FAST ONLINE ANOMALY DETECTION USING SCAN STATISTICS

Ryan Turner and Zoubin Ghahramani

University of Cambridge
Department of Engineering
Trumpington Street, Cambridge CB2 1PZ, UK

Steven Bottone

Rockwell Collins Inc.
400 Collins Road NE,
Cedar Rapids, IA 52498, USA

ABSTRACT

We present methods to do fast online anomaly detection using scan statistics. Scan statistics have long been used to detect statistically significant bursts of events. We extend the scan statistics framework to handle many practical issues that occur in application: dealing with an unknown background rate of events, allowing for slow natural changes in background frequency, the inverse problem of finding an unusual lack of events, and setting the test parameters to maximize power. We demonstrate its use on real and synthetic data sets with comparison to other methods.

1. INTRODUCTION

Scan statistics are a powerful method for detecting unusually high rates of events, also called anomalies. Scanning for bursts of events has many applications in diverse fields such as telecommunications, epidemiology, molecular biology, astronomy, quality control, and reliability [1, 2]. In monitoring and control of communication networks, scan statistics can be used to monitor the occurrence of events in time, a *point process*, such as status messages, alarms, and faults. We are not looking for outliers, but rather unusual bursts in events. In an online application for events occurring in time, the number of events which have occurred in the *scanning window* $[t - w, t]$, where t is the current time and w is the scanning window size, are compared with the number of events expected to have occurred in that window under normal conditions. If that number of events is large compared to what is expected, then an alert of an abnormal condition can be given. Scan statistics can be used to compute the distribution of events under normal conditions (the null hypothesis, H_0) to determine what is a significantly large number (the critical value) in the scanning window, while properly controlling the false positive rate (FPR), which is the probability of exceeding the critical value for any scanning window of size w in the larger time interval $[0, T)$ under H_0. A key advantage of scan statistics is that they allow for computationally simple implementation; therefore, it is possible to monitor many processes at once with a small computational burden.

In the usual treatment of scan statistics the times of events occurring in the interval $[0, T)$ are assumed to be generated by a Poisson process under H_0.[1]

This paper's contribution addresses four practical problems with scan statistics: finding the optimal window size to optimize power (Section 3.1), controlling the FPR in the presence of an unknown background rate (Section 3.2), detecting an unusual lack of events (Section 3.3), and allowing for slow natural changes in the background rate (Section 3.4). We present a very fast method for updating the estimated background rate. Finally, in Section 4 we test our methods on synthetic and real-world data sets from meteorology and geology.

2. SCAN STATISTICS

Assuming that λ is known, the scan statistic is defined as follows. Let $X_1, X_2, \ldots, X_N$ denote the ordered values of the events occurring in the interval $[0, T)$ and let $Y_t(w)$ be the number of points (X's) in the interval $[t - w, t]$.[2] The scan statistic S_w is then defined as the maximum number of points to be found in any subinterval of $[0, T)$ of length w. That is,

$$S_w := \max_{w \le t \le T} Y_t(w) = \max_t \sum_{i=1}^{N} \mathbb{I}\{t - w \le X_i \le t\} . \quad (1)$$

A related statistic is W_k, the minimum subinterval of $[0, T)$ containing k points

$$W_k := \min_{0 \le w \le T} \{w \colon S_w \ge k\} = \min_{1 \le i} (X_{i+k-1} - X_i) . \quad (2)$$

The distributions of these statistics are related by $P(S_w \ge k) = P(W_k \le w)$. Equivalently, S_w and W_k are inverses: $S_{W_k} = k$ for $k \le N$.[3] The key trick in scan statistics is

[1] In a Poisson process with rate λ over $[0, T)$, the number of events is given by Poisson(λT). The inter-arrival times are iid distributed according to Exponential(λ). Conditional on there being k events, the times of the events are distributed uniformly in $[0, T)$.

[2] Extensions of scan statistics exist in discrete time and on circular data, such as time of year, but we do not focus on them here.

