

**Empirical Calibration of Time Series Monitoring Methods
Using Receiver Operating Characteristic Curves**

Jacqueline Cohen
Samuel Garman
Wilpen Gorr*

H. John Heinz III School of Public Policy and Management
Carnegie Mellon University
Pittsburgh, PA 15213

September 1, 2008

Abstract

Time series monitoring methods, such as the Brown and Trigg methods, have the purpose of detecting pattern breaks (or “signals”) in time series data reliably and in a timely fashion. Traditionally, researchers have used the average run length statistic (ARL) on results from generated signal occurrences in simulated time series data to calibrate and evaluate these methods, with a focus on timeliness of signal detection. This paper investigates the receiver operating characteristic (ROC) framework, well-known in the diagnostic decision making literature, as an alternative to ARL analysis for time series monitoring methods. ROC analysis traditionally uses real data to address the inherent tradeoff in signal detection between the true and false positive rates when varying control limits. We illustrate ROC analysis using time series data on crime at the patrol district level in two cities and use the concept of Pareto frontier ROC curves and reverse functions for methods such as Brown’s and Trigg’s that have parameters affecting signal-detection performance. We compare the Brown and Trigg methods to three benchmark methods, including one commonly used in practice. The Brown and Trigg methods collapse to the same simple method on the Pareto frontier and dominate the benchmark methods under most conditions. The worst method is the one commonly used in practice.

Keywords: time series monitoring, ROC curve, average run length statistic, exponential smoothing, structural breaks, step jumps, outliers

1. Introduction

Time series monitoring methods have the purpose of detecting structural breaks or other unexpected pattern changes in time series data reliably and as soon as possible after those changes occur. In the signal processing literature (e.g., Swets, 1986, 1988), a structural break is the “signal” (or “positive”) we wish to detect and all other variation, for example, from time trend, seasonality, and random error, is considered “noise”. Thus time series monitoring concerns a binary classification problem: either there is an indication that a signal exists or not.

Brown (1959, 1963) developed a time series monitoring method to call attention to changes in product demand that result in degraded forecast performance. It is based on the simple cumulative sum of errors (CUSUM) measure with a single smoothing parameter. Trigg (1964) proposed an improved method that has two smoothing parameters and is widely-used today. These methods use one-step-ahead forecast errors from exponential smoothing as the basis for a continuous statistical decision variable analogous to the z-statistic of hypothesis testing. If the decision variable exceeds a chosen control limit (i.e., is a “signal trip”) then there is an indication of a signal. Generally a signal trip requires further investigation to determine if managerial or other action is required in response.

This paper addresses the Brown and Trigg methods and additional research that has evaluated them and improved their implementation, including Batty, 1969; Golder & Settle, 1976; McKenzie, 1978; Gardner, 1983 and 1985; and McClain, 1988. These researchers used idealized, simulated time series data and the average run length (ARL) statistic for method parameter selection, evaluation, and comparison. Run length is the number of periods after a chosen period until a signal trip. We desire ARL to be large between signal trips for time series undergoing no pattern changes and small (or timely) after a pattern change (or signal).

Unfortunately making the latter smaller (which is desirable) also makes the former smaller (which is undesirable), leading to the necessity of making a tradeoff. None of the preceding papers address this tradeoff. Depending on the application domain, the benefits and costs of detecting signals, the prevalence of signals, and resources available for monitoring, one may accept a smaller ARL between signal trips for time series undergoing no pattern changes (i.e., more false positives) if the likelihood and timeliness of a trip after a signal are sufficiently improved.

This paper introduces an alternative to the ARL statistic and simulated data; namely, the receiver operating characteristic (ROC) framework which uses real data (e.g., Swets, 1988; Swets, Dawes, & Monahan, 2000). First developed to determine the accuracy of radar in World War II for detecting enemy aircraft, ROC analysis is widely used for assessing the accuracy of diagnostic systems; for example, for medical imaging, weather forecasting, information retrieval, materials testing, and polygraph lie detection (Swets, 1988).

ROC analysis directly addresses the tradeoff in signal detection between true and false positive rates, which is another form of the tradeoff in ARLs discussed above. The true positive rate (or power) is the fraction of signals correctly detected by monitoring and the false positive rate (or Type I error in hypothesis testing) is the fraction of non-signal time series data points incorrectly classified as signal (see Section 3 for a further discussion of this terminology). It is desirable to have a high true positive rate and a low false positive rate, but when increasing the true positive rate by decreasing control limits or appropriately changing monitoring method parameters, the false positive rate also increases. Thus it is necessary to balance these two rates based on associated utilities. Fortunately, with the objective estimation or judgmental assessment of a single ratio of utilities and an estimate of the prevalence of signals, it is possible to use the

graphical representation—the ROC graph—of true and false positive rates from a test data set to determine the optimal tradeoff point (e.g., Metz, 1978). ROC analyses use real data for calibration of diagnostic tests to reflect field conditions and thus facilitate tradeoff analysis.

Section 2 provides a literature review of time series monitoring methods and Section 3 is a brief literature review of the ROC framework. Section 4 provides details on our research design employing crime data and the application of alternative methods to detect large crime changes in individual patrol districts within cities. Section 5 provides results for ROC analysis of the crime data and Section 6 concludes the paper.

2. Simple time series monitoring methods

Three of the time series monitoring methods used in this paper—Standardized Forecast Errors, the Brown method, and the Trigg method—are based on one-step-ahead forecast errors made using exponential smoothing. The purpose of the forecast is to estimate the current mean of a time series accounting for drift in model parameters given no structural break or other unexpected pattern change currently, even though one may have just occurred. To make contact with management applications, we call these “business-as-usual” forecasts. When a forecast error or series of errors in the same direction is sufficiently large, a potential pattern change may be present, calling for investigation and managerial attention. Exponential smoothing will adapt with lag after a structural break, but the forecast errors in the interim period allow for detection.

To facilitate discussion of these time series monitoring methods, we introduce notation for them here:

t = time index ranging from 1 to T

T = period for which statistic is calculated

y_t = monitored value (such as forecast error) in t

\bar{y}_T = mean of (y_1, \dots, y_T)

s_T = standard deviation of (y_1, \dots, y_T)

e_t = forecast error for period t made from model estimated from data up to time $t - 1$

e_0 = Initialization value used in MAD = $\sum_{t=1}^6 |e_t| / 6$

$CUSUM_T^k = \sum_{t=T-k+1}^T e_t$, Cumulative Sum of Errors for k periods

α = Smoothing parameter (Trigg numerator)

β = Smoothing parameter (Trigg and Brown denominator)

Exponentially Smoothed Error, $E_T = \sum_{t=1}^T \alpha(1-\alpha)^{T-t} e_t$ (2)

Exponentially Smoothed Mean Absolute Deviation, $MAD_T = (1-\beta)^T |e_0| + \sum_{t=1}^T \beta(1-\beta)^{T-t} |e_t|$ (3)

The k -period Brown method is calculated in period T by summing forecast errors in period T and the $k - 1$ preceding periods and normalizing it by the exponentially smoothed mean absolute deviation in period T, MAD_T :

k - period Brown = $\left| CUSUM_T^k / MAD_T \right|$

As originally proposed by Brown, his method did not use a k -period CUSUM, but a CUSUM from the beginning of the series. Trigg (1964) pointed out that unless intervention is taken after a time series pattern change to reset the CUSUM, the tracking signal may continue to issue false alarms because the cumulative sum will not “forget” the large forecast errors that the signal generated. This was a primary impetus for the development of Trigg’s monitoring method which incorporates exponential smoothing to decrease the weight on older forecast errors. Other CUSUM methods, however, calculate CUSUM in a data window of one or more periods in length as opposed to a sum from the start of the series, thus addressing the reset issue, and

monitor CUSUMs with different window lengths simultaneously (see for instance Harrison & Davies, 1964; Golder & Settle, 1976). We adopt the CUSUM window approach in implementing Brown's method, but we simply use alternative window lengths individually.

The Trigg time series monitoring method replaces the CUSUM numerator of Brown's method with the exponentially smoothed forecast errors, E_T :

$$\text{Trigg} = |E_T / \text{MAD}_T|$$

Smoothing the numerator allows the signal to place decreasing weight on errors as their distance from the current period increases thus allowing the method to reset itself, albeit with lag, after detecting a structural break. Note that k-period Brown's method with a window length of one period is a special case of Trigg with the numerator smoothing constant, $\alpha=1$.

