A new adaptive procedure for multiple window scan statistics

Tung-Lung Wu *, Joseph Glaz
Department of Mathematics and Statistics, Mississippi State University, USA
Department of Statistics, University of Connecticut, USA

A R T I C L E   I N F O
Article history:
Received 1 July 2013
Received in revised form 25 August 2014
Accepted 6 September 2014
Available online 16 September 2014

Keywords:
Adaptive procedure
Clusters
Multiple window scan statistics
Sequential
Window size
Two-dimensional

A B S T R A C T
Scan statistics have been widely applied to test for unusual cluster of events in many scientiﬁc areas. It has been of practical interest on how to select the window size of a scan statistic. An adaptive procedure for multiple window scan statistics is proposed and the distributions are studied for independent identically distributed Bernoulli trials and uniform observations on \((0, 1)\) in one-dimensional case. The idea of the procedure is to select the window sizes sequentially. An initial window size is chosen and the subsequent window sizes are then determined, depending on the value of the current scan statistic at each stage. The power of scan statistics based on the adaptive procedure is compared with power of standard scan statistics. Numerical results and applications for disease clusters detection are given to illustrate our procedure.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

The literature of scan statistics has grown rapidly in the past decade. Many applications involving scan statistics in one or two dimensions have been found in many areas. In one-dimensional case, the scan statistic has been applied by Hoh and Ott (2000) and Lachenbruch et al. (2005) to locate susceptibility genes and to monitor manufactured blood products, respectively. In two-dimensional case, a spatial scan statistic has been studied in detail by Kulldorff (1997) and applied in epidemiology. In systems reliability, Barbour et al. (1996) and Yamamoto and Akiba (2005) studied the reliability of 2-dimensional consecutive \(k\)-out-of-\(n\) system and 2-dimensional rectangular \(k\)-within-consecutive-\((r, s)\)-out-of-\((m, n)\):F system, respectively. When the components can only have two outcomes, success and failure, the reliability of a 2-dimensional system is equivalent to a 2-dimensional scan statistic problem. Other examples are such as image analysis (Rosenfeld, 1978), signal detection in a sensor network (Song et al., 2012) and DNA sequence analysis (Karwe and Naus, 1997). Two books by Glaz et al. (2009) and Glaz and Balakrishnan (1999) provide an excellent overview for recent developments and advances in scan statistics.

The generalized likelihood ratio test will reject the null hypothesis of homogeneity against a cluster alternative if the scan statistic is too large. To apply a fixed window scan statistic, one needs to specify in advance the window size, while the cluster size is usually unknown. Loader (1991) and Nagarwalla (1996) generalized the fixed window scan statistic to a variable window scan statistic where all windows sizes in a given interval are considered. The price for this generalization is that the distribution of a variable window scan statistic can only be obtained using Monte Carlo simulation which is usually time-consuming. In another slightly different direction, Naus and Wallenstein (2004) and Wu et al. (2013) studied multiple window scan statistics, where a fixed number of window sizes are given, and provided the approximate and exact distributions.

In this paper, we propose an adaptive procedure for multiple window scan statistics without specifying the window sizes except for the initial one. Following the procedure, the data-dependent window sizes are determined sequentially according...
to the value of the current fixed window scan statistic at each stage. The procedure continues until a certain criterion is met. To illustrate the statistic of interest, consider a conditional continuous 2-window scan statistic. Let the initial window size be $r_1$ and the implementation of an adaptive procedure relies on the following probability

$$P(S_1(r_1) \geq s_1 \mid f(S_1(r_1))) \geq s_2),$$

where $S_1(r)$ is the conditional scan statistic with window size $r$ given the total number of points equal to $N$. $f$ can be any suitably chosen real-valued function and $s_1$ and $s_2$ are chosen to achieve a given significance level. Throughout this paper, the distribution of an adaptive scan statistic is of a similar form in (1).

In Section 2, the general adaptive procedure is described. The approximations of the distributions of the adaptive 2-stage scan statistics for one-dimensional case are also given, including conditional continuous scan statistics and conditional and unconditional discrete scan statistics. The extension to two-dimensional case is given in Section 3. A simple algorithm to find the two-dimensional scan statistic for a given data set is provided. The power comparison between adaptive scan statistics and classic fixed window scan statistics is given in Section 4. In Section 5, applications for disease clusters detection are given to illustrate our procedure for both one and two-dimensional cases. Summary and discussion are given in Section 6.