[3] One should also note the edge cases of S_w and W_k: $W_k = \infty$ for $k > N$, $W_1 = 0$ and $S_0 = 1$ for $N \ge 1$, and $S_T = N$.

controlling the FPR by accounting for the overlapping multiple comparisons that are a result of the rolling scan window while maintaining more power than simple Bonferroni correction.

For a Poisson process with mean rate λ per unit time over the interval $[0, T)$, [3] (see also [4]) gives the following approximation for the distribution $P(S_w \geq k \,|\, \mu, L)$, where $\mu := \lambda w$ and $L := T/w$ (also equal to $P(W_k \leq w \,|\, \mu, L)$). Let $p(k; \mu)$ be the probability of exactly k events occurring for a Poisson distribution with mean μ and $F(k; \mu)$ the cumulative distribution function (CDF) for the Poisson, then

$$P(S_w \geq k \,|\, \mu, L) \approx 1 - Q_2 (Q_3/Q_2)^{L-2}\,, \tag{3}$$

$$\begin{aligned} Q_2 &:= F(k-1,\mu)^2 - (k-1)p(k;\mu)p(k-2;\mu) \\ &\quad - (k-1-\mu)p(k;\mu)F(k-3;\mu)\,, \end{aligned}$$

$$Q_3 := F(k-1,\mu)^3 - A_1 + A_2 + A_3 - A_4\,, \tag{4}$$

where

$$\begin{aligned} A_1 &:= 2\,p(k;\mu)F(k-1;\mu) \\ &\quad \times [(k-1)F(k-2;\mu) - \mu F(k-3;\mu)]\,, \\ A_2 &:= 0.5\,p(k;\mu)^2\,[(k-1)(k-2)F(k-3;\mu) \\ &\qquad - 2(k-2)\mu F(k-4;\mu) + \mu^2 F(k-5;\mu)]\,, \\ A_3 &:= \sum_{i=1}^{k-1} p(2k-i;\mu)F(i-1;\mu)^2\,, \\ A_4 &:= \sum_{i=2}^{k-1} p(2k-i;\mu)p(i;\mu) \\ &\quad \times [(i-1)F(i-2;\mu) - \mu F(i-3;\mu)]\,. \end{aligned}$$

To test the null hypothesis, H_0, that the background rate $\lambda = \lambda_0 = \text{constant}$ at the significance level α, find the smallest k, which we call k_{Crit}, such that

$$P(S_w \geq k_{\text{Crit}} \,|\, \mu_0, L) \leq \alpha\,, \tag{5}$$

where $\mu_0 := \lambda_0 w$. For an online test with fixed w, if at time t (the current time) the number of points, k, occurring in the time interval of length w ending at t, $[t-w, t]$, is $\geq k_{\text{Crit}}$, then the null hypothesis is rejected at significance level α and an alert may be given indicating that the rate of events has likely increased. An equivalent alternative test is to determine the length of time separating the most recent k_{Crit} points, $W_{\text{Crit}} := X_i - X_{i-k_{\text{Crit}}+1}$, where $X_i = t$. If $W_{\text{Crit}} \leq w$ then an alert may be given.

3. FOUR PROBLEMS WITH SCAN STATISTICS

In this section we address four practical problems with scan statistics. Firstly, we would like to make the choice of w less arbitrary. Second, we want to estimate λ while accounting for the estimation error. Third, define a test to look for an unusual lack of points. Finally, we look at updating our estimate of λ online.

3.1. The Window Size Problem

The window size w in scan statistics is typically treated as arbitrary and ignored in the literature. We can get much different results depending on the window size, so it is unsatisfactory to have it set arbitrarily. Smaller windows will be quicker to alert while larger windows will detect smaller changes in rate. We show how to set the window size based on the rate change we would like to detect.

We wish to maximize the power of detection for an alternative hypothesis where there is an abrupt change in the rate of the Poisson process from λ_0 to $\lambda_1 = c\lambda_0$, where $c > 1$. We would like to choose the w that minimizes the expected *time to detention* (TTD): $\mathbb{E}\,[\text{TTD}]$.

