Online change detection: Monitoring land cover from remotely sensed data *

Yi Fang, Auroop R. Ganguly †, Nagendra Singh, Veeraraghavan Vijayaraj, Neal Feierabend ‡, David T. Potere §
Oak Ridge National Laboratory
Computational Sciences and Engineering Division
Bethel Valley Road, Oak Ridge, TN 37831
{fangy,gangulyar,singhn,vijayarajv}@ornl.gov

Abstract

We present a fast and statistically principled approach for land cover change detection. The approach is illustrated with a geographic application that involves analyzing remotely sensed data to detect changes in the normalized difference vegetation index (NDVI) in near real time. We use the Wal-Mart land cover change data set as a nontraditional way to monitor and validate known cases of NDVI change. A reference distribution has been justified to fit the available data. An adaptive metric based on the exponentially weighted moving average (EWMA) of normal scores derived from p-values is tracked for new or streaming data, leading to alarms for large or sustained changes. A heuristic algorithm based on the property of the metric is proposed for change point detection. The proposed framework performed well on the validation dataset.

1. Background

Change detection is the process of identifying differences in the state of a feature or phenomenon by observing it at different times[6]. This process has been widely used in quality control, network intrusion detection, and financial analysis[8]. In geographic applications, change detection plays a critical role in land use/land cover change analysis such as monitoring deforestation or vegetation phenology. Coupled with recent advances in sensor technology, a huge amount of land cover information is now available and accessible. Real-time detection of land cover change has been identified as a critical research area and the demand is expected to grow significantly in the future given potential applications for high-priority domains like disaster response, urban planning, deforestation, intelligence analysis, and warfare scenario assessment.

1.1. Related Work

Change detection has been extensively studied in the context of time series analysis and forecasting. The standard approaches include various smoothing techniques, the Box-Jenkins ARIMA modeling, innovation and outlier analysis, and more recently wavelet-based methods[8]. Previous researchers have applied these techniques to land cover change detection[12]. However, fast and statistically principled approaches in this context are not frequently encountered in the literature. From a statistical point of view, a change point is a point in time where the observations exhibit one combination of distribution/parameters up to that point and another combination after that point. Thus, the change detection problem is two-fold: one is to decide if there is any change and another is to locate the change point if change has occurred. The method presented here is close in spirit to the approaches reported in [10, 2] for monitoring network counts and Chinese websites, with the important difference being that one of our major foci here is on the identification of change points. A preliminary version of our work was developed for linking sensor and cyber networks in the context of security applications[5].

1.2. Our Contributions

We approach the online land cover change detection problem by proposing a statistically principled framework. The Wal-Mart land cover change validation data set[14]

*This manuscript has been authored by employees of UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. Accordingly, the United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

†Corresponding author. Auroop R. Ganguly; Tel.: +1-865-241-1305; Email: gangulyar@ornl.gov; Address: Oak Ridge National Laboratory, 1 Bethel Valley Road, MS 6085, Oak Ridge, TN 37831

‡Neal Feierabend is currently at Mississippi State University. Email: neal@nealf.com.

§David T. Potere is currently at Princeton University. Email: dpotere@princeton.edu.

has been used as a nontraditional data source to validate our framework. While our method works for any baseline model that can capture normal behavior, in the case study presented here we justify using the Gaussian distribution. An alarm will be generated based on the exponentially weighted moving average (EWMA) Statistical Process Control (SPC) chart. We provide a heuristic approach to identify change points based on the property of the "severity metric" [10, 2].