2. One-dimensional case

Let $N(t)$ be a Poisson process with intensity $\lambda$ on $[0, 1)$. For $0 \leq r < 1$, let $S(r, t) = N(t + r) - N(t)$ denote the number of events that have occurred in the interval $[t, t + r)$, where $r$ is the window size. The unconditional continuous scan statistic is defined as

$$S(r) = \sup_{0 \leq t < 1-r} S(r, t).$$

Given the total number of events $N(1) = N$, the dependency on $\lambda$ is removed. The $N$ arrival times are independent uniformly distributed on $[0, 1)$ and (2) becomes a conditional continuous scan statistic, denoted by $S_1(r)$. As the number of events is usually known, the conditional continuous scan statistic (uniform observations) has been widely studied and therefore is the focus in the continuous cases in this paper. The discrete version of conditional and unconditional scan statistics can be defined similarly.

Given a function $f$, the adaptive procedure for a one-dimensional multiple window scan statistic is described as follows:

(i) choose an initial window size $r_1$;
(ii) obtain the value of the scan statistic, say, $S_1(r_1) = s_1$;
(iii) choose $r_2 = f(s_1)$ as the next window size;
(iv) obtain the value of the scan statistic, say, $S_1(r_2) = s_2$;
(v) repeat steps (iii) and (iv) until finished.

In this paper, we consider two types of adaptive procedures. The first one is the $k$-stage adaptive procedure where the procedure will stop until $k$ window sizes have been specified. The distribution of the resulting scan statistic is

$$P(S_1(r_1) \geq s_1 \mid S_1(r_2) \geq s_2 \mid \ldots \mid S_1(r_k) \geq s_k),$$

where $r_j = f(S_1(r_{j-1})), j = 2, \ldots, k$. The special case $k = 2$ is given in (1) and will be studied in detail for both continuous and discrete cases. The second adaptive procedure is to repeat the step (v) until a certain criterion is satisfied. For example, a $p$-value is calculated for the fixed window scan statistic at each stage. We continue the procedure as long as the successive $p$-values are non-increasing (Hoh and Ott, 2000). In other words, the procedure will stop if the $p$-value starts increasing. We call this the minimum $p$-value procedure.

2.1. Continuous case: uniform observations

Given a function $f$, we derive an approximation for the distribution in (1) when the observations are taken from the uniform distribution on $[0, 1)$. Throughout this section, the specific function $f(S_1(r_1)) = r_1 S_1(r_1)/E(S_1(r_1))$ is discussed, where $E(S_1(r_1))$ is the expected value of $S_1(r_1)$. For simplicity, the second window size $f(S_1(r_1))$ induced by the value of the first scan statistic is denoted by $r_2$ throughout this paper.

A 2-stage procedure. Given $N(1) = N$, the $N$ points are randomly distributed according to the uniform distribution on $[0, 1)$. An adaptive 2-stage procedure for a continuous scan statistic is described as follows:

(i) choose an initial window size $r_1$;
(ii) obtain the value of the scan statistic, say, $S_1(r_1) = s_1$;
(iii) choose $r_2 = r_1 S_1/E(S_1(r_1))$ as the next window size;
(iv) obtain the value of the scan statistic, say, $S_1(r_2) = s_2$.

Based on the above procedure, the probability we are interested in is given by

$$P(S_1(r_1) < s_1, S_1(r_2) < s_2) = \sum_{x=0}^{s_1-1} P(S_1(r_2) < s_2 | S_1(r_1) = x) P(S_1(r_1) = x).$$

Note that $r_2$ depends on $r_1$. First we assume $r_2 < r_1$. The probability $P(S_1(r_1) = x)$ has been obtained by many authors either exactly or approximately. For example, the exact probabilities for limited values of $r_1$ are tabulated in Neff and Naus (1980).
However, in the paper, we use the finite Markov chain imbedding technique in Fu et al. (2010) to approximate \( P(S_N(r_1) = x) \). The idea is to approximate a (conditional) continuous scan statistic by a (conditional) discrete one, where distributions of scan statistics are interpreted as waiting time distributions of patterns and can be obtained exactly. Fu et al. (2010) have also shown that the approximation performs well for moderately small discretization step size. We use the argument in Naus and Wallenstein (2004) to approximate the conditional probability \( P(S_N(r_2) \geq s_2 | S_N(r_1) = x) \), i.e.

