zou, Z., Kulldorff, M. and Feuer, E. J. (2009). Weighted normal spatial scan statistic for heterogeneous population data. *Journal of the American Statistical Association* **104** 886–898. MR2750223

Ishioka, F., Kawahara, J., Mizuta, M., Minato, S.-i. and Kurihara, K. (2019). Evaluation of hotspot cluster detection using spatial scan statistic based on exact counting. *Japanese Journal of Statistics and Data Science* **2** 241–262. MR3969147

Izakian, H. and 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.

Jaccard, P. (1901). Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. *Bull Soc Vaudoise Sci Nat* **37** 241–272.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc., New Jersey. MR0999135

Jung, I. (2019). Spatial scan statistics for matched case-control data. *PloS One* **14** e0221225.

Jung, I. and Cho, H. J. (2015). A nonparametric spatial scan statistic for continuous data. *International Journal of Health Geographics* **14** 30.

Jung, I., Kulldorff, M. and Klassen, A. C. (2007). A spatial scan statistic for ordinal data. *Statistics in Medicine* **26** 1594–1607. MR2359161

Jung, I., Kulldorff, M. and Richard, O. J. (2010). A spatial scan statistic for multinomial data. *Statistics in Medicine* **29** 1910–1918. MR2758462

Kenett, R. S. and Pollak, M. (1996). Data-analytic aspects of the Shiryayev-Roberts control chart: Surveillance of a non-homogeneous Poisson process. *Journal of Applied Statistics* **23** 125–138. MR1395487

Kim, A. Y. and Wakefield, J. (2018). SpatialEpi: Methods and Data for Spatial Epidemiology R package version 1.2.3.

Kim, R. S. J., Kepner, J. V., Postman, M., Strauss, M. A., Bahcall, N. A., Gunn, J. E., Lupton, R. H., Annis, J., Nichol, R. C., Castander, F. J. et al. (2002). Detecting clusters of galaxies in the sloan digital sky survey. i. monte carlo comparison of cluster detection algorithms. *The Astronomical Journal* **123** 20.

Kleinman, K. (2015). rsatscan: Tools, Classes, and Methods for Interfacing with SaTScan Stand-Alone Software R package version 0.3.9200.

Knox, E. and Bartlett, M. (1964). The detection of space-time interactions. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **13** 25–30.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics-Theory and Methods* **26** 1481–1496. MR1456844

Kulldorff, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **164** 61–72. MR1819022

Kulldorff, M., Fang, Z. and Walsh, S. J. (2003). A tree-based scan statistic for database disease surveillance. *Biometrics* **59** 323–331. MR1987399

Kulldorff, M., Huang, L. and Konty, K. (2009). A scan statistic for continuous data based on the normal probability model. *International Journal of Health Geographics* **8** 58.

Kulldorff, M. and Nagarwalla, N. (1995). Spatial disease clusters: detection and inference. *Statistics in Medicine* **14** 799–810.

Kulldorff, M., Athas, W. F., Feurer, E. J., Miller, r. A. and Key, C. R. (1998). Evaluating cluster alarms: a space-time scan statistic and brain cancer in Los Alamos, New Mexico. *American Journal of Public Health* **88** 1377–1380.

Kulldorff, M., Huang, L., Pickle, L. and Duczmal, L. (2006). An elliptic spatial scan statistic. *Statistics in Medicine* **25** 3929–3943. MR2297401

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34** 1–14.

Lee, J., Gangnon, R. E. and Zhu, J. (2017). Cluster detection of spatial regression coefficients. *Statistics in Medicine* **36** 1118–1133. MR3621013

Lee, M. and Jung, I. (2019). Modified spatial scan statistics using a restricted likelihood ratio for ordinal outcome data. *Computational Statistics & Data Analysis* **133** 28–39. MR3926464

Lee, J., Sun, Y. and Chang, H. H. (2020). Spatial cluster detection of regression coefficients in a mixed-effects model. *Environmetrics* **31** e2578.

Lee, J., Gangnon, R. E., Zhu, J. and Liang, J. (2017). Uncertainty of a detected spatial cluster in 1D: Quantification and visualization. *Stat* **6** 345–359. MR3722934

Lee, J., Kamenetsky, M. E., Gangnon, R. E. and Zhu, J. (2021). Clustered spatio-temporal varying coefficient regression model. *Statistics in Medicine* **40** 465–480. MR4194595

Li, X.-Z., Wang, J.-F., Yang, W.-Z., Li, Z.-J. and Lai, S.-J. (2011). A spatial scan statistic for multiple clusters. *Mathematical Biosciences* **233** 135–142. MR2856850

Li, M., Shi, X., Li, X., Ma, W., He, J. and Liu, T. (2019). Sensitivity of disease cluster detection to spatial scales: an analysis with the spatial scan statistic method. *International Journal of Geographical Information Science* **33** 2125–2152.

