ORNL-12879

Computer Science and Mathematics Division
Mathematical Sciences Section

A METHOD FOR DETECTING CHANGES IN LONG TIME SERIES

Darryl J. Downing
William F. Lawkins
Max D. Morris
George Ostrouchov

Mathematical Sciences Section
Computer Science and Mathematics Division
Oak Ridge National Laboratory
P. O. Box 2008
Oak Ridge, Tennessee 37831-6367

Date Published: September 1995

Research supported by the Applied Mathematical Sciences Research Program of the Office of Energy Research, U.S. Department of Energy.

Prepared by the
Oak Ridge National Laboratory
Oak Ridge, Tennessee 37831
managed by
Lockheed Martin, Inc.
for the
U.S. DEPARTMENT OF ENERGY
under Contract No. DE-AC05-96OR22464

Contents

1 Introduction 1
2 Statistical Change Point Problems 2
3 Statistical Model 4
4 Basic Method 4
5 Reference Values for S 8
6 Computational Issues 9
7 Trends 10
8 Example: A Computer-Generated Time Series 11
9 Example: A Physical Time Series 15
10 Extension to Multivariate Series 19
11 References 21

List of Figures

1	Window arrangement for change point detection procedure.	6
2	Computer-Generated Time Series (a), and Results of Proposed Analysis on Original Data (b) and First-Order Differences (c) . .	13
3	Computer-Generated Time Series (a), and Results of CUSUM Analysis of Original Data (b) and First-Order Differences (c). . .	16
4	Sunspot Time Series (a), and Frequency Polygon of Data Values (b).	18
5	200-Point Neighborhoods Around Four Most Significant Changes in Sunspot Data.	20

A METHOD FOR DETECTING CHANGES IN LONG TIME SERIES

Darryl J. Downing

William F. Lawkins

Max D. Morris

George Ostrouchov

Abstract

Abstract Modern scientific activities, both physical and computational can result in time series of many thousands or even millions of data values. Here we describe a statistically motivated algorithm for quick screening of very long time series data for the presence of potentially interesting but arbitrary changes. The basic data model is a stationary Gaussian stochastic process, and the approach to detecting a change is the comparison of two predictions of the series at a time point or contiguous collection of time points. One prediction is a "forecast", i.e. based on data from earlier times, while the other a "backcast", i.e. based on data from later times. The statistic is the absolute value of the log-likelihood ratio for these two predictions, evaluated at the observed data. A conservative procedure is suggested for specifying critical values for the statistic under the null hypothesis of "no change".

1. Introduction

By "time series", we generally mean a sequence of data values indexed sequentially in time, usually spaced at equal time increments. Statistical methods for the analysis of time series data have generally been developed under the assumption that the series to be analyzed consists of tens, hundreds, or perhaps thousands of such data values. Analytical techniques have been proposed for inferring the structural form (or model) of a time series, estimation of model parameters, and prediction of values for time frames not yet observed. Another more specialized form of analysis is the detection or location of changes in the series, i.e. time frames at which the basic structure of the generating process appears to be altered.

In recent years, improvements in measurement and computational technologies have made possible the collection or generation of time series data sets which are very much larger than those of interest a few decades ago. Examples include time series generated in speech recognition, weather monitoring, and satellite imaging problems. In addition to these physical measurement systems, computer models have been developed which simulate physical systems and often produce output in the form of time series. Computer models which simulate molecular dynamics, engineered structures under stress, and global climate are examples of current interest to the DOE.

Modern scientific activities, both physical and computational, can result in time series of many thousands or even millions of data values. Because of the sheer quantity, particularly in computational activities, animated graphic techniques are often used to generate a "movie" for visual analysis. Of major interest to the scientific investigator is the question of where changes may occur in the time series, where the nature of these changes may not be well defined. A related problem is that of detecting "outliers" or "transients" which may not be associated with a permanent change in the state of the system, but are in some sense meaningfully different from other output values nearby in time.

For very long time series, however, visual inspection by animated graphics can be unsatisfactory due to the length of the presentation which must be viewed by the investigator. The purpose of the research reported here is to develop a statistically motivated algorithm to screen very long time series data for the presence of potentially interesting changes and outliers. If successful, such an algorithm would be of considerable value, since it would allow physical and computational scientists to concentrate most of their effort on inspecting relatively short data segments of potentially greater scientific interest.

