DISCRETE SCAN STATISTICS AND THE GENERALIZED LIKELIHOOD RATIO TEST

MICHAEL GENIN

Communicated by Marius Iosifescu

The scan statistic is used to detect statistically significant clusters of events. In the continuous one-dimensional case, the test based on scan statistic is equivalent to the generalized likelihood ratio test. We show that this result remains true for one and two dimensional discrete scan statistics. The binomial and Poisson models are considered.

AMS 2010 Subject Classification: 62F03.

Key words: scan statistics, generalized likelihood ratio test.

1. INTRODUCTION

In many areas it is necessary to decide whether a certain accumulation of events is "normal" or not. In public health, epidemiology services seek the factors that explain the clusters of cancer or birth defects. Biologists seek clusters palindromes in DNA sequences to find clues to the origin of replication of some viruses. In quality control, one wonders about the clusters of defective products. The decision is then made according to the probability of observing such a cluster under the null hypothesis of a "normal" situation. If this probability is small, it is reasonable to assume the presence of a deviation from the normal situation and then decisions must be taken.

Scan statistics are used to determine how significant a cluster of events is. More specifically, scan statistics are random variables used as test statistics to check the null hypothesis that the observations are independent and identically distributed (i.i.d.) from a specified distribution, against an alternative hypothesis which supports the existence of some cluster of events. Many works are devoted to scan statistics. For an overview, include the monographs of Glaz and Balakrishnan [5], Balakrishnan and Koutras [1], Fu and Lou [4] and more recently the monograph of Glaz, Pozdnyakov and Wallenstein [6]. Scan statistics have been widely used in several fields of application such as cosmology [3], reliability theory [2], epidemiology and public health [8].

In Naus 1966 [7], the authors show that, for continuous scan statistics and a scanning window of fixed size, the generalized likelihood ratio test (GLRT) is used to reject $\mathcal{H}_0$ whenever the scan statistic is greater than its quantile of order $1-\alpha$ where α is the type 1 error. This result is taken for granted in the case of discrete scan statistics but, as far as we know, there is no proof in the literature. In this article, we formalize this result in case of the one-dimensional and two-dimensional discrete scan statistics for both binomial and Poisson models. Section 2 is devoted to the one-dimensional case and Section 3 to the two-dimensional case. Section 4 provides a conclusion about the obtained results.

2. ONE-DIMENSIONAL DISCRETE SCAN STATISTIC AND THE GLRT

Let N be a positive integer and $\{X_i\}$, $1 \leq i \leq N$ be a sequence of i.i.d. nonnegative integer random variables from a specified distribution (Bernoulli, binomial, Poisson, etc). Let m be a positive integer such that $1 \leq m \leq N$. For $1 \leq t \leq N-m+1$, let

$$\nu_t = \nu_t(m) = \sum_{i=t}^{t+m-1} X_i \tag{1}$$

be the number of observed events in the scanning window of size m.

The *one-dimensional discrete scan statistic* is defined as the maximum number of events occurring in any window of size m within $\{1, \ldots, N\}$,

$$S_m = S(m, N) = \max_{1 \leq t \leq N-m+1} \nu_t. \tag{2}$$

The statistic S_m is used to test $\mathcal{H}_0$ assuming that the X_i's are i.i.d. from a specified distribution against an alternative hypothesis $\mathcal{H}_1$ which supports the existence of some cluster of events.

For the binomial model, the null hypothesis assumes that the X_i's are i.i.d., $X_i \sim \mathcal{B}(n, p_0)$, $n > 0$, with p_0 the probability of success. The alternative hypothesis presumes the existence of a window of size m, $[a, a+m-1]$, $a \in [0, N-m+1]$, where the X_i's are i.i.d. as binomial $\mathcal{B}(n, p_1)$ with $p_1 > p_0$ if $i \in [a, a+m-1]$ and $p_0 = p_1$ otherwise.

