Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computers and Chemical Engineering 28 (2004) 1635–1647
Computers & Chemical Engineering
www.elsevier.com/locate/compchemeng

On-line outlier detection and data cleaning

Hancong Liu [a], Sirish Shah [a,*], Wei Jiang [b]

[a] *Department of Chemical and Materials Engineering, University of Alberta, Edmonton, Alberta, Canada T6G 2G6*
[b] *Department of Systems Engineering and Engineering Management, Stevens Institute of Technology, Hoboken, NJ 07030, USA*

Received 24 November 2003; accepted 20 January 2004

Available online 21 March 2004

Abstract

Outliers are observations that do not follow the statistical distribution of the bulk of the data, and consequently may lead to erroneous results with respect to statistical analysis. Many conventional outlier detection tools are based on the assumption that the data is identically and independently distributed. In this paper, an outlier-resistant data filter-cleaner is proposed. The proposed data filter-cleaner includes an on-line outlier-resistant estimate of the process model and combines it with a modified Kalman filter to detect and "clean" outliers. The advantage over existing methods is that the proposed method has the following features: (a) a priori knowledge of the process model is not required; (b) it is applicable to autocorrelated data; (c) it can be implemented on-line; and (d) it tries to only clean (i.e., detects and replaces) outliers and preserves all other information in the data.

Keywords: Data preprocessing; Outlier detection; Breakdown point; Data filter-cleaner; Time series analysis

1. Introduction

Outliers are observations that deviate significantly from the majority of observations. They may be generated by a different mechanism corresponding to normal data and may be due to sensor noise, process disturbances, instrument degradation, and/or human-related errors. It is futile to do data based analysis when data are contaminated with outliers because outliers can lead to model misspecification, biased parameter estimation and incorrect analysis results. The majority of outlier detection methods are based on an underlying assumption of identically and independently distributed (i.i.d.) data, where the location (e.g., the mean) and the scatter (e.g., variance/covariance) are the two most important statistics for data analysis in the presence of outliers (see Rousseeuw & Leroy, 1987 and references therein). Among these methods, the Hampel identifier is regarded as one of the most robust and efficient outlier identifiers (Davies & Gather, 1993; Perarson, 2002). It is well known that significant autocorrelation exists in regularly sampled data from the process industry. For autocorrelated data, following the pioneering work of Fox (1972), there is a substantial amount of research on maximum likelihood based outlier detection methods assuming known process models (e.g., Bianco, Garcia, Martinez, & Yohai, 1996; Bianco, Garcia, Ben, Martinez, & Yohai, 2001; Chen & Liu, 1993; Tsay, 1988, 1996). Martin and Thomson (1982) proposed a data cleaning method using a modified Kalman filter which is based on an estimated autoregressive (AR) model. However, in practice, it is hard to know the exact process models and outlier detection method without a process model is still an open area of research. Most existing outlier detection methods are essentially off-line operations and it is generally hard to filter outliers and simultaneously keep a track of a changing process model.

For process monitoring purposes, on-line data-based analysis, such as, on-line PCA, on-line PLS-based monitoring, and on-line controller performance monitoring, needs "clean" data to provide reliable detection and diagnosis results. Data preprocessing is a necessary pre-requisite step prior to process and performance monitoring. One important purpose of data preprocessing is to sort or sieve all data and to remove and replace outliers with their expected values. The first step in data preprocessing is to detect outliers. The detected outliers are considered as missing values, and the data are reconciled or estimated. In view of this point, data

* Corresponding author. Tel.: +1-780-492-5162; fax: +1-780-492-2881.
E-mail address: sirish.shah@ualberta.ca (S. Shah).

0098-1354/$ – see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compchemeng.2004.01.009

preprocessing is different from data filtering. Data filtering changes the data structure by not only removing outliers but also reducing data variations. Data preprocessing is generally more difficult than filtering because it simultaneously requires identification of the data structure and estimation of the noise level so as to remove the outliers and retain "good" data. An efficient on-line data cleaner is an important ingredient of data preprocessing. Recently, Nounou and Bakshi (1999) proposed an on-line data filtering method without the use of process models. They applied wavelet thresholding to data in a moving window of dyadic length to filter data on-line and a finite impulse response (FIR) median hybrid (FMH) filter to reconcile outliers and process shifts. The given filter focuses mainly on data filtering for multiscale process data contaminated with outliers, but not on data preprocessing since the filtered data may lose much original data information.