Liu, Y., Liu, Y. and Zhang, T. (2018). Wald-based spatial scan statistics for cluster detection. *Computational Statistics & Data Analysis* **127** 298–310. MR3820325

Liverani, S., Hastie, D. I., Azizi, L., Papathomas, M. and Richardson, S. (2015). PReMiuM: An R Package for Profile Regression Mixture Models Using Dirichlet Processes. *Journal of Statistical Software* **64** 1–30.

Loche, R., Giron, B., Abrial, D., Cucala, L., Charras-Garrido, M. and De-Goer, J. (2016). graphscan: Cluster Detection with Hypothesis Free Scan Statistic R package version 1.1.1.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research* **27** 209–220.

Marchette, D. (2012). Scan statistics on graphs. *Wiley Interdisciplinary Reviews: Computational Statistics* **4** 466–473.

Meyer, S., Held, L., Höhle, M. et al. (2017). Spatio-Temporal Analysis of Epidemic Phenomena Using the R Package surveillance. *Journal of Statistical Software* **77** 1–55.

Minato, S.-i. (1993). Zero-suppressed BDDs for set manipulation in combinatorial problems. In *Proceedings of the 30th International Design Automation Conference* 272–277.

Mo, H. and White, S. D. (1996). An analytic model for the spatial clustering of dark matter haloes. *Monthly Notices of the Royal Astronomical Society* **282** 347–361.

Molinari, N., Bonaldi, C. and Daurés, J.-P. (2001). Multiple temporal cluster detection. *Biometrics* **57** 577–583.

Moraga, P. (2017). SpatialEpiApp: A Shiny Web Application for the Analysis of Spatial and Spatio-Temporal Disease Data R package version 0.3.

Murray, A. T. and Estivill-Castro, V. (1998). Cluster discovery techniques for exploratory spatial data analysis. *International Journal of Geographical Information Science* **12** 431–443.

Murray, A. T., Grubesic, T. H. and Wei, R. (2014). Spatially significant cluster detection. *Spatial Statistics* **10** 103–116.

Myers, N., Mittermeier, R. A., Mittermeier, C. G., Da Fonseca, G. A. and Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature* **403** 853.

Neill, D. B. (2011). Fast Bayesian scan statistics for multivariate event detection and visualization. *Statistics in Medicine* **30** 455–469.

Neill, D. B. (2012). Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74** 337–360.

Neill, D. B. and Cooper, G. F. (2010). A multivariate Bayesian scan statistic for early event detection and characterization. *Machine Learning* **79** 261–282.

Neill, D., Moore, A. and Cooper, G. (2005). A Bayesian spatial scan statistic. *Advances in Neural Information Processing Systems* **18** 1003–1010.

Openshaw, S., Charlton, M., Wymer, C. and Craft, A. (1987). A mark 1 geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information System* **1** 335–358.

Ord, J. K. and Getis, A. (1995). Local spatial autocorrelation statistics: distributional issues and an application. *Geographical Analysis* **27** 286–306.

Otani, T. and Takahashi, K. (2020). rflexscan: The Flexible Spatial Scan Statistic R package version 0.3.1.

Prates, M. O., Assunção, R. M. and Costa, M. A. (2012). Flexible scan statistic test to detect disease clusters in hierarchical trees. *Computational Statistics* **27** 715–737.

Prates, M. O., Kulldorff, M. and Assuncao, R. M. (2014). Relative risk estimates from spatial and space–time scan statistics: are they biased? *Statistics in Medicine* **33** 2634–2644.

Priebe, C. E., Conroy, J. M., Marchette, D. J. and Park, Y. (2005). Scan statistics on enron graphs. *Computational & Mathematical Organization*

Theory **11** 229–247.

Prim, R. C. (1957). Shortest connection networks and some generalizations. *Bell System Technical Journal* **36** 1389–1401.

Ross, S. M. (2014). *Introduction to Probability Models*. Academic Press.

Rosychuk, R. J., Huston, C. and Prasad, N. G. (2006). Spatial event cluster detection using a compound Poisson distribution. *Biometrics* **62** 465–470.

Sherman, L. W. and Weisburd, D. (1995). General deterrent effects of police patrol in crime "hot spots": A randomized, controlled trial. *Justice quarterly* **12** 625–648.

Sierksma, G. and Hoogeveen, H. (1991). Seven criteria for integer sequences being graphic. *Journal of Graph Theory* **15** 223–231.

Soltani, A. and Aboukhamseen, S. (2015). An alternative cluster detection test in spatial scan statistics. *Communications in Statistics-Theory and Methods* **44** 1592–1601.

Stohlgren, T. J., Binkley, D., Chong, G. W., Kalkhan, M. A., Schell, L. D., Bull, K. A., Otsuki, Y., Newman, G., Bashkin, M. and Son, Y. (1999). Exotic plant species invade hot spots of native plant diversity. *Ecological Monographs* **69** 25–46.

Streit, R. (2010). *Poisson Point Processes Imaging, Tracking, and Sensing*. Springer, New York.