Trigg's initial specification did not include separate smoothing parameters for the numerator and the denominator. Golder & Settle (1976) and Gardner (1983) follow Trigg's initial specification and require both of the time series monitoring method's smoothing parameters to be the same as the smoothing parameter used in the simple exponential smoothing method generating forecast errors. McKenzie (1978) and Gardner (1985) allow the time series monitoring method parameters to differ from the forecast smoothing parameter, but with the numerator and denominator smoothing parameters being equal. McClain (1988) finds that the time series monitoring method generally performs better if both parameters vary and recommends low values for the denominator smoothing parameter. Our findings in Section 5 generally confirm those of McClain.

We examine three other time series monitoring methods in addition to those of Brown and Trigg: Percent Change, Standardized Forecast Error, and Standardized Observed values. Percent Change is a common method used in practice for analyzing performance measures over

time for seasonal time series. For example, the New York City Police Department and Baltimore City management reports use Percent Change for monthly time series data (see New York City Police Department Crime Statistics, 2008 and Baltimore CitiStat Reports and Maps, 2008). This method uses Percentage Change from the same month in the previous year. In this case y_t refers to the observed time series values (e.g., number of crimes per month) instead of forecast error:

$$\text{PercentChange} = 100(y_T - y_{T-12})/y_{T-12}$$

This method has both a positive direction and a negative direction control limit. Percent Change is not defined in cases where the denominator is zero, so we simply define percent change to be $100y_T$ in these cases. This is equivalent to assuming the denominator is one.

Standardized forecast error in period T is the difference between the forecast error in T and the historical mean up to T and then normalized by the historical standard deviation. In this case $y_t = e_t$:

$$\text{Standardized ForecastError} = |(y_T - \bar{y}_T)/s_T|$$

The standardized observations method has the same form except y_t refers to the observed time series values.

Previous research has used ARL as a measure for comparing time series monitoring methods, but it has some inadequacies. Golder & Settle (1976, p 491) lament the high variability of the ARL: "... no suitable alternative suggests itself and, despite this variability, it is felt that this measure will still provide a useful indicator of performance." With simulated data, however, the researcher merely needs to increase the number of replications to obtain a precise ARL, but the variability of the ARL is problematic for application to real data, which generally are limited in sample sizes. McClain (1988, p 566) furthermore points out that ARL is overly-affected by

long runs and therefore is not an appropriate measure if one is interested in detecting a step jump within a certain number of periods. A time series monitoring method with a lower ARL may actually detect a lower percentage of step jumps within a given number of periods than a signal with a higher ARL. McClain (1988) claims that some previous studies therefore had erroneously found CUSUM to be superior to Trigg, when in fact Trigg is better at detecting jumps quickly.

In general ARL-based analysis is impractical with real data and needs simulated data which is highly idealized, for example, with very long runs of no pattern changes after a single pattern change. In contrast, the crime time series data analyzed in Section 5 often has more than one pattern change or outlier within relatively short time periods. Thus an issue with ARL-based results is whether a recommended method which functioned well in a simplified, simulated world remains the right choice for the complex real world.

ARL-based studies also take as given that all signals are eventually detected regardless of the length of time between the simulated pattern change and the initial signal trip. With simulated data this may be a tenable assumption. Real data, such as the crime frequencies in this study, that do not fit this idealized state may produce pattern changes that are difficult to detect. Therefore, we must consider the possibility that a time series monitoring method will produce false negatives. McClain (1988, p 566) argues against using the ARL statistic as a criterion for time series monitoring. Instead he recommends the approach of specifying a maximum run length as most appropriate for management: if the run length is too long after a step change has occurred, then it is too late for managers to take corrective action. Any cases exceeding the maximum run length are considered false negatives, thus penalizing the monitoring method at hand.

3. ROC analysis

The ROC curve originated in the theory of detectability by Peterson, Birdsall, and Fox (1954) who conceived of the task as discriminating between “signal plus noise” from “noise alone” (Swets, 1986). Because noise is a random variable, the two alternatives just stated are statistical hypotheses and the theory of statistical decision making in regard to errors (Wald, 1950) applies. A discrimination cannot be made perfectly because noise can mimic the signal. Hence, with two alternative events and two diagnostic alternatives, the primary data for accuracy determination is contained in a two-by-two contingency table (Swets, 1988). See Table 1 for notation and definitions. For the event, “Positive” refers to occurrence of the signal whereas “Negative” is its absence. Similarly, for the diagnosis or test, “Test Positive” is a signal trip and “Test Negative” is no signal trip after an observation and application of a classifier or test.

True positive, false positive, false negative, and true negative rates are widely, though inconsistently, used as measures of prediction or classification accuracy. For example, *contrary* to the use of terminology in this paper, criminology research has focused on errors of prediction or classification with the false positive rate (sometimes called "1-positive predictive power") and false negative rate (sometimes called "1-negative predictive power") defined by the proportions of incorrect predictions found among all *predicted* positives and negatives respectively (e.g., Monahan, 1981; Farrington & Tarling, 1985). Alternatively, the usage of terms in this paper—corresponding to a large body of literature in medical testing, epidemiology and psychology—focuses on the accuracy of detecting among all *actual* positives and negatives (as in Table 1). The true positive rate (also called "sensitivity") is the proportion of positive outcomes that are correctly detected among actual positives: $TPR = TP / (TP + FN)$. The false positive rate (also called "1-specificity") is the proportion of negative outcomes that are incorrectly identified

as positives among all actual negatives: $FPR = FP / (FP + TN)$. For example, see Pepe (2000) and Kleinman & Abrams (2006). Hart, Webster & Menzies (1993) illustrate how lack of standardized usage for accuracy measures has sometimes produced misleading results. For example, Monahan (1981) defined the false positive rate relative to all predicted positives, while Otto's 1992 update of Monahan defined the false positive rate relative to actual negatives and overstated the improvement in predictive accuracy between earlier and later studies.

A plot of TPR versus FPR, as defined in Table 1, is the ROC graph. Fig. 1 is an example from our research, a ROC curve for a time series monitoring method. A ROC curve dominates any points from alternative classifiers that fall below the curve. The dashed line represents chance classifiers that randomly assign events to the positive diagnosis (Fawcett, 2003). The tradeoff between true and false positive rates is evident in the ROC curve: To obtain a higher TPR one must decrease the control limit value of the diagnostic test statistic, making it necessary to accept a higher FPR.

Determination of the ROC curve is independent of prevalence, P , of positives (or signals) but a sample of real data is needed that is representative of signal plus noise and of noise alone (each of the two columns of Table 1). A major issue in the derivation of ROC curves is the adequacy of truth in sample data; that is, knowing for certain whether every event is positive or negative (Swets, 1988). In practice, the “gold-standard” of truth is rarely perfect itself, but rather is simply much more accurate than the diagnostic test (such as Trigg’s method for time series monitoring). For example, the gold standard for medical imaging is analysis of tissue from biopsy or autopsy. Problems with this gold standard include “...imagery and pathology are not perfectly correlated in space or time... The image interpreter and pathologist may look at different locations, and the pathological abnormality may not have been present when the image

was taken.” (Swets, 1988, p 1290). Likewise, gold standard determinations of signals by crime analysts noted in Section 4 are not perfect, but they nevertheless incorporate assessments of how large structural breaks in crime series data must be to merit detailed crime analysis and possible police interventions.

A simple assessment of benefits enables the analyst to find the *optimum* tradeoff using a ROC curve (Metz, 1978; DeNeef & Kent, 1993). Let U_{TP} , U_{FN} , U_{FP} , and U_{TN} be the utilities of a true positive, false negative, false positive, and true negative respectively, corresponding to Table 1.

