Approximation for the Distribution of Three-dimensional Discrete Scan Statistic

Alexandru Amarioarei, Cristian Preda

Approximation for the Distribution of Three-dimensional Discrete Scan Statistic

Alexandru Amărioarei[1,2,3], Cristian Preda[1,2]

[1]Laboratoire de Mathémathiques Paul Painlevé, UMR 8524, Université de Sciences et Technologies de Lille 1, France
[2]INRIA Nord Europe/Modal, France
[3]National Institute of R&D for Biological Sciences, Bucharest, Romania

March 18, 2013

1. Introduction

Let T_1, T_2, T_3 be positive integers, $\mathcal{R} = [0, T_1] \times [0, T_2] \times [0, T_3]$ be a rectangular region and $\{X_{ijk} | 1 \leq i \leq T_1, 1 \leq j \leq T_2, 1 \leq k \leq T_3\}$ be a family of independent and identically distributed integer valued random variables from a specified distribution. In practice, X_{ijk} can be interpreted as the number of events that occur in the elementary subregion $r_{ijk} = [i-1, i] \times [j-1, j] \times [k-1, k]$. For each $j \in \{1, 2, 3\}$, consider the positive integers m_j such that $2 \leq m_j \leq T_j - 1$, and define the random variables

$$Y_{i_1 i_2 i_3} = \sum_{i=i_1}^{i_1+m_1-1} \sum_{j=i_2}^{i_2+m_2-1} \sum_{k=i_3}^{i_3+m_3-1} X_{ijk}, \quad 1 \leq i_j \leq T_j - m_j + 1, \tag{1.1}$$

as the number of events occurring in the rectangular region

$$\mathcal{R}(i_1, i_2, i_3) = [i_1 - 1, i_1 + m_1 - 1] \times [i_2 - 1, i_2 + m_2 - 1] \times [i_3 - 1, i_3 + m_3 - 1].$$

The three dimensional discrete scan statistic is defined as the maximum number of events in any rectangle $\mathcal{R}(i_1, i_2, i_3)$ within the region $\mathcal{R}$,

$$S_{m_1,m_2,m_3}(T_1, T_2, T_3) = \max_{\substack{1 \leq i_j \leq T_j - m_j + 1 \\ j \in \{1,2,3\}}} Y_{i_1 i_2 i_3}. \tag{1.2}$$

The distribution of scan statistics,

$$\mathbb{P}\left(S_{m_1,m_2,m_3}(T_1, T_2, T_3) \leq n\right), \quad n \in \{1, 2, \ldots, m_1 m_2 m_3\}$$

is used with success in astronomy (Darling and Waterman [1986]), image analysis and reliability theory (Boutsikas and Koutras [2000]) and many other domains. For an overview of the potential application of scan statistics one can refer to the monographs of Glaz, Naus and Wallenstein [2001] and more recently the one of Glaz, Pozdnyakov and Wallenstein [2009].
From a statistical point of view, the scan statistic $S_{m_1,m_2,m_3}(T_1, T_2, T_3)$ is used for testing the null hypothesis of randomness that X_{ijk}'s are independent and identically distributed according to some specified distribution. Under the alternative hypothesis there exists one cluster location where the X_{ijk}'s have a larger mean than outside the cluster. As an example, in the Poisson model, the null hypothesis, H_0, assumes that X_{ijk}'s are i.i.d. with $X_{ijk} \sim Pois(\lambda)$ whereas the alternative hypothesis of clustering, H_1, assumes the existence of a rectangular subregion $\mathcal{R}(i_0, j_0, k_0)$ such that for any $i_0 \leq i \leq i_0 + m_1 - 1$, $j_0 \leq j \leq j_0 + m_2 - 1$ and $k_0 \leq k \leq k_0 + m_3 - 1$, X_{ijk} are i.i.d. Poisson random variables with parameter $\lambda' > \lambda$. Outside the region $\mathcal{R}(i_0, j_0, k_0)$, X_{ijk} are i.i.d. distributed according to the distribution specified by the null hypothesis. The generalized likelihood ratio test rejects H_0 in favor of the local change alternative H_1, whenever $S_{m_1,m_2,m_3}(T_1, T_2, T_3)$ exceeds the threshold τ determined from $\mathbb{P}\left(S_{m_1,m_2,m_3}(T_1, T_2, T_3) \geq \tau | H_0\right) = \alpha$ and where α represents

the significance level of the testing procedure (Glaz, Naus and Wallenstein [2001, Chapter 13]).
Since there are no exact formulas available for the distribution of three dimensional scan statistics, approximation methods are necessary. For the Bernoulli model, Glaz, Guerriero and Sen [2010] propose four approximation formulas: one Markov like product type approximation and three Poisson type approximations that extends the special case described by Darling and Waterman [1986] when $n = m_1 m_2 m_3$.
The advantage of the method described in this paper is that it can be used for any distribution of the random field and provides accurate approximations and sharp error bounds. The methodology used to obtain the approximation and the error bounds is presented in Section 2. In Section 3 we describe adapt the importance sampling algorithm developed by Naiman and Priebe [2001] to estimate the simulation error. A simulation study is conducted in Section 4 for considered Bernoulli, binomial and Poisson models. Concluding remarks are given in Section 5.