For the Poisson model, the null hypothesis assumes that the X_i's are i.i.d., $X_i \sim \mathcal{P}(\lambda_0)$. The alternative hypothesis presumes the existence of a window of size m, $[a, a+m-1]$, $a \in [0, N-m+1]$, where the X_i's are i.i.d. as Poisson $\mathcal{P}(\lambda_1)$ with $\lambda_1 > \lambda_0$ if $i \in [a, a+m-1]$ and $\lambda_0 = \lambda_1$ otherwise.

In what follows, we assume that the value of m is known. We show that, for both binomial and Poisson models, the GLRT rejects $\mathcal{H}_0$ in favor of $\mathcal{H}_1$ when S_m is greater than its quantile of order $1-\alpha$ where α is the type 1 error.

The binomial model. Let $X_1, X_2, \ldots, X_N$ be a sequence of independent random variables binomially distributed as $\mathcal{B}(n, p_k)$ where k is such that

$$k = \begin{cases} 0 & \text{if} \quad 1 \le i < t \quad \text{or} \quad t+m \le i \le N \\ 1 & \text{if} \quad t \le i < t+m. \end{cases}$$

It is assumed that the parameters p_k are known. One wants to check the null hypothesis $\mathcal{H}_0$ that the X_i's are independent and identically distributed (i.i.d.) as $\mathcal{B}(n, p_0)$:

$$\mathcal{H}_0 : p_0 = p_1, \tag{3}$$

against the alternative hypothesis $\mathcal{H}_1$ which supports a cluster of events of length m where the X_i's are i.i.d. as $\mathcal{B}(n, p_1)$:

$$\mathcal{H}_1 : p_1 > p_0 \tag{4}$$

PROPOSITION 1. *The generalized likelihood ratio test rejects $\mathcal{H}_0$ in favor of the alternative hypothesis $\mathcal{H}_1$ when the unidimensional discrete scan statistic with scanning window of fixed length m exceeds a threshold τ determined from $\mathbb{P}(S(m,N) > \tau | \mathcal{H}_0) = \alpha$, where α corresponds to the type 1 error.*

Proof. Let $L_{\mathcal{H}_0}$ be the likelihood function under $\mathcal{H}_0$. It is expressed as

$$L_{\mathcal{H}_0}(x_1, \ldots, x_N) = L_{\mathcal{H}_0} = \prod_{i=1}^{N} \binom{n}{x_i} p_0^{x_i} (1-p_0)^{n-x_i},$$

where $x_i \in \{0, \ldots, n\}$ and $\forall i = 1, \ldots, N$.

One can remark that $\mathcal{H}_1$ can be expressed as a function of t:

$$\mathcal{H}_1 = \bigcup_{t=1}^{N-m+1} \mathcal{H}_1(t),$$

where $\mathcal{H}_1(t)$ corresponds to an alternative of a cluster of length m starting at the tth position, for all $t \in \{1, \ldots, N-m+1\}$.

Thus, the likelihood function under $\mathcal{H}_1$, $L_{\mathcal{H}_1}(t)$, can be expressed as

$$L_{\mathcal{H}_1}(t) = \left(\prod_{i=1}^{t-1} \binom{n}{x_i} p_0^{x_i} (1-p_0)^{n-x_i} \right) \left(\prod_{i=t}^{t+m-1} \binom{n}{x_i} p_1^{x_i} (1-p_1)^{n-x_i} \right) \times \left(\prod_{i=t+m}^{N} \binom{n}{x_i} p_0^{x_i} (1-p_0)^{n-x_i} \right).$$

Hence, the likelihood ratio $LR(t,m)$ is defined as

$$LR(t,m) = \frac{\left(\prod_{i=1}^{t-1}\binom{n}{x_i}p_0^{x_i}(1-p_0)^{n-x_i}\right)\left(\prod_{i=t}^{t+m-1}\binom{n}{x_i}p_1^{x_i}(1-p_1)^{n-x_i}\right)}{\prod_{i=1}^{N}\binom{n}{x_i}p_0^{x_i}(1-p_0)^{n-x_i}} \times \frac{\left(\prod_{i=t+m}^{N}\binom{n}{x_i}p_0^{x_i}(1-p_0)^{n-x_i}\right)}{\prod_{i=1}^{N}\binom{n}{x_i}p_0^{x_i}(1-p_0)^{n-x_i}},$$

which can be simplified to

$$LR(t,m) = \frac{\prod_{i=t}^{t+m-1}\binom{n}{x_i}p_1^{x_i}(1-p_1)^{n-x_i}}{\prod_{i=t}^{t+m-1}\binom{n}{x_i}p_0^{x_i}(1-p_0)^{n-x_i}}.$$

