Health Monitoring Techniques Using Scan Statistics

Sotiris Bersimis, Athanasios Sachlas, and Markos V. Koutras

Contents

Abstract

Scan statistics appeared in the statistics literature about half a century ago, and since then many papers suggesting either extensions and modifications or applications into various research fields have been published. Scan statistics are mainly used to detect clusters of events in time or space. In the last two decades several researchers have proposed techniques or systems for the surveillance of public health or other healthcare processes. In this paper, we shall present a systematic review of health monitoring techniques which exploit scan statistics in order to set up early warning systems detecting potential threats for public health.

S. Bersimis · A. Sachlas · M. V. Koutras (✉)
Department of Statistics and Insurance Science, University of Piraeus, Piraeus, Greece
e-mail: sbersim@unipi.gr; asachlas@unipi.gr; mkoutras@unipi.gr

J. Glaz, M. V. Koutras (eds.), *Handbook of Scan Statistics*,
https://doi.org/10.1007/978-1-4614-8414-1_54-1

Keywords

Biosurveillance · Continuous monitoring · Disease outbreaks identification · Health monitoring · Public health · Scan statistics

Introduction

Healthcare industry, due to its substantial role in achieving high standards of living, should always maintain a superior quality of services. On the other hand, government health authorities are interested in exploiting public health data for maintaining safety and well-being for the general population. The information carried in past health data is valuable in setting up policies for the risk management of outbreaks (significant increases in the number of occurrences of specific disease events). Thus, the continuous monitoring of a healthcare process is very crucial for securing an effective management in both private and public sector. During the last decades, a large number of statistical techniques have been exploited to establish efficient monitoring schemes for healthcare processes. Prevailing approaches in this field are the ones that make use of scan statistics and statistical process monitoring (SPM) (see, e.g., Hanslik et al. 2001, Burkom 2003, Sabhnani et al. 2005).

In many statistical applications, the scientists have to assess the significance of the occurrence of clusters of events in time or space. The scientists are especially interested to determine whether, under the assumption that the events are distributed independently and uniformly over time or space, an observed cluster of events has occurred by chance or not. Applications of scan statistics have been recorded in many areas of science and technology including, among others, geology, geography, medicine, minefield detection, molecular biology, photography, quality control and reliability theory, radio-optics (Glaz et al. 2001, Balakrishnan and Koutras 2011, and references therein), seismology (Taylor et al., 2010), environment (So et al., 2013), suicide commitment (Cheung et al., 2013), human health (Imanishi et al., 2015), and economic activity (Bersimis et al., 2014).

According to Tsui et al. (2008), the most common disease spread monitoring methods can be categorized into temporal, spatial, and spatiotemporal surveillance techniques. In those techniques the use of scan statistics has been extensively practiced during the last decades.

Focusing exclusively on the temporal dimension, we are interested in observing trends or subsequences of similar values in a sequence of outcomes associated to crucial health data. From a probabilistic point of view, according to Koutras and Alexandrou (1995), the discrete scan statistics are enumerating variables based on a moving window in a linearly ordered sequence of binary outcomes (success or failure). For example, the maximum number of successes contained in a moving window of length $m = 5$ over the sequence *SSFFSSFSFSSSSFS* of length $n = 15$ is 4. This is perhaps the simplest example of a one-dimensional scan statistic.

Besides the one-dimensional scan statistics, the two- and three-dimensional scan statistics have also been defined. The first one is used when we are interested in

Table 1 An example of the two-dimensional scan statistic

	1	2	3	4	5
1	S	S	F	S	S
2	S	F	S	S	S
3	F	S	F	F	S
4	S	F	F	F	F
5	S	S	F	S	S

identifying events of a specific type in a geographical area while the latter when we take time into account as well.

Table 1 shows a 5 × 5 grid containing S and F's. Applying the two-dimensional scan statistic with a moving window of size 2 × 2, it arises that $S_{2,2;5,5} = 4$ (there are four S's in the highlighted cells).

The exact locations of the disease events in the region under study may be either available or not. Usually, in the latter case, the sum of events in subareas is known. Thus, a common approach is to divide the area into rectangular subareas and to exploit the sum of the individual events. These are known as grid-based methods.

Two- or higher-dimensional scan statistics have three main features: the geometry of the area under study, the probability distribution generating the events under the null hypothesis, and the morphology (shape and size) of the scanning window (Koutras and Alexandrou, 1995).

Assume that we wish to monitor a geographic region in order to identify disease outbreaks. The basic three aspects of monitoring are (i) the determination of regions with exceptionally high (or low) percentages of disease incidents, (ii) the determination of statistical significance of high or low numbers of incidents, and (iii) the evaluation of factors resulting in the presence of very high or very low number of incidents. The main use of the scan statistic in the case of risk management of outbreaks is to provide evidence for significant increases in the number of occurrences of specific events in both space and time; thus scans will be used to identify clusters of high-frequency events in a specific area and time interval.

Note that Kulldorff introduced a different version of the spatial scan statistic, which does not make use of a grid (see, e.g., Kulldorff and Nagarwalla 1995, Kulldorff 1997, and Kulldorff et al. 2005). Scan statistics have become very popular in practice because on one hand they are relatively simple to apply and on the other hand the SaTScan software (Kulldorff, 1997) is available for anyone interested using them.

According to Chen et al. (2010), Kulldorff's space scan statistics (Kulldorff, 1997) are mainly used for retrospective surveillance (i.e., to test whether disease events are randomly distributed over space and time for a specific region during a specific time period). Kulldorff's prospective version of time-space scan statistics (Kulldorff, 2001) is mainly used for prospective surveillance with repeated time periodic analyses.

According to Wagner et al. (2011), “The main goals of spatial scanning are to identify the locations, shapes, and sizes of potential clusters (i.e., pinpointing those areas which are most relevant), and to determine whether each of these potential clusters is more likely to be a “true” cluster (requiring further investigation by public health officials) or simply a chance occurrence (which can safely be ignored).”

In this paper we review the most significant health monitoring techniques, which use scan statistics. The structure of the paper is as follows: in section “Scan Statistics for Linear Sequences and Grids” the notion of scan statistic applied either in a sequence of random variables (e.g., temporal scan statistic) or in a grid of random variables (spatial scan statistics) is presented, while in section “Kulldorff’s Scan Statistic and its Extensions” the widely used Kulldorff’s spatiotemporal scan statistic and its extensions/modifications is described. In section “Monitoring Health Processes Through Scan Statistics”, we review scan statistics or scan-based systems that can be exploited for monitoring health processes. The material covered in this section was split in three parts according to the specific areas/purpose where the techniques are going to be applied for: (a) public health, disease outbreaks, (b) health organizations, and (c) biosurveillance systems. In section “Retrospective Analysis Through Scan Statistics”, we present the use of scan statistics for retrospective analysis. Finally, in the last section, we provide some concluding remarks and directions for further research in the area under review.

