Using Polynomial Smoothing and Data Bounding for the Detection of Adverse Process Changes in a Chemical Process

Paul R. Sebastian,* David E. Booth, and Michael Y. Hu

Graduate School of Management, College of Business Administration, Kent State University,
Kent, Ohio 44242

Received December 2, 1993[⊗]

This paper focuses upon the problem of detecting outliers in a time series used to model a production process in the chemical industry. Significant deviations from the underlying time series pattern, i.e. outliers, indicate an adverse process change or out-of-control situation relative to the model. The underlying process is modeled using either least squares moving polynomial fit smoothing based upon the Savitzky–Golay algorithm[21] or data bounding. This makes any outliers in the original data more salient when compared to the smoothed graph. Thus outliers can be detected earlier while the process output is still within standard control limits and product specifications. The proposed algorithms improve upon and complement the conventional control chart, particularly with interdependent observations. The process control capabilities of these methods were successfully tested on an autocorrelated data set taken from a chemical production process with known adverse process changes and assigned causes. These algorithms should be of assistance to the chemical engineer or industrial chemist involved in process and quality control.

A. INTRODUCTION

Researchers and practitioners in manufacturing and chemical processing have long been looking for ways to detect significant adverse process changes as they occur. The idea is to discover these adverse process changes while they are still relatively minor, before substandard product is produced. These out-of-control process situations may be recognized through the detection of statistical outliers.[3,17,24] For over 50 years the most common means of process control has been the conventional control chart.[14] In effect, it models the process as a time series, a sequential arrangement of observations. The conventional control chart (e.g. X-chart or $\overline{X}$-chart) is based upon the assumption that the observations are independent[2] and identically distributed (IID) about the targeted process mean (μ_0 or TV) at any point and that the distribution is normal when the process is in statistical control, i.e. stable conditions.[25] Independence implies that there is no particular pattern in the data.

Points outside of three standard deviations of the targeted process mean (μ_0 or TV) are usually considered to be outliers. If such exist, the process is said to be "out of control". In other words, there is a significant adverse process change. If the process is functioning as designed, i.e. output within certain control limits, the process is "in control".

Controlling the Process by Utilizing Time Series Models. In reality, the IID assumption of the conventional control chart often does not hold. In those cases there is an interdependence among the observations, i.e. the points are autocorrelated.[1,4]

Probably most process time series exhibit a characteristically repetitive pattern which can be mathematically modeled by a more complex autoregressive moving average [ARMA(p,q)] model.[7] For example, ARMA(1,1) and other time series models have been empirically found in some cases to be appropriate for modeling a process time series.[4,18] An outlier would then be any point that deviates significantly from the underlying process model or time series pattern, thus indicating an out-of-control situation with respect to the process model.

The goal is to discover outliers indicative of an adverse process change as early as possible while minimizing the frequency of false alarms. For example, in the case of a drifting process mean due to a gradually deteriorating part in the process, outliers would be identified while the output is still within conventional control chart limits, well before any substandard product is produced. Then the operator may make adjustments in a timely manner. Of course, a sudden malfunction in the process, such as the fracture of a part, may immediately yield poor quality product.

Objectives of this Research. With this introduction, the reader may better appreciate the objectives of this research: (1) to develop outlier detection methods that are relatively simple in concept, flexible, and adaptable to the process environment; (2) to develop techniques that can be used to modify the existing control chart methodology, leading to an earlier detection of outliers indicative of an adverse process change, while reducing the probability of a false alarm by using a more realistic process model.

B. APPLICATIONS TO INDUSTRIAL PROCESS CONTROL AND SOME PRIOR RESEARCH

Booth[4,26] used the generalized M estimator (GM) procedure of Denby and Martin[9] to fit a time series model to a set of observations from a stable production process. The resulting model is then used as the basis for detecting changes in the process.

The GM method models the process more accurately and gives information on the type of change, which is helpful in finding the cause of the problem, by identifying additive and innovative outliers. This is extremely useful information in determining how to return the production process to the in-control state. However, the GM procedure is limited to those processes that can be modeled as a pth-order autoregressive [AR(p)] time series. Booth and Isenhour[6] successfully used the GM method for the early detection of poisoning of a platinum catalyst used for the oxidation of ammonia to nitric acid. Booth[5] also used the method for the early discovery of nuclear material losses.

Prasad et al.[20] recently tested Chen and Liu's[8] joint estimation procedure, an extension of the GM idea with

[⊗] Abstract published in *Advance ACS Abstracts,* April 1, 1994.

0095-2338/94/1634-0881$04.50/0

satisfactory results on chemical process data sets with known outliers. It is more discriminating in identifying outlier types and can deal with any type of time series model, not only AR(p).

C. TWO NEW ALGORITHMS FOR THE DETECTION OF OUTLIERS

The previously mentioned outlier identification methods are sophisticated and complex. An ARMA type model must be identified, the parameters estimated, and complex mathematics used. True, most of this is done by computer, but some practitioners may be reluctant to accept, use, and have confidence in a method which they do not understand, despite the fact that they can be trained. Thus an excellent method may not be used.

We have developed outlier detection methods which also utilize the time series pattern to model the underlying process but are simple in concept and flexible. The concept is that forms of data bounding and smoothing filter out much of the random errors, noise, and outliers in the original data. What remains should be essentially the underlying time series pattern. The smoothed or adjusted curve represents the underlying model. In contrast to this adjusted graph, the outliers in the original data should be more obvious and thus easier to detect both quantitatively and qualitatively. Although especially appropriate for autocorrelated data sets, these methods can be used on any time series.

With the proper fine tuning of the parameters of the two recommended methods and consequently the sensitivity, the user can choose the method which is more suitable to the particular process conditions and environment.

First we shall describe the two outlier detection methods which have been computerized. Then the results of testing them on an industrial chemical data set will be shown. The methods are polynomial smoothing and data bounding for process control. The former is a commonly used smoother. Data bounding is essentially a means of adjusting data points that lie beyond certain limits, thus smoothing significant peaks and troughs.

In each algorithm, the concept of outliers is operationalized, i.e. defined so that it can be measured. Each method has its own independent nomenclature which shall be defined in subsequent sections, although N_{drop}, Z_3, and Z_4 are common to both of them. The parameters N_{drop}, Z_1, Z_2, Z_3, Z_4, m (the number of moving points), and the degree of polynomial fit are chosen by the user. These algorithms can also be used on standard deviation and range chart data when variability must also be controlled.