Then the expected utility from a detection test is:

$$E(U) = P \cdot [TPR \cdot U_{TP} + (1 - TPR) \cdot U_{FN}] + (1 - P) \cdot [FPR \cdot U_{FP} + (1 - FPR) \cdot U_{TN}] \quad (4)$$

Applying the chain rule to $E(U)$, with the assumption that the ROC curve is strictly monotonic and continuously differentiable, yields the following optimality criterion for the slope of the ROC curve:

$$dTPR/dFPR = [(1 - P)/P] \cdot [(U_{TN} - U_{FP}) / (U_{TP} - U_{FN})]. \quad (5)$$

Call $(U_{TN} - U_{FP}) / (U_{TP} - U_{FN})$ the “benefit ratio”. $U_{TN} - U_{FP}$ is the benefit of avoiding a false positive decision given a negative event and $U_{TP} - U_{FN}$ is the benefit of avoiding a false negative decision given a positive event. The purpose of screening and diagnosis is to identify positive events as quickly as possible for treatment (e.g., identifying breast or prostate cancer at an early stage for treatment), making the benefit of avoiding a false negative larger than avoiding a false positive. The latter merely incurs the cost of further diagnosis (e.g., biopsy). Thus $U_{TN} - U_{FP} > U_{TP} - U_{FN}$. If we set $U_{TP} - U_{FN} = 1$, then domain experts merely have to determine how many more times is it valuable to avoid a false negative than a false positive.

With ROC analysis, we can also explicitly handle the issue of timeliness, addressed by the ARL statistic, by specifying the maximum acceptable run length after a signal for a signal trip to occur, as discussed at the end of Section 2.

4. Research design

First, we describe the crime data and business-as-usual forecast method used in our case study. Note that others also have examined crime data monitoring: Anderson et al. (1996) used ARIMA modeling and control charts, Gorr & McKay (2005) used Trigg's method, and Rogerson (2005) used CUSUM. Next, we describe how we elicited expert judgments from crime analysts to identify outliers and step jumps and produce test data for ROC analysis. This is followed by procedures we used to generate ROC curves, and finally by a description of resampling methods used to assess sampling errors in our ROC curves.

4.1 Data

The test data for this research is based on monthly crime frequencies from January 1991 to December 2000 and for police car beats (or patrol districts) in Pittsburgh, Pennsylvania and Rochester, New York (Cohen & Gorr, 2005). A car beat is the territory assigned to a single patrol unit: Pittsburgh has 42 car beats and Rochester has 38. These data are highly disaggregated and micro-scale. The series tend to have noisy variation about mean levels and seasonal effects, but they also have step jumps in levels sometimes attributable to a single cause such as a gang rivalry, a serial criminal, or police crackdown. The objectives of police for such data are to identify step jump increases as quickly as possible, diagnose their causes, and then suppress the causes to bring crime back to its base level or lower. In the process, police interventions may suppress a sustained step jump, resulting in what appears to be an outlier.

Also, in real time it is not possible to distinguish the first observation of a step jump from an outlier observation. Finally, using time series monitoring to identify one-time spikes in criminal activity that have already occurred and passed is valuable because police have the opportunity to analyze the circumstances that allowed for or led to such a spike. Therefore, we must recognize that the detection of both outliers and step jumps is important.

Included with the crime data are business-as-usual forecast errors from deseasonalized simple exponential smoothing. As shown to be the most accurate seasonality method for these data (Gorr et al., 2003), the forecasts used multiplicative seasonality estimated using classical decomposition from city-wide data, rather than from individual car beats. These forecasts are from a rolling-horizon design made one month ahead over 60 months from 606 time series with the smoothing parameter optimized over a grid and seasonality re-estimated at each forecast origin using historical data relative to the forecast origin. Forecasts were made for January 1996 to December 2000 and for nine crime types—larceny, simple assault, disorderly conduct, robbery, burglary (Pittsburgh only), criminal mischief, motor vehicle theft, 911 drug calls for service (Pittsburgh only), and 911 shot fired calls for service (Pittsburgh only).

We extracted a sample of 30 time series from the original 606 for coding by crime analysts to produce test data for ROC analysis. Twenty of these time series were selected systematically by us to have at least one potential step jump. We identified these time series by visually examining the 20 percent of series in each city and crime type with the highest crime frequency. The remaining 10 in the sample of 30 time series were randomly chosen from the data set that remained after removing the 20 that were selected by visual inspection.

4.2 Expert Coding of Step Jumps and Outliers

We generated time series plots from the 30 sample series to be evaluated by the three police crime analysts working for the Pittsburgh Bureau of Police. The goal was to have the analysts identify crime series pattern changes—step jumps or outliers—that were “worthy of further investigation and possible intervention” as judged by them. The plots, such as the example in Fig. 2, show the actual crime frequency per month on the vertical axis and month along the horizontal axis. The frequencies are from the years 1996-2000, which corresponds to the five-year period in which business-as-usual forecasts were made. We used the first six months of each time series as the burn-in period for the methods and as historical information for the analysts. Hence no data on potential step jumps and outliers were collected for these months. These plots display the correct month of observation, but not the year, crime type, car beat, nor city. The sample size is too small to adequately control for these other features of the crime series. Segregating the time series into subgroups based on similarity of some substantive criterion, such as crime type or average crime level, and calibrating time series monitoring methods separately for these subgroups might lead to improved performance over the results reported in this paper. Different methods may also prove superior for different subgroups.

We asked the crime analysts to independently review the time series plots and annotate months that they believed were step jumps or outliers of relevance. Before providing the analysts with the sample of time series, we first met with them to discuss the research topic and their task, including some example time series charts. Based on this meeting we developed a set of instructions to guide coding in which we provided methods and a tool for visually inspecting time series data for outliers and step jumps (see <http://www.forecasters.org/ijf/data.htm> for the instructions). We then observed the analysts as they annotated a few example series that were not part of the 30 sampled series to verify that the task had been communicated clearly. Finally, the

analysts completed their coding tasks individually at their own paces over a period of a few weeks.

Relying on the analysts' coded series, we combined the judgments using the following rules:

1. If two or three of the three analysts classified a month to be a step jump or outlier then we classified the month to be a true step jump or outlier, respectively.
2. If two consecutive months were judged to be a step jump, each month by a different analyst, then we classified the earlier month as the true step jump point.

The first rule takes into account that the analysts had no previous experience with the task and thus did not have the benefit of feedback from each other, field officers, or management on the quality of their assessments. Hence, we adopted the majority rule policy to eliminate misclassification "errors" by the analysts in regard to desirable magnitude of change in patterns needed to prompt further analysis. Note that many other ROC studies have used experts for gold standard coding and majority rule (e.g., Kazui et al., 1996 and Fiszman et al., 2000). The second rule also follows the majority rule policy. This rule is conservative: the earlier a pattern change is coded in test time series data, the more difficult it is to detect it in a timely fashion. Of the 1620 months evaluated in the sample, there were 44 step jumps (2.7 percent) and 79 outliers (4.9 percent). All three analysts concurred on 9 of these step jumps and 49 of the outliers.

4.3 ROC Analysis

One typically generates a ROC curve by varying the control limit of a classification method and tabulating an FPR, TPR point for each limit for plotting. However, time series monitoring methods, such as Brown's and Trigg's, have smoothing parameters that further affect attainable TPR and FPR values. Thus each unique combination of smoothing parameter values

and control limits gives rise to a point in a ROC plot. We use a grid of control limits and parameter values and generate the corresponding ROC points. The northwest frontier of the ROC points is the Pareto frontier ROC curve for a time series monitoring method that dominates all other sampled grid points. Kupinski & Anastasio (1999) also examine ROC curves generated by varying multiple parameters of a classification model, focusing on cases where a grid search procedure is computationally impractical. Accompanying the Pareto frontier curve must be a reverse function that yields the smoothing parameters and control limits for any point on the curve.

Thus ROC analysis proceeds with a complete enumeration grid search over a range of parameter values and control limits for each method. The first step in the process varies the parameters of each method and records the periods in which there is an exception report. For the Trigg and Brown methods, we searched parameters over the ranges in Table 2. We restricted β to values less than or equal to α because the denominator, which is a measure of variability used for normalization, should not adjust to new data more quickly than the numerator (McClain, 1988). The control limit value for all methods and all parameter sets was allowed to range from 0 to the maximum value needed for the method to never issue a signal trip in .01 increments. This assured the ROC curves ranged from (0, 0) to (1, 1).

We specified one to four-month maximum run length windows for step jump detection beginning in the month of the analyst-identified jump. If the automated method generated a signal trip in a designated window, then the step jump was deemed a true positive; if not, then a false negative jump occurred. If multiple signal trips occurred in this window then they were counted as a single true positive. Results reported below are for the four month window unless otherwise stated. A time series monitoring method had to issue a signal trip in the month that an

outlier occurred to be classified as true positive; otherwise it was a false negative. This differential treatment is warranted due to the nature of the two event types. Outliers are single period, large spikes, so the monitoring methods should react immediately and then return to normal levels. Step jumps are sustained and possibly more subtle, so the time series methods may lag in detection and issue trips in several periods.