2. Data and Domain Description

Our goal is to detect land cover changes in real time. By "real time" we mean, the components involved in our procedure are updated almost at the same rate as the data is received. To accomplish this goal, we analyze time series of the normalized difference vegetation index (NDVI) values from the study area. NDVI is a simple spectral ratio derived from remote sensing imagery that is correlated with vegetation health[16]. The NDVI value is a direct indicator of green leaf biomass and green leaf area index [3]. Time series of NDVI values indicate seasonal and temporal profiles of vegetation activity. In this study, a time series of 16-day NDVI composites produced by the Global Land Cover Facility at the University of Maryland was used. Wal-Mart stores in the US are spatially distributed in a wide variety of environments with a majority of the stores being built on undeveloped vegetated land. The stores have a long lifespan in the US, the first store opened in 1962 and currently they are expected to grow at the rate of one store per day. In addition, each store may have significant impact on the vegetation in the surrounding area. All of these factors make the Wal-Mart data set ideal for validating land cover change events within the NDVI record[14]. In this paper, three Wal-Mart stores were used. One is located in a desert environment in Apple Valley, California, another is in forested land in Brewer, Maine, and the third one is in an area of mixed vegetation in Fayetteville, North Carolina. 20 locations were sampled from each region, representing the Wal-Mart construction area (4 locations), the bordering area (8 locations), and the background area (8 locations), respectively.

3. Theoretical Framework

3.1. Building a Reference Model for Difference Time Series

A key step of the approaches developed and utilized by previous researchers [10, 2] is to build a baseline model that captures normal behavior. However, it is often hard to find a generic baseline model in many real applications. For example, in our case, the NDVI time series may show completely different characteristics in different spatial regions. While some of the time series have strong seasonal effects, others do not exhibit seasonality. The predictive distribution justified for one region may not be appropriate or adequate in another region. In this study, we monitor the difference between time series pairs instead of monitoring the original NDVI time series.

If there is no change in vegetation in a region, the mean of the difference between time series pairs fluctuates around zero for regions that are close and have similar vegetation and land cover. When a Wal-Mart store is constructed in a certain area, the difference between the NDVI at the Wal-Mart location and at the background begins to deviate. This change can be detected based on the mean shift.

We arbitrarily choose 3 locations (Wal-Mart, bordering region, background) for our analysis for the Wal-Mart store at Fayetteville and obtained the NDVI time series difference for each location. We removed the mean-variance dependence from our data through a cubic-root transformation. Our analysis is based on the transformed difference scores $d_1, d_2, ..., d_i$. We assume that the differences are independent identically distributed (i.i.d.) with the Gaussian distribution, conditional on their means $\mu_1, \mu_2, ..., \mu_i$. If no change occurs, $\mu_t = 0$. Thus, the baseline model for the difference time series is

$$p(d_1, ..., d_i|\mu_t = 0) = \prod_{t=1}^{i} p(d_i|\mu_t = 0) \qquad (1)$$

We validate this assumption by the quantile-quantile (QQ) plot of the residuals against the theoretical quantiles of the standard normal distribution. Figure 1 shows the QQ plot for each difference, which indicates good agreement with the assumption.

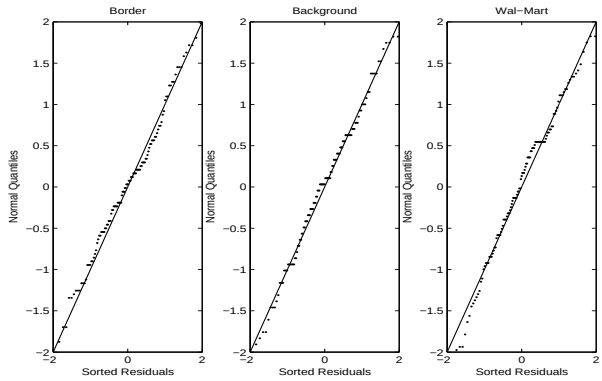

Figure 1. QQ plots of the difference time series between nearby points in space shows good agreement with the Gaussian distributional assumptions