2. Methodology

In order to approximate the distribution of $S_{m_1,m_2,m_3}(T_1, T_2, T_3)$ we use a similar approach as in Haiman and Preda [2006]. The key idea is to observe that we can write the scan statistic random variable as a maximum of 1-dependent stationary sequence of random variables. A sequence $(Z_k)_{k\geq 1}$ is m-dependent, $m \geq 1$, if for any $h \geq 1$ the σ-fields generated by $\{Z_1, \ldots, Z_h\}$ and $\{Z_{h+m+1}, \ldots\}$ are independent. The method is based on the following result developed in Haiman [1999, Theorem 4] and improved in Amarioarei [2012, Theorem 2.6]:
Let $(Z_k)_{k\geq 1}$ be a strictly stationary 1-dependent sequence of random variables and for $x < \sup\{u|\mathbb{P}(Z_1 \leq u) < 1\}$, let

$$q_m = q_m(x) = \mathbb{P}(\max(Z_1, \ldots, Z_m) \leq x). \tag{2.1}$$

Theorem 2.1. *For all x such that $q_1(x) \geq 1-\alpha \geq 0.9$, the following approximation formula holds:*

$$\left| q_m - \frac{2q_1 - q_2}{[1 + q_1 - q_2 + 2(q_1 - q_2)^2]^m} \right| \leq mF(\alpha, m)(1 - q_1)^2 \tag{2.2}$$

with

$$F(\alpha, m) = 1 + \frac{3}{m} + \left[\frac{\Gamma(\alpha)}{m} + K(\alpha)\right](1 - q_1) \tag{2.3}$$

where $\Gamma(\alpha) = L(\alpha) + E(\alpha)$,

$$K(\alpha) = \frac{\frac{11-3\alpha}{(1-\alpha)^2} + 2l(1+3\alpha)\frac{2+3l\alpha-\alpha(2-l\alpha)(1+l\alpha)^2}{[1-\alpha(1+l\alpha)^2]^3}}{1 - \frac{2\alpha(1+l\alpha)}{[1-\alpha(1+l\alpha)^2]^2}} \tag{2.4}$$

$$L(\alpha) = 3K(\alpha)(1 + \alpha + 3\alpha^2)[1 + \alpha + 3\alpha^2 + K(\alpha)\alpha^3] + \alpha^6 K^3(\alpha) + 9\alpha(4 + 3\alpha + 3\alpha^2) + 55.1 \tag{2.5}$$

$$E(\alpha) = \frac{\eta^5 [1 + (1-2\alpha)\eta]^4 [1 + \alpha(\eta - 2)] [1 + \eta + (1-3\alpha)\eta^2]}{2(1-\alpha\eta^2)^4 [(1-\alpha\eta^2)^2 - \alpha\eta^2(1 + \eta - 2\alpha\eta)^2]} \tag{2.6}$$

and where $\eta = 1 + l\alpha$ with $l = l(\alpha) > t_2^3(\alpha)$ and $t_2(\alpha)$ the second root in magnitude of the equation $\alpha t^3 - t + 1 = 0$.

In this section we obtain an approximation formula for the distribution of scan statistic defined by Eq.(1.2) in three steps as follows.

Let assume that $L_j = \frac{T_j}{m_j - 1}$, $j \in \{1, 2, 3\}$, are positive integers and define for each $k \in \{1, 2, \ldots, L_3 - 1\}$ the random variables

$$Z_k = \max_{\substack{1 \le i_1 \le (L_1-1)(m_1-1) \\ 1 \le i_2 \le (L_2-1)(m_2-1) \\ (k-1)(m_3-1)+1 \le i_3 \le k(m_3-1)}} Y_{i_1 i_2 i_3}. \tag{2.7}$$

The set of random variables $\{Z_1, \ldots, Z_{L_3-1}\}$ forms a 1-dependent stationary sequence. Indeed, from Eq.(2.7) and the independence of X_{ijl} we observe that for any $k \ge 1$, $\sigma(\cdots, Z_k)$ and $\sigma(Z_{k+2}, \cdots)$ are included in $\sigma(\{X_{ijl} | 1 \le i \le T_1, 1 \le j \le T_2, 1 \le l \le (k+1)(m_3-1)\})$ and $\sigma(\{X_{ijl} | 1 \le i \le T_1, 1 \le j \le T_2, (k+1)(m_3-1)+1 \le l\})$, respectively, which are independent (see Fig. 1).