In this paper, an on-line filter-cleaner is proposed based on a filter-cleaner developed by Martin and Thomson (MT) (1982), from here onwards denoted as the revised MT filter-cleaner. Filtering implies use of past and current data to estimate the current point and its variation. Cleaning on the other hand is concerned with detection and replacement of outliers. The term "filter-cleaner" combines both of these functions. Unlike the original MT filter-cleaner, the revised MT filter-cleaner can work without priori knowledge of the exact underlying model by capturing the dynamics of the process data on-line. The revised MT filter-cleaner resorts to a high breakdown decorrelation approach for outlier detection. It is demonstrated that the proposed filter-cleaner is efficient in outlier detection and data cleaning for autocorrelated and even nonstationary process data. An important requirement of the revised MT filter-cleaner which is crucial to data preprocessing is that it should only clean outliers and otherwise preserve all correct data information. However, being a filter, there is also a small risk of cleaning "good outliers".

The paper is organized as follows: Preliminaries in the area of outlier detection and data cleaning are presented in Section 2. This is followed by the introduction of the new method to construct a decorrelation model for the revised MT filter-cleaner in Section 3. The revised MT filter-cleaner is proposed in Section 4. In Section 5, the robustness of this filter-cleaner for a broad class of autocorrelated and even nonstationary process data is analyzed, and its performance is compared with the Hampel identifier, followed by concluding remarks in Section 6.

2. Preliminaries

Many outlier detection methods assume that the underlying data model is i.i.d.. For univariate process data, parameter estimation from contaminated data has been investigated extensively in the statistical literature (see Barnett & Lewies, 1994). The location and the scatter are two of the most useful parameters for describing or characterizing data mean and variation. Traditionally, the sample mean $\bar{X}$ and variance S^2 of a sample $X_N = \{x_i\}_{i=1}^N$ give good estimation for data location and scatter if outliers do not contaminate the sample, i.e.,

$$\bar{X} = \frac{\sum_{i=1}^N x_i}{N}, \quad S^2 = \frac{\sum_{i=1}^N (x_i - \bar{X})^2}{n-1}$$

However, when a dataset contains outliers, even a single out-of-scale observation, the sample mean may deviate significantly. To measure the robustness of an estimator against outliers, Hampel (1971) introduced the concept of the breakdown point. The breakdown point is the smallest percentage of contaminated data (outliers) that can cause an estimator to take arbitrary large aberrant values. Generally, the larger breakdown point an estimator has, the more robust it is. It can be seen that the sample mean has a breakdown point of $1/N$ since a single large observation can make the sample mean and variance cross any bound. Thus, to robustly estimate the location and the scatter, the median and the median absolute deviation (MAD) are often recommended

$$\begin{aligned}\text{median}(X_N) &= \frac{x_{[(N+1)/2]:N} + x_{[N/2]+1:N}}{2},\\ \text{MAD}(X_N) &= \text{median}(|x_1 - \text{median}(X_N)|, \ldots,\\ &\qquad |x_N - \text{median}(X_N)|).\end{aligned}$$

where [.] is the function to "round-down" to the nearest integer and $x_{1:N}, \ldots, x_{N:N}$ are the order statistics of the sample X_N. The median estimator has a breakdown point of 50%. Hampel (1974) suggested an identifier using the median to estimate data location, and the MAD to estimate data standard deviation, i.e., x is identified as an outlier if

$$|x - \text{median}(X_N)| \geq g(N, \alpha_N)\,\text{MAD}(X_N), \tag{1}$$

where g is a function related to the number of data points and a specified type I error (for details, see Davies & Gather, 1993). The Hampel identifier is often considered extremely effective in practice (Perarson, 2002).