4.4 ROC Sample Variability

The ROC curves that we described above and analyze below summarize time series monitoring methods' performance on a sample of real data. We should not expect to get identical ROC curves from a different sample of thirty time series, even if we used the same sample selection process. In Section 5 we thus use a resampling technique to explore the variability of ROCs for the Trigg method and its optimal parameters derived from our full test data sample. We generated 100 new samples of time series by randomly choosing 30 time series from our initial sample *with replacement* (e.g., see Moise et al., 1985). Others have used bootstrap and jackknife resampling techniques for calculating ROC related standard errors and confidence intervals (e.g., McNeil & Hanley, 1984; Mossman, 1995; Obuchowski & Lieber, 1998).

5. Results

The results demonstrate the benefits of ROC analysis with clear winners and some surprises. We proceed by first illustrating the process of obtaining a Pareto frontier ROC curve for the Trigg method and then we analyze its reverse functions. After that we compare the five time series monitoring methods of this paper and discuss the optimal method and tradeoff point for implementation by police. Lastly are results from resampling to investigate sampling error in ROC curve estimates.

Fig. 3 illustrates the construction of the Trigg Pareto frontier ROC curve. Plotted are all points from the three-dimensional grid of smoothing parameter and control limit values (the gray-shaded area actually consists of these points), along with the Pareto frontier. Evidently, the choice of smoothing parameters greatly affects ROC performance of the Trigg method. Note that the performance of the frontier is very good, with a large area under the ROC curve.

In ROC analysis a 45-degree line from (0,0) to (1,1) represents random classification of cases to the positive and negative diagnosis. However, the detection window after a step jump used in our research increases the probability that a step is detected. Hence the line for random detection of step jumps follows a binomial distribution line, with number of trials equal to window length, that is bowed up above the 45-degree line and similar to the southeast boundary in Fig.2.

Fig. 4 has reverse functions to provide optimal parameter values for the frontier Trigg ROC curve. The chart has parameter values for sample points along the Trigg frontier, as a function of FPR over a range relevant for practice (0.05 to 0.25), and fitted trend lines for each parameter. Given a frontier ROC point, such as (0.16, 0.88) from Fig. 8 below, one can look up the corresponding parameters for implementing Trigg: the numerator smoothing constant is 1.00, the denominator smoothing constant is 0.07, and the control limit is 1.57.

There are interesting and some surprising patterns in Fig. 4. First, as expected, the control limit decreases as FPR increases and more liberal positions taken. Surprising, though, is that the numerator smoothing factor of the Trigg method is 1.0 or slightly less around 0.9. Thus for the crime test data, instead of a smoothed average of forecast errors, it is best to just use the most recent forecast error to trigger exceptions (or to also keep a relatively small fraction of history). Trigg's method, for the most part, thus collapses to Brown's method with a CUSUM window of

width one. Most of the benefit from smoothing takes place in the denominator, which scales or normalizes the forecast errors. For conservative settings, where a low FPR is desired, it is necessary to increase the adaptability (i.e., the denominator smoothing constant) of this scaling while keeping the control limit relatively high. While the Trigg method may not fire as often, under conditions of changing variance the denominator adapts more quickly, which reduces false positives. Note that the scatter for the denominator in Fig. 4 is much higher for low FPR rates than for high rates. The reverse functions for the 1-period Brown frontier model in Fig. 5 are similar to those for Trigg in Fig. 4 except that the trends monotonically decrease.

Next, Fig. 6 compares four of the five alternate time series monitoring methods. We did not include the Brown frontier method because it is nearly the same as the Trigg frontier method. The Random-with-detection-window method is random detection with a four-month detection window for step jumps. The frontier Trigg method is best for the relevant range of FPR values up to about 0.26 after which Standardized Forecast Errors is better up to about 0.36. The latter method performs surprisingly well except for the important range of FPR values from 0.14 through 0.26, the interval that contains the optimal tradeoff point for the crime analysts of our study (see Fig. 8 below). The Percentage Change method, often used in practice, is heavily dominated by all other methods and therefore should not be used. Noteworthy is the fact that the best performing methods make use of forecast errors for time series monitoring.

Fig. 7 shows ROC frontier curves for Brown with CUSUM window lengths of one, three, and five months. The one-month-window Brown method, which is simply Trigg with $\alpha=1$, dominates the longer window lengths.

Fig. 8 shows the optimal tradeoff point, from condition (5). The three police crime analysts who participated in our research assessed the benefit ratio to be 1÷10 for step jumps in

crime time series. We do not have a precise estimate of prevalence, but use the estimate of 0.076 from our full sample of 30 time series. (An estimate obtained by pooling our 10 random series and an additional 25 random series coded by a single crime analysts is high, 0.103, but for illustration purposes we are using the more conservative estimate from our 30 time series. Use of the higher estimate for prevalence would move the optimum to the right in Fig. 8, to a higher FPR and TPR point.) Given the mild concavities in the estimated Trigg frontier ROC curve, it is somewhat difficult to apply the resulting optimality criterion exactly, but we nevertheless obtain sufficient information to choose Trigg parameters. One optimum is the straight line in Fig. 8 where for the frontier Trigg ROC curve we find the optimal FPR and TPR roughly to be the point (0.16, 0.88). An alternate optimum is approximately (0.22, 0.93). Analysts could make a selection between these two points based on additional criteria, such as expected workload estimates from each optimum.

Fig. 9 has Trigg frontier ROC curves with maximum run lengths of one to four months for step jump detection. As expected, the longer the maximum run length, the greater the area under the ROC curve. The effect is sizeable, but performance remains quite good when reducing the window length for step jumps. The movement in the ROC curve is directly attributable to missing more step jumps rather than outliers. For example the number of step jumps detected at a FPR of approximately 0.16 goes from 36, to 29, to 22, to 15 for step jump windows of 4, 3, 2, and 1 respectively, while the number of outliers detected stays constant at 72. Note that all methods compared would have had similar results, shifting ROC curves southeast as the maximum run length window shortens.

Next, Fig. 10 has results from the resampling procedure. We generated 100 new samples of 30 time series by randomly choosing series from the initial sample with replacement. We then

created ROC curves for each of these new samples by varying the control limit for Trigg with $\alpha=1$ and $\beta=.07$. We broke the FPR axis up into intervals of width 0.02 and tabulated the extremes, quartiles, and median for each interval of bootstrapped TPR values. As expected, because we have a relatively small sample of signals in the test data (44 step jumps and 79 outliers), there is sampling error evident, although the inter-quartile range is relatively small. As a further analysis of sampling error, from the same 100 ROC curves we plotted the TPR and FPR pairs at the optimal tradeoff control limit of 1.57. The resulting points are plotted in Fig. 10. We point out cautiously, due to the small size of the sample, that the performance of the Trigg tracking signal at the optimally chosen parameters appears robust. Had we found substantial performance variability this would have been cause for concern.

6. Conclusion

The ROC framework—widely used in diagnostic decision making—provides a valuable alternative to the traditional ARL approach to calibrating and evaluating time series monitoring methods. ARL analysis stresses timeliness of signal detection and uses simulated time series data and simulated structural breaks. In contrast, ROC analysis stresses the tradeoff between true and false positive rates and uses real data coded for signal occurrences. Decreasing control limits and changing smoothing parameters to increase true positive rates necessarily increases false positive rates as well. The more prevalent signals and the more resources made available for monitoring in an organization, the higher the true and false positive rates accepted as optimal.

Both the tradeoff and timeliness of monitoring are important for societal applications: both need to be assessed for each domain and setting. ROC analysis directly assesses timeliness of detection by specifying the maximum acceptable run length after a signal occurs, as

recommended in the literature. If a signal is not detected within a detection window, it is classified as a false negative, the worst kind of error, and thus is given heavy weight in making tradeoffs. While ARL analysis has the benefit of presenting simulated time series with known signals for validation by researchers, it is difficult to conceive of a simulation that could present the decision maker with valid information for making the tradeoff inherent in monitoring. ROC analysis incurs the cost of collecting and coding time series data for signals (instead of carrying out simulations). Coding signals requires judgment—not only must pattern breaks be detectable, but they must also be worthy of additional investigation after a signal trip. It is not possible to be certain that gold standard coding for signals is true, as it is for simulated data.