Using (3) we can solve for the CDF on TTD:

$$\begin{aligned} P(\text{TTD} \leq t) &= P(\text{Alert}|\text{Observed in } [0,t)) \qquad (6) \\ &= P(S_w \geq k_{\text{Crit}} \,|\, \mu = \lambda_1 w, L = t/w) \\ P(\text{TTD}/w \leq L) &= \max(0, 1 - Q_2 e^{\log(Q_3/Q_2)(L-2)})\,, \\ &= \max(0, 1 - e^{-r(L-L_0)})\,, \qquad (7) \end{aligned}$$

$$L_0 := \log(a)/r,\; a := Q_2^3/Q_3^2,\; r := \log(Q_2/Q_3)\,, \tag{8}$$

where Q_2 and Q_3 have been computed using λ_1 and k_{Crit}. $p(\text{TTD})$ is in the form of a shifted exponential distribution. Therefore, $\mathbb{E}\,[\text{TTD}] = w(\log(a)+1)/r$.

The problem is reduced to the following optimization:

$$\begin{aligned} \min \mathbb{E}\,[\text{TTD}] = w(\log(a)+1)/r \quad &\text{wrt} \quad k_{\text{Crit}} \in \mathbb{N}, w \in \mathbb{R}^+ \\ \text{st} \quad P(S_w \geq k_{\text{Crit}} \,|\, \mu_0, L) &\leq \alpha\,. \end{aligned} \tag{9}$$

Increasing w will lower the TTD if the increase is small enough that the inequality does not require k_{Crit} to increase too. Therefore, we can make (9) an equality constraint since the optima will always occur when (9) is an equality. We can implement the joint optimization on k_{Crit} and w in a nested way. In the outer loop we can do a binary search on k_{Crit} that minimizes $\mathbb{E}\,[\text{TTD}]$ using the appropriate w. In the inner loop, we find the appropriate w given k_{Crit} using a bisection search to solve $P(S_w \geq k_{\text{Crit}} \,|\, \mu_0, L) = \alpha$.

Alternatively, the same solution is approximately given by setting w such that: $c\lambda_0 w = \lambda_1 w = k_{\text{Crit}}$. This implementation uses bisection search on w in the outer loop and binary search on k_{Crit} in the inner loop.

We have presented an optimization routine that can be used to set the window in a principled manner. The computational burden is small since the routine only needs to be run when configuring the test.

3.2. The Background Rate Problem

Formulation in [4] assumes known background rate λ. In most real-world applications the true background rate is unknown and must be estimated by $\hat{\lambda}$ from a period of time,

the *training period* $[0, T_{\text{Train}}]$, where the system is assumed to be in a normal state. Underestimating the true rate can lead to a FPR much higher than α per *test period* of length T. Therefore, we must account for the estimation error of λ during training. A common choice for estimating λ is the maximum likelihood estimate (MLE) $\hat{\lambda} = N/T_{\text{Train}}$, where N is the number of events in $[0, T_{\text{Train}}]$. However, it is hard to control for the estimation error using the MLE. Given a procedure for estimating λ, such as the MLE, we can calculate the false positive rate in test $\text{FPR}(\lambda)$:

$$\text{FPR}(\lambda) = \int \text{FPR}(\hat{\lambda}|\lambda) p(\hat{\lambda}|\lambda) d\hat{\lambda}\,, \tag{10}$$

where $\text{FPR}(\hat{\lambda}|\lambda)$ is the false positive rate in a test period of time T if we plug in $\hat{\lambda}$ to the scan statistic if the true background rate is λ. In the hypothesis testing framework we want to control the FPR of our statistic in the worst-case, meaning we want to control the quantity $\max_\lambda \text{FPR}(\lambda) \leq \alpha$. Consequently, we bound $\text{FPR}(\lambda)$ to remove the requirement of knowing the true rate λ by simplifying (10):