2. Statistical Change Point Problems

Within the statistical literature, the detection of change in an ordered sequence of observations, including time series, is often referred to as the "change point problem." Some of the earliest statistical work in change point problems can be credited to Page (1954, 1955, 1957) who developed change point detection schemes for the situation in which observations are statistically independent over time, the pre- and post-change distributions are known, and all that must be determined is the time of change. This corresponds in some cases to CUSUM procedures, and was an important contribution to the development of control charts. Chernoff and Zacks (1964) also examined data assumed to be independent realizations, where each distribution is Gaussian and the change is a shift of known direction but unknown magnitude in the mean. Hinkley (1970,1971) derived the maximum likelihood estimate of the time of an unknown change in the mean under a model of independent Gaussian measurements. James et al. (1987) compared the performance of some of these and other procedures developed under models of independent Gaussian observations. Most work described in the literature relies on the assumption of independent observations; an exception is Box and Tiao (1965), who considered detection of a shift in the mean of a nonstationary integrated moving average process. Three examples of more recent treatments of change point detection methods are given by Yao (1988), Gordon and Smith (1990), and Barry and Hartigan (1993).

A recent book by Brodsky and Darkhovsky (1994) presents a summary of work done on change point problems. They present an argument (pp 15-16) showing that for an arbitrary change in a stochastic process model, a transformation of the process exists so that in the transformed process, the change affects the expectation. They interpret this as justification for concentrating effort on procedures which detect changes in the mean, but this argument may be primarily of theoretical value.

Most of the methods which have been described seem to be inappropriate for our purposes for at least one of the following reasons:

1. The changes for which these methods have been developed are generally changes in the mean of the stochastic process only; the variance of the data is assumed to be constant throughout the time series, and all observations are usually modeled as mutually independent random variables.

2. Existing methods developed for analysis of all data simultaneously generally require algorithms with computational complexity of $O(N^2)$ or greater, where N is the length of the time series.

3. Control chart methods are ordinarily designed for sequential operation as the data are being generated, where the aim is to detect changes as quickly as possible after their occurrence.

Because we are interested in more general classes of change than simple changes in mean over time (e.g. deviations from established trends, changes in variability, et cetera), (1.) seems an unacceptable restriction for our purposes. For practical reasons, (2.) is very undesirable because our intended applications will often involve quite large values of N. Finally, although control chart techniques may be of some use for our problem, (3.) is an unnecessary restriction in our setting and may limit the performance characteristics of control charts relative to what may be done with access to more data. Because of these problems with existing methods, we have begun development of a different approach, as described below.

3. Statistical Model

Our basic data model is a stationary Gaussian stochastic process. Letting y_i represent the data variate generated at the i th time step, a stationary Gaussian model is then completely defined by:

$$\mu = E\left[y_i\right]$$

$$B(d) = Cov\left[y_i, y_j\right], \quad d = |i - j|$$

A change point is defined as any time at which the mean or covariance function changes, and a potential outlier is one or a very few contiguous data points which would have very low probability under models which fit the surrounding data well.

We use a Gaussian process for computational convenience, and because it provides a realistic class of models for many interesting time series. However, some series may require transformation before analysis, e.g. series which are always positive and relatively more variable when series values are relatively great may benefit from a logarithmic or similar transformation. This statistical model is usually not based on physical or mathematical knowledge about how the data are generated. Rather, it may more realistically be said to reflect idealized "uncertainty" about the time series. Given this interpretation, a Bayesian view of the issue of change point and outlier detection may be most appropriate.

4. Basic Method

Let y_P denote the collection or vector of data values generated in a contiguous "window" of time steps, i.e. $P = [i, j]$, $i < j$ represents the window comprised of the ith, $i + 1$st, $i + 2$nd, …, jth time steps. A "prediction" or "forecast" of y_P, based on data which are located before P in time,can be constructed via the commonly used conditional approach to Gaussian time series. For specified parameter values, a commonly used forecast based on data from a conditioning window C with $C < P$ (i.e. C occurring before P in time) is the mean of the conditional Gaussian density

$$\Phi_E\left(Y_P|y_C\right)$$

sometimes called the predictive density in Bayesian applications. (The notation capital Y is used to denote the random vector of which y is the corresponding realization.) A fully Bayesian approach could begin with placing a subjective prior on the process parameters and generating a corresponding (generally) non-Gaussian predictive density for y_P. Here we shall take a more expedient route (similar in spirit to the approach described in Currin et al. (1991)), estimating the parameters from an estimation window of data$E < P$, not necessarily the same as C, resulting in a predictive Gaussian density

$$\Phi_E\left(Y_P|y_C\right)$$

A measure of the quality of this prediction is its evaluation at the observed data value, i.e.