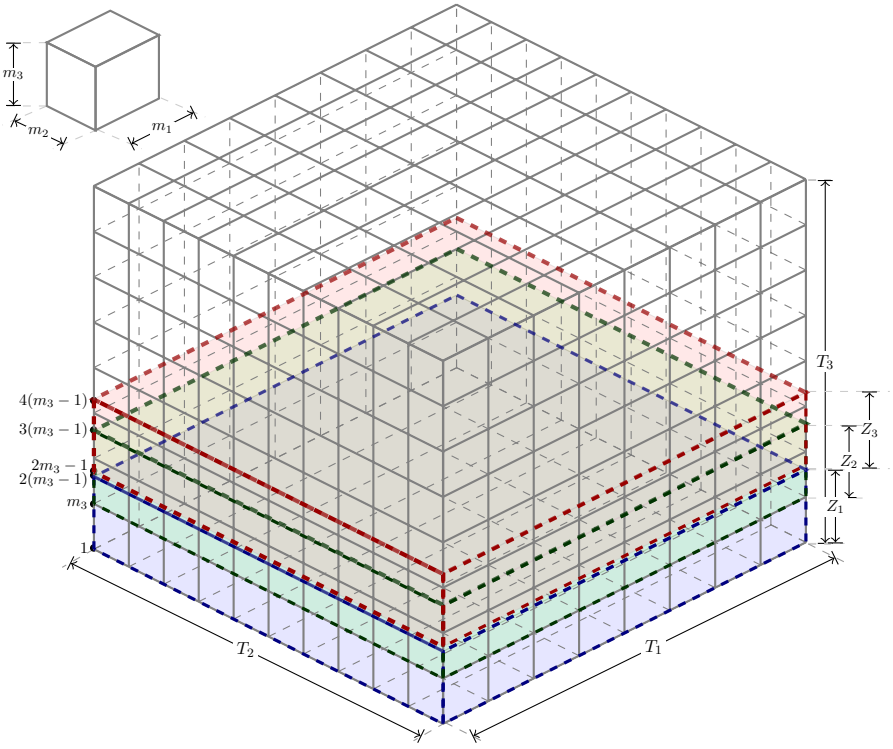

FIGURE 1. Illustration of Z_k emphasizing the 1-dependence

Notice that from Eq.(1.2) and Eq.(2.7) we have

$$S(L_1, L_2, L_3) = S_{m_1,m_2,m_3}(T_1, T_2, T_3) = \max_{1 \le k \le L_3 - 1} Z_k. \tag{2.8}$$

Take for $s \in \{2, 3\}$,

$$Q_s = Q_s(n) = \mathbb{P}\left(\bigcap_{k=1}^{s-1} \{Z_k \le n\}\right) = \mathbb{P}\left(\max_{\substack{1 \le i_1 \le (L_1-1)(m_1-1) \\ 1 \le i_2 \le (L_2-1)(m_2-1) \\ 1 \le i_3 \le (s-1)(m_3-1)}} Y_{i_1 i_2 i_3} \le n\right). \tag{2.9}$$

Notice that in the notation of Eq.(2.1) we have $Q_s = q_{s-1}$. For n such that $Q_2(n) \ge 1 - \alpha_1 \ge 0.9$ we apply the result in Theorem 2.1 to obtain the first step approximation

$$\mathbb{P}\left(S(L_1, L_2, L_3) \le n\right) \approx \frac{2Q_2 - Q_3}{\left[1 + Q_2 - Q_3 + 2(Q_2 - Q_3)^2\right]^{(L_3-1)}}, \tag{2.10}$$

with an error bound of $(L_3-1)F(\alpha_1, L_3-1)(1-Q_2)^2$. Observe that Q_2 and Q_3 represents the distribution of the scan statistics over the rectangular subregions $[1,T_1]\times[1,T_2]\times[1,2(m_3-1)]$ and $[1,T_1]\times[1,T_2]\times[1,3(m_3-1)]$, respectively (see also Fig 1). To simplify the results of the presentation, in what follows we abbreviate the approximation formula by

$$H(x,y,m)=\frac{2x-y}{[1+x-y+2(x-y)^2]^{m-1}}. \tag{2.11}$$

In order to evaluate the approximation in Eq.(2.10) it is necessary to find approximations for Q_2 and Q_3. Thus, the second step consists in applying Theorem 2.1 for each Q_s. We define, as in Eq.(2.7), for $s\in\{2,3\}$ and $l\in\{1,2,\ldots,L_2-1\}$ the sequences

$$Z_l^{(s)}=\max_{\substack{1\le i_1\le(L_1-1)(m_1-1)\\(l-1)(m_2-1)+1\le i_2\le l(m_2-1)\\1\le i_3\le(s-1)(m_3-1)}} Y_{i_1i_2i_3}, \tag{2.12}$$

which are strictly stationary, 1-dependent and satisfy

$$Q_s=\mathbb{P}(S(L_1,L_2,s)\le n)=\mathbb{P}\left(\max_{1\le l\le L_2-1} Z_l^s\le n\right). \tag{2.13}$$

Set for $t,s\in\{2,3\}$,

$$Q_{ts}=Q_{ts}(n)=\mathbb{P}\left(\bigcap_{l=1}^{t-1}\{Z_l^{(s)}\le n\}\right)=\mathbb{P}\left(\max_{\substack{1\le i_1\le(L_1-1)(m_1-1)\\1\le i_2\le(t-1)(m_2-1)\\1\le i_3\le(s-1)(m_3-1)}} Y_{i_1i_2i_3}\le n\right). \tag{2.14}$$

If the condition $Q_{2s}(n)\ge 1-\alpha_{2s}\ge 0.9$ is fulfilled, then using Theorem 2.1, we find, for $s\in\{2,3\}$, the approximation

$$|Q_s-H\left(Q_{2s},Q_{3s},L_2\right)|\le(L_2-1)F(\alpha_{2s},L_2-1)(1-Q_{2s})^2. \tag{2.15}$$

The last step involves the evaluation of Q_{ts} in Eq.(2.15). For $s,t\in\{2,3\}$ and $j\in\{1,2,\ldots,L_1-1\}$ let consider the following random sequences

$$Z_j^{(ts)}=\max_{\substack{(j-1)(m_1-1)+1\le i_1\le j(m_1-1)\\1\le i_2\le(t-1)(m_2-1)\\1\le i_3\le(s-1)(m_3-1)}} Y_{i_1i_2i_3}. \tag{2.16}$$