Modified Standard Deviation (MSD) and Modified Process Average (MPA). The following formula is common to the two algorithms:

$$\mathrm{MSD} = [\sum(X_t - \mathrm{MPA})^2/(N - N_{\mathrm{drop}} - 1)]^{1/2} \qquad (1)$$

Outliers on either tail of the distribution across the entire time period may greatly inflate the standard deviation estimate because of the squared term. To obtain a more realistic estimate, a number (N_{drop}) of extreme points on both tails are trimmed from the total number of points (N) in the data set for the MSD calculation. Therefore we use the term modified standard deviation, which is thus a trimmed estimator.

When possible, however, typical in-control process data (i.e. a training set) should be used for the MSD calculation and thus no points would have to be trimmed. Then, as the method is used, this estimate may be employed directly if the standard deviation can be assumed to be relatively constant. However, it is not always the case that such training sets are available.

N_{drop} is chosen by the user according to the estimated variability in the data and the desired sensitivity to outliers. When the variability is high, more outliers are expected. Since increasing N_{drop} decreases MSD, which in turn narrows the control limits in the algorithms, the number of points dropped is one of the ways for controlling the sensitivity of each of the two suggested outlier detection methods. However, dropping points will affect the estimate of the standard deviation. If MSD is large, some outliers may be undetected. On the other hand, if too many points are dropped, MSD will be small and the method will be oversensitive (i.e. cause excessive false alarms). As a typical starting point, 10% of the observations are trimmed (5% on each tail).

The modified process average (MPA) is a simple mean with the same points trimmed as in the MSD calculation, of which it is part. The center line of our modified control chart is set equal to the expected process mean as planned or designed, i.e. a fixed target value (TV). If no information is available regarding a specific target value, then MPA would be the center line.

Both computer programs (the implementations we developed for the two recommended methods) identify and count the number of outliers. The programs indicate with a "*+" next to the identified outliers which values are beyond the planned process mean or target value TV ± Z_3(MSD), where $Z_3 = 3$ is the default value. This is based upon the observation that approximately 99.7% of the points of a normal distribution lie within three standard deviations of the mean. This range is often referred to as the process capability. Clearly, Z_3 and other parameters should be adjusted as the desired probabilities of a false alarm (α) and of overlooking an outlier (β) change.

In the case where the standard deviation is already known from a previous data set taken when the process was in control, MSD may be set equal to that value. Then, if $Z_3 = 3$ is selected, a "*+" indicates that the conventional control chart limits [TV ± 3(MSD)] are also violated, where MSD would be the in-control standard deviation. Thus the user is able to compare the results with the traditional control chart.

The computer program also provides Z_x and Z_y, the number of MSDs away from TV for each raw data point and smoothed value, respectively. This feature should help the user to fine-time the sensitivity and judge whether any suspicion of an undetected outlier or false alarm is true.

Least Squares Moving Polynomial Fit Smoothing—Using the Savitzky–Golay Approach for Process Control. The smoothing method we employed is used extensively in chemical spectroscopy to filter random errors, noise, and outliers. Maximum likelihood smoothers are also used for this purpose.[28,29] Savitzky and Golay[21] initially devised and Steinier et al.[22] and Madden[15] later improved 11 tables of convolution coefficients for calculating the smoothed data values plus the first five derivatives.

This approach gives the user a wide variety for choices for the degree of polynomial fit (quadratic, cubic, quartic, and quintic) as well as the odd number of moving points ($m = 2s + 1$) at equal time intervals. The variable s is the number of observations on each side of the central point and varies from 2 to 12 in the tables.

The least squares polynomial fit is made across the first m points. Only the central point ($X_{m-s} = X_t$) is adjusted as Y_t. The central point then moves to the next observation at time $t + 1$ and the process is repeated.

The method leaves unsmoothed the first s points and the final s points of the time series. They are dealt with by using either the conventional derivative end point manipulation procedure[32] or the linear regression end point manipulation procedure to smooth the end points not covered by the polynomial smoothing algorithm. The latter procedure consists of a simple linear regression over a minimum of five points, but only the previously unsmoothed s end points are actually adjusted. The resulting smoothed time series is a model of the production process. Once the smoothing has been completed, the following procedure is used.

Outlier Detection Procedure for Process Control. This is the part of our method that operationalizes the concepts of outliers and out-of-control processes. The importance of this approach is that it can be used with any smoothing technique if properties other than those of the Savitzky–Golay method are desired. In other words, this algorithm identifies outliers and thus an adverse process change in a more consistent and objective manner. Polynomial smoothing and other techniques smooth or adjust data but do not in themselves identify outliers.

(1) Primary Operationalization Condition. X_t is operationalized as an outlier if there is a marked deviation [Z_2(MSD)] between it and the adjusted time seies or model (i.e. Y_t) in the appropriate direction. In other words, if X_t is part of a peak above the center line (TV), the outlier criterion is $X_t > Y_t + Z_2$(MSD). If X_t is part of a trough below TV, the outlier criterion is $X_t < Y_t - Z_2$(MSD). Z_2 is chosen according to the desired sensitivity, where $Z_2 = 0.5$ is the usual starting point.

It may happen that Y_t and other parts of the adjusted graph or model do not accurately represent the underlying process. Then one may miss bona fide outliers (type II error) and falsely identify others (type I error). Obviously, no method can specify a model that gives an exact representation of the process. As George Box has said, "All models are wrong, but some are useful".[27] To obtain greater robustness with respect to model specification, two other criteria are introduced.

(2) Secondary Operationalization Condition. An outlier is also operationalized if X_t is beyond the limits, TV ± Z_3(MSD). $Z_3 = 3$ is the usual starting point. This criterion increases sensitivity and decreases β, the probability of overlooking an outlier. Thus a bona fide outlier may still be detected if an observation does not meet the primary operationalization condition.

The TV ± Z_3(MSD) limits, rather similar to those of a conventional control chart but often tighter, are based on the overall distribution. If the distribution is approximately normal when the process is in control, the probability that X_t is beyond TV ± 3(MSD)—yet by chance part of the original distribution—would be less than 0.01. This secondary operationalization condition should be able to eventually detect a trend if previous outliers have been overlooked. When a false alarm is very costly, Z_3 may be increased to a high value such as 3.5 or 4.