\[
P(S_N(r_2) \geq s_2 | S_N(r_1) = x) \approx \sum_{i=s_2}^{x-1} b(i-1; x-1, \frac{r_2}{r_1}) + P(S_{N-1} \geq s_2).
\]

where \( b(s; x, p) \) is the binomial probability of \( s \) successes in \( x \) trials with probability of success \( p \). Note that, in their paper, Naus and Wallenstein (2004) used the right hand side of (5) to approximate \( P(S_N(r_2) \geq s_2 | S_N(r_1) = x) \). Their argument neglecting the event \( S_N(r_1) > x \) provides a better approximation for \( P(S_N(r_2) \geq s_2 | S_N(r_1) = x) \) instead. The procedure sometimes requires to compute \( E(S_N(r_1)) \approx \sum_{x=1}^{m} P(S_N(r_1) \geq x) \). A small value of \( m \) relative to \( r_1 \) and \( N \) usually provides a very good approximation.

If \( r_2 > r_1 \), we propose to approximate the conditional probability \( P(S_N(r_2) \geq s_2 | S_N(r_1) = x) \) by

\[
P(S_N(r_2) \geq s_2 | S_N(r_1) = x) \approx \sum_{i=s_2}^{N-x} b(i-x; N-x, \frac{r_2-r_1}{1-r_1}) + P(S_{N-s_2+1} \geq s_2).
\]

We will make use of (4)–(6) to select critical values \( s_1 \) and \( s_2 \) and implement the test at a given significance level \( \alpha \). Table 1 provides the numerical comparison of approximation (6) and simulation for three sets of different values of \( r_1, r_2, s_2 \) and \( x \), which are chosen to produce reasonable tail probabilities. The approximation (6) performs quite well, especially for small probabilities. We observe that the binomial probability provides a good approximation for the conditional probability in (6) when \( N \) is small and \( s_2 \) is large. For example, in Table 1, the probability 0.01055 under the case \( N = 10 \) is solely approximated by the binomial probability because the second term of the approximation is zero. Table 2 provides critical values for selected adaptive 2-stage scan statistics with approximate significance levels close to 0.01 and 0.02. Note that the approximation of \( P(S_N(r_1) \geq s_1 \text{ or } S_N(r_2) \geq s_2) \) involves (5) and (6) depending on \( r_1 > r_2 \) and \( r_2 > r_1 \), respectively. All simulations are performed with 100,000 replicates.

### 2.2. Discrete case: Bernoulli trials

We describe an adaptive procedure for multiple window discrete scan statistics here and derive an approximation of the distributions of the adaptive 2-stage scan statistics under Bernoulli trials. Throughout this section, the function \( f(x) = x \) is discussed.

A 2-stage procedure. We first study the unconditional case. Let \( S_n(r) \) denote the fixed window discrete scan statistic of window size \( r \) in a sequence of \( n \) independent Bernoulli trials \( \{X_i\}_{i=1}^n \) with probability of success \( p \). An adaptive 2-stage procedure for a discrete scan statistic follows the same steps as in the continuous case with the function \( f \) replaced by \( f(S_n(r_1)) = S_n(r_1) \):

(i) choose an initial window size \( r_1 \);
(ii) obtain the value of the scan statistic, say, \( S_n(r_1) = s_1 \);
(iii) choose \( r_2 = s_1 \) as the next window size;
(iv) obtain the value of the scan statistic, say, \( S_n(r_2) = s_2 \).

Following the procedure, the probability that we need to approximate is

\[
P(S_n(r_1) < s_1, S_n(r_2) < s_2) = \sum_{x=0}^{s_1-1} P(S_n(r_2) < s_2 | S_n(r_1) = x)P(S_n(r_1) = x).
\]
where the hypergeometric function. The value of  

Note again that since  

Two-dimensional case (conditional continuous scan statistic)  

The accuracy of the approximation (8) is shown in Table 3.  

<table>
<thead>
<tr>
<th>n</th>
<th>r₁</th>
<th>s₁</th>
<th>s₂</th>
<th>Approximation</th>
<th>Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>25</td>
<td>7</td>
<td>6</td>
<td>0.016371</td>
<td>0.016152</td>
</tr>
<tr>
<td>500</td>
<td>15</td>
<td>6</td>
<td>4</td>
<td>0.016449</td>
<td>0.014410</td>
</tr>
<tr>
<td>100</td>
<td>20</td>
<td>5</td>
<td>4</td>
<td>0.038674</td>
<td>0.038587</td>
</tr>
</tbody>
</table>

The conditional scan statistics are often used in retrospective investigations. The probability of success  plays an important role in unconditional scan statistics. The dependency on the probability  is removed given the total number of successes  over the region  

Given a window size  and the total number of successes  , we denote by  the conditional discrete scan statistic. To implement an adaptive 2-stage procedure for a conditional discrete scan statistic, we need to evaluate  

The probability  can be obtained exactly using Theorem 2.1 in Fu et al. (2010). Using a similar argument in continuous case, the conditional probability can be approximated via  

The accuracy of the approximation (8) is shown in Table 3.  