$$\Phi_E\left(Y_P|y_C\right)$$

The basic idea behind our approach to detecting a change point is the comparison of two predictions y_P, one a "forecast" as described above, and the other a "backcast" that is based on data from windows which are subsequent to P in time. Formally, for a window P of data, we also select windows:

$$C_1, \quad P, \quad E_1 < P$$

and

$$C_2, \quad P, \quad E_2 > P$$

The statistic is then defined as the log of the ratio of the larger to the smaller evaluated predictive densities, or equivalently.

$$S(P) = |\log \Phi_{E1}(Y_P|y_{C1}) - \log \Phi_{E2}(y_P|y_{C2})| \tag{1}$$

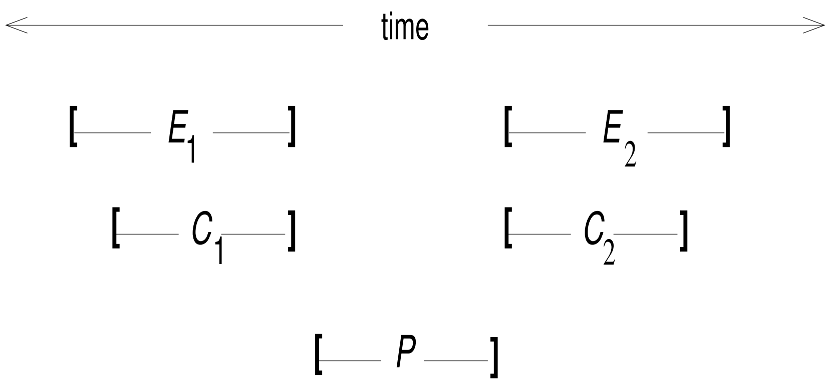

Figure 1: Window arrangement for change point detection procedure.

Without the absolute value, this is the statistic associated with the likelihood ratio test where one hypothesis is that the distribution of Y_P corresponds to the parameters estimated from E_1, given y_{C_1}, and the other is that the distribution of Y_P corresponds to the parameters estimated from E_2, given y_{C_2}.

Figure 1 displays graphically the relationships which exist between these windows. Although in principle C_1 could appear anywhere before P in time, the predictor is most strongly influenced by the conditioning data if the two windows are close together since correlations between data in the two windows are usually strongest in this case. Also, it is reasonable that the estimation windows should be relatively large, since parameter estimates should be reliable. Finally, because our approach assumes that the forecast and backcast should be equally reliable when changes or outliers do not occur, symmetry of size and placement of the estimation windows and conditioning windows around P is desirable.

When the time series is "well behaved" without change points or outliers, values of S should be small since both forecast and backcast would be expected to predict about equally well. On the other hand, if a change point exists in $E_1(E_2)$, i.e. the mean or covariance function changes somewhere in this window, the forecast (backcast) should suffer due to loss of accuracy in the process parameters. Similarly, if an outlier exists in $C_1(C_2)$, the forecast (backcast) should also lose accuracy. Both predictions will be relatively poor if the change point or outlier is actually in P, but this would not necessarily affect the comparative form of S. Hence a large value of $S(P)$ should be interpreted as evidence for a change or outlier in at least one of the five data windows.

The procedure described above is used by "sliding" the arrangement of 5 windows through the entire time series, evaluating S at each time-step increment. The result is a second time series consisting of the values of S generated. Intervals of time over which S is relatively large are flagged as containing potential change points or outliers, and subsequently reviewed graphically by the investigator to characterize their meaning in the context of the problem.

The approach described here for detecting changes is different from that which motivates control chart methodology. In control chart applications, the goal is generally to detect changes or outliers sequentially, as the data are actually generated. A common approach to control charting for correlated data is to examine predictive residuals – in our case, the difference between predictions of the data in P based on data appearing before it, and their observed values, eg. Alwan and Roberts (1988). An "out of control" signal is issued if the absolute value of the predictive residual exceeds some specified value. In contrast to this, we are not limited to using data from before P, but have access to later data from which our backcast is formed. If a change occurs in the series just before P, the backcast should more reliably predict y_P than the forecast. Similarly, if a change occurs in the series just after P, the forecast should more reliably predict y_P than the backcast. In either case, $S(P)$ will tend to be larger than it would be if no changes occur within any of the windows. Using data on both sides of a potential change should produce a more effective detection method than one based simply on data preceding the potential change. In addition, we conjecture that using two predictions in this manner should make this method somewhat less dependent on the Gaussian assumption. Any prediction method based on the Gaussian assumption may be prone to erroneous predictions when changes are not actually present; this can seriously damage the effectiveness of a control chart procedure. However, by comparing two such predictions, some of this effect may be "canceled out" – $S(P)$ may be less sensitive to non-Gaussian behavior than a statistic based on a single prediction.