(3) Screening and Elimination Condition. To decrease α, the probability of a false alarm, an observation (X_t) is ruled out as a possible outlier if it is within TV ± Z_4(MSD), since it is close to the center line (TV). This condition thus may negate the primary operationalization for some points. Z_4 is chosen according to the desired sensitivity, where $Z_4 = 1$ is the usual starting point.

Data Bounding for Process Control. This algorithm is based upon the deviation of each data point (X_t) from the average ($\bar{X}_i$) of its adjacent points, where $i = 1$ refers to the past two adjusted points and the next two raw data points and $i = 2$ refers to the average of the proceding three adjusted points. If the data point (X_t) is beyond $\bar{X}_1 \pm Z_1$(MSD) OR $\bar{X}_2 \pm Z_1$(MSD), an *adjustment* may be made. Ideally, the points should be at equal time intervals.

The adjusted or smoothed point (Y_t) is the original point (X_t) pulled downward, if it is part of a peak above the center line (TV), or upward, if part of a trough below TV, by an amount equal to Z_1(MSD). Z_1 is chosen according to the desired sensitivity, where $Z_1 = 1$ is the usual starting point. $Z_1 < 1$ may be appropriate when overlooking an outlier (type II error) is crucial and some false alarms (type I errors) are tolerable. The adjusted time series represents the underlying chemical process model.

(1) The Primary Operationalization Condition is the same as the adjustment criterion. An outlier and thus an out-of-control situation is operationalized as a data point (X_t) which is outside of the average of the four adjacent points OR the average of the three preceding adjusted points, $\bar{X}_i \pm Z_1$(MSD), in the appropriate direction as in the preceding paragraph. The second part of the OR condition handles the final end point and increases sensitivity, especially to consecutive outliers. Two more criteria are added to give greater robustness in case Y_t and other parts of the adjusted graph or model do not accurately represent the underlying process.

(2) Secondary Operationalization Condition. Alternatively, an outlier is identified if the data point is outside of the limits associated with the target value TV ± Z_3(MSD), where $Z_3 = 3$ is the usual starting point. This alternate condition increases sensitivity to outliers not detected under the primary condition, thus decreasing β. It should be able to eventually detect a trend if previous outliers have been overlooked. When a false alarm is very costly, Z_3 may be set equal to a high value such as 3.5 or 4.

(3) Screening and Elimination Condition. At the same time, it is very unlikely that X_t would be an outlier if it is near the center line (TV), i.e. within TV ± Z_4(MSD), where $Z_4 = 1$ is the usual starting point. This criterion decreases α, the false alarm probability.

Fine Tuning the Sensitivity. Although reasonable default values for the parameters are already set in the algorithms, further sensitivity adjustments are often necessary, depending upon the process conditions, organizational needs, product specifications, and the relative costs of a false alarm vs overlooking an outlier. A few principles and guidelines should be helpful in this empirical procedure.

When the data have considerable variability and thus are more likely to contain outliers, MSD is relatively large and the Z_i(MSD) limits may be so wide that some outliers could remain undetected because of possible increases in β and insensitivity. In contrast, when the variability and consequently the standard deviation is small, the algorithm may be oversensitive if any points are trimmed, giving a greater frequency of false alarms, i.e. an increase in α. The percentage of points trimmed (%drop) must be adjusted accordingly.

If it is very costly to pass poor quality product at any stage in the process, high sensitivity (low β) is essential. Then management may tolerate a greater risk (α) or frequency of false alarms. In general, sensitivity would be increased if the cost of not detecting an outlier is high in comparison to the cost of a false alarm.

The user can increase the sensitivity in up to four different ways, depending upon the algorithm employed. Table 1 summarizes them for each algorithm.

Table 1. Increasing the Sensitivity of Each Algorithm

algorithm	means of increasing sensitivity
data bounding	(1) decreasing Z_1 narrows $\bar{X}_i \pm Z_1$(MSD) (2) decreasing Z_3 narrows TV ± Z_3(MSD) (3) decreasing Z_4 narrows TV ± Z_4(MSD) (4) increasing %drop decreases MSD above
polynomial smoothing	(1) decreasing Z_2 narrows $Y_t \pm Z_2$(MSD) (2) decreasing Z_3 narrows TV ± Z_3(MSD) (3) decreasing Z_4 narrows TV ± Z_4(MSD) (4) increasing %drop decreases MSD above

One way to increase sensitivity is to increase the percentage of points trimmed (%drop) and thus N_{drop}, which decreases MSD and consequently the different X_i, TV, or $Y_t \pm Z_j$(MSD) limits. Decreasing Z_1, Z_2, Z_3, and Z_4 similarly decreases four different control limits for data bounding adjustments and outlier detection in both algorithms.

If the product tolerance or specification limits are much wider than the estimated process capability [TV ± 3(MSD)], a lower sensitivity may be used to reduce the number of false alarms. The approximately standardized equivalents of the original data (Z_x corresponds to X_i) and also the smoothed data points (Z_y to Y_t) may help the user judge whether the fine tunings of the parameters are reasonable and evaluate suspicions of a false alarm or an overlooked outlier.

After the modified standard deviation (MSD) is established, the other parameters should be initialized by using, if available, a data set with known outliers due to assigned causes from the process to be controlled, that is, a previous data set among which a few of the measures were taken when the process was out of control due to a problem that was later solved. This may serve as a training set.

"A Guide for Parameter Estimation If Information Is Limited", the computer program, and other details of implementation are available from the first author. The ideal is to have training sets, i.e. previous data sets for the process—one which is in control (functioning as designed) to obtain an accurate MSD and another which has known outliers. However, according to Alwan and Roberts,[1] it is difficult to ascertain when the process is truely in control due to, for example, fluctuations in the process mean. Thus having a training set is very helpful but not indispensable.

Presently, we are developing *a computerized search program* to find the best combination of parameters to fit a data set with known outliers. Basically, the program is a systematic empirical search as the algorithm is iteratively run for each combination of parameters according to the range and the interval between trials that the user chooses for each parameter.

For each parameter combination, the computer implementation provides the total number of outliers identified (TOI), the number of actual outliers identified with assigned causes (NAI), the number of known outliers missed with assigned causes (NOM, i.e. the number of known outliers less NAI), and the apparent number of false alarms (NFA, i.e. TOI − NAI). Of course, NAI should be maximized while NOM and NFA should be minimized according to the user's discretion.