$$\begin{aligned}
&\text{FPR}(\lambda) \\
&= \underbrace{\text{FPR}(\lambda|\hat{\lambda} < \lambda)}_{\leq 1} P(\hat{\lambda} < \lambda) + \underbrace{\text{FPR}(\lambda|\hat{\lambda} \geq \lambda)}_{\leq \alpha} P(\hat{\lambda} \geq \lambda) \\
&\leq P(\hat{\lambda} < \lambda) + \alpha P(\hat{\lambda} \geq \lambda) = \beta + \alpha(1-\beta) \qquad (11) \\
&\leq \beta + \alpha,\ \forall_\lambda\,. \qquad (12)
\end{aligned}$$

This means the natural way to bound the FPR is by using the upper end of a one-sided confidence interval on λ, with coverage $1-\beta$, for $\hat{\lambda}$. If the coverage is not exact then the weaker bound (12) must be used instead of (11). We can now provably control the FPR and do not have to reserve residual concern on our results due to estimation error in $\hat{\lambda}$.

3.3. The Low End Problem

Typically, scan statistics focus on finding unusual bursts of events, which is desirable when events are viewed as bad things. However, in some applications the absence of an event might be cause for concern. For instance, if an event is a network synchronization, a long period without one would be justification for alert. For this purpose we define an analogous scan statistic to (2)

$$\tilde{W}_k := \max_{1 \leq i} (X_{i+k+1} - X_i)\,. \tag{13}$$

For simplicity we focus on the $k = 0$ case; we use the longest inter-arrival time as the test statistic in this case.

In order to use $\tilde{W}_k$, we must compute its sample distribution. To do this we first consider the case where N is known, then we will marginalize N out as a second step in the analysis. We also consider $T = 1$ for the time being without loss of generality as the time units can always be rescaled. In the following analysis we make use of the fact that conditional on N the events are uniformly distributed. In the case of $N = 0$ the distribution is trivial:

$$p(\tilde{W}_0 \geq w | N = 0) = 1,\ w \in [0,1]\,. \tag{14}$$

For higher N,

$$\begin{aligned}
p(\tilde{W}_0 \geq w | N = 1) &= 1,\ w \in [0, 1/2], \\
&\quad 2(1-w),\ w \in [1/2, 1] \qquad (15) \\
p(\tilde{W}_0 \geq w | N = 2) &= 1,\ w \in [0, 1/3], \\
&\quad 1 - (3w-1)^2,\ w \in [1/3, 1/2], \\
&\quad 3(1-w)^2,\ w \in [1/2, 1] \qquad (16)
\end{aligned}$$

Consistent with these equations we find the following bounds:

$$p(\tilde{W}_0 \geq w | N) \leq (N+1)(1-w)^N \tag{17}$$

$$p(\tilde{W}_0 \geq w | N) \geq \min(1, 1 - ((N+1)w - 1)^N) \tag{18}$$

The upper bound, (17), is quite tight in the tail region, usually around $p \leq 0.3$, and is exact for $w \geq 1/2$. Now we must marginalize out N as it is not known a priori:

$$\begin{aligned}
p(\tilde{W}_0 \geq w) &= \sum_{N=0}^{\infty} p(\tilde{W}_0 \geq w | N) p(N) \qquad (19) \\
&\leq \sum_{N=0}^{\infty} (N+1)(1-w)^N \lambda^N e^{-\lambda} / N! \qquad (20) \\
&= e^{-w\lambda} \sum_{N=0}^{\infty} (N+1)/N! (\lambda(1-w))^N e^{-\lambda(1-w)} \\
&= e^{-w\lambda} \mathbb{E}\left[N+1\right],\ N \sim \text{Poisson}(\lambda(1-w)) \\
&= e^{-w\lambda}(1 + \lambda(1-w))\,. \qquad (21)
\end{aligned}$$

For the case of general T we have: $p = e^{-w\lambda}(1+\lambda(T-w))$. We can bound the FPR below α in false alarms in T by only alerting when $p \leq \alpha$. The bound is quite tight in the tail of small α and is an effective test for low end problems.