5. Reference Values for S

Here we will assume that the conditioning windows are of equal size, the estimation windows are of equal size, and the window arrangement is symmetric about the center of P as depicted in Figure 1. Consider the "signed" version of $S(P)$, i.e. without the absolute value:

$$\begin{aligned} S_{sign}(P) &= \log \Phi_{E1}(y_P | y_{C1}) - \log \Phi_{E2}(y_P | y_{C2}) \\ &= \frac{1}{2}\Big[\log \left|\hat{\Sigma}_{E_1}(P|C_1)\right| - \log \left|\hat{\Sigma}_{E_2}(P|C_2)\right| \\ &\quad + \Big(y_P - \hat{\mu}_{E_1}(P|C_1)\Big)' \hat{\Sigma}_{E_1}(P|C_1)^{-1} \Big(y_P - \hat{\mu}_{E_1}(P|C_1)\Big) \\ &\quad - \Big(y_P - \hat{\mu}_{E_2}(P|C_2)\Big)' \hat{\Sigma}_{E_2}(P|C_2)^{-1} \Big(y_P - \hat{\mu}_{E_2}(P|C_2)\Big)\Big] \end{aligned} \tag{2}$$

where $\hat{\mu}$ and $\hat{\sum}$ and denote estimates, based on data indicated by their subscripts, of the conditional mean vectors and variance matrices indicated by their arguments. Under the "null hypothesis" of no change in process parameters, and assuming that the parameter estimates are consistent, the two determinants are asymptotically equal as n_E, the common width of the estimation windows, increases:

$$\lim_{n_E \to \infty} S_{sign}(P) = \frac{1}{2}\Big[Q(P|C_1) - Q(P|C_2)\Big] \tag{3}$$

where $Q(P|C_1)$ and $Q((P|C_2)$ are the quadratic forms shown in (2) with true parameter values substituted for estimates associated with E_1 and E_2, respectively. Under the model of no change, each quadratic form has a $\chi^2_{n_P}$ distribution, where n_P is the width of P. Hence,

$$E\left[Q(P|C_1)\right] = E\left[Q(P|C_2)\right] = n_P$$

$$Var\left[Q(P|C_1)\right] = Var\left[P|C_2)\right] = 2n_P$$

It follows immediately that the distribution of $S_{sign}(P)$ is symmetric about zero. Complete specification of the distribution is difficult because the two quadratic forms are correlated. However, an upper limit on the variance is $2n_P$, which would be achieved only when the covariance of $Q(P|C_1)$ and $Q(P|C_2)$ is as small as possible. Hence, an upper bound on "k standard deviations" of S_{sign} is $k\sqrt{2n_P}$, and deviations of more than this from the mean of S_{sign} correspond to values of S greater than $k\sqrt{2n_P}$. Conservative values of k can be selected based on Chebyshev's inequality; the probability α critical value by this bound is $\sqrt{\frac{2n_P}{\alpha}}$.

6. Computational Issues

Parameter Estimation: In the early stages of this research, we attempted to use reasonable parametric forms for B and apply the method of maximum likelihood to estimate process parameters within each of E_1 and E_2 but this appeared to be too computationally demanding for our purposes. We have also used simple closed-form estimators which are popular in are time series literature, e.g.

$$\hat{\mu}_E = \frac{1}{n_E} \sum_E y_i$$

$$\hat{B}_E(d) = \frac{1}{n_E - d} \sum_E (y_i - \hat{\mu}_{low})(y_{i+d} - \hat{\mu}_{high}) \tag{4}$$

where

$$\hat{\mu}_{low} = \frac{1}{n_E - d} \sum_{E_{low}} y_i, \quad \hat{\mu}_{high} = \frac{1}{n_E - d} \sum_{E_{high}} y_i$$

and E_{low} and E_{high} are subwindows of E, containing its first and last $n_E - d$ values, respectively. Although somewhat less precise than MLE's derived under an appropriate parametric model of covariance, these estimators can be quickly computed and easily updated as the windows slide over the data stream. However, a disadvantage of these estimators is that they can result in estimated variance matrices which are not positive definite, especially when C and/or P are large windows and the output is a smoothly varying function of time.