Scan Statistics for Linear Sequences and Grids

The scan statistic was first proposed by Naus (1965b) for establishing a decision rule for the multiple hypothesis testing problem (Wagner et al., 2011). For a point process on an interval $[a, b]$, a window $[t, t + w]$ of fixed size $w < b - a$ shifts along the interval by varying t from $t = a$ to $t = b - w$. The maximum number of points in the window is recorded over all possible values of t. For example, let us consider the sequence $SSFFSSFSFSSSSFS$ where $a = 1$ and $b = 15$. We then slide a window $[t, t + 4]$ of size $w = 4$ over the sequence by varying t from $t = 1$ to $t = 11$. The maximum number of successes contained in the moving window over this sequence is 4. In one dimension, the exact distribution of the test statistic is known only for special cases (see, e.g., Ebneshahrashoob et al. 2005 and Wu 2013). Much of the literature has been focused in finding good approximations (see, e.g., Chen and Glaz 1997 and Glaz et al. 2001).

Nagarwalla (1996) proposed a scan statistic with a variable window, whose size does not need to be chosen a priori. The test statistic based on the variable window scan statistic is a generalized likelihood ratio test for a uniform null distribution against an alternative of nonrandom clustering which allows for clusters of variable width.

In higher dimensions the statistical theory becomes more complex. Naus (1965a) obtained distributional bounds for a two-dimensional scan statistic on a square with uniform underlying measure.

Suppose that we are interested in identifying the appearance of specific events in a geographical area $R = [0, L_1] \times [0, L_2]$, $L_1, L_2 > 0$. To study the frequency and the location of the occurrence of the events, we divide R into $n_1 \times n_2$ cells using $n_1 + 1$ horizontal lines and $n_2 + 1$ vertical lines (n_1 and n_2 are two positive integers). Each cell corresponds to a sub-area of R with size $h_1 \times h_2$, where $h_1 = L_1/n_1$ and $h_2 = L_2/n_2$.

Let Y_{ij} for $1 \leq i \leq n_1$ and $1 \leq j \leq n_2$ be a random variable, counting the number of the occurrences of the event under study, in the sub-area $[(i-1) \times h_1, i \times h_1] \times [(j-1) \times h_2, j \times h_2]$. Chen and Glaz (1996) defined the statistic

$$S(i_1, i_2) = \sum_{i=i_1}^{i_1+m_1-1} \sum_{j=i_2}^{i_2+m_2-1} Y_{ij}$$

for $1 \leq i_1 \leq n_1 - m_1 + 1$ and $1 \leq i_2 \leq n_2 - m_2 + 1$ (where m_1 and m_2 are two positive integers, which set the limits of a moving rectangular window). The random variable $S(i_1, i_2)$ tracks the total number of events occurrences in the neighboring area $[(i_1 - 1) \times h_1, (i_1 + m_1 - 1) \times h_1] \times [(i_2 - 1) \times h_2, (i_2 + m_2 - 1) \times h_2]$. The same authors defined the two-dimensional scan statistic as the maximum of $S(i_1, i_2)$ for $1 \leq i_1 \leq n_1 - m_1 + 1$ and $1 \leq i_2 \leq n_2 - m_2 + 1$, i.e.

$$S_{m_1,m_2;n_1,n_2} = max\{S(i_1, i_2) : 1 \leq i_1 \leq n_1 - m_1 + 1, 1 \leq i_2 \leq n_2 - m_2 + 1\};$$

needless to say $S_{m_1,m_2;n_1,n_2}$ provides the maximum number of events in neighboring areas of size $(m_1 h_1) \times (m_2 h_2)$. If no confusion arises about the size of the rectangular region we shall suppress the indices n_1, n_2 and the above scan statistic will be denoted by S_{m_1,m_2}. In the special case $n_1 = n_2 = n$ and $m_1 = m_2 = m$, the simplified notation $S_{m,m}$ will be used.

Based on the approximations for the tail probabilities of the two-dimensional scan statistic, derived in Chen and Glaz (1996), one can test the null hypothesis that the observed events in a two-dimensional region are randomly distributed. In Chen and Glaz (1996), the approximations for the tail probabilities of $S_{m,m}$ are derived for Y_{ij} for $1 \leq i, j \leq n - m + 1$ being independent and identically distributed random variables according to the Bernoulli, binomial, or Poisson distributions. No exact results are available for the distribution of $S_{m,m}$. For the Bernoulli model, Darling and Waterman (1986) derived a Poisson approximation for $P(S_{m,m} = m^2)$. Moreover, they derived a Poisson approximation for higher-dimensional cubes as well. Fu and Koutras (1994) studied Poisson approximations for the Bernoulli model for the special case of $n_1 = n_2 = n$, $m_1 = m_2 = m$, and $k = m^2$, while Sheng and Naus (1996) derived accurate product-type approximations and Bonferroni-type inequalities for the cdf of $S_{m,m;n_1,n_2}$ for the special case of $m_1 = m_2 = m$ and $k = m^2$. Chen and Glaz (1996) restricted to the case of a square region with a square moving window of size $m \times m$, i.e., $n_1 = n_2 = n$ and $m_1 = m_2 = m$, and they derived approximations and inequalities for any value of k for the binomial and the Poisson models. The quasi-Poisson model can be used to handle overdispersion (Zhang et al., 2012).

Kulldorff's Scan Statistic and its Extensions

In this section we discuss Kulldorff's scan statistic and its extensions. More specifically, we will deal with the two types of Kulldorff's scan statistic: the space scan statistic, used for retrospective syndromic surveillance, and the time-space scan statistic, used for prospective syndromic surveillance.

Kulldorff's Spatial Scan Statistic and Extensions

Kulldorff and Nagarwalla (1995) presented a method useful in detecting and carrying out inference for spatial clusters of a disease, based on the likelihood ratio method. Their test can detect clusters of any size, located anywhere in the region under study. It is not restricted to clusters that coincide with administrative or political regions (e.g., prefectures). The test can be used for spatially aggregated data when exact geographic coordinates are known for each disease event. The method was illustrated on a data set describing the occurrence of leukemia in Upstate New York.

Kulldorff (1997) proposed a spatial scan statistic that uses a circular window on the map. This method lets the center of the circle move continuously over the area under study so that in different locations the window includes different sets of adjacent census areas. If the centroid of a census area is contained in the window, then the entire area is included in the window. For each circle centroid, the radius of the circular window changes continuously from zero to a maximum radius, which is defined by the practitioner. In such a way, the circular window is flexible both in location and size.

Conditioning on the observed total number of cases N, Kulldorff (1997) defined the spatial scan statistic as the maximum likelihood ratio over all possible circles Z, namely:

$$S = \frac{\max_Z\{L(Z)\}}{L_0} = \max_Z \left\{ \frac{L(Z)}{L_0} \right\},$$

where $L(Z)$ is the likelihood for circle Z and L_0 is the likelihood under the null hypothesis. Since the likelihood ratio $\frac{L(Z)}{L_0}$ is maximized over all the circles, it determines the cluster that is the most likely one. Then, the p-value for identifying whether the cluster is statistically significant or no is calculated through Monte Carlo techniques (details may be found in Kulldorff 1997).