Future research would benefit from ROC analysis of time series monitoring methods in other types of agencies or organizations that have different time series patterns in their performance measures and different cost tradeoffs. Nevertheless, we believe that the crime time series data used in this paper—which are highly disaggregated at the car beat level and characterized by level and seasonality, are very noisy, and have step jumps—are representative of much service delivery or product sales time series. Nevertheless it would be very desirable to code large test data samples and see if method performance varies substantially over product or service families. Perhaps with a sufficient number of such studies, it would be possible to provide product or industry type guidelines on how best to select and calibrate time series monitoring systems.

The Brown and Trigg time series monitoring methods have long been compared in the literature, seeking a champion method, but our crime test data set and ROC analysis provides evidence that they both collapse to the same simplified method: use the most recent forecast error as the numerator and smooth the MAD of forecast errors for the denominator with a

relatively low smoothing constant. There was no benefit of CUSUM windows longer than one period for Brown's method and no benefit of substantially smoothing the numerator forecast errors for the Trigg method. Moreover, our research provides evidence that use of a good business-as-usual forecast method and the resulting forecast errors to construct a stochastic decision variable is critical for time series monitoring. Even the simple Standardized Forecast Error method is a good time series monitoring method.

Acknowledgements

Funding for this research was provided by Grants No. 98-IJ-CX-K005 and 2001-CX-0018 awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. Points of view in this document are ours and do not necessarily represent the U.S. Department of Justice. Any errors in this paper are the responsibility of the authors. We wish to thank Jan De Gooijer for handling this paper as editor. We also thank the associate editor and three referees for their insightful and thorough comments. We are grateful to Sgt. Mona Wallace, Detective Deborah Gilkey, and Detective Peg Sherwood of the Pittsburgh Bureau of Police's Crime Analysis unit for their efforts and suggestions in carrying out this research. We wish to thank Professors Robyn Dawes, Michael DeKay, and Daniel Neill of Carnegie Mellon University and Dr. Ned Levine of Ned Levine & Associates for their insightful comments on our research. Versions of this paper were presented at the International Symposium on Forecasting in 2006 and 2007.

References

- Anderson, E. A. & Diaz, J. (1996). Using process control chart techniques to analyze crime rates in Houston, Texas, *Journal of the Operational Research Society*, 47, 871–881.
- Baltimore CitiStat Reports and Maps (2008), Internet site accessed 2/2/2008, <http://www.ci.baltimore.md.us/news/citistat/reports.html>.
- Batty, M. (1969). Monitoring an exponential smoothing forecasting system, *Operational Research Quarterly*, 20, 319–325.
- Brown, R. G. (1959). *Statistical forecasting for inventory control*, New York: McGraw-Hill.
- Brown, R. G. (1963). *Smoothing, forecasting and prediction of discrete time series*, Englewood Cliffs, N.J.: Prentice-Hall.
- Cohen, J. & Gorr, W. L. (2005). *Development of Crime Forecasting and Mapping Systems for Use by Police*, National Institute of Justice, Grant 2001-IJ-CX-0018 final report.
- DeNeef, P. & Kent, D. L. (1993). “Using treatment-tradeoff preferences to select diagnostic strategies: Linking the ROC curve to threshold analysis”, *Medical Decision Making*, 13, 126–132.
- Farrington, D. P. & Tarling, R. (1985). Criminological prediction: an introduction, in Farrington, D. P. & Tarling, R. (eds.), *Prediction in Criminology*. Albany: State University of New York Press.
- Fawcett, T. (2003). ROC graphs: Notes and practical considerations for researchers. Technical Report HPL-2003-4, HP Laboratories. Available: http://www.hpl.hp.com/personal/Tom_Fawcett/papers/ROC101.pdf.

- Fiszman, M., Chapman, W. W., Aronsky, D., Evans, R. S. & Haug, P. J. (2000). Automatic detection of acute bacterial pneumonia from chest x-ray reports, *Journal of the American Medical Informatics Association* 7, 593-604.
- Gardner, E. S. (1983). Automatic monitoring of forecast errors, *Journal of Forecasting*, 2, 1-21.
- Gardner, E. S. (1985). CUSUM vs smoothed-error forecast monitoring schemes - some simulation results, *Journal of the Operational Research Society*, 36, 43-47.
- Golder, E. R. & Settle, J. G. (1976). Monitoring schemes in short-term forecasting, *Operational Research Quarterly*, 27, 489-501.
- Gorr, W. L., Olligschlaeger, A. & Thompson, Y. (2003). Short-term forecasting of crime, *International Journal of Forecasting*, 19, 579-594.
- Gorr, W. L. & McKay, S. A. (2005). Application of time series monitoring methods to detect time series pattern changes in crime mapping systems. In: Wang, F. (Ed.) *Geographic information systems and crime analysis*. Hershey, PA; Idea Group Pub., pp. 171-182.
- Harrison, P. J. & Davies, O. L. (1964). The use of cumulative sum (Cusum) techniques for the control of routine forecasts of product demand, *Operations Research*, 12, 325-333.
- Hart, S. D., Webster, C. D., & Menzies, R. J. (1993), A note on portraying the accuracy of violence predictions, *Law and Human Behavior*, 17, 695-700.
- Kazui, S.; Naritomi, H.; Yamamoto, H.; Sawada, & Yamaguchi, T. T. (1996). Enlargement of spontaneous intracerebral hemorrhage, *Stroke*, 27, 1783-1787
- Kleinman, K. P. & Abrams, A. M. (2006). Assessing surveillance using sensitivity, specificity and timeliness, *Statistical Methods in Medical Research*, 15, 445-464.
- Kupinski, M. A. & Anastasio, M. A. (1999). Multiobjective genetic optimization of diagnostic

- classifiers with implications for generating receiver operating characteristic curves, *IEEE transactions on medical imaging*, 18, 675–685.
- McClain, J. O. (1988). Dominant time series monitoring methods, *International Journal of Forecasting*, 4, 563–572.
- McKenzie, E. (1978). Monitoring of exponentially weighted forecasts, *Journal of the Operational Research Society*, 29, 449–458.
- McNeil, B. J. & Hanley, J. A. (1984). Statistical approaches to the analysis of receiver operating characteristic (ROC) curves, *Medical Decision Making*, 4, 137–150.
- Metz, C. E. (1978). Basic principles of ROC analysis, *Seminars in Nuclear Medicine*, 8, 283–298.
- Moise, A., Clement, B., Ducimetiere, P. & Bourassa, M. G. (1985). Comparison of receiver operating curves derived from the same population: A bootstrapping approach, *Computers and Biomedical Research*, 18, 125–131.
- Monahan, J. (1981). The clinical prediction of violence, DHHS Publication No (ADM) 81-921. Rockville (MD): US Department of Health and Human Services.
- Mossman, D. (1995). Resampling techniques in the analysis of non-binormal ROC data, *Medical Decision Making*, 15, 358–366.
- New York City Police Department Crime Statistics (2008), Internet site accessed 2/2/2008, <http://www.nyc.gov/html/nypd/html/pct/cspdf.html>.
- Obuchowski, N. A. & Lieber, M. L. (1998). Confidence intervals for the receiver operating characteristic area in studies with small samples, *Academic Radiology*, 5, 561–571.
- Otto, R. K. (1992). Prediction of dangerous behavior: A review and analysis of second generation research, *Forensics Reports*, 5, 103-133.

- Pepe, M. S. (2000). Receiver operating characteristic methodology. *Journal of the American Statistical Association*, 95, 308-311.
- Peterson, W. W, Birdsall, T.G., & Fox, W.C. (1954). The theory of detectability, *Transactions of the IRE Professional Group on Information Theory*, 4, 171–212.
- Rogerson, P. A. (2005). Geographic surveillance of crime frequencies in small areas. In: Wang, F. (Ed.) *Geographic information systems and crime analysis*. Hershey, PA; Idea Group Pub., pp. 153–170.
- Swets, J. A. (1986). Indices of discrimination or diagnostic accuracy: their ROCs and implied models, *Psychological Bulletin*, 99, 100–117.
- Swets, J. A. (1988). Measuring the Accuracy of Diagnostic Systems, *Science*, 240, 1285–1293.
- Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Better Decisions through Science, *Scientific American*, 283, 82–87
- Taylor, F. W. (1911). Shop Management. Project Gutenberg EBook available from <http://www.gutenberg.org/dirs/etext04/shpmsg10.txt>.
- Trigg, D. W. (1964). Monitoring a forecasting system, *Operational Research Quarterly*, 15, 271–274.
- Wald, A. (1950). *Statistical decision functions*. New York: Wiley.