Hence, the logarithm of the likelihood ratio $LLR(t,m)$ can be expressed as

$$\begin{aligned} LLR(t,m) &= \log\prod_{i=t}^{t+m-1}\binom{n}{x_i}p_1^{x_i}(1-p_1)^{n-x_i} - \log\prod_{i=t}^{t+m-1}\binom{n}{x_i}p_0^{x_i}(1-p_0)^{n-x_i}, \\ &= \sum_{i=t}^{t+m-1} x_i \log\left(\frac{p_1}{p_0}\right) + (n-x_i)\log\left(\frac{1-p_1}{1-p_0}\right). \end{aligned}$$

Denoting $C_1 = \log\left(\frac{p_1}{p_0}\right)$ and $C_2 = \log\left(\frac{1-p_1}{1-p_0}\right)$. The $LLR(t,m)$ can be expressed as

$$LLR(t,m) = C_1 \sum_{i=t}^{t+m-1} x_i + C_2 \sum_{i=t}^{t+m-1} (n-x_i).$$

For $1 \le t \le N-m+1$, let

$$\nu_t = \sum_{i=t}^{t+m-1} x_i,$$

be the number of observed events in the window $[t, t+m-1]$. Hence, the $LLR(t,m)$ can be written as follows

$$LLR(t,m) = C_1\nu_t + C_2(mn - \nu_t).$$

For fixed m and since $C_1 > 0$ and $C_2 < 0$, $LLR(t,m)$ is a monotically increasing function of ν_t. Consequently, the GLRT rejects $\mathcal{H}_0$ for a value of ν_t as large as possible, *i.e.* the unidimensional discrete scan statistic with scanning window of fixed length m, S_m defined in Eq.(2). □

The Poisson model. Let $X_1, X_2, \ldots, X_N$ be a sequence of independent random variables Poisson distributed as $\mathcal{P}(\lambda_k)$ where k is such that

$$k = \begin{cases} 0 & \text{if} \quad 1 \le i < t \quad \text{ou} \quad t+m \le i \le N \\ 1 & \text{if} \quad t \le i < t+m. \end{cases}$$

It is assumed that the parameters λ_k are known. One wants to check the null hypothesis $\mathcal{H}_0$ that the X_i's are i.i.d. as $\mathcal{P}(\lambda_0)$:

$$\mathcal{H}_0 : \lambda_0 = \lambda_1 \tag{5}$$

against the alternative hypothesis $\mathcal{H}_1$ which supports a cluster of events of length m where the X_i's are i.i.d. as $\mathcal{P}(\lambda_1)$

$$\mathcal{H}_1 : \lambda_1 > \lambda_0 \tag{6}$$

PROPOSITION 2. *The generalized likelihood ratio test rejects $\mathcal{H}_0$ in favor of the alternative hypothesis $\mathcal{H}_1$ when the unidimensional discrete scan statistic with scanning window of fixed length m exceeds a threshold τ determined from $\mathbb{P}(S(m,N) > \tau|\mathcal{H}_0) = \alpha$, where α corresponds to the type 1 error.*

Proof. Let $L_{\mathcal{H}_0}$ be the likelihood function under $\mathcal{H}_0$. It is expressed as follows

$$L_{\mathcal{H}_0}(x_1,\ldots,x_N) = L_{\mathcal{H}_0} = \prod_{i=1}^{N} \frac{e^{-\lambda_0}\lambda_0^{x_i}}{x_i!},$$

where $x_i \in \{0,\ldots,n\}$ and $\forall i = 1,\ldots,N$.