Kulldorff's Spatiotemporal Scan Statistic and Extensions

A space-time scan statistic, useful for evaluating space-time cluster alarms was presented in Kulldorff et al. (1998). This scan statistic is defined through a cylindrical window with a circular geographic base. Its height corresponds to time. The base of the cylinder is centered around one of the various possible centroids, which are located throughout the area under study, with the radius constantly varying in size. The window then moves both in space and time so that for every possible location and size it also visits every possible time interval. In fact, we have an infinite number of overlapping cylinders, which differ in size and shape, and cover the entire area of interest. Each cylinder reflects a possible cluster and the space-time scan statistic accounts for the preselection bias and multiple testing inherent in a cluster alarm. The method was illustrated on brain cancer cluster alarms in Los Alamos, New Mexico.

Kulldorff (2001) proposed a method for regular time periodic disease surveillance to detect currently "active" geographical clusters of disease, using a space-time scan statistic. The method tests the statistical significance of such clusters adjusting for the multitude of possible geographical locations and sizes, time intervals, and time periodic analyses. The method was illustrated on thyroid cancer among men in New Mexico, 1973–1992.

Boscoe et al. (2003) proposed a technique for visualizing the results of Kulldorff's spatial scan statistic and related cluster detection methods that provide a greater degree of informational content. By combining likelihood ratio and relative risk, it is possible to identify sub-clusters of higher (or lower) relative risk among broader regional excesses or deficits. The result is a map with a nested or contoured appearance. The technique was applied to prostate cancer mortality data in counties within the contiguous United States during the period 1970–1994. The resulting map shows both broad and localized patterns of excess and deficit, which complements a choropleth map of the same data.

Kulldorff et al. (2003) proposed a tree-based scan statistic, by which the surveillance can be conducted with minimum prior assumptions about the group of occupations/drugs that increase risk and which adjusts for the multiple testing inherent in the many potential combinations. Kulldorff et al. (2006) explored an elliptic version of the spatial scan statistic, using a scanning window of variable location, shape (eccentricity), angle, and size, with and without an eccentricity penalty.

Duczmal et al. (2006) proposed a modification of the simulated annealing spatial scan statistic that incorporates the concept of "non-compactness" in order to penalize clusters that are very irregular in shape.

Kulldorff et al. (2007) presented multivariate scan statistics for disease surveillance. They first presented an extension of the spatial and space-time scan statistic that incorporates multiple data sets into a single likelihood function. In this way, a signal is generated whether it occurs in only one or in multiple data sets. More specifically, the authors defined the combined log likelihood as the sum of the individual log likelihoods for those data sets for which the observed case count exceeds the expected. They also presented an extension, by combining likelihoods from different data sets in order to adjust for covariates.

Jung et al. (2007) proposed a spatial scan statistic for ordinal data. This test statistic is based on a likelihood ratio test, and it was evaluated by the aid of Monte Carlo hypothesis testing. The proposed method was illustrated using prostate cancer grade and stage data from the Maryland Cancer Registry.

Huang et al. (2007) proposed a spatial scan statistic based on an exponential model which is capable of handling both uncensored and censored continuous survival data. This scan statistic performs well for different survival distribution functions including the exponential, gamma, and lognormal distributions. The cluster detection method was illustrated using survival data for men diagnosed with prostate cancer in Connecticut from 1984 to 1995.

Cucala (2008) proposed a new technique for identifying clusters in temporal point processes. This relies on the comparison between all the m-order spacings, and it does not rely on any alternative hypothesis.

Cucala (2009) proposed a spatial method for identifying clusters in spatial point processes. It relies on a specific ordering of events and on the definition of area spacings (i.e., intervals between consecutive ordered points) which have the same distribution as one-dimensional spacings. Then the spatial clusters are detected using a scan statistic adapted to the analysis of one-dimensional point processes. He applied these methods to a data set describing the spatial distribution of the larynx cancers and the lung cancers recorded between 1973 and 1984 in the Chorley Ribble area in Lancashire, UK.

Monitoring Health Processes Through Scan Statistics

In this section we present several cases where scan statistics have been used to monitor health processes. A plausible categorization of health monitoring would consider three different levels (see Bersimis et al. 2018b): individuals level (personalized level), health organizations level (e.g., primary health care units, hospitals, etc.), and community (public health) level. However, to the best of our knowledge, there exist no references in the bibliography about the problem of health monitoring at individuals level.

Using Scan Statistics for Public Health Surveillance

In this subsection we review the literature on the use of scan statistics for public health surveillance. We have divide the review into three categories: (i) spatial and spatiotemporal scan statistics, (ii) combination of scan statistics with control charts, and (iii) Bayesian scan statistics.

Spatial and Spatiotemporal Scan Statistics

Thirty years ago, Wallested et al. (1989) proposed a test for time-space clustering based on a scan statistic defined as the maximum number of events in a 365-day period in each of several geographic units. The authors argued that, unlike other time-space clustering method, the scan statistic allows the calculation of attributable risk and effect size measures.

Duczmal and Assuncao (2004) proposed a new graph-based strategy for the detection of spatial clusters of arbitrary geometric form in a map of geo-referenced populations and cases. Their test statistic was based on the likelihood ratio test previously introduced by Kulldorff and Nagarwalla (1995) for circular clusters. The new technique of adaptive simulated annealing was exploited for treating the problem of finding the local maxima of a certain likelihood function over the subspace of the connected subgraphs of the graph associated to the regions of interest.

Ozdenerol et al. (2005) examined the spatial and population characteristics of low birth-weight babies using two different cluster estimation techniques: Kulldorff's spatial scan statistic and Rushton's spatial filtering technique across spatial filters (circle) of increasing size. They concluded that the two methods give different clusters in terms of population characteristics and geographic area within clusters. The simultaneous use of the two methods provides more details about the population and spatial features of each cluster.

Kleinman et al. (2005) proposed a model-based method for adjusting the space-time scan statistic so that it accommodates both temporal and geographical variation in syndromic event rates and examined its impact on syndromic surveillance. In addition, they presented an example of syndromic surveillance of lower respiratory infection (LRI). This pertains to bioterrorism surveillance given that a case of anthrax in the initial phase would include symptoms that would probably cause it to be classified into the LRI syndrome.

Kulldorff et al. (2005) proposed a prospective space-time permutation scan statistic for the early detection of disease outbreaks which uses only case numbers, and does not require the availability of additional population-at-risk data. It makes minimal assumptions about the time, geographical location, or size of the outbreak, and it adjusts for natural purely spatial and purely temporal variation. The new method was evaluated using daily analyses of hospital emergency department visits in New York City. Four of the five strongest signals were likely local precursors to citywide outbreaks due to rotavirus, norovirus, and influenza. The number of false signals was at most modest.

Kulldorff et al. (2006) explored an elliptic version of the spatial scan statistic, using a scanning window of variable location, shape (eccentricity), angle and size, with and without an eccentricity penalty. The method was applied to breast cancer mortality data from Northeastern United States and female oral cancer mortality in the United States.

Naus and Wallenstein (2006) proposed monitoring scan statistic methods in the case of temporal-Bernoulli and temporal-exponential data. More specifically, this approach uses a window whose width is constant over the review period without the need to simulate p-values. The P-scan, as it is called, exploits an approximation formula which was available for the calculation of the p-values in the constant background case to deduce the respective p-value for the nonconstant background case.