Table 1.
Contingency Table

	Positive	Negative	
Test positive	TP	FP	
Test negative	FN	TN	
	TP + FN	FN + TN	N= TP + FP + FN +TN

$$\text{Prevalence (P)} = (TP + FN)/N$$

$$\text{True Positives} = TP$$

$$\text{False Positives} = FP$$

$$\text{False Negatives} = FN$$

$$\text{True Negatives} = TN$$

$$\text{True Positive Rate (TPR)} = TP/(TP+FN)$$

$$\text{False Positive Rate (FPR)} = FP/(FP + TN)$$

Table 2 .

Ranges of parameters for optimization of the Trigg and Brown methods

	Parameter	Range	Increment
Numerator smoothing factor (Trigg)	α	$0.05 \leq \alpha \leq 1$	0.01
Denominator smoothing factor (Trigg and Brown)	β	$0.05 \leq \beta \leq \alpha \leq .5$	0.01
Periods for CUSUM (Brown)	k	$1 \leq k \leq 5$	1

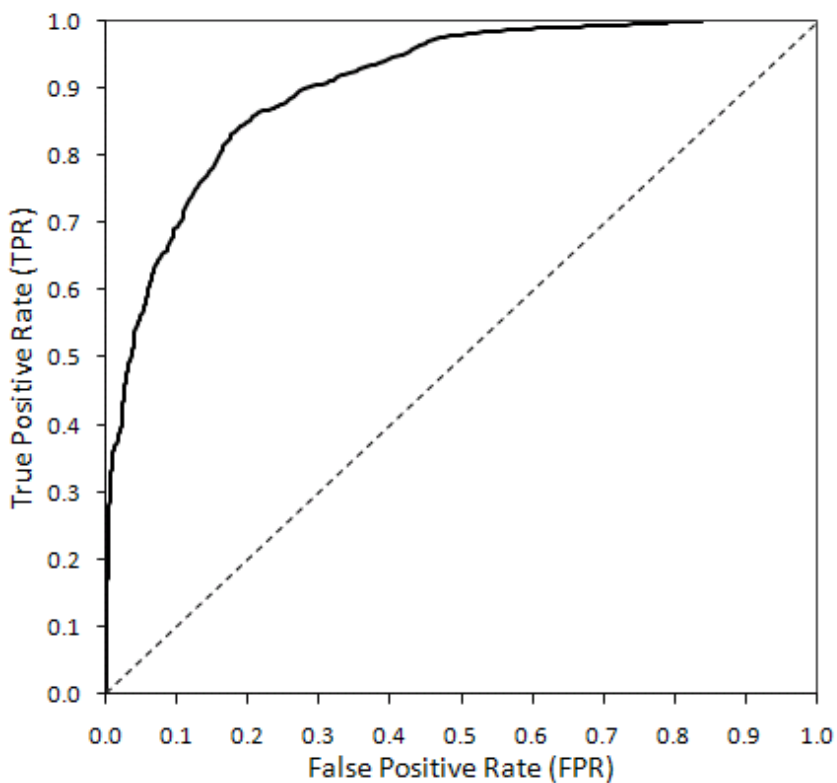


Fig. 1. Sample ROC curve.

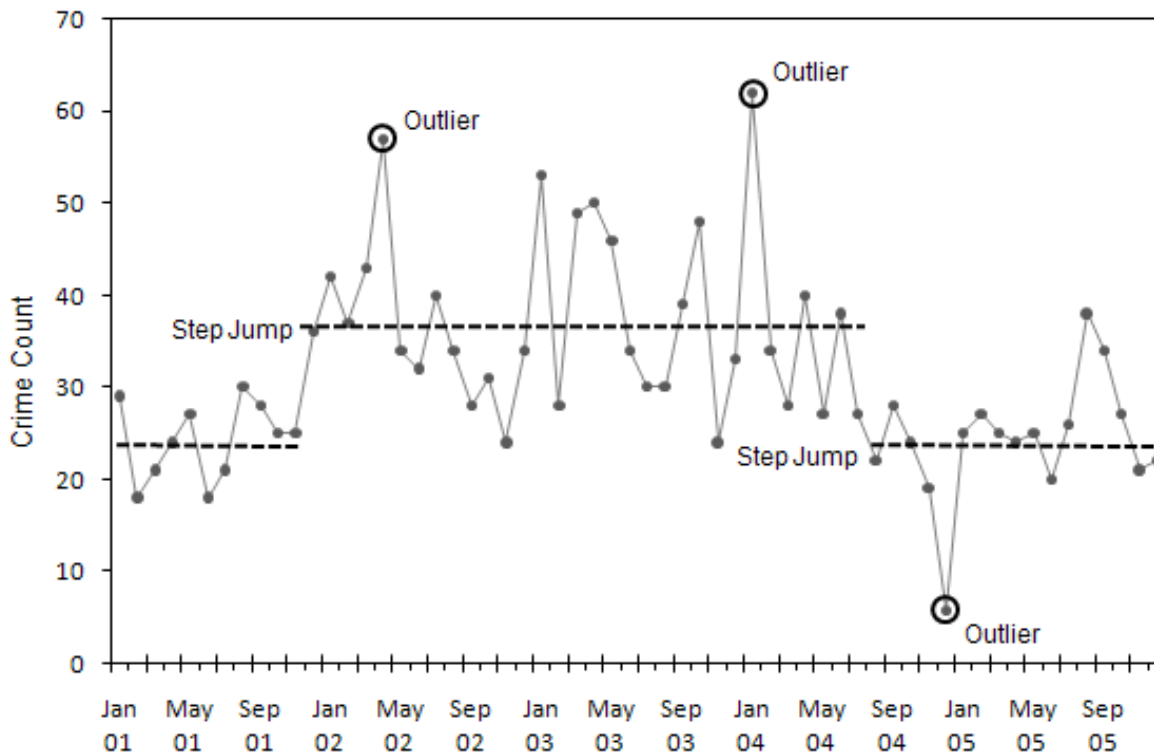


Fig. 2. Sample crime time series coded for pattern changes.

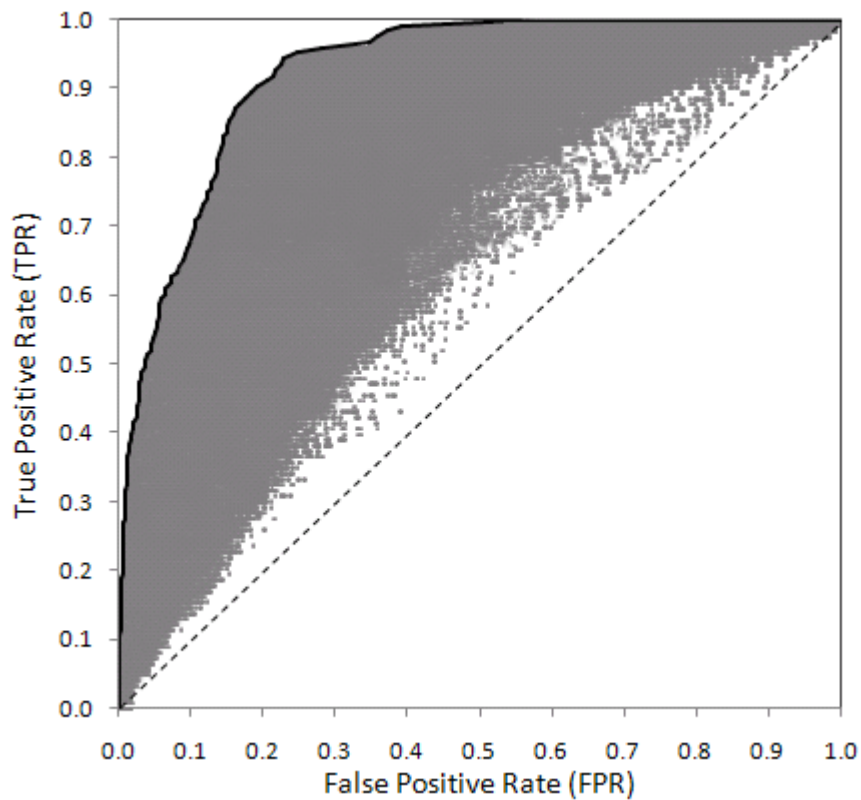


Fig. 3. Trigg Pareto frontier ROC curve and dominated grid search points.