3.4. The Online Problem

In many applications we only want to identify a sudden change in background rate. For instance, in a networking application a sudden spike in packets on a router might mean a failure on another route; while a gradual increase in rate might simply be the result of a network gaining more users. We would like a method that can distinguish between these two scenarios.

We extend scan statistics to allow the test to ignore gradual enough changes in λ over time that a domain expert determines them to be irrelevant. A first step in dealing with a null hypothesis that has a changing background rate is to

explicitly estimate the background rate. For computational simplicity we consider a kernel intensity estimator (KIE), the rate analog of kernel density estimation (KDE). The KIE without edge correction is

$$\hat{\lambda}(t) = \sum_{i=1}^{N} k\left(\frac{t-t_i}{u}\right), \tag{22}$$

where u is the *bandwidth*, the time scale over which the rate naturally changes without concern. We denote the smoothing kernel by $k(\cdot)$ and refer to the time of the ith event with t_i. When exponential kernels are used,

$$k(x) = ue^{-ux},\ x \geq 0\,, \tag{23}$$

the online updating can be made very efficient even for large N. Given we know the estimate $\hat{\lambda}(t)$ we can update after a period Δt:

$$\begin{aligned}\hat{\lambda}(t+\Delta t) &= \sum_{i=1}^{N} ue^{-u(t+\Delta t-t_i)} = e^{-u\Delta t}\sum_{i=1}^{N} ue^{-u(t-t_i)} \\ &= e^{-u\Delta t}\hat{\lambda}(t)\,. \end{aligned} \tag{24}$$

If a new event has occurred at Δt we must add $k(0)$ at the end: $\hat{\lambda}(t+\Delta t) = e^{-u\Delta t}\hat{\lambda}(t) + u$. This estimator is highly biased toward low rate estimates near $t = 0$. We must apply an *edge correction* to remove this bias. [5] recommends using the correction

$$\hat{\lambda}(t) = \sum_{i=1}^{N} k\left(\frac{t-t_i}{u}\right) \Big/ \underbrace{\int_A k\left(\frac{t-\tau}{u}\right) d\tau}_{=:Z}\,, \tag{25}$$

where A is the region observed, usually $[0, t]$. This estimator is unbiased in the case when the true rate is constant. In the case of an exponential kernel $Z = 1 - e^{-ut}$. The update equations with edge correction are

$$\hat{\lambda}(t_i) = \hat{\lambda}(t_i^-) + \frac{u}{1-e^{-ut_i}}\,, \tag{26}$$

$$\hat{\lambda}(t_i^-) = \hat{\lambda}(t_{i-1}+\Delta t),\ \Delta t = t_i - t_{i-1}\,, \tag{27}$$

$$\hat{\lambda}(t+\Delta t) = \hat{\lambda}(t)\frac{1-e^{-ut}}{e^{u\Delta t}-e^{-ut}}\,, \tag{28}$$

where the time t_i^- represents the time immediately before the ith event. Note that as $t \to \infty$ these equations approach those without edge correction, (24), as expected. This setup is computationally trivial, each update is $\mathcal{O}(1)$ in computation and memory as opposed to $\mathcal{O}(N)$ for exact calculation for general KIE; at the same time, we attain the desired behavior ignoring gradual changes.

It is also possible to do efficient online updating using boxcar kernels implemented using a queue data structure rather than manipulating exponentials. More general smoothing kernels can be created by averaging over exponentials or boxcar kernels smoothers of different bandwidths u, which maintains both the unbiasedness property and the tractability of a simple exponential.

The methods in this section find a background rate by smoothing events over time. It is worth noting that if multiple draws of the point processes are observed in parallel we can smooth across point processes. This was done by [1] for scan statistics on spatial count data.