3.2. Alarms for Single Large Change or Small but Sustained Changes

To detect small mean shift in the difference time series $d_1, d_2, ..., d_i$, we follow the approach of [10] and monitor the EWMA of normal score $q_t = \Phi^{-1}(P(d_t < d_t^{obs}))$, where Φ^{-1} is the inverse standard Gaussian cumulative distribution function and $P(d_t < d_t^{obs})$ is called tail probability, or p-values. If there is no change present and the baseline model is correct, the p-values will be approximately a uniform (0,1) distribution and q_t will be approximately normal (0,1), which allows standard control chart technology to be applied. The change detection is based on thresholding the severity metric z_t,

$$z_t = \lambda q_t + (1 - \lambda) z_{t-1} \tag{2}$$

where weight $\lambda \in (0, 1]$. This is also known as Q-charting in quality control[15]. Either a single large change or small but sustained changes over time can cause z_t to exceed a threshold. Thus, magnitude and duration are both incorporated into the severity metric z_t, and the weight between them is controlled by λ. In our case, we are only interested in small but sustained changes since Wal-Mart stores are not built in one day. Hence we tend to choose small values of λ.

It can be shown that z_t approximately follows a normal distribution $N(0, \lambda/(2-\lambda))$ if the baseline model is correct and the process is in control[13]. An alarm is generated, indicating a large change or sustained smaller changes, if $|z_t| > M\sqrt{\lambda/(2-\lambda)}$, where M is a user-defined parameter. Insight into the choice of M has been provided by [11, 17].

3.3. Tracking the Change Origin

As we have discussed in Section 1.1, change detection should include the ability to locate the change time point, in addition to generating alarms when change occurs. This process is called change point analysis[4]. The process is important both in our application, as well as in the development of our proposed theoretical framework, where we may need to identify the change point to be able to optimally re-adjust the severity metric after an alarm has been generated[5]. Change point analysis is often viewed as a parameter estimation problem and the methods used are primarily based on either likelihood ratios, or nonparametric methods, and/or Bayesian approaches[4]. However, these approaches are usually not appropriate for the online change detection environment since they typically require large volumes of data for parameter estimation. Moreover, most of these methods are limited by their assumptions on the model settings.

Here we propose a heuristic algorithm for our case study. The algorithm is fast, needs to retrieve only few past data points and locates the change point with satisfactory accuracy.

Since Wal-Mart stores are not built in a day, the shift of normal score q_t is not sudden but gradual. We assume the shift to be linear with slope parameter β. That is, if the change point is $k+1$ and the time point to generate an alarm is i, then the expectation of q_t is $E(q_t) = E(q_k) + \beta(t-k)$, for $i > t > k$. Once a Wal-Mart store is constructed at time point j ($j > i$), then $E(q_t)$, for $t > j$, is expected to remain invariant. For simplicity, our discussion is limited to the case of $\beta > 0$ (z_i exceeds the upper limits). The following property of the severity metric indicates that z_t will increase in a statistical sense if there are linear changes occurring in the means.

Property 1 *(see Appendix A for proof): In the case of linear changes in $E(q_t)$ with slope parameter β, then we have $E(z_t - z_{t-1}) = \beta(1 - (1-\lambda)^{t-k})$, for $i > t > k$.*

Based on Property 1, we propose the following algorithm for estimating the ground-breaking dates of the Wal-Mart stores. In fact, this property holds for any monotonic change function. Thus, the given algorithm is applicable to a wide range of cases.

```
Pseudo_code for change point identification
y := z_i; n := 0;    % Initial state
while (y > z_max) and (n < n_max)    % While not good
                                       enough and time remains
  yc := z_{i-1}          % Check the previous time point
  if yc < y
  then y := yc; i := i - 1;  % Decision point to
                               move to the prior time point
  elseif rand < e^{(y-yc)/(T_0 α^n)}    % rand is a random draw
                                          from [0,1] uniform
    then y := yc; i := i - 1;
  n := n + 1;        % Next iteration
return i;        % Return the change point
```