However, the i.i.d. assumption is often violated in reality, especially for auto- and cross-correlated chemical process data. For highly autocorrelated data, process data can be characterized by a time series model. For example, suppose $\{y_t\}$ is the observed series and $\{x_t\}$ is the underlying outlier-free series. Assume that $\{x_t\}$ is Gaussian with mean μ and autocorrelation coefficient function given by

$$\rho_k = \frac{E[(x_t - \mu)(x_{t+k} - \mu)]}{\sqrt{E[(x_t - \mu)(x_t - \mu)]E[(x_{t+k} - \mu)(x_{t+k} - \mu)]}},$$

where $E[\cdot]$ denotes the expectation operator. Suppose $\mu = 0$ and $\{x_t\}$ follows a general autoregressive integrated moving average (ARIMA) model (Box, Jenkins, & Reinsel, 1994), i.e.,

$$\Phi(q^{-1})(1 - q^{-1})^d x_t = \Theta(q^{-1})a_t, \tag{2}$$

where a_t is an independent Gaussian noise with mean 0 and variance σ_a^2, and q^{-1} is the backshift operator. Here,

$\Phi(q^{-1})$ and $\Theta(q^{-1})$ are polynomials in q^{-1} with degrees m and n, respectively, and are parameterized as $\Phi(q^{-1}) = 1-\phi_1 q^{-1}-\phi_2 q^{-2}-\cdots-\phi_m q^{-m}$ and $\Theta(q^{-1}) = 1-\theta_1 q^{-1}-\theta_2 q^{-2}-\cdots-\theta_n q^{-n}$. We assume that all of the zeros of the determinants $|\Phi(q^{-1})|$ and $|\Theta(q^{-1})|$ are on or outside the unit circle. The model (2) is also called ARIMA(m, d, n) model. The additive outlier (AO) model is defined as

$$y_t = x_t + v_t, \tag{3}$$

where y_t are the observed data and v_t are additive outliers with unknown distribution. It has been shown that additive outliers are most detrimental for model parameter estimation (e.g., Chang, Tiao, & Chen, 1988; Chen, & Liu, 1993).

Most of the research literature assumes a known process data model whose outliers are identified and cleaned off-line. Based on a robust Kalman filter introduced by Masreliez and Martin (1997), Matin and Thomson (1982) proposed a filter-cleaner where a psi-function Ψ is applied to the innovations to prevent outliers from the data having undue influence on the Kalman filter predictions. The MT filter-cleaner provides a good alternative for data preprocessing by detecting gross errors. However, the MT filter-cleaner is essentially an off-line scheme which is based on the pre-estimated time series model. Moreover, the breakdown point of the MT filter-cleaner is low when the order of model becomes high. In this paper, we propose a moving-window-based MT filter-cleaner which can capture the dynamic changes in the process data in an on-line fashion. The proposed filter-cleaner also has a fixed breakdown point for any model order.

3. Construction of a prewhitening model for autocorrelated data contaminated with outliers

Outlier detection is difficult when process data is contaminated with outliers and the underlying model is unknown. It is necessary to prewhiten process data first in order to remove outliers. Here an AR(p) model is used to decorrelate process data.

For illustration, Fig. 1shows the plot of the observations $\{y_t\}_{t=1}^{300}$ generated from an AR(2) model contaminated by outliers, which is given by

$$x_t = 0.7x_{t-1} + 0.2x_{t-2} + a_t, \qquad y_t = x_t + v_t$$

where a_t is i.i.d. with $N(0, 1)$ and v_t is an additive outlier. The outliers, v_t, are randomly generated with magnitudes of -5 or 5 with the probability of 0.08 at time t.

It is obvious that the data sequence is autocorrelated and outliers are present. In order to decorrelate this data sequence, we estimate an AR(2) model from the data sequence. The least-square estimation can be obtained by minimizing

$$\sum_{t=3}^{300}(y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2})^2.$$