We observe that $\left(Z_j^{(ts)}\right)_{j\ge1}$ forms 1-dependent stationary sequences and

$$Q_{ts}=\mathbb{P}\left(S(L_1,t,s)\le n\right)=\mathbb{P}\left(\max_{1\le j\le L_1-1} Z_j^{(ts)}\le n\right). \tag{2.17}$$

Put for $r,t,s\in\{2,3\}$

$$Q_{rts}=Q_{rts}(n)=\mathbb{P}\left(\bigcap_{j=1}^{r-1}\{Z_j^{ts}\le n\}\right)=\mathbb{P}\left(\max_{\substack{1\le i_1\le(r-1)(m_1-1)\\1\le .i_2\le(t-1)(m_2-1)\\1\le i_3\le(s-1)(m_3-1)}} Y_{i_1i_2i_3}\le n\right) \tag{2.18}$$

Then, under supplementary condition that $Q_{2ts}\ge 1-\alpha_{3ts}\ge 0.9$, we apply the result in Theorem 2.1 to obtain

$$|Q_{ts}-H\left(Q_{2ts},Q_{3ts},L_1\right)|\le(L_1-1)F(\alpha_{3ts},L_1-1)(1-Q_{2ts})^2. \tag{2.19}$$

Combining the Eqs.(2.10), (2.15) and (2.19) we obtain an approximation formula for the distribution of the scan statistic depending on the eight quantities Q_{rts}, that we propose to evaluated by simulation. Note that in the above approximations, at

each step we consider different values for α. In the next section we show how to choose these values.

Remark 2.2. *If T_1, T_2 and T_3 are not multiples of m_1-1, m_2-1 and m_3-1, respectively, then let consider $L_j = \left\lfloor \frac{T_j}{m_j-1} \right\rfloor$ for $j \in \{1,2,3\}$. Based on the inequalities*

$$\mathbb{P}(S(L_1+1, L_2+1, L_3+1) \leq n) \leq \mathbb{P} \leq \mathbb{P}(S(L_1, L_2, L_3) \leq n), \tag{2.20}$$

we can approximate $\mathbb{P} = \mathbb{P}(S_{m_1,m_2,m_3}(T_1,T_2,T_3) \leq n)$ by linear interpolation (see Table 3).

2.1. **Computing the approximation error.** To simplify the presentation and the derivation of the approximation formulae, it is convenient to introduce the following notations for $s,t \in \{2,3\}$:

$$\begin{aligned}
&\alpha_3 = 1-Q_3,\ \alpha_{23} = 1-Q_{23},\ \alpha_{233} = 1-Q_{233},\\
&\gamma_{ts} = H(Q_{2ts}, Q_{3ts}, L_1),\ \gamma_s = H(\gamma_{2s}, \gamma_{3s}, L_2),\\
&F_1 = F(\alpha_3, L_3-1),\ F_2 = F(\alpha_{23}, L_2-1),\ F_3 = F(\alpha_{233}, L_1-1).
\end{aligned}$$

It is not hard to see that $Q_3 \leq Q_2$, $Q_{23} \leq Q_{22}$ and $Q_{233} \leq Q_{2ts}$, so that the choice for the thresholds α_3, α_{23} and α_{233} becomes natural. Based on the mean value theorem in two dimensions, one can easily verify that for $m \geq 6$ and $y_i \leq x_i$, $i \in \{1,2\}$ we have the inequality:

$$|H(x_1,y_1,m) - H(x_2,y_2,m)| \leq (m-2)\left[|x_1-x_2| + |y_1-y_2|\right]. \tag{2.21}$$

In what follows we use the result from Eq.(2.21) without restrictions. This is in agreement with the numerical values considered in Section 4. We begin by observing that applying Eq.(2.21) into Eq.(2.10) we obtain

$$\begin{aligned}
|\mathbb{P} - H(\gamma_2,\gamma_3,L_3)| &\leq |\mathbb{P} - H(Q_2,Q_3,L_3)| + |H(Q_2,Q_3,L_3) - H(\gamma_2,\gamma_3,L_3)|\\
&\leq (L_3-1)F_1(1-Q_2)^2 + (L_3-2)\left[|Q_2-\gamma_2| + |Q_3-\gamma_3|\right],
\end{aligned} \tag{2.22}$$

where for simplicity we used the notation $\mathbb{P} = \mathbb{P}(S(L_1,L_2,L_3) \leq n)$. In the same manner, one can see that for $s \in \{2,3\}$ we have

$$\begin{aligned}
|Q_s - \gamma_s| &\leq |Q_s - H(Q_{2s},Q_{3s},L_2)| + |H(Q_{2s},Q_{3s},L_2) - H(\gamma_{2s},\gamma_{3s},L_2)|\\
&\leq (L_2-1)F_2(1-Q_{2s})^2 + (L_2-2)\left[|Q_{2s}-\gamma_{2s}| + |Q_{3s}-\gamma_{3s}|\right].
\end{aligned} \tag{2.23}$$

We notice that Eq.(2.19) can be rewritten as

$$|Q_{ts} - \gamma_{ts}| \leq (L_1-1)F_3(1-Q_{2ts})^2, \quad s,t \in \{2,3\}. \tag{2.24}$$