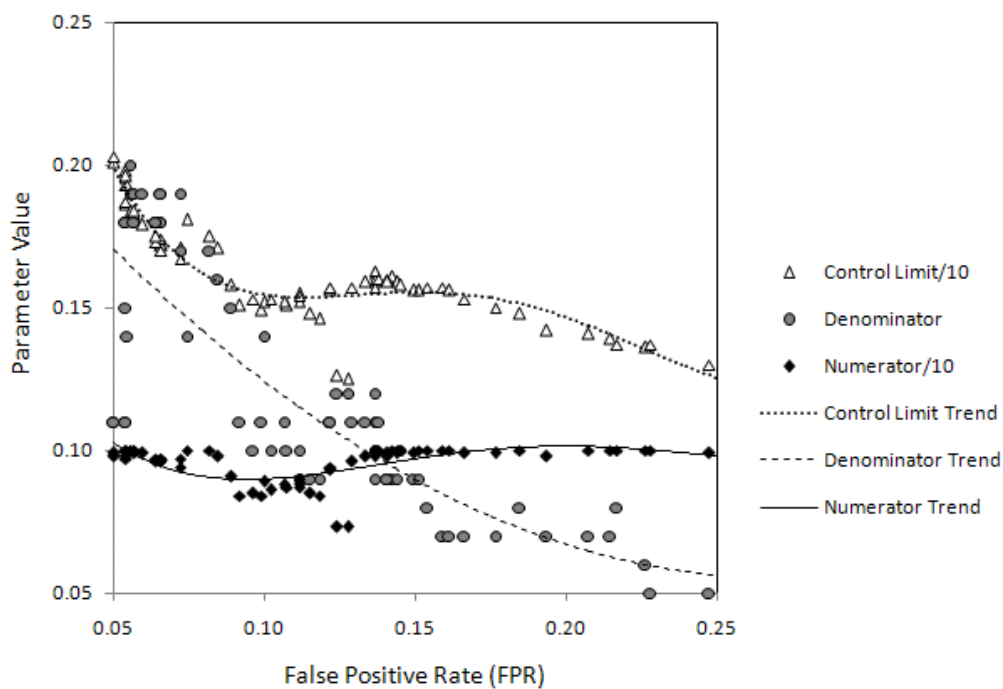


Fig. 4. Reverse functions for frontier Trigg ROC curve.

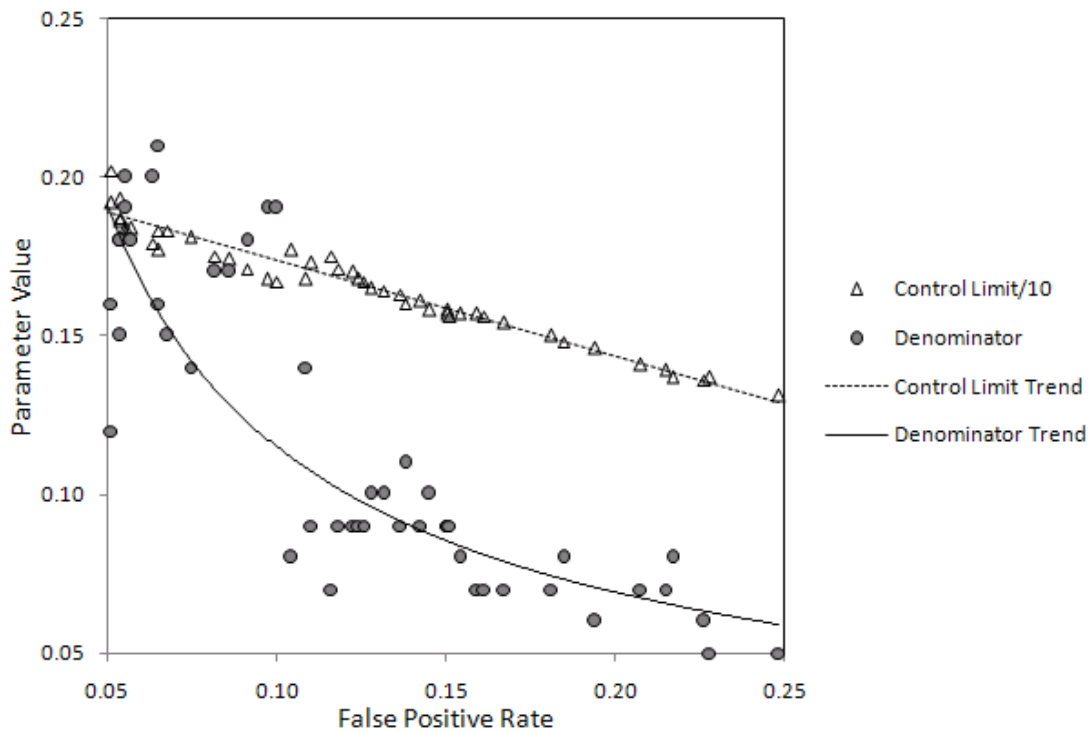


Fig. 5. Reverse functions for frontier 1-period Brown ROC curve.

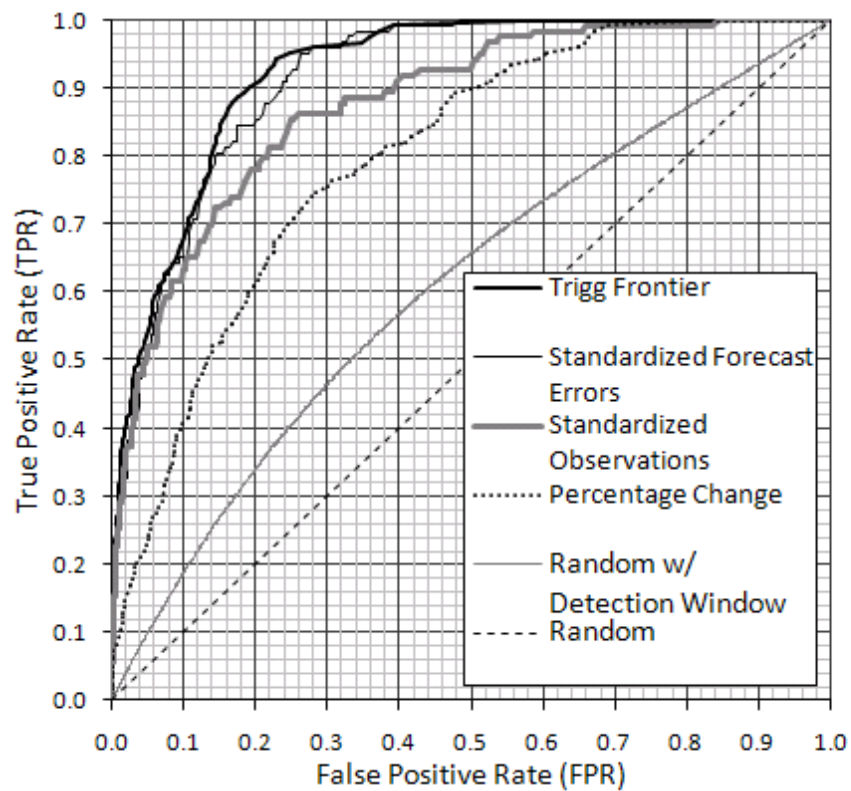


Fig. 6. ROC curves for alternative time series monitoring methods.