### 3. Two-dimensional case (conditional continuous scan statistic)  

In two-dimensional case, we consider a Poisson process on a rectangular region  . Without loss of generality, let  . Let  be the number of points covered by a rectangle of size  with lower left corner located at  , i.e.  . Then the two-dimensional continuous scan statistic is defined as  

It is known that given the total number of points  in the unit square  , where  are the  random points, where  are independent uniformly distributed on  . For simplicity, we let  and the two-dimensional conditional scan statistic is denoted by  .  

As in Section 2, the function  can be chosen according to the design. One can choose, for example,  . Both the -stage and the minimum -value procedures for one-dimensional case can be directly extended to the two-dimensional case. For example, the tail probability of an adaptive 2-stage two-dimensional scan statistic is given by  

The distribution of an adaptive two-dimensional scan statistic will be assessed through simulations. For continuous case, it can be very time-consuming to search the maximum number of points in any rectangle of size  over the region  . A simple algorithm concerning the extreme sets introduced by Alm (1999) is used to find the scan statistic within finite number of searches. We describe a simple algorithm to examine all extreme sets consisting of two types of rectangles. First, two random samples  and  are generated from the uniform distribution on  .
Parameters are chosen so that the significance level is close to 0.048. To be specific, tic with a smaller initial windowsize 54 and twofixed windows can statistic of windowsizes 10 and 28. The parameters are choosen to achieve the significance level around 0.068. The resulting probabilities are $P(S_{90}(100) \geq 20$ or $S_{90}(23) \geq 7|H_0) \approx 0.068$, $P(S_{90}(23) \geq 8|H_0) \approx 0.068$ and $P(S_{90}(25) \geq 13|H_0) \approx 0.068$. Table 5 compares another adaptive 2-stage scan statistic with a smaller initial windowsize 54 and two fixed window scan statistics of window sizes 10 and 28. The parameters are choosen so that the significance level is close to 0.048. To be specific, $P(S_{90}(54) \geq 15$ or $S_{90}(r_2) \geq 5|H_0) \approx 0.0481$, $P(S_{90}(10) \geq 5|H_0) \approx 0.0486$ and $P(S_{90}(28) \geq 8|H_0) \approx 0.0495$. The cluster sizes range from 5 to 40.

Although the adaptive scan statistics are not always the best among the scan statistics considered, overall they perform well. In Table 4, when the cluster size is small or large ($d = 5, 10$ or 100), the adaptive scan statistic performs the best, while its performance for moderate $d$ is satisfactory. This suggests that the classic fixed window scan statistic will perform poorly if the window size is mis-specified, and an adaptive $k$-stage scan statistic is a good alternative when the cluster size is unknown.

4.2. Two-dimensional conditional continuous case

Power comparison of adaptive scan statistics with classic fixed window scan statistics is carried out under various pulse alternatives (see Naus and Wallenstein, 2004 for one-dimensional case). The pulse alternative is given that the $N$ random points in the unit square are taken from the following density function:

$$ f(x) = \begin{cases} 
\theta & \text{for } x \in [b, b + w) \times [b, b + w), \\
\frac{1}{1 - w^2 + \theta w^2} & \text{for } x \in [0, 1) \times [0, 1) \setminus [b, b + w) \times [b, b + w).} 
\end{cases} $$

Numerical results for comparison of adaptive 2-stage scan statistics with fixed window scan statistics are given in Tables 6–9. The power simulations are carried out in a similar way to Naus and Wallenstein (2004) for various pulse sizes and

<table>
<thead>
<tr>
<th>Table 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power comparison of $S_{90}^{\text{adaptive}}(100)$, $S_{90}(23)$ and $S_{90}(56)$.</td>
</tr>
<tr>
<td>$d$</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power comparison of $S_{90}^{\text{adaptive}}(54)$, $S_{90}(10)$ and $S_{90}(28)$.</td>
</tr>
<tr>
<td>$d$</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>15</td>
</tr>
</tbody>
</table>

Then, $X_i = (X_{ij}, X_{ij})$, $i = 1, \ldots, n$, are the $N$ random points in the unit square. Let $B = \{[X_{ij}, X_{ij} + w) \times [X_{ij}, X_{ij} + w) : 0 \leq X_{ij} - X_{ij} \leq w, 0 \leq X_{ij} - X_{ij} \leq w, i, j = 1, \ldots, n\}$ be the collection of extreme sets associated with $S_{90}(w)$. We only need to examine the rectangles in $B$, and the maximum number of points in those rectangles is the scan statistic.