The expression of the likelihood function under $\mathcal{H}_1$, $L_{\mathcal{H}_1}(t)$, is the following

$$L_{\mathcal{H}_1}(t) = \left(\prod_{i=1}^{t-1} \frac{e^{-\lambda_0}\lambda_0^{x_i}}{x_i!}\right)\left(\prod_{i=t}^{t+m-1} \frac{e^{-\lambda_1}\lambda_1^{x_i}}{x_i!}\right)\left(\prod_{i=t+m}^{N} \frac{e^{-\lambda_0}\lambda_0^{x_i}}{x_i!}\right).$$

Hence, the likelihood ratio $LR(t,m)$ is defined as

$$LR(t,m) = \frac{\prod_{i=t}^{t+m-1} \frac{e^{-\lambda_1}\lambda_1^{x_i}}{x_i!}}{\prod_{i=t}^{t+m-1} \frac{e^{-\lambda_0}\lambda_0^{x_i}}{x_i!}},$$

and its logarithm, $LLR(t,m)$,

$$\begin{aligned} LLR(t,m) &= \textstyle\sum_{i=t}^{t+m-1} \left[(-\lambda_1 + x_i\log(\lambda_1) - \log(x_i!)) - (\lambda_0 + x_i\log(\lambda_0) - \log(x_i!))\right], \\ &= m(\lambda_0 - \lambda_1) + \textstyle\sum_{i=t}^{t+m-1} x_i \log\left(\frac{\lambda_1}{\lambda_0}\right). \end{aligned}$$

Denoting $C = \log\left(\frac{\lambda_1}{\lambda_0}\right)$. The expression of the $LLR(t,m)$ is now the following

$$LLR(t,m) = m(\lambda_0 - \lambda_1) + C \sum_{i=t}^{t+m-1} x_i.$$

For $1 \leq t \leq N - m + 1$, let

$$\nu_t = \sum_{i=t}^{t+m-1} x_i,$$

be the number of observed events in the window $[t, t+m-1]$. Hence, the $LLR(t,m)$ can be written as follows

$$LLR(t,m) = m(\lambda_0 - \lambda_1) + C\nu_t.$$

For fixed m and since $C > 0$ and $m(\lambda_0 - \lambda_1) < 0$, $LLR(t, m)$ is a monotically increasing function of ν_t. Accordingly, the GLRT rejects $\mathcal{H}_0$ for a value of ν_t as large as possible, *i.e.* the unidimensional discrete scan statistic with scanning window of fixed length m, S_m defined in Eq.(2). □

3. TWO-DIMENSIONAL DISCRETE SCAN STATISTIC AND THE GLRT

Let N_1, N_2 be positive integers, $\mathcal{R} = [0, N_1] \times [0, N_2]$ be a rectangular region and $\{X_{ij}\}$, $1 \leqslant i \leqslant N_1$, $1 \leqslant j \leqslant N_2$, be a family of i.i.d. nonnegative integer random variables from a specified distribution (Bernoulli, binomial, Poisson, etc). In practice, the X_{ij}'s represent the number of events occurring in the elementary square sub-region $[i-1, i] \times [j-1, j]$.

Let m_1, m_2 be positive integers such that $1 \leqslant m_1 \leqslant N_1$, $1 \leqslant m_2 \leqslant N_2$. For $1 \leqslant t \leqslant N_1 - m_1 + 1$ and $1 \leqslant s \leqslant N_2 - m_2 + 1$, let

$$\nu_{ts} = \nu_{ts}(m_1, m_2) = \sum_{i=t}^{t+m_1-1} \sum_{j=s}^{s+m_2-1} X_{ij} \tag{7}$$

be the number of events observed in the scanning window located on the rectangular sub-region $[t, t+m_1-1] \times [s, s+m_2-1]$ within $\mathcal{R}$ (see Fig. 1).