Tango, T. (2008). A spatial scan statistic with a restricted likelihood ratio. *Japanese Journal of Biometrics* **29** 75–95.

Tango, T. (2016). On the recent debate on the space-time scan statistic for prospective surveillance. *Statistics in Medicine* **35** 1927.

Tango, T. and Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics* **4** 11.

Tango, T., Takahashi, K. and Kohriyama, K. (2011). A space–time scan statistic for detecting emerging outbreaks. *Biometrics* **67** 106–115.

Tonini, M., Tuia, D. and Ratle, F. (2009). Detection of clusters using space–time scan statistics. *International Journal of Wildland Fire* **18** 830–836.

Turnbull, B. W., Iwano, E. J., Burnett, W. S., Howe, H. L. and Clark, L. C. (1989). Monitoring for clusters of disease; Application to leukemia incidence in upstate New York Technical Report, Cornell University Operations Research and Industrial Engineering.

Valles, G. (2014). AMOEBA: A Multidirectional Optimum Ecotope-Based Algorithm R package version 1.1.

Veloso, B. M., Correa, T. R., Prates, M. O., Oliveira, G. F. and Tavares, A. I. (2017). MAD-STEC: a method for multiple automatic detection of space-time emerging clusters. *Statistics and Computing* **27** 1099–1110.

Wakefield, J. and Kim, A. (2013). A Bayesian model for cluster detection. *Biostatistics* **14** 752–765.

Waller, L. A., Carlin, B. P., Xia, H. and Gelfand, A. E. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association* **92** 607–617.

Wan, Y., Pei, T., Zhou, C., Jiang, Y., Qu, C. and Qiao, Y. (2012).

ACOMCD: A multiple cluster detection algorithm based on the spatial scan statistic and ant colony optimization. *Computational Statistics & Data Analysis* **56** 283–296.

Wang, T.-C., Hsu, T.-C. and Phoa, F. K. H. (2016). SNscan: Scan Statistics in Social Networks R package version 1.0.

Wang, T.-C. and Phoa, F. K. H. (2016). A scanning method for detecting clustering pattern of both attribute and structure in social networks. *Physica A: Statistical Mechanics and its Applications* **445** 295–309.

Wang, B., Phillips, J. M., Schreiber, R., Wilkinson, D., Mishra, N. and Tarjan, R. (2008). Spatial scan statistics for graph clustering. In *Proceedings of the 2008 SIAM International Conference on Data Mining* 727–738. SIAM.

Wieland, S. C., Brownstein, J. S., Berger, B. and Mandl, K. D. (2007). Density-equalizing Euclidean minimum spanning trees for the detection of all disease cluster shapes. *Proceedings of the National Academy of Sciences* **104** 9404–9409.

Woodall, W. H., Zhao, M. J., Paynabar, K., Sparks, R. and Wilson, J. D. (2017). An overview and perspective on social network monitoring. *IISE Transactions* **49** 354–365.

Wu, T.-L. and Glaz, J. (2015). A new adaptive procedure for multiple window scan statistics. *Computational Statistics & Data Analysis* **82** 164–172.

Xu, J. and Gangnon, R. E. (2016). Stepwise and stagewise approaches for spatial cluster detection. *Spatial and Spatio-temporal Epidemiology* **17** 59–74.

Yamada, I. and Rogerson, P. (2008). *Statistical detection and surveillance of geographic clusters*. Chapman and Hall/CRC.

Yan, P. and Clayton, M. K. (2006). A cluster model for space–time disease counts. *Statistics in Medicine* **25** 867–881.

Yin, P. and Mu, L. (2018). A hybrid method for fast detection of spatial disease clusters in irregular shapes. *GeoJournal* **83** 693–705.

Zhang, Z., Assunção, R. and Kulldorff, M. (2010). Spatial scan statistics adjusted for multiple clusters. *Journal of Probability and Statistics, Article ID 642379* **2010** 1–11.

Zhang, T. and Lin, G. (2009). Spatial scan statistics in loglinear models. *Computational Statistics & Data Analysis* **53** 2851–2858.

Zhang, T. and Lin, G. (2014). Family of power divergence spatial scan statistics. *Computational Statistics & Data Analysis* **75** 162–178.

Zhang, T., Lin, G. et al. (2017). Asymptotic properties of spatial scan statistics under the alternative hypothesis. *Bernoulli* **23** 89–109.

Zhang, T., Zhang, Z. and Lin, G. (2012). Spatial scan statistics with overdispersion. *Statistics in Medicine* **31** 762–774.

Zhang, L. and Zhu, Z. (2012). Spatial multiresolution cluster detection method. *arXiv preprint arXiv:1205.2106*.

Zhou, Y., Cheng, H. and Yu, J. X. (2009). Graph clustering based on structural/attribute similarities. *Proceedings of the VLDB Endowment* **2** 718–729.

Zhou, R., Shu, L. and Su, Y. (2015). An adaptive minimum spanning tree test for detecting irregularly-shaped spatial clusters. *Computational Statistics & Data Analysis* **89** 134–146.