Finally, in order to find the approximation error it is sufficient to determine bounds for $1-Q_2$ and $1-Q_{2s}$. It can be easily checked that

$$1-Q_{2s} \leq 1-\gamma_{2s} + |Q_{2s}-\gamma_{2s}| \leq \delta_{2s} \tag{2.25}$$

where

$$\delta_{2s} = 1-\gamma_{2s} + (L_1-1)F_3(1-Q_{22s})^2. \tag{2.26}$$

Similarly, we can write

$$1-Q_2 \leq 1-\gamma_2 + |Q_2-\gamma_2| \leq \delta_2, \tag{2.27}$$

with

$$\delta_2 = 1-\gamma_2 + (L_2-1)F_2\delta_{22} + (L_2-2)(L_1-1)F_3\left[(1-Q_{222})^2 + (1-Q_{232})^2\right]. \tag{2.28}$$

Substituting Eqs.(2.23), (2.24), (2.25) and (2.27) in Eq.(2.22) we derive the formula for the approximation error

$$E_{app} = (L_3-1)F_1\delta_2^2 + (L_3-2)(L_2-1)F_2\left(\delta_{22}^2+\delta_{23}^2\right) + \\ + (L_3-2)(L_2-2)(L_1-1)F_3\left[\sum_{t,s\in\{2,3\}}(1-Q_{2ts})^2\right]. \tag{2.29}$$

2.2. **Computing the simulation errors.** Since, from our knowledge, there are no exact formulas available for the computation of Q_{rts} we propose to evaluate them by simulation. It is obvious that the simulation error appears from two terms: first, from the approximation formula in Eq.(2.22) and second, from the error bound in Eq.(2.29).
Usually, between the true and the estimated value we have a relation of the form

$$\left|Q_{rts}-\hat{Q}_{rts}\right| \le \beta_{rts}, \quad r,t,s\in\{2,3\} \tag{2.30}$$

where $\hat{Q}_{rts}$ are the simulated values corresponding to Q_{rts}. Provided a simulation error bound β_{rts} as in Eq.(2.30), let denote the simulated values by

$$\hat{Q}_{ts} = H(\hat{Q}_{2ts},\hat{Q}_{3ts},L_1),$$
$$\hat{Q}_{s} = H(\hat{Q}_{2s},\hat{Q}_{3s},L_2).$$

From Eq.(2.21) one obtains

$$\left|H(\gamma_2,\gamma_3,L_3) - H\left(\hat{Q}_2,\hat{Q}_3,L_3\right)\right| \le (L_3-2)\left[\left|\gamma_2-\hat{Q}_2\right| + \left|\gamma_3-\hat{Q}_3\right|\right]. \tag{2.31}$$

Observe that the differences in the right hand term in Eq.(2.31) can be bounded by

$$\left|\gamma_s-\hat{Q}_s\right| = \left|H(\gamma_{2s},\gamma_{3s},L_2) - H\left(\hat{Q}_{2s},\hat{Q}_{3s},L_2\right)\right| \\ \le (L_2-2)\left[\left|\gamma_{2s}-\hat{Q}_{2s}\right| + \left|\gamma_{3s}-\hat{Q}_{3s}\right|\right]. \tag{2.32}$$

In the same way we can write for $t,s\in\{2,3\}$

$$\left|\gamma_{ts}-\hat{Q}_{ts}\right| = \left|H(\gamma_{2ts},\gamma_{3ts},L_1) - H\left(\hat{Q}_{2ts},\hat{Q}_{3ts},L_1\right)\right| \\ \le (L_1-2)\left[\left|\gamma_{2ts}-\hat{Q}_{2ts}\right| + \left|\gamma_{3ts}-\hat{Q}_{3ts}\right|\right] \\ \le (L_1-2)\left[\beta_{2ts}+\beta_{3ts}\right]. \tag{2.33}$$

Combining Eqs.(2.33), (2.32) and (2.31) we get the simulation error corresponding to the approximation formula

$$E_{sf} = (L_1-2)(L_2-2)(L_3-2)\left(\sum_{r,t,s\in\{2,3\}}\beta_{rts}\right). \tag{2.34}$$

In order to obtain the simulation error corresponding to the approximation error bound in Eq.(2.29) we follow the lines of Section 2.1. With the following notations

$$u_{rts} = 1-\hat{q}_{rts}+\beta_{rts},$$
$$u_{ts} = 1-\hat{q}_{ts}+(L_1-2)(\beta_{2ts}+\beta_{3ts}),$$
$$u_{s} = 1-\hat{q}_{s}+(L_1-2)(L_2-2)(\beta_{22s}+\beta_{32s}+\beta_{23s}+\beta_{33s}),$$
$$\bar{\delta}_{2s} = u_{2s}+(L_1-1)F_3u_{22s}^2,$$
$$\bar{\delta}_{2} = u_{2}+(L_2-1)F_2\bar{\delta}_{22}+(L_2-2)(L_1-1)F_3(u_{222}^2+u_{232}^2),$$

the error can be expressed as

$$E_{sapp} = (L_3-1)F_1\bar{\delta}_2^2 + (L_3-2)(L_2-1)F_2\left(\bar{\delta}_{22}^2+\bar{\delta}_{23}^2\right) + \\ + (L_3-2)(L_2-2)(L_1-1)F_3\left(\sum_{t,s\in\{2,3\}} u_{2ts}^2\right). \tag{2.35}$$

The total simulation error is obtained by adding the two terms from Eq.(2.34) and Eq.(2.35)

$$E_{sim} = E_{sf} + E_{sapp}. \tag{2.36}$$

To evaluate Eq.(2.36), one needs to find suitable values for the bounds β_{rts}. If $ITER$ is the number of iterations used in the Monte Carlo simulation algorithm for the estimation of Q_{rts} then, one can consider, for example, the naive bound provided by the Central Limit Theorem with a 95% confidence level

$$\beta_{rts} = 1.96\sqrt{\frac{\hat{Q}_{rts}(1-\hat{Q}_{rts})}{ITER}}. \tag{2.37}$$

This bound has been used with some success for the two dimensional case (see Haiman and Preda [2006]). As the authors pointed out, the main contribution to the total error is due to the simulation error, especially for small sizes of the window scan with respect to the scanning region. Our numerical study shows that Eq.(2.37) is not feasible for the three dimensional case, the simulation error being to large with respect to the approximation error. Thus, for the simulation of $\hat{Q}_{rts}$, we use an importance sampling technique introduced in Naiman and Priebe [2001]. Next section illustrates how to adapt theirs algorithm to our problem.