4. Power comparison

Two scenarios, one for one-dimensional unconditional discrete case and the other for two-dimensional conditional continuous case, are studied to show the performance of the adaptive procedure.

4.1. One-dimensional unconditional discrete case

The power comparison of adaptive scan statistics and classic fixed window scan statistics is given for one-dimensional unconditional discrete case here. A pulse alternative, proposed by Wallenstein et al. (1994), is considered. Let $p_i$ be the probability of success of $i$th trial, $i = 1, \ldots, n$. The hypotheses are given by

$H_0 : p_i = p_0, \quad i = 1, \ldots, n,$

$vs. \quad H_1 : p_i = p_0, \quad i = 1, \ldots, \tau - 1, \tau + d, \ldots, n,$

$p_i = p_1 > p_0, \quad i = \tau, \ldots, \tau + d - 1,$

where $\tau$ is unknown and $d$ is the pulse (cluster) size.

Let $p_0 = 0.1(H_0)$ and $p_1 = 0.3(H_1)$. The powers are evaluated based on simulations, each with 100,000 runs. Table 4 compares the adaptive 2-stage scan statistic with an initial window size 100, denoted by $S_{90}^{\text{adaptive}}(100)$, and two fixed window scan statistics of window sizes 23 and 56. The cluster size ranges from 5 to 100. The parameters are choosen to achieve the significance level around 0.068. The resulting probabilities are $P(S_{90}(100) \geq 20$ or $S_{90}(23) \geq 7|H_0) \approx 0.068$, $P(S_{90}(23) \geq 8|H_0) \approx 0.068$ and $P(S_{90}(56) \geq 13|H_0) \approx 0.068$. Table 5 compares another adaptive 2-stage scan statistic with a smaller initial window size 54 and two fixed window scan statistics of window sizes 10 and 28. The parameters are choosen so that the significance level is close to 0.048. To be specific, $P(S_{90}(54) \geq 15$ or $S_{90}(r_2) \geq 5|H_0) \approx 0.0481$, $P(S_{90}(10) \geq 5|H_0) \approx 0.0486$ and $P(S_{90}(28) \geq 8|H_0) \approx 0.0495$. The cluster sizes range from 5 to 40.

Although the adaptive scan statistics are not always the best among the scan statistics considered, overall they perform well. In Table 4, when the cluster size is small or large ($d = 5, 10$ or 100), the adaptive scan statistic performs the best, while its performance for moderate $d$ is satisfactory. This suggests that the classic fixed window scan statistic will perform poorly if the window size is mis-specified, and an adaptive $k$-stage scan statistic is a good alternative when the cluster size is unknown.

4.2. Two-dimensional conditional continuous case

Power comparison of adaptive scan statistics with classic fixed window scan statistics is carried out under various pulse alternatives (see Naus and Wallenstein, 2004 for one-dimensional case). The pulse alternative is given that the $N$ random points in the unit square are taken from the following density function:
percentages of observations. The percentages are used to decide whether or not a random number from \([0, 1]\) was assigned to the pulse region. The purpose of this numerical study is to show the performances of adaptive scan statistics and fixed window scan statistics when the window size was mis-specified.