In order to avoid numerical problems associated with parameter estimation, we currently use a different covariance estimator:

$$\hat{B}_E(d) = \frac{1}{n_E} \sum_E (y_i - \hat{\mu}_E)(y_{i^*} - \hat{\mu}_E) \tag{5}$$

where

$$i^* = \begin{cases} i+d, & \text{if } i+d \in E \\ i+d-n_E, & \text{otherwise} \end{cases}$$

This "circular" estimator is not as appealing as (4), but is consistent as n_E increases relative to $n_C + n_P$. As a more practical matter, let $T\ [_Y E]$ represent the n_E-element vector whose i th element is the $i-1$st element of Y_E, $i = 2$, 3, ..., n_E, and whose first element is the n_E th element of Y_E. Then it can be shown that (5) results in estimated covariance matrices which are positive definite so long as there are not exact linear dependencies among Y_E, $T[Y_E]$, $T^2[Y_E]$, ..., $T^{n_C+n_P}[Y_E]$, and the n_E-element column vector of 1's.

Computational Effort: The procedure as described above requires calculation of $O(N)$. For each evaluation of $S(P)$, a large portion of the effort goes into calculating the determinant and inverse of the estimated conditional variance matrices of Y_P. For conditioning and prediction windows of width n_C and n_P, respectively, this effort is of $O((n_C + n_P)^3)$.

7. Trends

As with most "omnibus" tests designed to detect a condition which is not narrowly defined (e.g. "change" in our case), there are some relatively simple data patterns which can yield problems. One such pattern is response data which contain a simple linear trend in time, e.g.

$$y_1 = a + bi + \in_i \quad b \neq O$$

where $\in$ is now a realization of a stationary Gaussian process. As n_E becomes large relative to $n_C + n_P$, the estimated covariances of interest, $\hat{B}_E(d)$, all increase, and apparent correlations between values separated by d, $d \leq n_C + n_P$, all approach 1. As a result the variance matrices required to calculate likelihood become ill-conditioned, or even "numerically singular".

A simple solution to this problem, and one we advocate at this point, is transformation of y by a difference operator. In this case, the first order difference transformation of the time series, i.e. the series $y_{i+1} - y_i$, i.e. is stationary and in many cases leads to fewer numerical problems in our algorithm. Similarly, quadratic trends can be addressed by the second-order difference operator $y_{i+2} - 2y_{i+1} + y_i$, and so forth.

Finally, note that even though we have spoken of y as if it were a realization from a stochastic process Y, this is mainly conceptual. The actual data may be generated by a deterministic process,as with a non-stochastic computer model. A simple example of this is the above linear trend without a random component, i.e. an exact linear trend in time. In this case, first-differences are (numerically) even worse than the original series for our algorithm, since all estimated covariances will be exactly zero. However, in this case the situation is at least easy to detect; variance estimates of zero indicate that the data are constant if a difference transformation has not been made, or exactly follow an r th degree polynomial in time if r th differences are being analyzed. Since deterministic computer models may produce output which approaches a steady state reflected by such a simple function in time, implementations of this procedure should probably include a check for zero variances.

8. Example: A Computer-Generated Time Series

A simple nonlinear difference model which is often encountered in the biological literature as an empirical description of density-limiting population growth is defined by the equation:

$$y_{i+1} = \lambda y_i (1 + a y_i)^{-b}$$

where λ, a, $b > 0$; e.g. Hassell (1975). Here, we shall add a stochastic component to this model by letting a vary from time step to time step as an autoregressive Gaussian process. Specifically,our test model will be defined by:

$$\lambda = 1.1 \qquad b = 1.0$$
$$\mathrm{E}[a_i] = \mu_a \quad \mathrm{Var}[a_i] = a_a^2$$

$$\mathrm{Corr}[a_i, a_j] = \rho_a^{|i-j|}$$

Beginning with the initial value $y_1 = 1$, a series of 20,000 observations was generated. Changes were introduced by altering the parameters of the series a_i at the 5,000th, 10,000th, and 15,000th time points, as detailed in the following table:

Time Steps	μ_a	σ_a	ρ_a
0 - 4,999	0.100	0.001	0.5
5,000 - 9,000	0.101	0.101	0.5
10,000 - 14,999	0.101	0.002	0.5
15,000 - 20,000	0.101	0.002	-0.5

The series is plotted in Figure 2(a).

The analysis described above was performed on the series, using window sizes $n_E = 400$, $n_C = 10$, and $n_P = 10$. Values of S generated by analyzing the data and first order differences of the data are shown in Figures 2(b) and 2(c), respectively, where $S(P)$ is plotted against the center of P. Following the argument of Section 5, a conservative 0.05 upper critical value for S is 20.0, and a corresponding 0.01 value is 44.7.