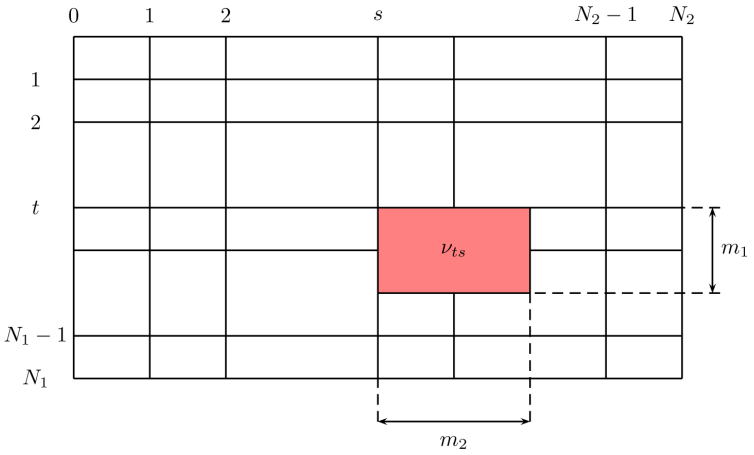

Fig. 1. The scanning process with $m_1 \times m_2$ scanning window in $\mathcal{R}$.

The *two-dimensional discrete scan statistic* is defined as the maximum number of events that occured in any $m_1 \times m_2$ rectangular window within the rectangular region $[0, N_1] \times [0, N_2]$,

$$S_{m_1,m_2} = S(m_1, m_2, N_1, N_2) = \max_{\substack{1 \leqslant t \leqslant N_1 - m_1 + 1 \\ 1 \leqslant s \leqslant N_2 - m_2 + 1}} \nu_{ts} \,. \tag{8}$$

The statistic S_{m_1,m_2} is used to test the null hypothesis ($\mathcal{H}_0$) assuming that the X_{ij}'s are i.i.d. from a specified distribution, against an alternative hypothesis ($\mathcal{H}_1$) which supports the existence of some cluster of events.

For the binomial model, the null hypothesis assumes that the X_{ij}'s are i.i.d., $X_{ij} \sim \mathcal{B}(n, p_0)$, $n > 0$, with p_0 the probability of success. The alternative hypothesis presumes the existence of a rectangular subregion of size $m_1 \times m_2$, $[a, a+m_1-1] \times [b, b+m_2-1]$, $a \in [0, N_1 - m_1 + 1]$, $b \in [0, N_2 - m_2 + 1]$, such that the random variables X_{ij}'s are independent and identically distributed as a binomial $\mathcal{B}(n, p_1)$ with $p_1 > p_0$ if $(i,j) \in [a; a+m_1-1] \times [b, b+m_2-1]$ and $p_0 = p_1$ otherwise.

For the Poisson model, assumes that the X_{ij}'s are i.i.d., $X_{ij} \sim \mathcal{P}(\lambda_0)$. The alternative hypothesis presumes the existence of a rectangular subregion of size $m_1 \times m_2$, $[a, a+m_1-1] \times [b, b+m_2-1]$, $a \in [0, N_1 - m_1 + 1]$, $b \in [0, N_2 - m_2 + 1]$, such that the random variables X_{ij}'s are independent and identically distributed as a Poisson $\mathcal{P}(\lambda_1)$ with $\lambda_1 > \lambda_0$ if $(i,j) \in [a; a+m_1-1] \times [b, b+m_2-1]$ and $\lambda_1 = \lambda_0$ otherwise.

In what follows, we assume that the values of m_1 and m_2 are known. We show that, for both binomial and Poisson models, the GLRT rejects $\mathcal{H}_0$ in favor of $\mathcal{H}_1$ when S_{m_1,m_2} is greater than its quantile of order $1-\alpha$ when α is the type 1 error.

The Binomial model. Let $[0, N_1] \times [0, N_2]$ be a rectangular region, $N_1, N_2 \in \mathbb{N}$. Let $\{X_{ij}\}$ be a family of independent random variables binomially distributed as $\mathcal{B}(n, p_k)$ where k is such that

$$k = \begin{cases} 0 & \text{if} \quad \{i,j\} \in [0,N_1] \times [0,N_2] \backslash [t, t+m_1-1] \times [s, s+m_2-1] \\ 1 & \text{if} \quad \{i,j\} \in [t, t+m_1-1] \times [s, s+m_2-1]. \end{cases}$$