This algorithm is inspired by simulated annealing (SA)[7] for global optimization problems. After the change point $k+1$, the severity metric z_t is expected to increase at each time step. When an alarm is generated at the time point i, we move "downhill" to the previous time point, which is similar in spirit to searching for global minimum of z_t. Since z_t does not strictly increase, but increases in a statistical sense (i.e., the increase is statistically significant but because of the stochastic nature there may be an occasional decrease), we allow "uphill" moves, which saves the method from being stuck due to random effects in a point analogous to the local minima. z_{max} is the threshold of the

expected value for z_k, which should be around zero. We choose $z_{max} = L\sqrt{\lambda/(2-\lambda)}$, where $L < M$. n_{max} is the maximum number of iterations of the algorithm. In our case study, since the groundbreaking date, or the change point, should not be too far from the time point we generate an alarm, we choose $n_{max} = 20$. $e^{\frac{y-yc}{T_0 \alpha^n}}$ is analogous to the transition probability function in SA, where T_0 and α ($0 < \alpha < 1$) are called "initial temperature" and "temperature reduction factor", respectively. Insight into the choice of T_0 and α has been provided by [9, 1].

3.4. Updating the Parameters

In our case study, the mean and variance of the NDVI difference need to be updated in online mode. If the process is in control, the mean does not need to be updated (kept at zero). When an alarm is generated (for example, when a Wal-Mart store is constructed), the mean is held invariant until it converges to a new state (i.e. the store is completely constructed). We update the mean for monitoring new changes.

The updated variance σ_{t+1} is given by

$$\sigma_{t+1}^2 = \frac{t}{t+1}\sigma_t^2 + \frac{t}{(t+1)^2}d_{t+1}^2 \tag{3}$$

Only the current variance σ_t and the new observation d_{t+1} are needed for the operation (see Appendix B for the detailed proof).

4. Experimental Results

We validated our approach on the data from the three regions introduced in Section 2. For each region, a reference location was taken from background (desert, forest or mixed vegetation) and an objective location was arbitrarily taken from the Wal-Mart area. The difference time series d_t for each region is included in Figure 2.

We choose $\lambda = 0.1$, $M = 3.5$ for the EWMA control chart. To evaluate the performance of EWMA control charts, the Average Run Length (ARL) is a widely used measure[13]. The in-control ARL (ARL_0) is the average number of samples taken before an out-of-control signal is given, if the process remains in control. Our choice of parameters results in $ARL_0 \cong 500$[13]. In other words, we will have a false alarm occurring once in about every 500 samples if the process is in control, indicating our EWMA chart has low false alarm rate if no change is present. We choose the following parameters for the heuristic change point algorithm: $L = 1$,$T_0 = 10$, $\alpha = 0.6$, and $n_{max} = 20$.

Figure 2 shows the EWMA control chart which monitors normal scores q_t. The approach works in online mode since the current severity metric, z_t, and the new data, d_{t+1}, are sufficient to update the old severity metric and hence produce the new severity metric z_{t+1}. Each step of the alarm generation, as well as the change point detection, algorithm only involves a bit of basic algebra. In general, each iteration of either of the algorithms at one time point can be completed in almost constant time. Thus, the worst case scenario for the change point detection algorithm to reach completion (or arrive at the change point) could be $O(n)$ where n is the number of the time points. This worst case occurs when the change point is at the origin, which is rarely the case. Therefore, the updates can be generally completed in real time and the change points can be detected in real-time or near real-time. We note that in this experiment we validated our approach on a single change (namely, construction of a Wal-Mart store), hence updating the estimated mean of the baseline model was not necessary.