The change caused by increasing the variance of a_i (10,000th time step) was most clearly detected in the analysis of both the original and differenced data. The change involving ρ_a (15,000th timestep) is more clearly seen in the analysis of the first differences; this is probably due to the special relationship between differencing and autoregressive series, even though the series a_i are not being directly analyzed here. The change caused by altering the expectation of a_i (5,000th time step) is least clear, although there are several “spikes” in S based

Figure 2: Computer-Generated Time Series (a), and Results of Proposed Analysis on Original Data (b) and First-Order Differences (c)

on the undifferenced data in the vicinity of i = 5,000. This is probably due to the dampening caused by the difference equation. Even though the mean of a changes abruptly, the change is more gradual in y. Larger values of n_C and n_P might help here, but would slow the algorithm down. Another possibility, where gradual changes are expected, would be to analyze, say, every 5th or 10th data value rather than the entire sequence.

Finally, analysis of the same time series was performed with a CUSUM statistic, to provide a limited comparison of how these two approaches might be expected to perform in at least this one demonstration case. Even thought the form of CUSUM test used is actually intended for detecting changes in mean level, it is of interest to see to what extent this computationally very easy technique might also detect the three changes embedded in this data set.

The one-sided CUSUM statistic for detecting an upward shift in the series is defined, for time index i, as

$$C_i^+ = \max(0, C_{i-1}^+ - \mu - k\sigma)$$

where μ and σ are the assumed mean and variance of the series up to time index i and $C_0 = 0$. In practice, these values are usually estimated from data; in this calculation, we estimated them using the preceding 400 data values (chosen to match n_E in the calculations described above), i.e. y_{i-400} through y_{i-1}. This is a slight departure from usual control chart practice, where the estimates are generally not changed, and are thought of as representing an "in control" process state. The one-sided CUSUM statistic for detecting a downward shift in the series is defined, for time index i, as

$$C_i^- = \max(0, C_{i-1}^- + \mu + k\sigma)$$

Finally, in this analysis, we shall use

$$C_1 = \max(C_i^-, C_i^+)$$

as the CUSUM statistic to indicate change, i.e. as a possible competitor for S. In this exercise,the CUSUM parameter k was set to the value of 1; other values were also tried with similar results.

Figure 3 displays the result of the CUSUM analysis. As in Figure 2, panel (a) is a plot of the data series, panel (b) is a plot of the CUSUM statistic calculated as described for these data, and panel (c) is the CUSUM series calculated on the first order differences of the data. While the change at $i = 10,000$ might be detected using this approach, the changes at $i = 5,000$ and 15,000 are not associated with particularly large values of C. It would appear, at least for this example, that use of the proposed method more clearly suggests the location of the 3 changes than this implementation of a CUSUM procedure.

9. Example: A Physical Time Series

The daily sunspot data set, a time series of 52,320 data values, is maintained and made available by the National Center for Atmospheric Research via ftp server (`ncardata.ucar.edu/datasets/ds834.0/dailydata`). The following description is provided with the data.

“In 1848 the Swiss astronomer Johann Rudolph Wolf introduced a daily measurement of sunspot number. His method, which is still used today, counts the total number of spots visible on the face of the sun and the number of groups into which they cluster, because neither quantity alone satisfactorily measures sunspot activity.”

“An observer computes a daily sunspot number multiplying the number of groups he sees by ten and then adding this product to his total count of individual spots. Results, however, vary greatly, since the measurement strongly depends on observer interpretation and experience and on the stability of the Earth’s atmosphere above the observing site. Moreover, the use of Earth as a platform form which to record these numbers contributes to their variability, too, because the sun rotates and the evolving spot groups are distributed unevenly across solar longitudes. To compensate for these limitations, each daily international number

Figure 3: Computer-Generated Time Series (a), and Results of CUSUM Analysis of Original Data (b) and First-Order Differences (c).

is computed as a weighted average of measurements made from a network of cooperating observatories."

"How do sunspot numbers in these tables compare with the largest values ever recorded? The highest daily count on record occurred December 24 and 25, 1957. On each of those days the sunspot number totaled 355. In contrast, during years near the minimum of the spot cycle, the count can fall to zero. Today, much more sophisticated measurements of solar activity are made routinely, but none has the link with the past that sunspot numbers have."

The dataset analyzed contains the daily sunspot numbers recorded from January 1, 1850 through March 31, 1993. A plot of the data over time is shown in Figure 4(a), and Figure 4(b) displays a frequency polygon of the values in the series. While the time scale of Figure 4(a) is such that local features cannot be identified, it does clearly demonstrate that the series has a strong periodic component; the dominant period is approximately 10 years, however previous investigators have indicated that the frequency of this pattern is not constant. A second characteristic which is clear from Figure 4 is that the relative frequency of 0's is quite high, while those of most of the other values less than 10 is relatively low. (All data are non-negative integers.) This may have something to do with the special use of the value 10 in computing the sunspot number, as indicated in the above description. This distribution of sunspot values is clearly skewed in a positive direction, and the time series plot show that segments of relatively large averages have relatively great variability as well.