4. EXPERIMENTS AND RESULTS

We evaluate our methods on two real and two synthetic data sets. For synthetic data we analyze the power of the scan statistic when the true rate function $\lambda(t)$ is 1) a step function that increases in rate and 2) a short lived pulse. We compare it to the CUSUM method, linear trend methods, and uniformity tests. Retrospective performance is evaluated by the use of receiver operator characteristic (ROC) curves; online performance is gauge by time to detection (TTD) curves. For real data we investigate changes in frequency of 1) global earthquakes with magnitude 5.0 or greater[4] and 2) occurrences of snow storms in Whistler, BC, Canada.[5] The ROC and TTD curves augment the results showing a bound on the FPR below α. ROC and TTD curves do not reward calibration and therefore only show that scan statistics do not lose anything by controlling FPR.

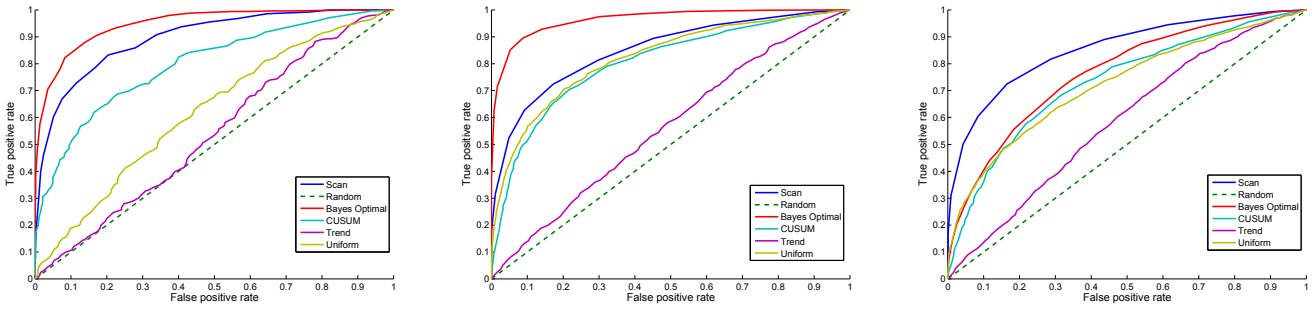
4.1. Synthetic Data

We first consider the step rate function with a $1.3\times$ rate increase at the step and a pulse function with a $2\times$ rate increase. We constructed the results using $N = 5,000$ samples from either a Poisson process with $\lambda_0 = 1$ or $\lambda_A(t)$; we randomly select between the two cases with even probability. In retrospective analysis, the method must provide a score to classify the observed points as either the high or low rate, after observing all the data. For this task the ROC curve can be used: a different threshold on the score will result in different false positive rates (FPR) and true positive rates (TPR). In the online case, the methods must provide a score after each event. For this task we use the time to detection curve. For each score threshold we plot the expected time until the first alert under H_0, the low rate, and H_A the high rate. Under both curves, we would ideally like to see the curves passing through the upper left corner.

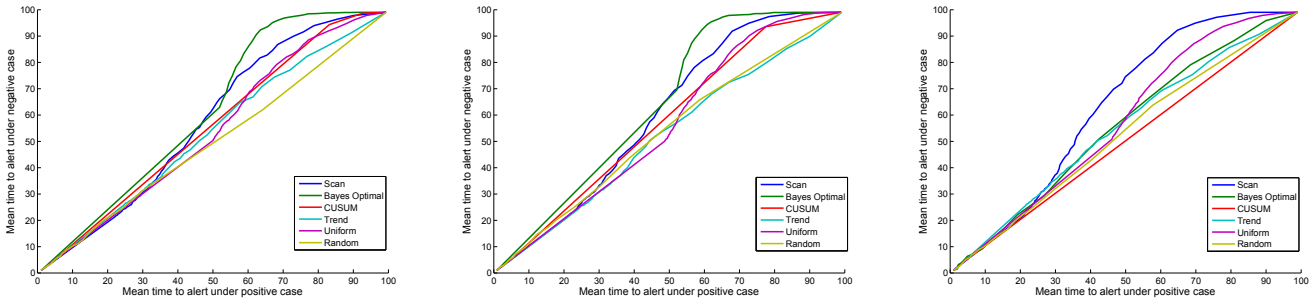
The CUSUM method [6] was originally developed to handle discrete time problems, and has been shown to have optimal power when the data follows a Brownian motion.