Duczmal and Buckeridge (2006) proposed a modification of the spatial scan statistic that takes into account the workflow, namely, the transportations of individuals between home and work. The objective was to detect clusters of disease in situations where exposure occurs in the workplace, but only home address is available for analysis. They described an extension of the usual Kulldorff's spatial scan statistic that uses workflow data to search for disease clusters resulting from workplace exposure.

Kulldorff et al. (2007) presented an extension of the spatial and space-time scan statistic that simultaneously incorporates multiple data sets into a single likelihood function, so that a signal is generated whether it occurs in only one or in multiple data sets. This is done by defining the combined log likelihood as the sum of the individual log likelihoods for those data sets for which the observed case count exceeds the expected. They also presented another extension, where the concept of combining likelihoods from different data sets is used to adjust for covariates.

Kedem and Wen (2007) merged a semi-parametric density ratio model with Kulldorff's scan procedure to obtain a fairly general cluster detection procedure. The idea is quite natural since the semi-parametric density ratio model is ideal for testing equidistribution given two or more samples. They illustrated their method on a real data set from Kulldorff's website which contains information for childhood leukemia and lymphoma in North Humberside, England, between 1974 and 1986.

Tango (2008) designed a new spatial scan statistic by modifying the likelihood ratio so that it scans only the regions with elevated risk. This statistic is free from the property of detection of a most likely cluster that is much larger than the true cluster by absorbing neighboring regions with non-elevated risk of disease occurrence. The circular spatial scan statistic, compared through Monte Carlo simulations with Kulldorff's original one, was shown to better identify the whole of, or a part of, the true cluster. The proposed circular spatial scan statistic was illustrated on mortality data from cerebrovascular disease in the areas of Tokyo Metropolis and Kanagawa prefecture in Japan.

Takahashi et al. (2008) proposed a flexibly shaped space-time scan statistic for early detection of disease outbreaks. The advantage of this scan statistic is that it is capable to detect noncircular areas.

Woodall et al. (2008) reviewed some prospective scan-based methods that are used in health-related applications to detect increased rates of mortality or morbidity and to detect bioterrorism or active disease clusters. They illustrated that the recurrence interval metric (i.e., the length of time for which the expected number of signals is 1) does not reflect some important aspects of the statistical performance of scan-based, and other, surveillance methods.

Tsui et al. (2008) reviewed existing data collection and surveillance systems and popular surveillance methods in healthcare and disease surveillance, which involve in healthcare, public health, and syndromic surveillance. According to the authors, the most common disease spread monitoring methods can be categorized into temporal, spatial, and spatiotemporal surveillance techniques. Among the popular health surveillance methods are the scan statistics developed by Kulldorff (1997) and the temporal and spatiotemporal techniques of Kulldorff (2001).

Neill (2009) evaluated the detection performance of 12 variants of spatial scan, using synthetic outbreaks injected into 4 real-world public health data sets.

Huang et al. (2009), motivated by growing demands to study the spatial heterogeneity of continuous measures in population data, such as mortality rates, survival rates, average body mass indexes, and pollution at state, county, and census tract levels, proposed a weighted normal scan statistic for investigating the clusters of the cells (geographic units such as counties) with unusual high/low continuous regional measures, where the weights reflect the uncertainty of the regional measures or sample size (number of observed cases) in the cells. Power, precision, the effect of the weights, and the sensitivity of the proposed test statistic to data from various distributions were investigated through intensive simulation. The method was applied to 1988–2002 stage I and II lung cancer survival data in Los Angeles County in order to search for clusters of geographic units with high/low survival rates in a short-term/long-term survival after diagnosis and to 1999–2003 breast cancer age-adjusted mortality rate data in the United States (collected by the Surveillance, Epidemiology, and End Results (SEER) program) in order to evaluate the clustering patterns of counties with high mortality rate.

Kulldorff et al. (2009) stated that temporal, spatial, and space-time scan statistics are commonly used to detect and evaluate the statistical significance of temporal and/or geographical disease clusters, without any prior assumptions on the location, time period, or size of those clusters. Scan statistics are mostly used for count data, such as disease incidence or mortality. However, if someone is interested in finding clusters with respect to a continuous variable (e.g., lead levels in children or low birth weight), he can use a scan statistic where the likelihood is calculated using the normal probability model. Other distributions may also be considered. The authors applied their method to find geographical clusters of low birth weight in New York City.

A novel working definition of an outbreak based on temporal and spatial clustering of molecular genotypes was introduced in the paper of Gallego et al. (2009). In order to account for time and space correlations within a set of disease counts the authors used the space-time permutation scan statistic defined by Kulldorff et al. (2005).

Zhang et al. (2010) proposed a sequential version of the spatial scan statistic to adjust for the presence of other clusters in the study region. The procedure removes the effect due to the more likely clusters on less significant clusters by sequential deletion of the previously detected clusters.

Jung et al. (2010) noted that, as a geographical cluster detection analysis tool, the spatial scan statistic has been developed for different types of data such as Bernoulli, Poisson, ordinal, exponential, and normal. Another interesting data type is the multinomial data. For example, one may wish to find clusters where the disease-type distribution is statistically significantly different from the rest of the study region when there are different types of diseases. They proposed a spatial scan statistic for such data, which is useful for geographical cluster detection analysis when categorical data with more than two attributes are available and no information exists for the attributes. The proposed method was applied to meningitis data consisting of five different disease categories to identify areas with distinct disease-type patterns in two counties in the United Kingdom. The performance of the method was evaluated through a simulation study.

Leibovici et al. (2011) presented a generic exploratory framework to detect clusters of bivariate or multivariate associations between categorical variables which describe one or more populations. The scan is performed in three steps: (1) a neighborhood of each point is delineated from a condition of sufficiency; (2) a local statistic based on co-occurrences counts for spatial association within this neighborhood is computed, and its value is assigned to the point; and (3) an assessment for hot-spots on the obtained statistical map is performed.

Tango et al. (2011) extended the space-time scan statistic of Kulldorff (2001) and the flexible space-time scan statistic of Takahashi et al. (2008). Their methods focuses on (i) comparing the observed number of cases with the unconditional expected number of cases, (ii) taking a time-to-time variation of Poisson mean into account, and (iii) implementing an outbreak model to capture localized emerging disease outbreaks more timely and correctly.

The paper of Chen et al. (2011) reviewed temporal, spatial, and spatial-temporal aberration detection techniques that can be used to facilitate the early detection of infectious disease outbreaks that can occur in nonrandom yet clustered distributions in GIS (geographic information systems)-based syndromic surveillance systems. The focus of the paper is on approaches, which are appropriate for prospective surveillance data.

A new "fast subset scan" approach for accurate and computationally efficient event detection in massive data sets was presented by Neill et al. (2013). The authors treated the detection of events as a search over subsets of data records, which looks for the subset which maximizes a score function. They proved that many commonly used functions (e.g., Kulldorff's spatial scan statistic and extensions) satisfy the "linear time subset scanning" property, enabling exact and efficient optimization over subsets. They also demonstrated that proximity-constrained subset scans substantially improve the timeliness and accuracy of event detection; their method manages to detect disease outbreaks 2 days faster than other competitive existing methods.