The data bounding algorithm parameters include Z_1, Z_3, Z_4, and %drop. Polynomial smoothing includes Z_2, Z_3, Z_4, %drop, the degree of fit, and the number of moving points. To decrease computer time, the user first chooses a wide range and interval or step for each parameter. Then a narrower range and interval are chosen for a finer tuning. Another strategy is to simply concentrate on the principal dimension, Z_1 or Z_2. At the same time, the user should use his/her intuition based upon experience as well as the relative costs of the type I (NFA) and type II (NOM) errors, desired sensitivity, process conditions, etc.

In essence, the fine tuning of the parameters, and consequently the sensitivity, is an empirically based procedure which also depends upon the reasonable judgment and experience of the user as well as any constraints such as cost. Previously mentioned guidelines, training sets, and any computerized search program can facilitate but cannot determine the optimum set of parameters. The default values and starting points we have given are reasonable for many if not most data sets. The user may then revise the parameters if necessary as s/he gains more experience in applying the outlier detection method to the particular process and its unique set of characteristics.

Once the parameters Z_1, Z_2, Z_3, Z_4, %drop, and MSD are established for the problem environment as appropriate for the particular algorithm (see Table 1), they should be confirmed for consistency with another data set or two from the same production process. Then the simplicity of the method as well as the concept becomes more apparent.

The parameters should then be reasonably constant unless there is a marked permanent change in the process. The process and thus the model may evolve over time—corrosion, wear, maintenance, minor process changes, a different source of raw material, etc. Therefore, periodic retuning of the sensitivity or recalibration of the parameters may be necessary.

A similar fine tuning should be used to find the best general polynomial smoothing fit for the process output time series. That is the best number of moving points and the degree of polynomial fit.

Handling Questionable Outliers. When a costly false alarm is suspected for an outlier identified within TV ± Z_3(MSD), the operator should look for an assigned cause. If none is found, then a false alarm is more likely.

If the specification limits are very wide or a small amount of substandard product is tolerable in case the process is indeed out of control, the suspected data point may be discounted as an outlier until it is confirmed with subsequent observations. Previous experience may also be a guide.

In any event, this questionable outlier could serve as an early warning that there may be an adverse process change or rudimentary problem that could worsen later. Management should be alert, and the operator should perform at least a cursory examination of the critical points in the process. The operator should also look for a trend and monitor the process more closely even though the process output is still well within conventional control chart limits and product specifications.

The supplementary control chart rules may be useful in resolving doubts regarding a possible false alarm, such as two out of three consecutive observations being beyond two MSDs.[10] False alarms can be eliminated, but at the cost of less sensitivity and a much greater risk (β) of overlooking some outliers.

D. TESTING THE ALGORITHMS ON A CHEMICAL PROCESS

The data set was taken from Grant and Leavenworth[12] to test our proposed outlier detection methods. The outliers are known, i.e. adverse process changes and their assigned causes. An $\bar{X}$-chart was used to control the acidity of a dyeing solution. The pH depends not only on the reagents added to the dye liquor but also on the organic contents of the particular variety or blend of wool being dyed. For best quality dyeing, the pH of the liquor must be controlled about the target value of 4.22.

Table 2. Least Squares Moving Polynomial Fit Smoothing: Detection of Outliers among Sample Averages of the Dye Liquor pH of Five Hussong Kettles ($\bar{X}_t$)[a,b]

period	pH	quad-5, $Z_2 = 0.5$, $Z_4 = 1.1$	quad-5, $Z_2 = 0.4$, $Z_4 = 1.1$	quad-7, $Z_2 = 0.5$, $Z_4 = 1.1$	quart-7, $Z_2 = 0.5$, $Z_4 = 1.1$	assigned cause
9	4.54	*+1	*+1	*+1	*+1	different wools
10	4.50					different wools
11	4.54	*+0	*+0	*+0	*+0	different wools
12	4.61	*+0	*+0	*+0	*+0	different wools
13	4.63	*+0	*+0	*+0	*+0	different wools
14	4.61	*+1	*+1	*+0	*+1	different wools
15	4.37					different wools
16	4.54	*+1	*+1	*+1	*+1	different wools
18	4.35		*1		*1	none or unknown
26	4.10		*2	*2		none or unknown
32	4.07		*2	*2		none or unknown
35	3.72	* + 2	* + 2	* + 2	* + 2	not neutralized
total (TOI)		7	10	9	8	
actual (NAI)		7	7	7	7	
false (NFA)		0	3	2	1	
missed (NOM)		2	2	2	2	

[a] * = detection of outlier within TV ± 3(MSD). *+ = detection of outlier beyond TV ± 3(MSD). Min = 3.72. Max = 4.63. Target value (TV) = 4.22. Modified process average (MPA) = 4.253. Median = 4.255. Modified standard deviation (MSD) = ±0.099. %drop = 20% (10). Z_3 = 3. N = 46. $-m$ = number of moving points. Total number of known outliers with assigned causes = 9. [b] 1 = outlier first detected as part of a significant peak above Y_t and TV. 2 = outlier first detected as a significant trough below Y_t and TV. 0 = outlier found only because of the TV ± 3(MSD) limits and does not meet the criteria of either 1 or 2.

An inspector obtained a sample from each of five Hussong kettles generally twice a day over a 23-day period. The sample average and the range were determined for each sample. We are using both of the methods described in this paper to detect outliers in the resulting time series based on the sample averages.

We used the 46 sample averages to calculate the modified process average (MPA) and the modified standard deviation (MSD) of the dye liquor pH data. The target value of 4.22 is the center line. It was found earlier that the series was autocorrelated according to an AR(1) model[4,20,26] and hence the use of standard control chart methods is suspect because of violations of the IID requirement.

Results. By applying polynomial smoothing to the dye liquor pH data, a quadratic equation fit with $m = 5$ moving points was used. The parameters of the outlier detection procedure were fine-tuned to $Z_2 = 0.5$, $Z_3 = 3$, and $Z_4 = 1.1$ with 20% or 10 of the points trimmed for the MSD calculation. As seen in the first column of Table 2 and Figure 1, outliers were identified at observations 9, 11, 12, 13, 14, 16, and 35.