[4] http://earthquake.usgs.gov/earthquakes/eqarchives/epic/epic_global.php/

[5] http://www.climate.weatheroffice.ec.gc.ca/ Whistler Roundhouse, id 1108906



(a) Step 1.3× (b) Pulse 2× (c) Random pulse 2×



(d) Step 1.3× (e) Pulse 2× (f) Random pulse 2×

Fig. 1. **Synthetic Data:** The top row shows ROC curves for scan statistics compared to other methods. The bottom row shows the TTD curves. The left column shows results when the positive case is generated from a step in the rate function from $\lambda = 1.0$ to $\lambda = 1.3$ at $t = 50$. The middle column shows when $\lambda = 2.0$ for $t \in [50, 60]$. The right column shows $\lambda = 2.0$ for 10 time units starting at uniformly distributed point in $[0, 90]$. In all cases the negative case is generated from a homogeneous Poisson process with $\lambda = 1.0$.

We can convert a point process problem into a CUSUM problem by binning the points using a small bin size; we have found that a bin size such that there will be 5 points in the bin under λ_0 works well in practice. A bin size giving an expectation near 15 points will become fairly close to normally distributed, matching the CUSUM's H_0 assumptions. Too large a bin size will lead to a slightly longer alerting and throws more information away. Once the points are binned we can also apply a rolling test for a linear trend, where we consider five bins at a time here. We can also apply a uniformity χ^2 test to measure how reasonable the homogeneous Poisson process assumption is.

We also compare to the Bayes' optimal solution: comparing the likelihood ratios assuming we know the different rate functions being considered. We compute the ratio using the true change time. Consequently, it provides a gold standard for the other methods to be compared to.

Figure 1 shows that the linear trend method (LTM) only does slightly better than random on the ROC and TTD curves. We show the area under the curve (AUC) scores in Table 1. The CUSUM has reasonably good performance, but not as good as the scan statistic. The Bayes' optimal method does better than the scan statistic when the time of the λ change is known in advance, giving it an (unfair) advantage over the other methods. When the pulse location is not known in advance, Fig. 1(c) and 1(f), the scan statistic dominates the other curves by a large margin on the ROC and TTD.

Table 1. Comparison on area under the curve (AUC) of the ROC curves on three tasks: step 1.3×, pulse 2×, and random pulse 2×. We separate Bayes' Optimal, which uses information about where the step occurs. The random classifier has AUC 0.5.

Method	Step 1.3×	Pulse 2×	R-pulse 2×
Bayes' Optimal	0.948	0.964	0.766
Scan Stats ⋆	**0.896**	**0.851**	**0.854**
CUSUM	0.797	0.804	0.732
Uniform	0.620	0.822	0.722
LTM	0.539	0.562	0.586

4.2. Snowfall Data

We consider the task of determining a long run change in large snow storm frequency in Whistler, BC, Canada during 1972–2008. We define the event of snow storm to be any day where it snows more than 30 cm, approximately the top 5% of snow days. Clearly the rate function will change as storm frequency depends on the time of year and will

be much higher in winter than summer. So we control for time of year by removing all but one of the months from each year of data, and consider each month independently. For example, we consider if major storms in January are becoming more likely.

We applied the online and offline scan statistics. In the offline case the rate was trained on the first 300 time steps or 6.5 years of data for any given month. We used an FPR period of $T = 35$ years and $\alpha = 0.05$. The window sizes were set automatically to optimize power for a rate increase of 10%. The online scan statistic did not find any significant bursts in any month. The offline method did not alert when controlling for the estimation error in λ. However, when the ordinary MLE was used the offline scan statistic alerted in the April and November data sets. This is likely due to there being few major snow storms in these months and the estimation error in their frequency will be larger, and it must be accounted for. It appears our more conservative offline method avoided spurious alerts that would have occurred had a normal scan statistic been used.