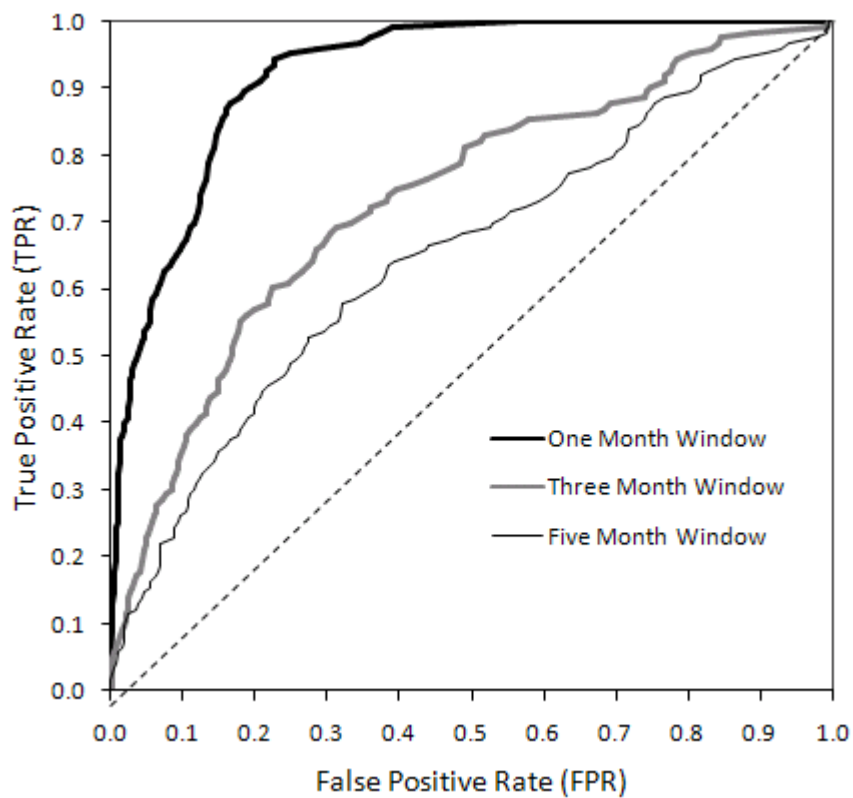


Fig. 7. ROC curves for alternative CUSUM window lengths in Brown's Method.

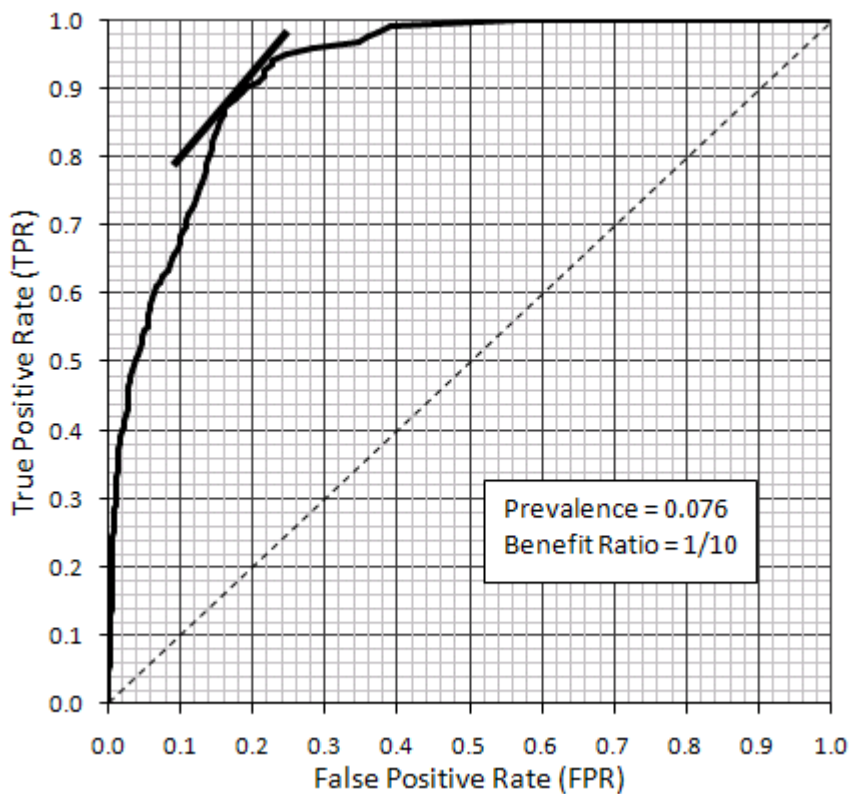


Fig. 8. Trigg ROC curve and optimal tradeoff point.

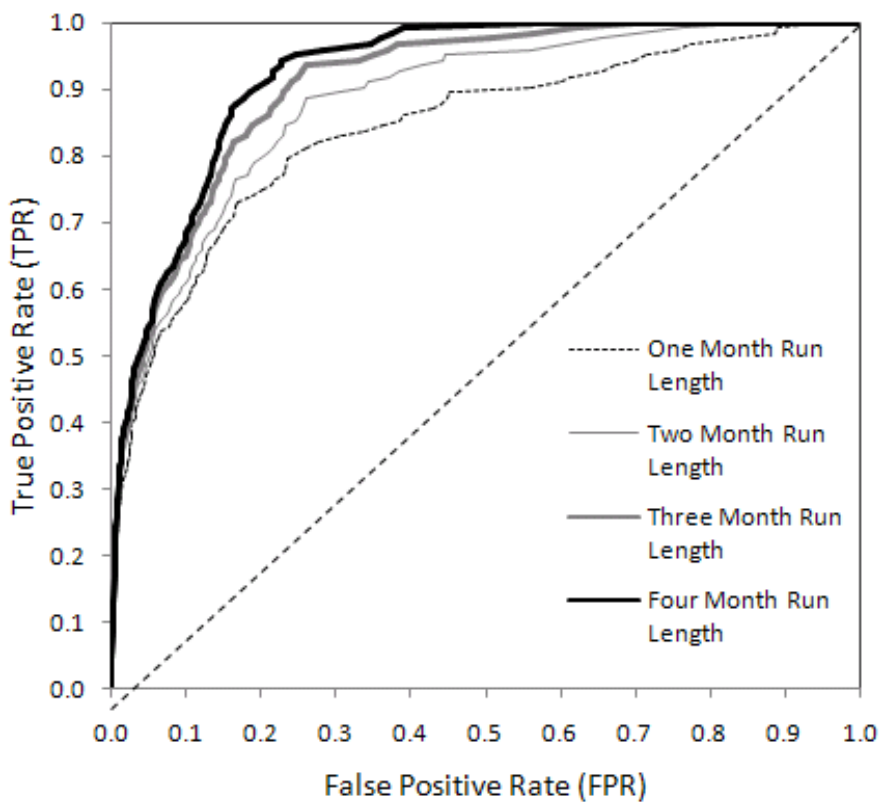


Fig. 9. Trigg ROC curves for alternative maximum run lengths.

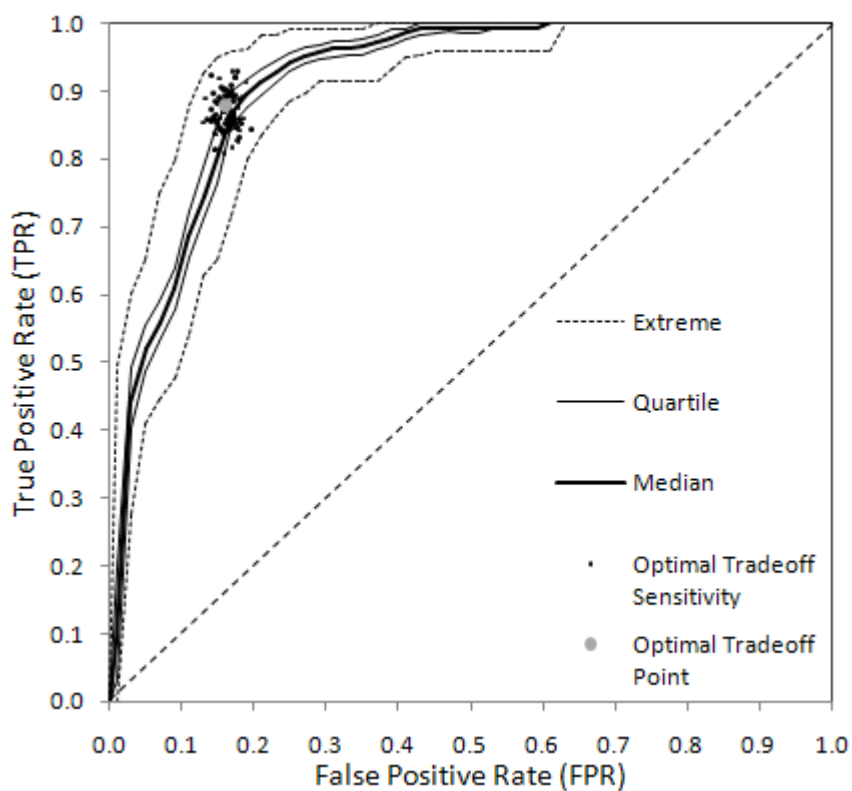


Fig. 10. Resampling results with extreme, quartile, and median true positive rates.