Using data bounding on the pH ($\bar{X}_t$) time series, we fine-tuned the sensitivity to $Z_1 = 1$, $Z_3 = 3$, and $Z_4 = 1$ with 10% or four of the points dropped for the MSD calculation. Outliers were identified at points 9, 10, 11, 12, 13, 14, 16, and 35 as shown in the first column of Table 3 and Figure 2.

Analysis. A new blend of different wools caused the acidity to drop significantly during periods 9–16. At point 35, two batches of improperly neutralized carbonized (baked with concentrated H_2SO_4) stock caused the pH to decrease. After the necessary fine tuning, *data bounding* detected all of these outliers having assigned causes except point 15 as seen in the first column of Table 3.

Only points 13 (4.63) and 35 (3.72) would have been detected by a conventional control chart using our calculated limits of TV ± 3(MSD) = 4.22 ± 3(0.133). Grant and Leavenworth[12] obtained the same results using TV ± $3s_{\bar{X}}$ = 4.22 ± 3(0.057), where the standard deviation is based on within sample variability. Our MSD gives between sample variability which is based upon the time series as advocated by Alwan and Roberts.[1] Neither our two proposed methods nor Grant and Leavenworth's control chart could detect point 15 as an outlier.

By increasing the sensitivity, i.e. decreasing Z_1 from 1 to 0.8, another outlier was found at point 32 as shown in column 2 of Table 3. This may be an early warning of an adverse process change which, although not serious at the time, could worsen later. An important outlier and out-of-control situation was indeed detected at point 35.

Polynomial smoothing detected the same known outliers as data bounding with the exception of point 10 as shown in Table 2. By the very nature of polynomial smoothing, the smoothed points and graph are very close to the original points in the case of consecutive outliers, thus making them difficult to detect. Therefore, the secondary operationalization of outliers was indispensable as a backup in the identification of points 11, 12, and 13. Thus polynomial smoothing still detected seven of the nine points with actual assigned causes. In any event, the crucial first point of the out-of-control run was identified.

By trimming 20% of the observations instead of the usual 10%, MSD decreased from ±0.133 to ±0.099, thus coming closer to the standard deviation calculated by Grant and Leavenworth (±0.057) which is based on the within sample variability. Consequently, the limits of the secondary operationalization condition were narrowed. The primary condition at times may miss one or more of a string of consecutive outliers. Furthermore, this 20% trim brought MPA from 4.263 (10% trim) to 4.253, which is closer to the median of 4.255 and the target value of 4.22.

In column 2 of Table 2, sensitivity was increased by decreasing Z_2 in quadratic fit smoothing ($m = 5$) from 0.5 in column 1 to 0.4. Thus the number of outliers increased to 10 with a greater frequency of false alarms, apparently points 18, 26, and 32.

Quadratic fit ($m = 7$) and quartic fit smoothing ($m = 7$) in columns 3 and 4 apparently did not give as good a fit as quadratic smoothing ($m = 5$).

It can be seen in Figure 1 that applying the conventional derivative end point manipulation procedure gives quartic equation fit smoothing some erratic and unreliable adjustments for the initial and final three end points. The adjustment is very high on both sides. The end point adjustments for quadratic smoothing are more reasonable. Thus the linear regression end point manipulation procedure should be

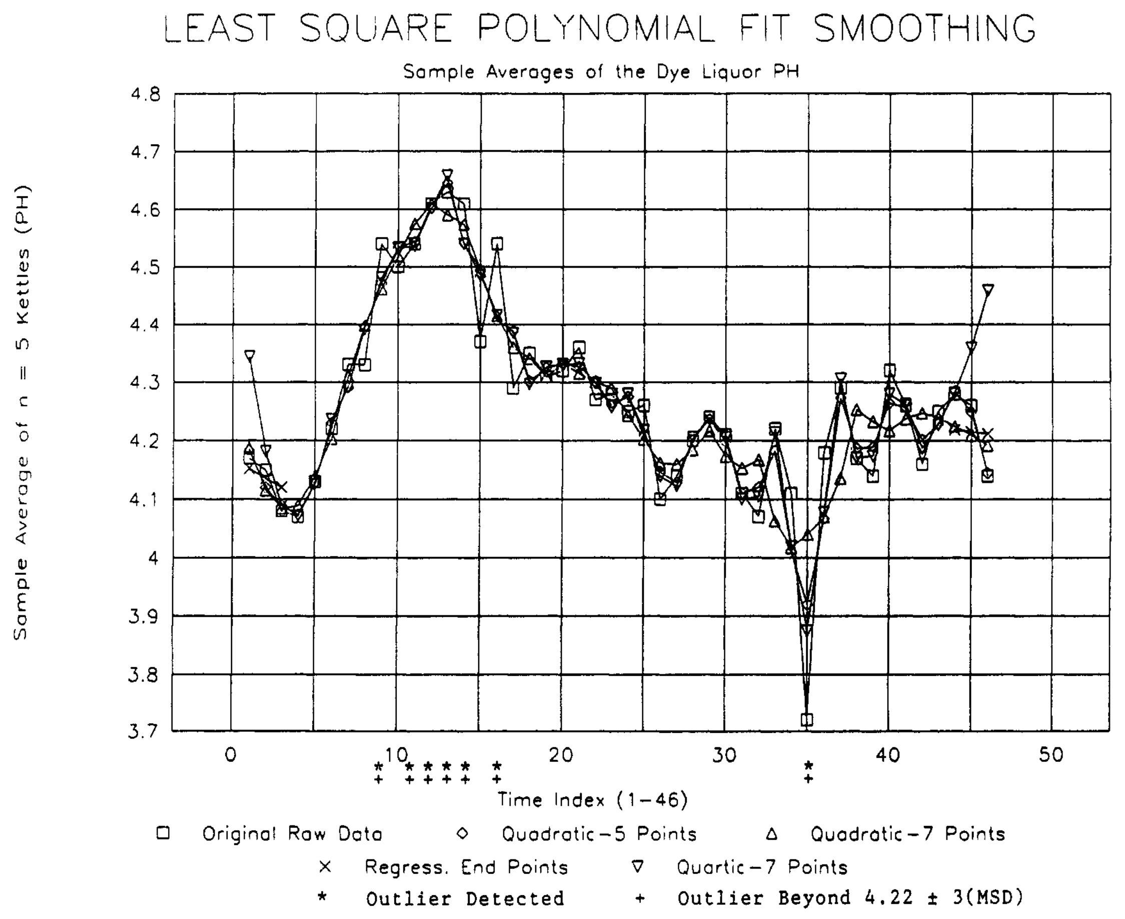

Figure 1.