4.3. Earthquake Data

Finally, we consider the frequency of global earthquakes during 1973–2010 with magnitude greater than 5.0 on the Richter scale. We use an FPR period of 40 years and a significance level of $\alpha = 0.05$. In the constant rate setup we train the background rate using the first 10 years of data. We find an MLE $\lambda_{\text{MLE}} = 4.16$ earthquakes per day, to bound the FPR we use the top of a 97.5% ($\beta = 0.025$) confidence interval to get $\lambda_{\text{CI}} = 4.23$ earthquakes per day. Using a window of 100 days we get a critical value of 510 events, optimal power for a 20% increase. If λ were known this would be an FPR of $\alpha = 0.025$, but since we only have $\hat{\lambda}$ we can only bound the FPR to $\alpha + \beta = 0.05$ by (12).

Figure 2 shows that the scan statistic alerts many times on the earthquake data suggesting there are bursts of quakes or that the rate is changing over time.

5. CONCLUSION

We have shown scan statistics are a useful tool for online anomaly detection. Four key practical issues have been addressed. First, we have presented a simple rule for optimizing the power, $\mathbb{E}\left[\text{points in } w \text{ under } H_A\right] = k_{\text{Crit}}$, by picking a window size w, an often ignored and arbitrary parameter. Second, the upper end of a confidence interval on λ can be used to provably control the FPR, solving the often ignored issue of finding the background event rate. Third, scan statistics have traditionally focused on unusual bursts of events, but we have shown how to detect an unusual absence of events. Fourth, kernel estimates of the intensity $\lambda(t)$ can be used for extremely fast updating a changing background; we can allow for slow changes in the background rate without setting off an alert. The allowable speed of the background changes can be tuned by the bandwidth u. In addition to controlling FPR we have shown it is still a powerful method with regard to ROC and quick time to detection, and demonstrated its use on real-world data sets.

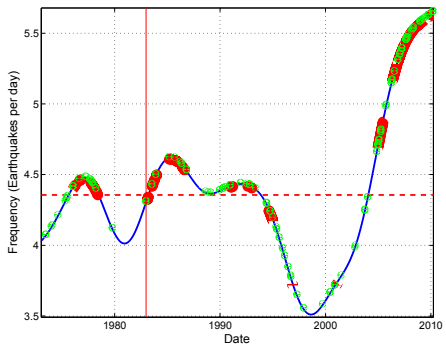


Fig. 2. **Earthquake data:** Estimated rate function (blue, solid) using KIE with Gaussian basis functions and edge correction. The MLE of the rate (red, dashed). The red portions of the rate function mark areas where the scan statistic alerts when no online updating of λ is used. The green dots mark alerts with online updating using KIE with exponential kernels and a bandwidth of 10 years, signaling that the rate is expected to naturally change on that time scale without case of for alert. The vertical red line marks the end of training for the offline scan statistic.

Acknowledgements

We thank Carl Edward Rasmussen and Jeroen Janssens for advice and feedback on this work.

6. REFERENCES

[1] Daniel B. Neill and Gregory F. Cooper, "A multivariate Bayesian scan statistic for early event detection and characterization," *Machine Learning*, vol. 79, pp. 261–282, 2010.

[2] Carey E. Priebe, John M. Conroy, and David J. Marchette, "Scan statistics on Enron graphs," *Computational and Mathematical Organization Theory*, vol. 11, pp. 229–247, 2005.

[3] Joseph I. Naus, "Approximations for distributions of scan statistics," *Journal of the American Statistical Association*, vol. 77, no. 377, pp. 177–183, 1982.

[4] Joseph Glaz, Joseph Naus, and Sylvan Wallenstein, *Scan Statistics*, Springer, 1 edition, August 2001.

[5] Peter Diggle, "A kernel method for smoothing point process data," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 34, no. 2, pp. 138–147, 1985.

[6] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, no. 1/2, pp. 100–115, June 1954.