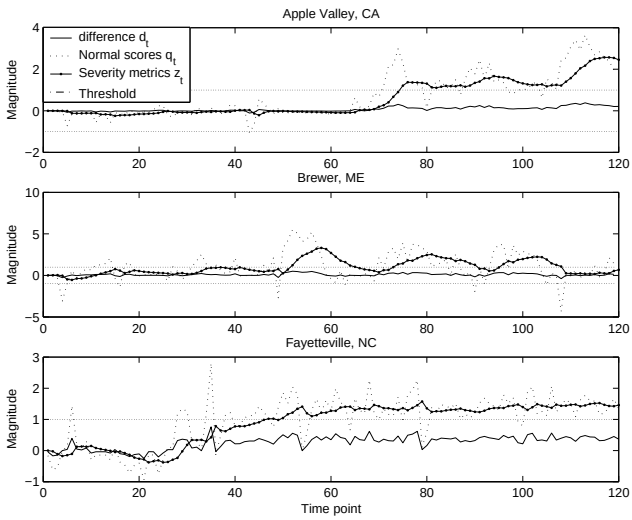

Figure 2. EWMA monitoring of normal scores

Figure 3 shows how the p-value evolves over time. As it has been discussed in Section 3.2, the p-values would follow a uniform (0,1) distribution if the reference distribution is correct and the process remains in control. We can see from Figure 3 that before the alarms were generated (solid line), p-values were approximately uniformly distributed, but after the alarm, this distribution became highly skewed compared to the distribution prior to the alarm. The dashed line indicates the change time point suggested by our approach.

Table 1 summarizes validation results for the three case study locations. The mode value from 100 runs are shown in the table. The row names, from top to bottom, denote the following: (a) Alarm: Time index when alarm was generated, (b) Change Point: Time when the original change

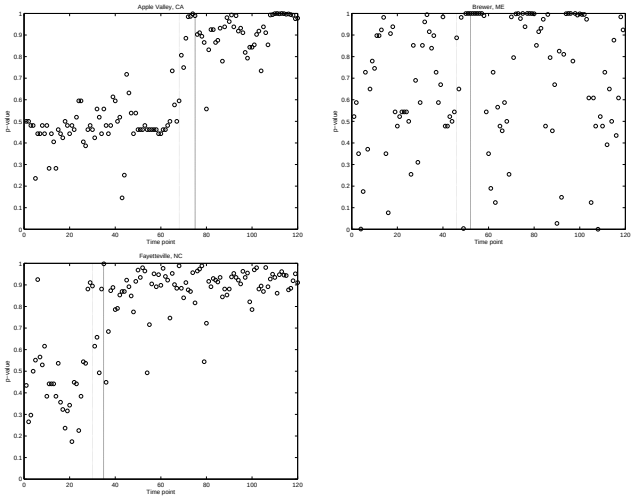

Figure 3. p-value distribution changes over time

was detected by our algorithm, (c) Store Opening: Time when the Wal-Mart store was opened, and (d) Groundbreaking: Time of groundbreaking, or the beginning of construction, for the Wal-Mart store. Alarms are generated when the sustained change in NDVI causes the severity metric to exceed a threshold and hence indicate that a change has been detected. Table 1 shows that an alarm was generated for each of the three cases. This is a validation of our proposed change detection approach since in each of the three cases a change in the NDVI is known to have occurred based on information about Wal-Mart store openings.

One further validation is that the time indices for the alarms are earlier than the actual store opening time in all three cases but later than the groundbreaking time in the one case where the latter is available. In addition, we would expect the detected change point to be on or after groundbreaking and the alarm to be generated after groundbreaking, possibly between the groundbreaking and the store opening dates. While we do not have information about the actual groundbreaking dates for the stores in ME and NC, if we were to assume identical time periods for construction (i.e., identical time gaps between groundbreaking and store opening), the groundbreaking time indices would be 51 and 30 for ME and NC respectively. This assumption, while imprecise, may be useful as a rough validity check. The experimental results shown in Table 1 are roughly consistent with our expectations.

5. Discussion

We present a statistical framework for land cover change analysis. It is fast, easy to implement, and storage efficient,

Table 1. Validation of alarms and change points

	CA	ME	NC
Alarm	75	52	35
Change Point	70	43	30
Store Opening	93	76	55
Groundbreaking	68	Missing	Missing

which satisfies online computation requirements. The approach is validated by the Wal-Mart land cover change validation data set[14]. In the case study, the approach did not truly perform real-time detection since data is available only every 16 days. However, the proposed approach updates information in real time and thus can be used to monitor daily (or more frequent) NDVI or other remotely sensed data.