Tango and Takahashi (2012) introduced a flexible spatial scan statistic implemented with a restricted likelihood ratio proposed by Tango (2008). This flexible spatial scan statistic was shown to abolish the constraint of the maximum of 30 nearest neighbors for searching the cluster candidates that the flexible spatial scan statistic proposed by Tango and Takahashi (2005) has. Furthermore, it has surprisingly much less computational time than the original flexible spatial scan statistic. In addition, it can detect clusters of any shape reasonably well.

Neill et al. (2013) presented new subset scan methods for multivariate event detection in massive space-time data sets. They extended the "fast subset scan" framework (Neill et al., 2013) from univariate to multivariate data, enabling computationally efficient detection of irregular space-time clusters even when the numbers of spatial locations and data streams are large. For two variants of the multivariate subset scan, they demonstrated that the scan statistic can be efficiently optimized over proximity-constrained subsets of locations and over all subsets of the monitored data streams, enabling timely detection of emerging events and accurate characterization of the affected locations and streams. Using their new algorithms, the authors performed an empirical comparison of the Subset Aggregation and Kulldorff multivariate subset scans on synthetic data and real-world disease surveillance tasks, demonstrating tradeoffs between the detection and characterization performance of the two methods.

Bhatt and Tiwari (2014) proposed a scan statistic for survival data based on the Weibull distribution in order to determine if there are geographical clusters of people with shorter survival time than the expected; should this happen, one might infer that there exists inadequate treatment or health practices in the regions where the scan statistic attains extreme values. Their scan statistic can also be used for other survival distributions, such as exponential, gamma, and lognormal.

Faires et al. (2014) investigated the efficiency of the temporal scan statistic in detecting Clostridium difficile infection (CDI) clusters and determining if there are significant differences in the rate of CDI cases by month, season, and year in a community hospital. CDI clusters were investigated using a retrospective temporal scan test statistic. Statistically significant clusters were compared to known CDI outbreaks within the hospital. A negative binomial regression model was used to identify associations between year, season, and month and the rate of CDI cases.

Jung and Cho (2015) proposed a nonparametric spatial scan statistic based on the Wilcoxon rank-sum test statistic and compared the performance of their method to parametric models via a simulation study under various scenarios.

A new spatial scan statistic was proposed by Cucala et al. (2017) for multivariate data indexed in space. As many other scan methods, it relies on a generalized likelihood ratio, but it also takes into account the correlations between variables. The resulting spatial scan test was proved to be more powerful than the independent version, whatever the level of correlation between variables. The method was applied to a data set recording the levels of pollutant metals in the area of Lille, France.

Combination of Scan Statistics with Control Charts

Sonesson (2007) combined ideas from space-time disease surveillance, public health surveillance, and Statistical Process Control (SPC) to detect emerging clusters, in the context of prospective surveillance. He demonstrated how the space-time scan statistics methods suggested by Kulldorff (2001) can be fitted into a general cumulative sum (CUSUM) framework.

The Ph.D dissertation of Fraker (2007) was devoted to the assessment of scan methods for monitoring of public health surveillance data. The author evaluated Kulldorff's temporal scan method and compared it with other surveillance schemes, such as the c-chart and the Poisson CUSUM chart.

Joner et al. (2008) investigated the properties of the scan statistic when used in prospective surveillance of the incidence rate under the assumption of independent Bernoulli observations. They compared the performance of the prospective scan statistic method with the Bernoulli-based CUSUM technique. They indicated that the latter tends to be more effective in detecting sustained increases in the rate, but the scan method is preferable in some applications due to its simplicity and the relatively small loss of efficiency.

Han et al. (2010) carried out a comparison of the performance of three detection methods: the temporal scan statistic, CUSUM, and exponential weighted moving average (EWMA) when the observations follow the Poisson distribution. Their simulation study revealed that the Poisson CUSUM and EWMA charts outperform, in general, the Poisson scan statistic methods.

Bayesian Scan Statistics

Neill and Cooper (2010) presented the multivariate Bayesian scan statistic (MBSS), a general framework for event detection and characterization in multivariate spatial time series data. The multivariate Bayesian scan statistic is an extension of the univariate Bayesian scan statistic (Neill et al., 2006) so as the event detection framework becomes functional for multiple data streams and multiple event types. The MBSS integrates prior information and observations from multiple data streams in a principled Bayesian framework, computing the posterior probability of each type of event in each space-time region. The MBSS uses a multivariate Gamma-Poisson model built from historical data and models the effects of each event type on each stream using expert knowledge or labeled training examples.

Jiang et al. (2010) developed a new Bayesian-network-based spatial scan statistic, called BNetScan, which models the relationships among the events of interest and the observable events using a Bayesian network. BNetScan is an entity-based Bayesian network that models the underlying state and observable variables for each individual in a population.

Neill (2011) extended the MBSS framework to enable detection and visualization of irregularly shaped clusters in multivariate data, by defining a hierarchical prior over all subsets of locations. Although a naive search over the exponentially many subsets would be computationally infeasible, the author demonstrated that the total posterior probability that each location has been affected can be efficiently computed, thereof enabling a rapid detection and visualization of irregular clusters. The performance of the method was tested on semisynthetic outbreaks injected into real-world emergency department data from Allegheny County, Pennsylvania.

Charras-Garrido et al. (2013) introduced a Bayesian disease mapping model, producing continuous estimations of the risks that requires a post-processing classification step to obtain clearly delimited risk zones. A risk partition model that produces a classification of the risk levels in a one-step procedure was also applied, and working with point data, they focused on the scan statistic clustering method.

Cançado et al. (2014) proposed extensions to the Poisson scan, namely, the zero-inflated Poisson (ZIP)-scan and the ZIP+EM-scan, in order to model excess zeroes (i.e., realizations zero outcomes in excess of the frequency predicted under the Poisson probability model) using the ZIP model. The ZIP+EM-scan is used when there are unknown structural zeros and the expectation-maximization (EM) algorithm is used to estimate them. Following the same procedures, they also derived a zero-inflated binomial (ZIB) scan.

Cançado et al. (2017) combining previous approaches (Hall 2000; Staubach et al. 2002; Cançado et al. 2014) proposed another extension of the spatial scan statistic, the Bayesian ZIB (BZIB) scan, which is a Bayesian methodology developed to detect spatial clusters for zero-inflated binomial data.

Scan Statistics Used at Health Organization Level

As far as hospital or health organizations' management is concerned, it is useful to analyze data both retrospectively and prospectively.

Ismail et al. (2003) developed a test for Poisson data based on scan statistics and applied it to monitor the occurrence of orthopedic wound infection and Methicillin-resistant Staphylococcus aureus colonization. They concluded that the test is sensitive in detecting changes in the process parameter which may not be detected by standard control chart methods.

Recently, Bersimis et al. (2017a) introduced a methodology for monitoring bivariate health processes when the intervention outcome may be classified as "absolutely successful," "with minor but acceptable complications," and "unsuccessful due to severe complications." The monitoring procedure uses appropriate two-dimensional scan rules.

A similar approach is used by Bersimis et al. (2017b, 2018a), in the context of monitoring performance and assessing competence in health services and clinical trials, respectively.

Biosurveillance Systems Exploiting Scan Statistics

Burkom (2003) presented a biosurveillance system that uses scan statistics with multiple, disparate data sources. More specifically, they employed Kulldorff's scan statistic as implemented in the SaTScan software.