Table 3. Data Bounding for Process Control: Detection of Outliers among Sample Averages of the Dye Liquor pH of Five Hussong Kettles $(\bar{X}_t)$[a,b]

period	pH	$Z_1 = 1, Z_4 = 1$	$Z_1 = 0.8, Z_4 = 1$	$Z_1 = 1, Z_4 = 1.3$	$Z_1 = 0.8, Z_4 = 1$	assigned cause
9	4.54	*3	*3	*+3	*3	different wools
10	4.50	*2	*2	*2	*2	different wools
11	4.54	*2	*2	*+2	*2	different wools
12	4.61	*2	*2	*+2	*2	different wools
13	4.63	*+3	*+3	*+3	*3	different wools
14	4.61	*3	*3	*+3	*3	different wools
15	4.37					different wools
16	4.54	*1	*1	*+1	*1	different wools
32	4.07		*3	*3	*2	none or unknown
35	3.72	* + 3	* + 3	* + 3	* + 3	not neutralized
total (TOI)		8	9	9	9	
actual (NAI)		8	8	8	8	
false (NFA)		0	1	1	1	
missed (NOM)		1	1	1	1	
%drop		10% (4)	10% (4)	20% (10)	5% (2)	
MPA		4.263	4.263	4.253	4.267	
MSD		±0.133	±0.133	±0.099	±0.143	

[a] * = detection of outlier within TV ± 3(MSD). *+ = detection of outlier beyond TV ● 3 (MSD). Min = 3.72. Max = 4.63. Target value (TV) = 4.22. $Z_3 = 3$. $N = 46$. Median = 4.255. Total number of known outliers with assigned causes = 9. [b] 1 = outlier first identified using the average of the four adjacent points—the preceding two adjusted points and the next two raw data points. 2 = outlier first identified using the three preceding adjusted points. 3 = outlier first detected according to both 1 and 2. 0 = outlier found only because of the TV ± 3(MSD) limits and does not meet the criteria of either 1 or 2.

consulted, at least in the case of quartic smoothing. Note the straight line plot of the three end points by regression on each side.

Previously, we tested the two methods on a process in which a precisely controlled depth is cut into each shaft for assembly to a gear hub. The three known outliers in the data set with assigned causes were detected by both methods. There were no consecutive outliers as in the dye liquor pH data set. However there was one apparent false alarm.[31]

E. DISCUSSION

These methods are simple in concept and more flexible than the conventional control chart and the more sophisticated procedures, but at the expense of some subjectivity in certain cases.

The adjusted graph of *polynomial smoothing*—by virtue of its mathematics and certain parameter values—may at times come very close to a true outlier, especially if two or more outliers are consecutive or part of a trend. With such

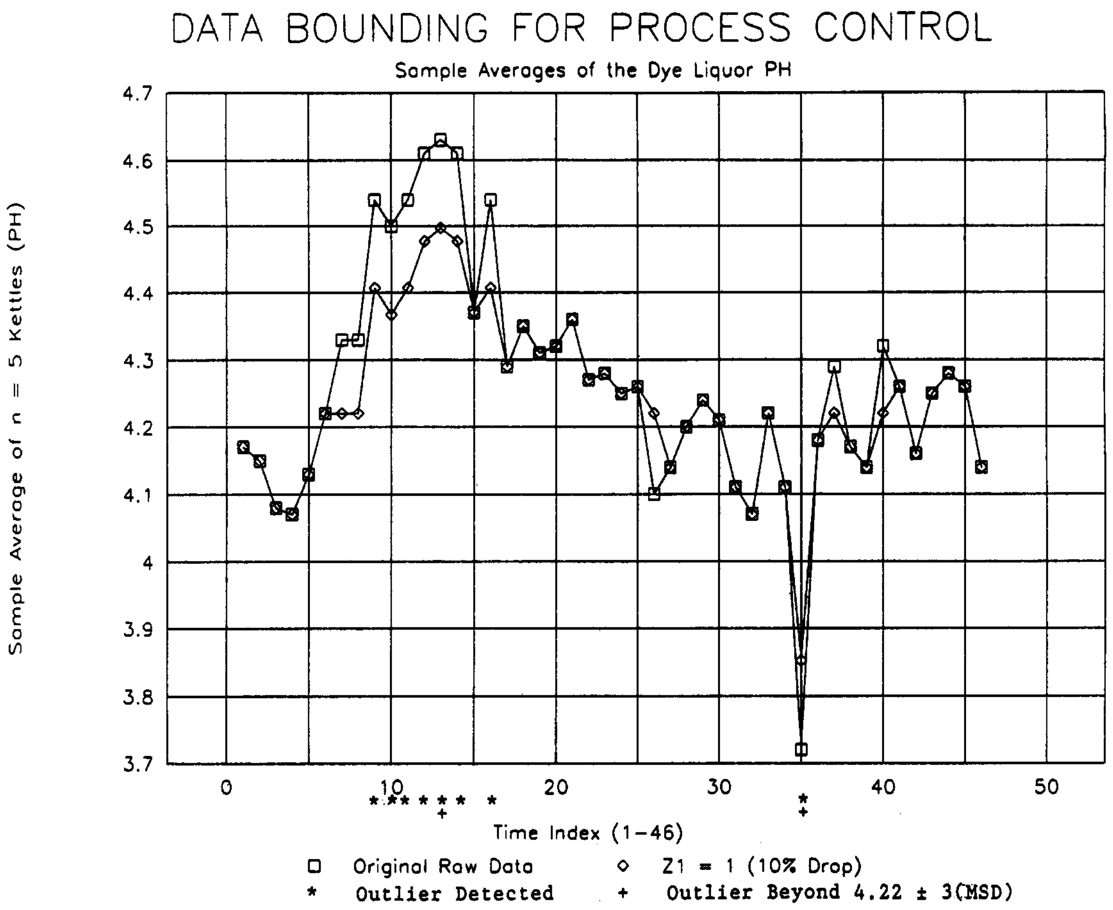

Figure 2.

a small deviation from the adjusted curve, an outlier would be virtually obscured from detection unless it is outside the TV ± Z_3(MSD) limits.