3. Simulation by importance sampling

In this section we present a simulation method for Q_{rts}, which gives an unbiased estimate whose variance is typically smaller then that of the naive hit or miss Monte Carlo approach. The method is an adaptation of the importance sampling algorithm developed in Naiman and Priebe [2001] to our problem. The main idea behind is to express the tail of the scan distribution as a Bonferroni upper bound (B) with some correction factor (ρ). Let define for $1 \le i_j \le N_j$, $j \in \{1,2,3\}$ the events $A_{i_1i_2i_3} = \{Y_{i_1i_2i_3} \ge \tau\}$. Then

$$\begin{aligned}\mathbb{P}\left(S_{m_1,m_2,m_3}(T_1,T_2,T_3) \ge \tau\right) &= \mathbb{P}\left(\bigcup_{i_1=1}^{T_1-m_1+1}\bigcup_{i_2=1}^{T_2-m_2+1}\bigcup_{i_3=1}^{T_3-m_3+1} A_{i_1i_2i_3}\right) \\ &= B\sum_{i_1=1}^{T_1-m_1+1}\sum_{i_2=1}^{T_2-m_2+1}\sum_{i_3=1}^{T_3-m_3+1} p_{i_1i_2i_3}I(i_1,i_2,i_3) \\ &= B\rho, \end{aligned} \tag{3.1}$$

where

$$\rho = \sum_{i_1=1}^{T_1-m_1+1}\sum_{i_2=1}^{T_2-m_2+1}\sum_{i_3=1}^{T_3-m_3+1} p_{i_1i_2i_3}I(i_1,i_2,i_3). \tag{3.2}$$

Under the null hypothesis (H_0), B is the Bonferroni upper bound given by

$$\begin{aligned} B &= \sum_{i_1=1}^{T_1-m_1+1}\sum_{i_2=1}^{T_2-m_2+1}\sum_{i_3=1}^{T_3-m_3+1} \mathbb{P}(A_{i_1i_2i_3}) \\ &= (T_1-m_1+1)(T_2-m_2+1)(T_3-m_3+1)\mathbb{P}(A_{111}), \end{aligned} \tag{3.3}$$

$p_{i_1i_2i_3}$ defines an uniform probability distribution over $\{1,\dots,T_1-m_1+1\}\times\{1,\dots,T_2-m_2+1\}\times\{1,\dots,T_3-m_3+1\}$,

$$p_{i_1i_2i_3}=\frac{\mathbb{P}(A_{i_1i_2i_3})}{\sum_{s_1=1}^{T_1-m_1+1}\sum_{s_2=1}^{T_2-m_2+1}\sum_{s_3=1}^{T_3-m_3+1}\mathbb{P}(A_{s_1s_2s_3})}=\frac{1}{(T_1-m_1+1)(T_2-m_2+1)(T_3-m_3+1)},\tag{3.4}$$

and $I(i_1,i_2,i_3)=\int\frac{1}{C(Y)}\frac{\mathbf{1}_{A_{i_1i_2i_3}}}{\mathbb{P}(A_{i_1i_2i_3})}d\mathbb{P}$ where $C(Y)$ represents the number of triples (i_1,i_2,i_3) such that $Y_{i_1i_2i_3}$ exceeds the threshold τ, that is

$$C(Y)=\sum_{i_1=1}^{T_1-m_1+1}\sum_{i_2=1}^{T_2-m_2+1}\sum_{i_3=1}^{T_3-m_3+1}\mathbf{1}_{A_{i_1i_2i_3}}.\tag{3.5}$$

Based on these identities the simulation algorithm (similar with the one described in Naiman and Priebe [2001, page 303]) can be written as follows:

Begin

Repeat for each k from 1 to $ITER$ (iterations number)

Step 1 Generate $T\in\{\tau,\dots\}$ according to the probabilities

$$p_T(t)=\frac{\mathbb{P}(Y_{111}=t)}{\sum_{s\geq\tau}\mathbb{P}(Y_{111}=s)},\quad t\geq\tau.$$

Step 2 Conditionally, given $T=t$, generate the triple (J_1,J_2,J_3) in the set $\{1,\dots,T_1-m_1+1\}\times\{1,\dots,T_2-m_2+1\}\times\{1,\dots,T_3-m_3+1\}$ uniformly.

Step 3 Conditionally, given T and (J_1,J_2,J_3), generate the set of random variables $\{\tilde{Y}_{i_1i_2i_3}|J_s\leq i_s\leq J_s+m_s-1,s\in\{1,2,3\}\}$, uniformly from the set of all the vectors of length $m_1\times m_2\times m_3$ over the set of values taken by $Y_{i_1i_2i_3}$ and whose sum is equal with T. Take the remaining $\tilde{Y}_{i_1i_2i_3}$ to be i.i.d. and distributed according to the null hypothesis distribution.