Sabhnani et al. (2005) described a biosurveillance system designed to detect anomalous patterns in pharmacy retail data. The system monitors national-level over-the-counter (OTC) pharmacy sales on a daily basis. Fast space-time scan statistics are used to detect disease outbreaks, and user feedback is incorporated to improve system's utility and usability. This system uses data from the National Retail Data Monitor (NRDM), developed and operated by the RODS (Real-time

Outbreak and Disease Surveillance) Laboratory at the University of Pittsburgh, and receives the OTC data from the national and local vendors. The data consists of daily store level sales of 9000 OTC products used for the symptomatic treatment of infectious diseases. The NRDM classifies individual product sales into 18 groups of similar products (e.g., baby/child electrolytes, cough/cold, thermometers, stomach remedies, internal analgesics, etc.).

Takahashi et al. (2010) developed the FleXScan software which analyzes spatial count data using Kulldorff's circular spatial scan statistic and the flexible spatial scan statistic developed by Tango and Takahashi (2005). In addition, it makes use of a spatial scan statistic with a restricted likelihood ratio proposed by Tango and Takahashi (2012). It is similar to the SaTScan software (2008) developed by Kulldorff and Information Management System Inc. FleXScan uses the Poisson model, where the number of events in an area is Poisson distributed according to a known underlying population at risk. It can also analyze data under the Binomial model. The data may be either aggregated at the census tract, zip-code, county or other geographical level. FleXScan can adjust for the underlying inhomogeneity of a background population and for any number of categorical covariates provided by the user.

Retrospective Analysis Through Scan Statistics

Hjalmars et al. (1996) tested a large set of childhood leukemia cases for the presence of geographical clusters, using the spatial scan statistic of Kulldorff (1997). The same statistic was also used by Kulldorff et al. (1997) to identify breast cancer clusters in the Northeast United States. The authors assumed that the number of deaths in each county was Poisson distributed.

Sabel et al. (2003) examined 1,000 cases of amyotrophic lateral sclerosis distributed throughout Finland who died between June 1985 and December 1995. Using a spatial-scan statistic (SaTScan), the authors examined whether there are significant clusters of the disease with respect to time of birth and time of death. Two significant, neighboring clusters were identified in southeast and south-central Finland with respect to the time of death. A single significant cluster was identified in southeast Finland at the time of birth, closely matching one of the clusters identified with respect to the time of death.

Sheehan and DeChello (2005) conducted an epidemiologic study to determine whether the observed variations in the proportion of breast cancers diagnosed at late stage are simply random or are statistically significant with respect to both geographical location and time. They performed space-time analyses, through SaTScan, assuming that the proportion of late stage cases follows a Bernoulli distribution.

Klassen et al. (2005) used the spatial scan statistic approach of Kulldorff (1997), in combination with predicted block group-level disease patterns from multilevel models, to examine geographic variation in prostate cancer grade and stage at diagnosis.

Costa and Kulldorff (2009) surveyed a wide variety of fields where spatial scan statistics can be used. These include early detection of disease outbreaks, epidemiology, psychology, veterinary medicine, brain imaging, demography, archeology, astronomy, criminology, ecology, forestry, geology, and history.

Huang et al. (2010) applied a space-time permutation scan statistic to microbiology data from patients admitted to an academic medical center in United States; the management of data was performed using the WHONET-SaTScan laboratory information software from the World Health Organization (WHO) Collaborating Centre for Surveillance of Antimicrobial Resistance.

Odoi et al. (2004) investigated the presence of local giardiasis clusters and investigated the extent to which livestock density and manure application on agricultural land might explain the "rural" effect. A spatial scan statistic was used to identify spatial clusters and geographical correlation analysis was practiced to explore associations of giardiasis rates with manure application on agricultural land and livestock density. All GIS manipulations and cartographic displays were performed in ArcView GIS. A spatial scan statistic implemented in SaTScan was used to test for the presence of giardiasis spatial clusters and identify their approximate locations.

Hsu et al. (2004) exploited the spatial scan statistic to examine the geographic excess of breast cancer mortality by race in Texas counties between 1990 and 2001. The test was conducted with a maximum scan window size set to 90% of the study period and a spatial cluster size equal to 50% of the population at risk. The next scan was conducted with the "purely spatial" option of SatScan to verify whether the excess mortality persisted further. Spatial queries were performed to locate the regions of excess mortality affecting multiple racial groups.

Scan statistics were also incorporated into the syndromic surveillance system presented by Lombardo et al. (2003), Heffernan et al. (2004), and Yih et al. (2004).

Discussion

Scan statistics play a critical role in several scientific areas. Perhaps, the most interesting field of application of scans is in process monitoring. Under this framework, healthcare, medicine, psychology, ecology, geology, criminology, architecture, reliability, and engineering are a few of the fields where scan statistics have been fruitfully applied. The utility of scan statistics in monitoring healthcare processes is highlighted by the large number of papers regarding this topic in bibliography (more than 80, in the last 20 years).

The aim of the present work was to review the main advances on the use of scan statistics in the healthcare sector. The papers were classified into four categories: (i) papers proposing the use of scan statistics for public health surveillance and detection of disease outbreaks, (ii) papers dealing with the use of scan statistics in health organizations, (iii) papers presenting biosurveillance systems which exploit scan statistics, and (iv) papers presenting specific applications of scan statistics.

The plausible categorization of health monitoring into three different levels (i.e., individuals level, health organizations level, and community level) provides a natural guide to future research. For example, our literature review revealed that no publications exist on health monitoring at individuals level. However, the use of scan statistics for monitoring biochemical markers or other patients' characteristics is crucial for patients' health and the public health as well. For example, scan statistics could be used for monitoring a patient's blood volume; a statistically significant decrease on the volume will result to an alert that will initiate a search for a tumor.

References

Balakrishnan N, Koutras MV (2011) Runs and scans with applications. Wiley, New York

Bersimis S, Chalkias C, Anthopoulou T (2014) Detecting and interpreting clusters of economic activity in rural areas using scan statistic and Lisa under a unified framework. Appl Stoch Models Bus Ind 30(5):573–587

Bersimis S, Sachlas A, Castagliola P (2017a) Controlling bivariate categorical processes using scan rules. Methodol Comput Appl Probab 19(4):1135–1149

Bersimis S, Sachlas A, Papaioannou T (2018a) Monitoring phase II comparative clinical trials with two endpoints and penalty for adverse events. Methodol Comput Appl Probab 20(2):719–738

Bersimis S, Sachlas A, Sparks R (2017b) Performance monitoring and competence assessment in health services. Methodol Comput Appl Probab 19(4):1169–1190

Bersimis S, Sgora A, Psarakis S (2018b) The application of multivariate statistical process monitoring in non-industrial processes. Qual Technol Quantit Manag 15(4):526–549

Bhatt V, Tiwari N (2014) A spatial scan statistic for survival data based on weibull distribution. Stat Med 33(11):1867–1876

Boscoe FP, McLaughlin C, Schymura MJ, Kielb CL (2003) Visualization of the spatial scan statistic using nested circles. Health Place 9(3):273–277

Burkom HS (2003) Biosurveillance applying scan statistics with multiple, disparate data sources. J. Urban Health 80(1):i57–i65