This suggests that other more outlier resistant smoothing methods—e.g. EDA Smoothers[23] or a LOWESS Smooth[13,30]—may be helpful. Such nonparametric smoothers give larger residuals between the outliers and the smoothed data which represents the underlying process model. This makes outliers easier to detect.

It should be noted that the last $s = (m - 1)/2$ end points are not adjusted by polynomial smoothing but by an adaptation which uses either simple regression or derivatives.

Data bounding may miss an outlier if there are two low points on one side of an outlier X_t and two high points (also outliers) on the other side as part of a trend, for example; the average of the four adjacent points ($\bar{X}_1$) could be almost equal to X_t, giving a very small deviation. Thus there would be no adjustment and, consequently, no outlier detection.

Similarly, the method may not make an outlier identification if the adjacent points are rather close to X_t, as in the case of actual outliers which are consecutive. To at least partially compensate for these two cases, the "three preceding adjusted points" clause (i.e. the OR condition) was added to the primary operationalization. Of course, the secondary operationalization condition [TV ± Z_3(MSD)] should be able to eventually detect any trend.

A misrepresentation of the underlying model is less likely with data bounding, which is more resistant to outliers. Polynomial smoothing does better in filtering out the random noise and may more closely model the process in some environments.

Robustness. We already mentioned that the secondary operationalization and the screening conditions add robustness in case parts of the adjusted graph or model do not accurately represent the underlying process. Furthermore, the fact that a certain number of extreme points are trimmed in calculating MSD should give our outlier detection methods a certain robustness with regard to any normality assumption that may have been necessary. Range data, for example, are often far from being normally distributed.

Since most if not all of the outliers which skew the distribution are trimmed for the calculations, the points that remain will better approximate normality. Then, for example, TV ± 3(MSD) will come closer to including 99.73% of the remaining data points as is characteristic of the normal distribution. Thus we have more justification for using Z values and have a better intuitive sense in choosing the Z_j parameters.

Of course, if the trimmed points and corresponding lower MSD cause a greater number of false alarms, the parameters must be adjusted accordingly. This would be the case, for example, if the underlying distribution were χ^2.

Earlier Outlier Detection. Because the primary operationalization condition in our case and some other time series based approaches[19,20] are able to detect outliers well within three standard deviations of the process mean, these methods often effectively result in narrower (i.e. tighter) control limits than those usually used for the conventional control chart. Outliers and the corresponding adverse process change or out-of-control process situation can thus be detected earlier while the process output is still within standard control chart limits and product specifications. For the same reason, the methods are especially useful for products with very strict and narrow specification limits which barely exceed process capability.

These methods are especially appropriate for autocorrelated data sets, i.e. interdependent observations, although the particular ARMA model need not be specified. However, they are not restricted to any specific set of time series models. Even if the data are independent, the algorithms can often detect the initial stages of an adverse process change earlier than a conventional control chart. Data bounding, for example, should be very effective in quickly detecting a drifting process mean, even though the particular sample happens to be within conventional control chart limits.

Real Time On-Line Process Control Capability. The results may give the impression of applying only to retrospective control. Since the algorithms are designed to handle end points, they also apply to real time on-line process control, i.e. as soon as the latest measurement is made.

This is illustrated by an additional run. We again tested both methods, but only on the first 35 data points. Thus observation 35, where the stock was improperly neutralized, became the current final observation. Both methods detected this final point as an outlier in all cases, thus demonstrating their real time on-line process control capability. The polynomial smoothing method identified this outlier in all cases by utilizing both the secondary operationalization condition and the linear regression end point procedure.

Furthermore, in data bounding, the reader may note in Table 3 that any outlier identified with a +*, *2, or *3 notation would have been detected if it were a final point. All previously identified outliers with assigned causes met at least one of the criteria except for point 16.

By identifying outliers earlier while the observations are still within standard control chart limits, we are in essence forecasting that the output will eventually lie beyond conventional limits as well as product specifications unless remedial measures are taken. The user may be able to draw further inferences from the graphs by observing the residuals ($y - \hat{y}$) between the raw data and the corresponding adjusted points or model. Certainly, increasing residuals would indicate a trend.

At worst, we obtain the same results as a control chart if greater reliance is placed upon the secondary operationalization condition. At best, we can detect adverse process changes earlier through the primary outlier operationalization condition, although we must be careful to see that the frequency of false alarms does not increase appreciably.

In other words, our methods are based upon the control chart but have added features which utilize the fact that many data sets are autocorrelated. Thus the practitioner, who is accustomed to the control chart, can more easily understand, accept, and identify with our methods.

Flexibility and Adaptability to the Process Environment. The two outlier identification methods have a quantitative base but yet have qualitative and graphical aspects to accommodate the needs and limitations of operating managers. Many of them may not accept a sophisticated method which they do not understand and thus would prefer a simpler one which may utilize their own experience and judgment.

Because the methods are both graphical and quantitative, the operating manager has the option of using his/her own analysis of the graphical output (original data vs smoothed or adjusted), personal judgment, and experience in conjunction with the computerized results. The confirmation and flexibility are especially important when a false alarm or overlooking an out-of-control situation (i.e. an outlier) is suspected.

Furthermore, these conceptually simple and intuitive algorithms allow the user to vary up to five parameters according to the desired sensitivity, the cost of overlooking an outlier vs the cost of a false alarm, process conditions, organizational needs, etc. With the proper fine tuning of the parameters, the typical user should be able to choose the outlier detection method which is better suited to the process environment.

If the user should choose to rely more on the secondary operationalization to conclude "out of control", s/he may use the primary operationalization as an early warning or incipient process change to be investigated for a possible problem. If left alone, a drifting process mean will later activate the secondary operationalization when the process is obviously out of control. If the primary operationalization (based upon the time series model) misses an outlier, the secondary operationalization (based upon the distribution) acts as a backup.

If an outlier is found by both the primary and secondary operationalization conditions, the practitioner can feel more assured that X_t is indeed an outlier and not a false alarm.

Summary. We have seen that the two methods used in this paper are effective but simple in concept, flexible, and adaptable to the process environment. Thus they may be more acceptable to practitioners than the more sophisticated methods mentioned in section B. The approach is clearly useful in a variety of applications, particularly the case of autocorrelated data and most likely with smoothers such as LOWESS to severely non-normal distribution situations. The exact conditions are being determined in current research.