We are extending our work to the multiple time series monitoring problem and taking into account the spatial dependence structure among the data. This can quickly get challenging in view of the increased computational complexity. We are exploring online dimensionality reduction techniques to summarize the statistics of multiple time series in real time. In addition, the proposed heuristic algorithm for change point detection has some attractive properties, which we are investigating in detail. A thorough comparison with traditional methods is needed. We are also justifying the theoretical foundation of the approach and exploring applications to other domains with streaming and evolving data.

Appendix A

In this appendix, we prove Property 1 given in Section 3.3.

EWMA has an alternative definition[13],

$$z_t = \lambda \sum_{j=0}^{t-1} (1-\lambda)^j q_{t-j} + (1-\lambda)^t z_0$$

Then, $z_t - z_{t-1} = \lambda q_t - \lambda^2 \sum_{j=0}^{t-2} (1-\lambda)^j q_{t-j-1} - \lambda(1-\lambda)^{t-1} z_0$

Thus, the expectation of $z_t - z_{t-1}$ is

$$E(z_t - z_{t-1}) = \lambda E(q_t) - \lambda^2 \sum_{j=0}^{t-2} (1-\lambda)^j E(q_{t-j-1}) - \lambda(1-\lambda)^{t-1} z_0$$

In the case of linear changes in the mean,

since $E(q_t) = E(q_k) + \beta(t-k)$, ($k+1$ is the time point when the change begins) and $E(q_k) = 0$ (the process is in control), it follows that

$$E(z_t - z_{t-1})$$

$$
\begin{aligned}
&= \lambda(z_0 + \beta(t-k)) - \lambda^2 \sum_{j=0}^{t-k-2} (1-\lambda)^j (z_0 + \beta(t-j \\
&\quad -1-k)) - \lambda^2 \sum_{j=t-k-1}^{t-2} (1-\lambda)^j z_0 - \lambda(1-\lambda)^{t-1} z_0 \\
&= \lambda z_0 - \lambda^2 \sum_{j=0}^{t-2} (1-\lambda)^j z_0 - \lambda(1-\lambda)^{t-1} z_0 + \lambda\beta(t-k) \\
&\quad -\lambda^2 \sum_{j=0}^{t-k-2} (1-\lambda)^j \beta(t-j-1-k) \\
&= \lambda\beta(t-k) - \lambda^2\beta \sum_{j=0}^{t-k-2} (1-\lambda)^j (t-k-1-j) \\
&= \lambda\beta(t-k) - \lambda\beta(t-k-1) + \beta((1-\lambda) - (1-\lambda)^{t-k}) \\
&= \beta(1-(1-\lambda)^{t-k})
\end{aligned}
$$

Appendix B

In this appendix, we prove the variance update formula (Eq.3) given in Section 3.4.

By definition, $\mu_t = \frac{1}{t} \sum_{j=1}^{t} d_j$

and $\sigma_t^2 = \frac{1}{t} \sum_{j=1}^{t} (d_t - \mu_j)^2 = \frac{1}{t} \sum_{j=1}^{t} d_j^2 - \mu_t^2$

Thus, $\sigma_{t+1}^2 = \frac{1}{t+1} \sum_{j=1}^{t+1} d_j^2 - \mu_{t+1}^2$

$$
\begin{aligned}
&= \tfrac{1}{t+1}(\sum_{j=1}^{t} d_j^2 + d_{t+1}^2) - (\tfrac{1}{t+1}(\sum_{j=1}^{t} d_j + d_{t+1}))^2 \\
&= \tfrac{1}{t+1}(t(\sigma_t^2 + \mu_t^2) + d_{t+1}^2) - (\tfrac{1}{t+1}(t\mu_t + d_{t+1}))^2
\end{aligned}
$$

In our case, $\mu_t = 0$, it follows that

$$
\begin{aligned}
\sigma_{t+1}^2 &= \tfrac{1}{t+1}(t\sigma_t^2 + d_{t+1}^2) - (\tfrac{1}{t+1} d_{t+1})^2 \\
&= \tfrac{t}{t+1}\sigma_t^2 + \tfrac{t}{(t+1)^2} d_{t+1}^2
\end{aligned}
$$