Cançado AL, da Silva CQ, da Silva MF (2014) A spatial scan statistic for zero-inflated poisson process. Environ Ecol Stat 21(4):627–650

Cançado AL, Fernandes LB, da Silva CQ (2017) A Bayesian spatial scan statistic for zero-inflated count data. Spat Stat 20:57–75

Charras-Garrido M, Azizi L, Forbes F, Doyle S, Peyrard N, Abrial D (2013) On the difficulty to delimit disease risk hot spots. Int J Appl Earth Obs Geoinf 22:99–105

Chen D, Cunningham J, Moore K, Tian J (2011) Spatial and temporal aberration detection methods for disease outbreaks in syndromic surveillance systems. Ann GIS 17(4):211–220

Chen H, Zeng D, Yan P (2010) Infectious disease informatics: syndromic surveillance for public health and bio-defense. Springer, Boston

Chen J, Glaz J (1996) Two-dimensional discrete scan statistics. Stat Probab Lett 31(1):59–68

Chen J, Glaz J (1997) Approximations and inequalities for the distribution of a scan statistic for 0-1 bernoulli trials. Adv Theory Pract Stat 1:285–298

Cheung YTD, Spittal MJ, Williamson MK, Tung SJ, Pirkis J (2013) Application of scan statistics to detect suicide clusters in australia. PLOS ONE (1):1–11

Costa MA, Kulldorff M (2009) Applications of spatial scan statistics: a review. Birkhäuser, Boston, pp 129–152

Cucala L (2008) A hypothesis-free multiple scan statistic with variable window. Biom J 50(2): 299–310

Cucala L (2009) A flexible spatial scan test for case event data. Comput Stat Data Analy 53(8): 2843–2850

Cucala L, Genin M, Lanier C, Occelli F (2017) A multivariate gaussian scan statistic for spatial data. Spat Stat 21:66–74

Darling R, Waterman M (1986) Extreme value distribution for the largest cube in a random lattice. SIAM J Appl Math 46(1):118–132

Duczmal L, Assuncao R (2004) A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. Comput Stat Data Anal 45(2):269–286

Duczmal L, Buckeridge DL (2006) A workflow spatial scan statistic Stat Med 25(5):743–754

Duczmal L, Kulldorff M, Huang L (2006) Evaluation of spatial scan statistics for irregularly shaped clusters. J Comput Graph Stat 15(2):428–442

Ebneshahrashoob M, Gao T, Wu M (2005) An efficient algorithm for exact distribution of discrete scan statistics. Methodol Comput Appl Probab 7(4):459–471

Faires MC, Pearl DL, Ciccotelli WA, Berke O, Reid-Smith RJ, Weese JS (2014) The use of the temporal scan statistic to detect methicillin-resistant staphylococcus aureus clusters in a community hospital. BMC Infect Dis 14(1):375

Fraker SE (2007) Evaluation of scan methods used in the monitoring of public health surveillance data. PhD thesis, Virginia Tech

Fu J, Koutras M (1994) Poisson approximations for 2-dimensional patterns. Ann Inst Stat Math 46(1):179–192

Gallego B, Sintchenko V, Wang Q, Hiley L, Gilbert GL, Coiera E (2009) Biosurveillance of emerging biothreats using scalable genotype clustering. J Biomed Inform 42(1):66–73

Glaz J, Naus J, Wallenstein S (2001) Scan statistics. Graduate texts in mathematics. Springer, New York/London

Hall DB (2000) Zero-inflated poisson and binomial regression with random effects: a case study. Biometrics 56(4):1030–1039

Han SW, Tsui K-L, Ariyajunya B, Kim SB (2010) A comparison of CUSUM, EWMA, and temporal scan statistics for detection of increases in poisson rates. Qual Reliab Eng Int 26(3):279–289

Hanslik T, Boelle P-Y, Flahault A (2001) The control chart: an epidemiological tool for public health monitoring. Public Health 115(4):277–281

Heffernan R, Mostashari F, Das D, Karpati A, Kulldorff M, Weiss D et al (2004) Syndromic surveillance in public health practice, New York city. Emerg Infect Dis 10(5):858–864

Hjalmars U, Kulldorff M, Gustafsson G, Nagarwalla N (1996) Childhood leukaemia in Sweden: using GIS and a spatial scan statistic for cluster detection. Stat Med 15(7–9):707–715

Hsu CE, Jacobson H, Mas FS (2004) Evaluating the disparity of female breast cancer mortality among racial groups-a spatiotemporal analysis. Int J Health Geogr 3(1):4

Huang L, Kulldorff M, Gregorio D (2007) A spatial scan statistic for survival data. Biometrics 63(1):109–118

Huang L, Tiwari RC, Zou Z, Kulldorff M, Feuer EJ (2009) Weighted normal spatial scan statistic for heterogeneous population data. J Am Stat Assoc 104(487):886–898

Huang SS, Yokoe DS, Stelling J, Placzek H, Kulldorff M, Kleinman K, O'Brien TF, Calderwood MS, Vostok J, Dunn J et al (2010) Automated detection of infectious disease outbreaks in hospitals: a retrospective cohort study. PLoS Med 7(2):e1000238

Imanishi M, Newton AE, Vieira AR, Gonzales-Aviles G, Kendall Scott ME, Manikonda K, Maxwell TN, Halpin JL, Freeman MM, Medalla F et al (2015) Typhoid fever acquired in the united states, 1999–2010: epidemiology, microbiology, and use of a space-time scan statistic for outbreak detection. Epidemiol Infect 143(11):2343–2354

Ismail NA, Pettitt AN, Webster RA (2003) "online" monitoring and retrospective analysis of hospital outcomes based on a scan statistic. Stat Med 22(18):2861–2876

Jiang X, Neill DB, Cooper GF (2010) A Bayesian network model for spatial event surveillance. Int J Approx Reason 51(2):224–239

Joner MD Jr, Woodall WH, Reynolds MR Jr (2008) Detecting a rate increase using a bernoulli scan statistic. Stat Med 27(14):2555–2575

Jung I, Cho HJ (2015) A nonparametric spatial scan statistic for continuous data. Int J Health Geogr 14(1):30

Jung I, Kulldorff M, Klassen AC (2007) A spatial scan statistic for ordinal data. Stat Med 26(7): 1594–1607

Jung I, Kulldorff M, Richard OJ (2010) A spatial scan statistic for multinomial data. Stat Med 29(18):1910–1918

Kedem B, Wen S (2007) Semi-parametric cluster detection. J Stat Theory Pract 1(1):49–72

Klassen AC, Kulldorff M, Curriero F (2005) Geographical clustering of prostate cancer grade and stage at diagnosis, before and after adjustment for risk factors. Int J Health Geogr 4(1):1

Kleinman K, Abrams A, Kulldorff M, Platt R (2005) A model-adjusted space-time scan statistic with an application to syndromic surveillance. Epidemiol Infect 133(3):409–419

Koutras M, Alexandrou V (1995) Runs, scans and urn model distributions: a unified Markov chain approach. Ann Inst Stat Math 47(4):743–766

Kulldorff M (1997) A spatial scan statistic. Commun Stat Theory Methods 26(6):1481–1496