The most important feature of these methods, however, is the ability to identify outliers, i.e. adverse process changes, earlier than the traditional control chart, well within the usual three standard deviations from the target value. This feature is crucial when the process mean is drifting but the output is still within product specifications and conventional control chart limits. In another case, the process may be really out of control, but the sample taken happens to be within the usual control chart limits. In both cases, the standard control chart may not detect the adverse process change until substandard product is already produced.

Thus we propose that these methods are a useful addition to those already available to industrial chemists or chemical engineers involved in process and quality control as they confront the challenges of global competition in a very interdependent and more integrated world economy.

ACKNOWLEDGMENT

We thank the reviewers for their most helpful suggestions.

REFERENCES AND NOTES

(1) Alwan, L. C.; Roberts, H. V. Time Series Modeling for Statistical Process Control. *J. Bus. Econ. Stat.* **1988**, *6* (1), 87–95.

(2) Anderson, R. L. *Practical Statistics for Analytic Chemists*; Van Nostrand Reinhold: New York, 1987; p 24.

(3) Beckman, R. J.; Cook, R. D. Outlier..........s. *Technometrics* **1983**, *25* (2), 119–149.

(4) Booth, D. E. Some Applications of Robust Statistical Methods to Analytical Chemistry. Ph.D. Dissertation, University of North Carolina at Chapel Hill, 1984.

(5) Booth, D. E. A Method for Early Identification of Loss from a Nuclear Material Inventory. *J. Chem. Inf. Comput. Sci.* **1986**, *26* (1), 23–28.

(6) Booth, D. E.; Isenhour, T. L. A Method for Early Discovery of Poisoning in Catalytic Chemical Processes. *J. Chem. Inf. Comput. Sci.* **1985**, *25* (2), 81–84.

(7) Box, G. E. P.; Jenkins, G. M. *Time Series Analysis Forecasting and Control*, 2nd ed.; Holden-Day: San Francisco, CA, 1976.

(8) Chen, C.; Liu, L. M. Joint Estimation of Model Parameters and Outlier Effects in Time Series. *J. Am. Stat. Assoc.* **1993**, *88*, 284–297.

(9) Denby, L.; Martin, R. D. Estimation of the First Order Autoregressive Parameter. *J. Am. Stat. Assoc.* **1979**, *74*, 140–146.

(10) Duncan, A. J. *Quality Control and Industrial Statistics*, 5th ed.; Irwin: Homewood, IL, 1986; p 434.

(11) *Executat*, Student Edition; PWS-Kent: Boston, 1991; p 74.

(12) Grant, E. L.; Leavenworth, R. S. *Statistical Quality Control*, 5th ed.; McGraw-Hill: New York, 1980; p 94.

(13) Hardle, W. *Applied Non-Parametric Regression*; Cambridge University Press: New York, 1990; p 190.

(14) Kateman, G.; Pijpers, F. W. *Quality Control in Analytical Chemistry*; Wiley: New York, 1981; p 107.

(15) Madden, H. H. Comments on the Savitzky–Golay Convolution Method for Least-Squares Fit Smoothing and Differentiation of Digital Data. *Anal. Chem.* **1978**, *50*, 1383–1386.

(16) Makridakis, S.; Wheelwright, S. C.; McGee, V. E. *Forecasting: Methods and Applications*; Wiley: New York, 1983; p 380.

(17) Miller, J. C.; Miller, J. N. *Statistics for Analytical Chemistry*; Halsted: New York, 1984; p 59.

(18) Prasad, S. A Robust Monitoring Method for the Early Detection of Deviations in a Time Series Process with Applications in Operations Management. Unpublished Ph.D. Dissertation, Kent State University, Kent, OH, 1990; pp 91, 129.

(19) Prasad, S.; Booth, D.; Hu, M.; Deligonul, S. The Detection of Nuclear Material Losses by Means of a Robust Time Series Method. Submitted to *Decis. Sci.*, 1993.

(20) Prasad, S.; Booth, D.; Hu, M.; Deligonul, S. Monitoring the Quality of a Production Process Using a Robust Time Series Method. Submitted to the *Int. J. Qual. Reliab.*

(21) Savitzky, A.; Golay, M. J. E. Smoothing 2nd Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.* **1964**, *36*, 1627–1639.

(22) Steinier, J.; Termonia, Y.; Deltour, J. Comments on Smoothing and Differentiation of Data by Simplified Least Square Procedures. *Anal. Chem.* **1972**, *44*, 1906–1909.

(23) Velleman, P. F.; Hoaglin, D. C. *Applications, Basics and Computing of Exploratory Data Analysis*; Duxbury Press: Boston, MA, 1981; p 159.

(24) Youden, W. J.; Steiner, E. H. *Statistical Manual of the Association of Official Analytical Chemists*; AOAC: Washington, 1975; p 30.

(25) Wardell, D. G.; Moskowitz, H.; Plante, R. D. Control Charts in the Presence of Data Correlation. *Manage. Sci.* **1992**, *38*, 1084–1105.

(26) Booth, D. E.; Acar, W.; Isenhour, T. L.; Ahkam, S. Robust Time Series Models and Statistical Process Control. *Ind. Math.* **1990**, *40* (Part 1), 73–91.

(27) Box, G. E. P. Robustness in the Strategy of Scientific Model Building. In *Robustness in Statistics*; Launer, R. L., Wilkinson, G. N., Eds.; Academic Press: New York, 1979; pp 201–236.

(28) DeNoyer, L. K.; Dodd, J. G. Maximum Likelihood Smoothing of Noisy Data. *Am. Lab.* **1990**, *22* (3), 21–27.

(29) DeNoyer, L. K.; Dodd, J. G. Maximum Likelihood Deconvolution for Spectroscopy and Chromatography. *Am. Lab.* **1991**, *23* (12), 24D–24H.

(30) Dusoir, A. Personal communication, 1993.

(31) Sebastian, P. R.; Booth, D. E.; Hu, M. Y. Polynomial Smoothing and Data Bounding Applied to Industrial Process Control. *Proceedings of the 1993 Annual Meeting*; Decision Sciences Institute: Atlanta, GA, 1993.

(32) Heilborn, J., Ed. *Science and Engineering Programs, Apple II Edition*; Osborne/McGraw-Hill: Berkeley, CA, 1981; p 42, 45.