Suppose the total number of points is \(N = 50\). In Table 6, the initial window size for the 2-stage scan statistic is \(r_1 = 0.1\) chosen to be larger than the actual pulse sizes and the adaptive window size is \(r_2 = r_1 S_N(r_1) / E_s(S_N(r_1))\), where \(E_s(S_N(r_1))\) is the expected value of \(S_N(r_1)\) under the pulse alternative with \(w = 0.1\) and 20\% chance a random point will be assigned to the pulse region, while the window size of the fixed window scan statistic is 0.0235 chosen to be smaller than the actual pulse sizes. The critical values are chosen such that \(P(S_5(0.1) \geq 6)\) or \(S_5(0.0235) \geq 4\) \(\approx 0.0468\) and \(P(S_5(0.0235) \geq 3) \approx 0.0477\). In Table 7, the initial window size for the 2-stage scan statistic is \(r_1 = 0.06\) chosen to be smaller than the actual pulse sizes and \(r_2 = 3r_1 S_N(r_1)/5\), while the window size of the fixed window scan statistic is 0.157 chosen to be larger than the actual pulse sizes. The critical values are chosen such that \(P(S_5(0.06) \geq 5)\) or \(S_5(0.0235) \geq 4\) \(\approx 0.0308\) and \(P(S_5(0.0235) \geq 3) \approx 0.0309\). In Table 8, the initial window size for the 2-stage scan statistic is \(r_1 = 0.05\) and \(r_2 = 3r_1 S_N(r_1)/4\). The window sizes of two fixed window scan statistic are 0.44 and 0.072 chosen to be close to \(r_1\). The critical values are chosen such that \(P(S_5(0.05) \geq 5)\) or \(S_5(0.044) \geq 4\) \(\approx 0.0213\) and \(P(S_5(0.044) \geq 3) \approx 0.0228\). In Table 9, adaptive scan statistics with initial window sizes over a range of values on either side of the true pulse size and \(r_2 = 3r_1 S_N(r_1)/5\) are examined when the actual pulse size is 0.1 and the percentage is 10\% under \(H_a\).

Tables 6 and 7 show that the adaptive scan statistics provide a good power when the initial window size is either smaller or larger than the actual cluster size. This indicates that the proposed procedure can quickly adapt the observed scan statistic to search the actual cluster, when the cluster size has been mis-specified in advance. In Table 8, the initial window size of the adaptive scan statistic and the window sizes of two fixed window scan statistics are chosen to be similar. Overall, the adaptive scan statistic performs well except for the pulse size 0.08. However, the pulse size 0.08 is more close to the window size 0.072 of the fixed window scan statistic than others. Similar results can be found in Table 9. For example, while...
the window size 0.06 (adaptive) is further away from the true pulse size 0.1 than the window size 0.075 (fixed), both the adaptive and fixed window scan statistics perform similarly.

5. Applications

Applications for both one and two-dimensional cases are given to illustrate the procedure.

One-dimensional. We apply our procedure to Knox’s (1959) data containing 2 groups of children with oesophageal atresia and/or tracheo-oesophageal fistula. These two data sets are analyzed using adaptive 2-stage scan statistics. The first data set consists of 35 cases of oesophageal atresia and/or tracheo-oesophageal fistula occurred in Birmingham Children’s Hospital, UK, between 1950 and 1955 (2191 days). The study started on January 1, 1950. The number of days on which each case occurred is given below:

| 170 | 316 | 445 | 468 | 938 | 1034 | 1128 | 1233 | 1248 | 1249 |
| 1252 | 1259 | 1267 | 1305 | 1385 | 1388 | 1390 | 1446 | 1454 | 1458 |
| 1461 | 1491 | 1583 | 1699 | 1702 | 1787 | 1924 | 1974 | 2049 | 2051 |
| 2067 | 2075 | 2108 | 2151 | 2174 |

This data set has been analyzed by many researchers, for example Nagarwalla (1996) and Naus and Wallenstein (2004). Both of them were able to identify a cluster of length 16, beginning with day 1233. To apply our adaptive procedure, let the initial window size be \( r_1 = 1/6 \) (window of a year) and \( r_2 = r_1 S_{35}(1/6)/E(S_{35}(1/6)) \), where \( E(S_{35}(1/6)) \) is estimated to be 10.6. The thresholds \( s_1 \) and \( s_2 \) are chosen to set the significance level close to 0.02 and we find

\[
P(S_{35}(1/6) \geq 15 \text{ or } S_{35}(r_2) \geq 16) \approx 0.028,
\]

where \( r_2 = S_{35}(1/6)/63.6 \). The numerical values are evaluated based on 100,000 simulation runs. Naus and Wallenstein’s (2004) multiple window scan statistic includes our initial window size 1/6, so our procedure can identify the same cluster detected by their multiple procedure with respect to the window size 1/6. From the data, the adaptive window size \( r_2 = 16/63.6 \) is determined from the value of the first fixed window scan statistic \( S_{35}(1/6) = 16 \). We then find \( S_{35}(0.2516) = 18 \) which is above the threshold 16. The second window identifies 3 clusters of length 18, beginning with days 1034, 1233 and 1248.

Let \( U_1, \ldots, U_N \) be an ordered sample of size \( N \) from the uniform distribution on \( [0, 1] \). Given a window size \( r \), the multiple scan statistic (Chen and Glaz, 2005) is defined as

\[
\xi_{m,N} = \sum_{j=1}^{N-m} I_j,
\]

where

\[
I_j = \begin{cases} 
1 & \text{if } U_{(j+m)} - U_{(j)} \leq r, \\
0 & \text{otherwise}.
\end{cases}
\]

The multiple scan statistic should be examined if there are presence of multiple clusters. Our finding is highly significant as \( P(\xi_{18,35} \geq 3) \approx 0.0015 \).