Although the data are clearly not normally distributed, and our changepoint method is developed based on an assumption that a Gaussian stochastic process is an appropriate model for the observed process, it is of interest to see how well the method might be expected to work. The algorithm was applied to the square-root of the actual data.

This transformation seems reasonable given the apparent relationship between the mean and variability of the series. Additionally, some preliminary calculations indicated that this transformation approximately eliminated the local (blocks of

Figure 4: Sunspot Time Series (a), and Frequency Polygon of Data Values (b).

100 observations) skewness from the dataset. Window sizes of $n_E = 100$, $n_C = 10$, and $n_P = 10$ were used. No detrending was performed prior to analysis; the change point algorithm as used here analyzes segments of 210 contiguous points, which is a small interval relative to the major periodic pattern in the data.

For analysis of the transformed data, statistic values S exceeded 10 in 216 separate intervals of time. (The number of time steps at which the statistic was greater than 10 was much larger than this – here we've reduced each such interval to the single time point at which the statistic reached its greatest value.) About half of these may be attributed to the interesting behavior of the series when the values drop below approximately 10. After eliminating all those times which were within 15 time intervals (i.e. the width of the conditioning window plus half the width of the prediction window) of an observed data values of 10 or less, this left 99 intervals in which the statistic value exceeded 10. Figures 5(a) - 5(d) display segments of 200 time-steps around each of the 4 time points for which the statistic was largest. The change most apparent at the middle of Figure 5(a) may be one of variability, which seems to decrease dramatically at approximately day 38670. A change in variability, or perhaps serial correlation, may also be the feature detected in Figure 5(b). Figures 5(c) and 5(d) appear to involve changes in level of the time series. Of course, the scientific significance of these or other apparent changes in the time series would require expert interpretation. However, this example does serve to indicate that the method can successfully identify segments over which the series visually appears to change.

10. Extension to Multivariate Series

Current and near-future work in this area will concentrate on generalizing the above method to multivariate data, or cases in which several data values representing different quantities are calculated for each time step. The basic model here will be one of parallel, correlated time series. However, a full multivariate analysis of the output will likely not be practical because of the computational effort which would be required. Our efforts will be in trying to find a highly structured

Figure 5: 200-Point Neighborhoods Around Four Most Significant Changes in Sunspot Data.

(restricted) multivariate model which can support a quick analysis, but which has enough flexibility to capture many important kinds of changes which are not apparent in the individual data streams when examined separately.

11. References

[1] Alwan, L. C., and H. V.Roberts (1988). Time-Series Modeling for Statistical Process Control, Journal of Business and Economic Statistics, b, 87-95.

[2] Barry, D., and J.A. Hartigan (1993). "A Bayesian Analysis for Change Point Problems," Journal of the American Statistical Association, 88, 309-319.

[3] Brodsky, B.E., and B.S. Darkhovsky (1994). Nonparametric Methods in Change-Point Problems. Kluwer Academic Publishers, Dordrecht, The Netherlands

[4] Box, G.E.P., and G.C. Tiao (1965). "A Change in Level of a Nonstationary Time Series," Biometrika, 52, 181-192.

[5] Chernoff, H., and S. Zacks (1964). "Estimating the Current Mean of a Normal Distribution which is Subject to Changes in Time," Annals of Mathematical Statistics, 35, 999-1028.

[6] Currin, C., T. Mitchell, M. Morris, and D. Ylvisaker (1991). "Bayesian Prediction of Deterministic Functions, with Applications to the Design and Analysis of Computer Experiments," Journal of the American Statistical Association, 86, 953-963.

[7] Gordon, K., and A.F.M. Smith (1990). "Modeling and Monitoring Biomedical Time Series," Journal of the American Statistical Association, 85, 328-337.

[8] Hassell, M.P. (1975). "Density Dependence in Single-Species Populations," J. Anim. Ecol., 44,283-295.

[9] Hinkley, D.V. (1970). "Inference about the Change-Point in a Sequence of Random Variables," Biometrika, 57, 1-17.

[10] Hinkley, D.V. (1971). "Inference about the Change-Point from CUSUM Tests," Biometrika, 58,509-523.

[11] James, B., K.L. James, and D. Siegmund (1987). "Tests for a Change-Point," Biometrika, 74,71-83.