Step 4 Take $c_k=C(\tilde{Y}_k)$ the number of all triples (i_1,i_2,i_3) such that $\tilde{Y}_{i_1i_2i_3}\geq T$ and put $\hat{\rho}_k=\frac{1}{c_k}$.

End Repeat

Return $\hat{\rho}=\frac{1}{ITER}\sum_{k=1}^{ITER}\hat{\rho}_k$.

End

Clearly, $\hat{\rho}$ is an unbiased estimator for ρ with estimated variance

$$Var(\hat{\rho})\approx\frac{1}{ITER-1}\sum_{k=1}^{ITER}\left(\hat{\rho}_k-\frac{1}{ITER}\sum_{k=1}^{ITER}\hat{\rho}_k\right)^2.\tag{3.6}$$

For $ITER$ sufficiently large, as a consequence of CLT the error between the true and the estimated value of the tail $\mathbb{P}\left(S_{m_1,m_2,m_3}(T_1,T_2,T_3)\geq\tau\right)$, corresponding to a 95% confidence level, is given by

$$\beta=1.96B\sqrt{\frac{Var(\hat{\rho})}{ITER}}.\tag{3.7}$$

Notice that for the simulation of Q_{rts}, we substitute T_1, T_2 and T_3 in the above relations with $r(m_1-1)$, $t(m_2-1)$ and $s(m_3-1)$ respectively. Therefore, we obtain the corresponding values for β_{rts} as described by Eq.(3.7).

4. Numerical values for Binomial and Poisson models

In this section, for selected values of the parameters of the binomial and Poisson distributions, we evaluate the approximation introduced in Section 2 and provide the corresponding error bounds. We show the contributions of the approximation (Eq.(2.29)) and simulation (Eq.(2.36)) errors in the overall error.
For all our simulations we used the importance sampling algorithm with $ITER = 10^5$ replications. We compare our results with those existing in literature, see Glaz, Guerriero and Sen [2010] for Bernoulli model, and with the simulated value of the scan statistics obtained by scanning the whole region $\mathcal{R}$, denoted by $\hat{\mathbb{P}}(S \leq n)$. The scanning of $\mathcal{R}$ being more time consuming than the scanning of the subregions corresponding to Q_{rst}, we used 10^3 repetitions of the algorithm.
In Table 1, we compare the results obtained by our approximation with the product type approximation presented by Glaz, Guerriero and Sen [2010]. We observe that our approximation is very sharp.

Table 1. Approximation for $\mathbb{P}(S \leq n)$ in Bernoulli case: $m_1 = m_2 = m_3 = 5, T_1 = T_2 = T_3 = 60, ITER = 10^5$

n	$\hat{\mathbb{P}}(S \leq n)$	Glaz et al. Product type	Our Approximation	E_{app} Eq.(2.29)	E_{sim} Eq.(2.36)	Total Error
			$p = 0.00005$			
1	0.841806	0.841424	0.851076	0.011849	0.064889	0.076738
2	0.999119	0.999142	0.999192	0.000000	0.000170	0.000170
3	0.999997	0.999998	0.999997	0.000000	3×10^{-7}	3×10^{-7}
			$p = 0.0001$			
2	0.993294	0.993241	0.993192	0.000010	0.001367	0.001377
3	0.999963	0.999964	0.999963	0.000000	0.000005	0.000005
4	0.999999	0.999999	0.999999	0.000000	2×10^{-9}	2×10^{-9}

Table 2 presents the numerical results obtained by scanning the region $\mathcal{R}$ of size $60 \times 60 \times 60$ with two windows of the same volume but different sizes, first a cubic window of size $4 \times 4 \times 4$ and second a rectangular region of size $8 \times 4 \times 2$. We observe that the results are closely related, but significantly different.
In Table 3 we have included numerical values emphasizing the situation described by Remark 2.2. We consider the Bernoulli model of parameter $p = 0.0001$ over the region $\mathcal{R}$ of size $185 \times 185 \times 185$ and scan it with a cubic window of length 10. The second and forth columns gives the values corresponding to the bounds described in Eq.(2.20), while in the third column we presented the simulated values for $\mathbb{P}\left(S_{10,10,10}(185, 185, 185) \leq n\right)$.
In order to compare the binomial and Poisson models, in Table 4, we have evaluated the distribution of the scan statistics over a region of size $84 \times 84 \times 84$ scanned with a $4 \times 4 \times 4$ cubic window, in the two situations. In the first case we have a binomial random field with parameters m and p, that is $X_{ijk} \sim B(m, p)$, while in the second we considered that $X_{ijk} \sim P(\lambda)$, with $\lambda = mp$.
Notice that the contribution of the approximation error (E_{app}) to the total error is almost negligible in most of the cases with respect to the simulation error (E_{sim}).