The second data set consists of 63 cases gathered in the Newcastle hospitals over a 9-year period (3287 days). The number of days on which each case occurred is given below:

| 46 | 372 | 377 | 422 | 504 | 528 | 597 | 675 | 698 | 1056 |
| 1143 | 1292 | 1301 | 1317 | 1718 | 1724 | 1794 | 1834 | 1834 | 1856 |
| 1862 | 1865 | 1878 | 1909 | 1910 | 1912 | 1940 | 1962 | 2026 | 2029 |
| 2080 | 2128 | 2144 | 2197 | 2219 | 2260 | 2325 | 2329 | 2332 | 2336 |
| 2384 | 2461 | 2518 | 2528 | 2603 | 2626 | 2691 | 2692 | 2789 | 2819 |
| 2832 | 2923 | 3070 | 3071 | 3084 | 3096 | 3101 | 3148 | 3177 | 3182 |
| 3211 | 3254 | 3277 |

To analyze this data set, we choose an initial window of 2-year and \( r_1 = 2/9 \). The simulated expected value of \( S_{63}(2/9) \) is \( E(S_{63}(2/9)) \approx 20.6 \). The thresholds \( s_1 \) and \( s_2 \) are chosen to achieve a significance level close to 0.02 and we find

\[
P(S_{63}(2/9) \geq 28 \text{ or } S_{63}(r_2) \geq 29) \approx 0.0223.
\]

From the second data set, we find \( S_{63}(2/9) = 27 < 28 \), and the second window size is then given by \( r_2 = (27/20.6)(2/9) = 0.29126 \). We find that \( S_{63}(0.29126) = 32 \). Three clusters of length 32 are detected by the adaptive scan statistic, beginning with days 1718, 1794 and 1834. The finding for the second data set is also highly significant as we examine the multiple scan statistic which gives \( P(\xi_{32,63} \geq 3) \approx 0.0010 \).

Hoh and Ott (2000) suggested a sequential procedure to test a range of cluster widths until a \( p \)-value of the fixed window scan statistic starts increasing. Let the initial window size \( r_1 \) be 1/9. At the \( j \)th stage, the window size is set to be
Table 10

The adaptive scan statistic with Hoh and Ott’s procedure for Knox’s second dataset.

<table>
<thead>
<tr>
<th>i\th stage</th>
<th>( r_i )</th>
<th>( S_N(r_i) )</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1/9</td>
<td>17</td>
<td>0.0246</td>
</tr>
<tr>
<td>2</td>
<td>0.1458</td>
<td>20</td>
<td>0.0220</td>
</tr>
<tr>
<td>3</td>
<td>0.1888</td>
<td>26</td>
<td>0.0018</td>
</tr>
<tr>
<td>4</td>
<td>0.2670</td>
<td>31</td>
<td>0.0040</td>
</tr>
</tbody>
</table>

Fig. 1. Plot of the 58 cases of laryngeal cancer.

\[ r_j = r_{j-1} S_N(r_{j-1}) / E(S_N(r_{j-1})), j = 2, 3, \ldots \]  
Hoh and Ott’s procedure applied on the above adaptive scan statistic stops at the fourth stage with minimum p-value \( p_{\text{min}} = 0.0018 \). The summary of the test results is given in Table 10. The overall significance level associated with the test statistic \( p_{\text{min}} \) is 0.00001 based on a simulation of 100,000 runs.

Two-dimensional. We analyze a two-dimensional data set in Diggle et al. (1990) using both 2-stage and minimum \( p \)-value procedures. The data set contains the locations of 58 cases of cancer of the larynx in a district of Lancashire. An interesting phenomenon can be seen in the left panel of Fig. 1 that there were 4 cases of cancer located near an incinerator which is a possible source of pollution and a potential source of clusters of laryngeal cancer.