[12] Page, E.S. (1954). "Continuous Inspection Schemes," Biometrika, 1, 100-115.

[13] Page, E.S. (1955). "A Test for a Change in a Parameter Occurring at an Unknown Point," Biometrika, 42, 523-526.

[14] Page, E.S. (1957). On a Problem in which a Change in a Parameter Occurs at an Unknown Point," Biometrika, 44, 248-252.

[15] Yao, Y.-C. (1988). "Estimating the Number of Change-Points via Schwarz' Criterion," Statistics and Probability Letters, 6, 181-189. ORNL/TM-12879

ORNL-12879

INTERNAL DISTRIBUTION

1. C. K. Bayne
2. T. S. Darland
3. C. S. Daw
4–8. D. J. Downing
9. L. J. Gray
10. P. Kanciruk
11–15. W. F. Lawkins
16–20. m. R, Leuze
21–25. M. D. Morris
26. C. E. Oliver
27–31. G. Ostrouchov
32–36. S. A. Raby
37. R. L. Schmoyer
38. R. F. Sincovec
39. P. T. Singley
40. D. A. Wolf
41. K-25 Applied Tech. Library
42. Y-12 Technical Library
43. Laboratory Records - RC
44–45. Laboratory Records Dept.
46. Central Research Library
47. ORNL Patent Office

EXTERNAL DISTRIBUTION

48. Dr. Richard Beckman, Statistics Group A1, Los Alamos National Laboratory, MS F600 Los Alamos, NM 87545

49. Prof. Dennis Cox, Department of Statistics, Rice University Houston, TX 77251-1892

50. Dr. Robert Easterling, Statistics, Computing and Human Factors Division Sandia National Laboratories P. O. Box 5800 Albuquerque, NM 87185

51. Prof. Sherwood Ebey, Department of Mathematics, University of the South Sewanee, TN 37375

52. Dr. David Hall, Statistics, Systems Department, Pacific Northwest Laboratory, P. O. Box 999, Richland, WA 99352

53. Prof. J. A. Hartigan, Department of Statistics, Yale University Box 2179 - Yale Station New Haven, CT 06520

54. Dr. Dan Hitchcock, Office of Scientific Computing, ER-7, Applied Mathematical Sciences Office of Energy Research U. S. Department of Energy Washington, DC 20585

55. Dr. Fred Howes, Office of Scientific Computing, ER-7, Applied Mathematical Sciences Office of Energy Research, U. S. Department of Energy Washington, DC 20585

56. Prof. Mark Johnson, Department of Statistics, University of Central Florida, Orlando, FL 32816-037057

57. Dr. A. M. Liebetrau, Computational Sciences Department Battelle-Northwest, P. O. Box 999, Richland, WA 99352

58. Dr. J. Liljegren, Pacific Northwest Laboratories, P. O. Box 999 Richland, WA 99352
59. Dr. Michael McKay, Statistics Group A1, Los Alamos National Laboratory, MS F600, Los Alamos, NM 87545
60. Prof. Lisa Moore, Institute of Statistics and Decision Sciences Duke University, Durham, NC 27706
61. Dr. David Nelson, Director of Scientific Computing, ER-7, Applied Mathematical Sciences, Office of Energy Research, U. S. Department of Energy, Washington, DC 20585
62. Dr. David J. Pack, CSR Institute, 8889 Bourgade AV, Lenexa, KS 66219
63. Mr. Brent Pulsipher, Computational Sciences Department, Battelle Northwest, P. O. Box 999, K1-86, Richland, WA 99352
64. Dr. Jerome Sacks, NISS, P. O. Box 14162, Research Triangle Park, NC 27709-4162
65. Prof. A. F. M. Smith, Department of Mathematics, University of Nottingham, University Park, Nottingham NG7 2RD, England
66. Dr. Alan Solomon, P. O. Box 227, Omer 84965, Israel
67. Dr. Daniel L. Solomon, Department of Statistics, North Carolina State University, P. O. Box 5457, Raleigh, NC 27650
68. Prof. William Welch, Department of Statistics and Actuarial Science University of Waterloo Waterloo, Ontario N2L 3G1, Canada
69. Prof. Y. C. Yao, Department of Statistics, Colorado State University Fort Collins, CO 80523
70. Prof. Don Ylvisaker, Department of Mathematics, University of California, Los Angeles, CA 90024
71. Office of Assistant Manager for Energy Research and Development, Department of Energy, Oak Ridge Operations Office, P. O. Box 2001 Oak Ridge, TN 37831-8600
72. Office of Scientific and Technical Information, P. O. Box 62 Oak Ridge, TN 37830