It is assumed that the parameters p_k are known. One wants to verify the null hypothesis $\mathcal{H}_0$ that the X_{ij}'s are i.i.d. as $\mathcal{B}(n, p_0)$:

$$\mathcal{H}_0 : p_0 = p_1 \tag{9}$$

against the alternative hypothesis $\mathcal{H}_1$ which supports the existence of a cluster of events of size $m_1 \times m_2$ where the X_{ij}'s are i.i.d. as $\mathcal{B}(n, p_1)$:

$$\mathcal{H}_1 : p_1 > p_0 \tag{10}$$

PROPOSITION 3. *The GLRT rejects $\mathcal{H}_0$ in favor of $\mathcal{H}_1$ when the two-dimensional discrete scan statistic with scanning window of fixed size $m_1 \times m_2$ exceeds a threshold determined from $\mathbb{P}(S(m1, m_2, N_1, N_2) > \tau | \mathcal{H}_0) = \alpha$, where α is the type 1 error.*

Proof. The following proof has a similar reasoning to the one-dimensional

case. Let $L_{\mathcal{H}_0}$ be the likelihood function under $\mathcal{H}_0$. It is expressed as follows:

$$L_{\mathcal{H}_0}(x_{11},\ldots,x_{N_1N_2}) = L_{\mathcal{H}_0} = \prod_{i=1}^{N_1}\prod_{j=1}^{N_2}\binom{n}{x_{ij}}p_0^{x_{ij}}(1-p_0)^{n-x_{ij}},$$

where $x_{ij}\in\{0,\ldots,n\}$ and $\forall i\in\{1,\ldots,N_1\}$ and $\forall j\in\{1,\ldots,N_2\}$.

Under $\mathcal{H}_1$, the likelihood function is given by

$$L_{\mathcal{H}_1}(t,s)=\prod_{i=1}^{t-1}\prod_{j=1}^{s-1}\binom{n}{x_{ij}}p_0^{x_{ij}}(1-p_0)^{n-x_{ij}}\prod_{i=t}^{t+m_1-1}\prod_{j=s}^{s+m_2-1}\binom{n}{x_{ij}}p_1^{x_{ij}}(1-p_1)^{n-x_{ij}}$$
$$\times\prod_{i=t+m_1}^{N_1}\prod_{j=s+m_2}^{N_2}\binom{n}{x_{ij}}p_0^{x_{ij}}(1-p_0)^{n-x_{ij}}.$$

Hence, the likelihood ratio $LR(t,m)$ is defined as

$$LR(t,s,m_1,m_2) = \frac{\prod_{i=t}^{t+m_1-1}\prod_{j=s}^{s+m_2-1}\binom{n}{x_{ij}}p_1^{x_{ij}}(1-p_1)^{n-x_{ij}}}{\prod_{i=t}^{t+m_1-1}\prod_{j=s}^{s+m_2-1}\binom{n}{x_{ij}}p_0^{x_{ij}}(1-p_0)^{n-x_{ij}}},$$

and its logarithm, $LLR(t,s,m_1,m_2)$,

$$\begin{aligned}LLR(t,s,m_1,m_2) =\ & \log\textstyle\prod_{i=t}^{t+m_1-1}\prod_{j=s}^{s+m_2-1}\binom{n}{x_{ij}}p_1^{x_{ij}}(1-p_1)^{n-x_{ij}}\\ & -\log\textstyle\prod_{i=t}^{t+m_1-1}\prod_{j=s}^{s+m_2-1}\binom{n}{x_{ij}}p_0^{x_{ij}}(1-p_0)^{n-x_{ij}},\\ =\ & \textstyle\sum_{i=t}^{t+m_1-1}\sum_{j=s}^{s+m_2-1}\Big[x_{ij}\log\left(\frac{p_1}{p_0}\right)+(n-x_{ij})\log\left(\frac{1-p_1}{1-p_0}\right)\Big].\end{aligned}$$

Denoting $C_1=\log\left(\frac{p_1}{p_0}\right)$ and $C_2=\log\left(\frac{1-p_1}{1-p_0}\right)$. Then

$$LLR(t,s,m_1,m_2) = C_1\sum_{i=t}^{t+m_1-1}\sum_{j=s}^{s+m_2-1}x_{ij}+C_2\sum_{i=t}^{t+m_1-1}\sum_{j=s}^{s+m_2-1}(n-x_{ij}).$$