TABLE 2. Approximation for $\mathbb{P}(S \leq n)$ over the region $\mathcal{R}$ with windows of the same volume by different sizes: $T_1 = T_2 = T_3 = 60, p = 0.0025, ITER = 10^5$

n	$\hat{\mathbb{P}}(S \leq n)$	Our Approximation	E_{app} Eq.(2.29)	E_{sim} Eq.(2.36)	Total Error
		$m_1 = m_2 = m_3 = 4$			
5	0.961691	0.963506	0.000038	0.003622	0.003660
6	0.999006	0.999023	0.000000	0.000071	0.000071
7	0.999980	0.999980	0.000000	0.000001	0.000001
8	0.999999	0.999999	0.000000	2×10^{-9}	2×10^{-9}
		$m_1 = 8, m_2 = 4, m_3 = 2$			
5	0.969189	0.969110	0.000007	0.003387	0.003395
6	0.999297	0.999228	0.000000	0.000071	0.000071
7	0.999984	0.999984	0.000000	0.000001	0.000001
8	0.999999	0.999999	0.000000	2×10^{-9}	2×10^{-9}

TABLE 3. Approximation for $\mathbb{P}(S \leq n)$ based on Eq.(2.20): $m_1 = m_2 = m_3 = 10, T_1 = T_2 = T_3 = 185, L_1 = L_2 = L_3 = 20, ITER = 10^5$

n	$\mathbb{P}\left(S(L_1+1, L_2+1, L_3+1) \leq n\right)$	$\hat{\mathbb{P}}(S \leq n)$	$\mathbb{P}\left(S(L_1, L_2, L_3) \leq n\right)$
4	0.97524633	0.97465263	0.97491935
	(±0.00754004)	(±0.00618987)	(±0.00643099)
5	0.99931055	0.99935163	0.99938629
	(±0.00015833)	(±0.00014759)	(±0.00013490)
6	0.99998641	0.99998632	0.99998784
	(±0.00000272)	(±0.00000326)	(±0.00000230)

TABLE 4. Approximation for $\mathbb{P}(S \leq n)$ in Binomial and Poisson cases: $m_1 = m_2 = m_3 = 4, T_1 = T_2 = T_3 = 84, ITER = 10^5$

n	$\hat{\mathbb{P}}(S \leq n)$	Our Approximation	E_{app} Eq.(2.29)	E_{sim} Eq.(2.36)	Total Error
		Binomial :	$m = 10, p = 0.0025$		
10	0.726386	0.723224	0.007763	0.032197	0.039960
11	0.954605	0.955417	0.000123	0.003079	0.003202
12	0.993938	0.993906	0.000001	0.000331	0.000333
13	0.999289	0.999284	0.000000	0.000033	0.000033
14	0.999923	0.999921	0.000000	0.000003	0.000003
15	0.999992	0.999992	0.000000	3×10^{-7}	3×10^{-7}
		Poisson :	$\lambda = 0.025$		
10	0.713184	0.708481	0.009211	0.035294	0.044506
11	0.950947	0.950197	0.000143	0.003345	0.003488
12	0.993624	0.993452	0.000002	0.000365	0.000367
13	0.999218	0.999210	0.000000	0.000038	0.000038
14	0.999912	0.999911	0.000000	0.000003	0.000003
15	0.999990	0.999990	0.000000	3×10^{-7}	3×10^{-7}

Thus, the precision of the method will depend mostly on the number of iterations ($ITER$) used to estimate Q_{rts}.

The time required for the computations presented in this section was about two hours for each table on a computer of medium performances. The programs are written in MATLAB and are available from the authors.

5. Conclusions

In this article we derived an approximation for the three dimensional discrete scan statistic viewed as the maximum of a 1-dependent stationary sequence of random variables. We also provide the corresponding theoretical and simulation error bounds. In the three dimensional scan statistics framework, it is essential to reduce the variance of simulated values. For this purpose we used an importance sampling method. A simulation study for the binomial and Poisson models shows the accuracy as well as the limit of our method.

References

Amarioarei, A.: Approximation for the distribution of extremes of one dependent stationary sequences of random variables. arXiv:1211.5456v1(submitted)

Boutsikas, M.V., Koutras, M.: Reliability approximations for Markov chain imbeddable systems. *Methodol Comput Appl Probab* **2** (2000), 393–412.

Darling, R., Waterman, M.: Approximations for three dimensional scan statistic. *SIAM J. Appl Math* **46** (1986), 118–132.

Glaz, J., Naus, J., Wallenstein, S.: Scan statistic. *Springer* (2001).

Glaz, J., Pozdnyakov, V., Wallenstein, S.: Scan statistic: Methods and Applications. *Birkhauser* (2009).

Glaz, J., Guerriero, M., Sen, R.: Approximations for three dimensional scan statistic. *Methodol Comput Appl Probab* **12** (2010), 731–747.

Haiman, G.: First passage time for some stationary sequence. *Stochastic Processes and their Applications* **80** (1999), 231–248.

Haiman, G.: Estimating the distribution of scan statistics with high precision. *Extremes* **3** (2000), 349–361.

Haiman, G., Preda, C.: A new method for estimating the distribution of scan statistics for a two-dimensional Poisson process. *Methodology and Computing in Applied Probability* **4** (2002), 393–407.

Haiman, G., Preda, C.: Estimation for the distribution of two-dimensional scan statistics. *Methodology and Computing in Applied Probability* **8** (2006), 373–381.

Haiman, G.: Estimating the distribution of one-dimensional discrete scan statistics viewed as extremes of 1-dependent stationary sequences. *J. Stat Plan Infer* **137** (2007), 821–828.

Naiman, D., Priebe C.: Computing Scan Statistic p Values Using Importance Sampling, with Applications to Genetics and Medical Image Analysis. *J. Comp Graph Stat* **10** (2001), 296–328.

E-mail address: alexandru.amarioarei@inria.fr

E-mail address: cristian.preda@polytech-lille.fr