Kulldorff M (2001) Prospective time periodic geographical disease surveillance using a scan statistic. J R Stat Soc Ser A (Stat Soc) 164(1):61–72

Kulldorff M, Athas WF, Feurer EJ, Miller BA, Key CR (1998) Evaluating cluster alarms: a space-time scan statistic and brain cancer in Los Alamos, New Mexico. Am J Public Health 88(9): 1377–1380

Kulldorff M, Fang Z, Walsh SJ (2003) A tree-based scan statistic for database disease surveillance. Biometrics 59(2):323–331

Kulldorff M, Feuer EJ, Miller BA, Freedma LS (1997) Breast cancer clusters in the northeast united states: A geographic analysis. Am J Epidemiol 146(2):161–170

Kulldorff M, Heffernan R, Hartman J, Assunção R, Mostashari F (2005) A space-time permutation scan statistic for disease outbreak detection. PLoS Med 2(3):e59

Kulldorff M, Huang L, Konty K (2009) A scan statistic for continuous data based on the normal probability model. Int J Health Geogr 8(1):58

Kulldorff M, Huang L, Pickle L, Duczmal L (2006) An elliptic spatial scan statistic. Stat Med 25(22):3929–3943

Kulldorff M, Mostashari F, Duczmal L, Yih W, Kleinman K, Platt R (2007) Multivariate scan statistics for disease surveillance. Stat Med 26(8):1824–1833

Kulldorff M, Nagarwalla N (1995) Spatial disease clusters: detection and inference. Stat Med 14(8):799–810

Leibovici DG, Bastin L, Anand S, Hobona G, Jackson M (2011) Spatially clustered associations in health related geospatial data. Trans GIS 15(3):347–364

Lombardo J, Burkom H, Elbert E, Magruder S, Lewis SH, Loschen W, Sari J, Sniegoski C, Wojcik R, Pavlin J (2003) A systems overview of the electronic surveillance system for the early notification of community-based epidemics (essence II). J Urban Health 80(1):i32–i42

Nagarwalla N (1996) A scan statistic with a variable window. Stat Med 15(7-9):845–850

Naus JI (1965a) Clustering of random points in two dimensions. Biometrika 52(1/2):263–267

Naus JI (1965b) The distribution of the size of the maximum cluster of points on a line. J Am Stat Assoc 60(310):532–538

Naus J, Wallenstein S (2006) Temporal surveillance using scan statistics. Stat Med 25(2):311–324

Neill DB (2009) An empirical comparison of spatial scan statistics for outbreak detection. Int J Health Geogr 8(1):20

Neill DB (2011) Fast Bayesian scan statistics for multivariate event detection and visualization. Stat Med 30(5):455–469

Neill DB, Cooper GF (2010) A multivariate Bayesian scan statistic for early event detection and characterization. Mach Learn 79(3):261–282

Neill DB, McFowland E, Zheng H (2013) Fast subset scan for multivariate event detection. Stat Med 32(13):2185–2208

Neill DB, Moore AW, Cooper GF (2006) A Bayesian spatial scan statistic. In: Advances in neural information processing systems, pp 1003–1010

Odoi A, Martin SW, Michel P, Middleton D, Holt J, Wilson J et al (2004) Investigation of clusters of giardiasis using GIS and a spatial scan statistic. Int J Health Geogr 3(1):11

Ozdenerol E, Williams BL, Kang SY, Magsumbol MS (2005) Comparison of spatial scan statistic and spatial filtering in estimating low birth weight clusters. Int J Health Geogr 4(1):19

Sabel CE, Boyle P, Löytönen M, Gatrell AC, Jokelainen M, Flowerdew R, Maasilta P (2003) Spatial clustering of amyotrophic lateral sclerosis in finland at place of birth and place of death. Am J Epidemiol 157(10):898–905

Sabhnani MR, Neill DB, Moore AW (2005) Detecting anomalous patterns in pharmacy retail data. In: Data mining methods anomaly detection, vol 58

Sheehan TJ, DeChello LM (2005) A space-time analysis of the proportion of late stage breast cancer in Massachusetts, 1988 to 1997. Int J Health Geogr 4(1):15

Sheng K, Naus J (1996) Matching fixed rectangles in 2-dimension. Stat Probab Lett 26(1):83–90

So HC, Pearl DL, von Königslöw T, Louie M, Chui L, Svenson LW (2013) Spatiotemporal scan statistics for the detection of outbreaks involving common molecular subtypes: using human cases of escherichia coli o157:h7 provincial pfge pattern 8 (national designation ecxai.0001) in Alberta as an example. Zoonoses Public Health 60(5):341–348

Sonesson C (2007) A CUSUM framework for detection of space-time disease clusters using scan statistics. Stat Med 26(26):4770–4789

Staubach C, Schmid V, Knorr-Held L, Ziller M (2002) A bayesian model for spatial wildlife disease prevalence data. Prev Vet Med 56(1):75–87

Takahashi K, Kulldorff M, Tango T, Yih K (2008) A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring. Int J Health Geogr 7(1):14

Takahashi K, Yokoyama T, Tango T (2010) FleXScan user guide

Tango T (2008) A spatial scan statistic with a restricted likelihood ratio. Jpn J Biom 29(2):75–95

Tango T, Takahashi K (2005) A flexibly shaped spatial scan statistic for detecting clusters. Int J Health Geogr 4(1):11

Tango T, Takahashi K (2012) A flexible spatial scan statistic with a restricted likelihood ratio for detecting disease clusters. Stat Med 31(30):4207–4218

Tango T, Takahashi K, Kohriyama K (2011) A space-time scan statistic for detecting emerging outbreaks. Biometrics 67(1):106–115

Taylor SR, Arrowsmith SJ, Anderson DN (2010) Detection of short time transients from spectrograms using scan statistics. Bull Seismol Soc Am 100(5A):1940–1951

Tsui K-L, Chiu W, Gierlich P, Goldsman D, Liu X, Maschek T (2008) A review of healthcare, public health, and syndromic surveillance. Qual Eng 20(4):435–450

Wagner M, Moore A, Aryel R (2011) Handbook of biosurveillance. Elsevier Science, Burlington

Wallested S, Gould MS, Kleinmaw M (1989) Use of the scan statistic to detect time-space clustering. Am J Epidemiol 130(5):1057–1064

Woodall WH, Brooke Marshall J, Joner MD Jr, Fraker SE, Abdel-Salam A-SG (2008) On the use and evaluation of prospective scan methods for health-related surveillance. J R Stat Soc Ser A (Stat Soc) 171(1):223–237

Wu T-L (2013) On finite Markov chain imbedding and its applications. Methodol Comput Appl Probab 15(2):453–465

Yih WK, Caldwell B, Harmon R, Kleinman K, Lazarus R, Nelson A, Nordin J, Rehm B, Richter B, Ritzwoller D et al (2004) National bioterrorism syndromic surveillance demonstration program. Morb Mortal Weekly Rep 53:43–49

Zhang Z, Assunção R, Kulldorff M (2010) Spatial scan statistics adjusted for multiple clusters. J Probab Stat 2010:1–11

Zhang T, Zhang Z, Lin G (2012) Spatial scan statistics with overdispersion. Stat Med 31(8): 762–774