For $1\le t\le N_1-m_1+1$ and $1\le s\le N_2-m_2+1$, let

$$\nu_{ts} = \sum_{i=t}^{t+m_1-1}\sum_{j=s}^{s+m_2-1}x_{ij},$$

be the number of observed events in the window of size $[t,t+m_1-1]\times[s,s+m_2-1]$. Thus, the $LLR(t,s,m_1,m_2)$ can be written as

$$LLR(t,s,m_1,m_2) = C_1\nu_{ts}+C_2(m_1m_2n-\nu_{ts}).$$

For fixed m_1 and m_2 and since $C_1>0$ and $C_2<0$, $LLR(t,s,m_1,m_2)$ is a monodically increasing function of ν_{ts}. Consequently, the GLRT rejects $\mathcal{H}_0$ for a value of ν_{ts} as large as possible, *i.e.* the two-dimensional discrete scan statistic with scanning window of fixed size $m_1\times m_2$, S_{m_1,m_2} defined in Eq. (8). □

The Poisson model. Let $[0, N_1] \times [0, N_2]$ be a rectangular region, $N_1, N_2 \in \mathbb{N}$. Let $\{X_{ij}\}$ be a family of independent random variables Poisson distributed as $\mathcal{P}(\lambda_k)$ where k is such that

$$k = \begin{cases} 0 & \text{if} \quad \{i,j\} \in [0,N_1] \times [0,N_2] \backslash [t, t+m_1-1] \times [s, s+m_2-1] \\ 1 & \text{if} \quad \{i,j\} \in [t, t+m_1-1] \times [s, s+m_2-1]. \end{cases}$$

It is assumed that the parameters p_k are known. One wants to verify the null hypothesis $\mathcal{H}_0$ that the X_{ij}'s are i.i.d. as $\mathcal{P}(\lambda_0)$:

$$\mathcal{H}_0 : \lambda_0 = \lambda_1 \tag{11}$$

against the alternative hypothesis $\mathcal{H}_1$ which supports the existence of a cluster of events of size $m_1 \times m_2$ where the X_{ij}'s are i.i.d. as $\mathcal{P}(\lambda_1)$:

$$\mathcal{H}_1 : \lambda_1 > \lambda_0 \tag{12}$$

PROPOSITION 4. *The GLRT rejects $\mathcal{H}_0$ in favor of $\mathcal{H}_1$ when the two-dimensional discrete scan statistic with scanning window of fixed size $m_1 \times m_2$ exceeds a threshold determined from $\mathbb{P}(S(m1, m_2, N_1, N_2) > \tau | \mathcal{H}_0) = \alpha$, where α is the type 1 error.*

Proof. Let $L_{\mathcal{H}_0}$ be the likelihood function under $\mathcal{H}_0$:

$$L_{\mathcal{H}_0}(x_{11}, \ldots, x_{N_1 N_2}) = L_{\mathcal{H}_0} = \prod_{i=1}^{N_1} \prod_{j=1}^{N_2} \frac{e^{-\lambda_0} \lambda_0^{x_{ij}}}{x_{ij}!},$$

where $x_{ij} \in \{0, \ldots, n\}$ and $\forall i \in \{1, \ldots, N_1\}$ and $\forall j \in \{1, \ldots, N_2\}$.

The likelihood function under $\mathcal{H}_1$ is given by

$$L_{\mathcal{H}_1}(t,s) = \prod_{i=1}^{t-1} \prod_{j=1}^{s-1} \frac{e^{-\lambda_0} \lambda_0^{x_{ij}}}{x_{ij}!} \prod_{i=t}^{t+m_1-1} \prod_{j=s}^{s+m_2-1} \frac{e^{-\lambda_1} \lambda_1^{x_{ij}}}{x_{ij}!} \prod_{i=t+m_1}^{N_1} \prod_{j=s+m_2}^{N_2} \frac{e^{-\lambda_0} \lambda_0^{x_{ij}}}{x_{ij}!}.$$