For the adaptive 2-stage procedure, the initial square window size is set to be \( r_1 = 0.05 \) and the adaptive window size is set to be \( r_2 = r_1 S_{58}(r_1) / E(S_{58}(r_1)) \). From the data, we find \( S_{58}(r_1) = 5 \) and the adaptive window size is \( r_2 = 0.0893 \) with \( S_{58}(0.0893) = 7 \). The simulated p-value is 0.0032. The detected clusters are highlighted in the right panel of Fig. 1. The minimum \( p \)-value procedure stops at the third stage. At the 1st-stage, we find \( S_{58}(0.05) = 5 \) with \( p \)-value = 0.0031 and continue to the second stage. At the second stage, we find \( S_{58}(0.0893) = 7 \) with \( p \)-value = 0.0018 < 0.0031 and continue to the third stage. At the third stage, we find \( S_{58}(0.1563) = 10 \) with \( p \)-value = 0.0027 > 0.0018 and stop. The simulated global \( p \)-value is 0.0007.

In the analysis of the two-dimensional data, we do not use the information that 4 cases are near the incinerator, since we would most likely obtain a significant result using a small window size. It is interesting that the 2-stage procedure detects several clusters, but they are not near the incinerator. The minimum \( p \)-value procedure results in a smaller overall \( p \)-value.

6. Summary and discussion

We propose an adaptive procedure for multiple window scan statistics. The idea is to choose the next window size according to the value of the current fixed window scan statistic. The adaptive 2-stage scan statistics for both discrete and continuous cases are studied in detail and approximations of the distributions of the adaptive scan statistics are also given for one-dimensional case. Based on the 2-stage procedure, it is then easy to extend the procedure to k-stage by repeating the steps (iii) and (iv) in Section 2 until all \( k \) window sizes have been specified. In an adaptive k-stage (\( k \geq 3 \)) procedure, the test needs to be implemented based on simulations.

The choice of the function \( f \) determining the next window size is not the focus of this paper, nevertheless we provide a general idea to select the function \( f \). Ideally, the function \( f \) should lead the window size toward the size of a true cluster and give size values of almost all potential clusters. For example in discrete case, we select a decreasing function \( f(S_n(r_1)) = S_n(r_1) \) when we start with a large window size. Accordingly, if one starts with a small initial window size, then a suitable function \( f \) should be an increasing function, otherwise the scan statistic will never detect a cluster of size larger than the initial window size. In Section 4.2, the implementation of one of the adaptive scan statistics depends on the probability \( P(S_{50}(0.05) < 5 \) and \( S_{50}(r_2) < 7 \). By using \( f(S_n(r_1)) = 3r_1 S_n(r_1) / 4 \) with an initial window size 0.05, the detectable size
range is between 0 and 0.15. Thus, one may choose a function $f$ according to the desired detectable range. Another reasonable choice of the function is $f(S_n(r_1)) = S_n(r_1)/E_{a}(S_n(r_1))$ where $E_{a}(S_n(r_1))$ is the expected value of the fixed window scan statistic of window size $r_1$ under the targeted alternative hypothesis. This function adjusts the next window size to be smaller or larger according to the value of the current fixed window scan statistic which reflects the potential cluster size.

In essence, the adaptive scan statistic is similar to the multiple window scan statistic in Naus and Wallenstein (2004) (or the variable window scan statistic in Nagarwalla (1996)). The main difference is that the window sizes of an adaptive scan statistic are chosen sequentially and are data-dependent, while others have to be specified in advance. Ideally, the adaptive procedure will select the window size close to the cluster size, if existing. It has been illustrated in the applications. In analyzing Knox’s (1959) data, we employ an adaptive 2-stage scan statistic with the initial windows size $1/6$ for the first dataset. The initial window detects clusters of length 16 which coincides the findings in Naus and Wallenstein (2004), and furthermore the second adaptive window detects clusters of length 18. For the second data set gathered over a 9-year period, we choose the initial window size $r_1 = 2/9$. It turns out that the initial window is not able to detect any cluster. However, the second window adjusted by $S_{a3}(2/9) = 27$ detects three clusters of length 32 which are highly significant with simulated $p$-value = 0.009. The second data set is again analyzed using the minimum $p$-value procedure, the finding is highly significant with global significance level 0.00001.

Two adaptive procedures are introduced. In the two-dimensional application, the minimum $p$-value procedure produces a smaller overall $p$-value. However, it only tests whether there exist clusters, while the 2-stage procedure can also identify the location. In general, the minimum $p$-value procedure is recommended for testing the existence of clusters. However if one also needs to know the locations of clusters, then the 2-stage procedure should be used as the minimum $p$-value lacks the ability to locate the potential clusters.

The procedure proposed in this paper has demonstrated that the adaptive scan statistic is a good alternative to the classic fixed window scan statistic in one and two dimensions. The extension to high-dimensional multiple window scan statistics is straightforward.

References