Acknowledgment

This research was sponsored by the Laboratory Directed Research and Development program of the Oak Ridge National Laboratory (ORNL), managed by UT-Battelle, LLC for the U. S. Department of Energy under contract no. DEAC05-00OR22725. We would like to thank David Andrew Gerdes, Gabriel Kuhn, Shiraj Khan, Vladimir Protopopescu, Robert Patton, George Ostrouchov and Budhendra Bhaduri, all of ORNL, for their support and helpful comments. In addition, we would like to thank Diane Lambert, formerly at Bell Labs Research and currently at Google, as well as Deepak Agarwal of Yahoo! Research, for helpful discussions. Finally, we thank Kalyan Perumalla and Robert Perumalla of ORNL, as well as Yuerong Chen and Professor M.K. Jeong of the University of Tennessee at Knoxville, for reviewing the manuscript.

References

[1] E. Aarts and J. Korst. *Simulated annealing and Boltzman machines*. Wiley, Chicheste, 1989.

[2] D. Agarwal, J. Feng, and V. Torres. Monitoring massive streams simultaneously: a holistic approach. *Interface*, January 2006.

[3] F. Baret and G. Guyot. Potentials and limits of vegetation indices for lai and apar assessment. *Remote Sensing of Environment*, 35:161–173, 1991.

[4] J. Chen and A. Gupta. *Parametric Statistical Change Point Analysis*. Birkhauser, Boston, MA, 2000.

[5] A. R. Ganguly and Y. Fang. Online alarm generation in sensor-cyber networks. In *Session on Sensor-Cyber Networks for Homeland Defense, 9th ONR/GTRI Workshop on Target Tracking in Sensor Fusion, Analytical Predictions of Tracking Performance, Office of Naval Research and Georgia Tech Research Institute, Gatlinburg, TN*, 2006.

[6] A. Habib and R. Al-Ruzouq. Semi-automatic registration of multi-source satellite imagery with varying geometric resolutions. *Photogrammetric Engineering and Remote Sensing*, 71(3):325–332, March 2005.

[7] S. Kirkpatrick, C. D. Gelatt, and M. Vecchi. Optimization by simulated annealing. *Science, Number 4598, 13 May 1983*, 220, 4598:671–680, 1983.

[8] B. Krishnamurthy, S. Sen, Y. Zhang, and Y. Chen. Sketch-based change detection: Methods, evaluation, and applications. In *IMC '03: Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement*, 2003.

[9] v. P. Laarhoven and E. Aarts. *Simulated Annealing: Theory and Application*. Reidel, Dordrect, 1987.

[10] D. Lambert and C. Liu. Adaptive thresholds: monitoring streams of network counts online. *Journal of the American Statistical Association*, 101(473):78–88, March 2006.

[11] J. M. Lucas and M. S. Saccucci. Exponentially weighted moving average control schemes: properties and enhancements (with discussion). *Technometrics*, 32:1–29, 1990.

[12] J.-F. Mas. Monitoring land-cover changes: a comparison of change detection techniques. *International Journal of Remote Sensing*, 20(1):139–152, January 1999.

[13] D. C. Montgomery. *introduction to statistical quality control,the fourth edition*. John Wiley and Sons, New York, 2001.

[14] D. Potere, N. Feierabend, E. Bright, and A. Strahler. A new source for land cover change validation: Wal-mart from space. *Working Paper (http://opr.princeton.edu/papers/), Office of Population Research, Princeton University*, 2006-04.

[15] C. Quesenberry. Spc q-charts for start-up processes and short or long runs. *Journal of Quality Technology*, pages 213–224, 1991.

[16] J. W. Rouse, R. H. Haas, J. A. Schell, and D. W. Deering. Monitoring vegetation systems in the great plains with erts. In *Third ERTS Symposium, NASA SP-351*, volume 1, pages 309–317, 1973.

[17] S. B. Vardeman and J. M. Jobe. *Statistical Quality Assurance Methods for Engineers*. John Wiley and Sons, New York, 1998.