Hence, the likelihood ratio $LR(t, s, m_1, m_2)$ is defined as

$$LR(t,s,m_1,m_2) = \frac{\prod_{i=t}^{t+m_1-1} \prod_{j=s}^{s+m_2-1} \frac{e^{-\lambda_1} \lambda_1^{x_{ij}}}{x_{ij}!}}{\prod_{i=t}^{t+m_1-1} \prod_{j=s}^{s+m_2-1} \frac{e^{-\lambda_0} \lambda_0^{x_{ij}}}{x_{ij}!}},$$

and its logarithm, $LLR(t, s, m_1, m_2)$,

$$\begin{aligned} LLR(t,s,m_1,m_2) &= \textstyle\sum_{i=t}^{t+m_1-1} \sum_{j=s}^{s+m_2-1} [(-\lambda_1 + x_{ij} \log(\lambda_1)) - (\lambda_0 + x_{ij} \log(\lambda_0))], \\ &= m_1 m_2 (\lambda_0 - \lambda_1) + C \textstyle\sum_{i=t}^{t+m_1-1} \sum_{j=s}^{s+m_2-1} x_{ij}, \end{aligned}$$

where $C = \log\left(\frac{\lambda_1}{\lambda_0}\right)$.

For $1 \leq t \leq N_1 - m_1 + 1$ and $1 \leq s \leq N_2 - m_2 + 1$, let

$$\nu_{ts} = \sum_{i=t}^{t+m_1-1} \sum_{j=s}^{s+m_2-1} x_{ij},$$

be the number of observed events in the window of size $[t, t+m_1-1] \times [s, s+m_2-1]$. Hence, the $LLR(t, s, m_1, m_2)$ can be written as

$$LLR(t, s, m_1, m_2) = m_1 m_2(\lambda_0 - \lambda_1) + C\nu_{ts}.$$

For fixed m_1 and m_2, $LLR(t, s, m_1, m_2)$ is a monodically increasing function of ν_{ts}. Consequently, the GLRT rejects $\mathcal{H}_0$ for a value of ν_{ts} as large as possible, *ie* the two-dimensional discrete scan statistic with scanning window of fixed size$m_1 \times m_2$, S_{m_1,m_2} defined in Eq. (8). □

4. CONCLUSION

In this paper, we showed that the generalized likelihood ratio test rejects the null hypothesis of randomness in favor of an alternative hypothesis which supports the existence of a cluster when the discrete scan statistic exceeds its quantile of order $1 - \alpha$. We proved this result for one-dimensional and two-dimensional discrete scan statistics for both binomial and Poisson models.

REFERENCES

[1] N. Balakrishnan and M.V. Koutras, *Runs and scans with applications.* Wiley, 2001.

[2] A. Barbour, O. Chryssaphinou and M. Roos, *Compound poisson approximation in systems reliability.* Naval Research Logistics **43** (1996), *2*, 251–264.

[3] R. Darling and M.S. Waterman, *Extreme value distribution for the largest cube in a random lattice.* SIAM Journal on Applied Mathematics **46** (1996), *1*, 118–132.

[4] J. Fu and W. Lou, *Distribution theory of runs and patterns and its applications.* World Scientific Publishing Co., 2003.

[5] J. Glaz and N. Balakrishnan, *Scan Statistics and Applications.* Statistics for Industry and Technology. Birkhauser Boston, 1999.

[6] J. Glaz, V. Pozdnyakov and S. Wallenstein, *Scan Statistics : Methods and Applications.* Statistics for Industry and Technology. Birkhauser Boston, 2009.

[7] J. Naus. *Power comparison of two tests of non-random clustering.* Technometrics **8** (1966), *3*, 493–517.

[8] J.F. Viel, P. Arveux, J. Baverel and J.Y. Cahn, *Soft-tissue sarcoma and non-hodgkinXs lymphoma clusters around a municipal solid waste incinerator with high dioxin emission levels.* American Journal of Epidemiology **152** (2000), *1*, 13-9.

Received 9 December 2013

Université Lille 2
EA2694, CERIM, Faculté de Médecine,
1, Place de Verdun, F-59045 Lille
Cedex, France
michael.genin@univ-lille2.fr