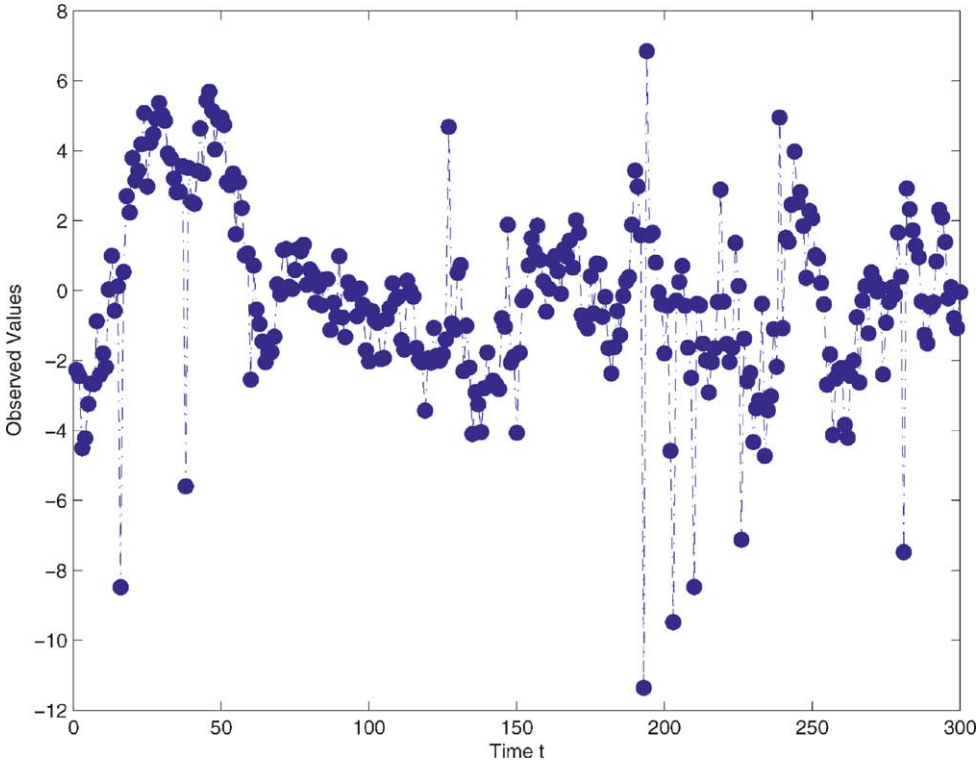

Fig. 1. Autocorrelated data contaminated with outliers.

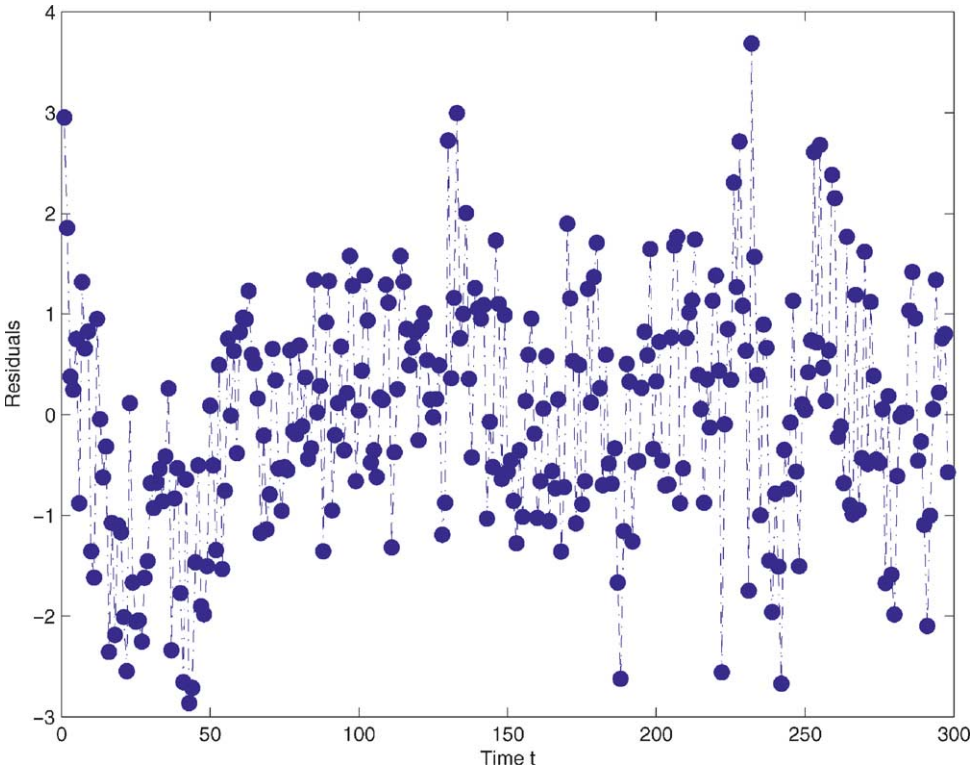

Fig. 2. Residuals based on least square estimation.

The estimated model is

$$x_t = 0.3698x_{t-1} + 0.2883x_{t-2} + a_t, \quad (4)$$

which is far away from the real data model. It can be observed from Fig. 2 that the residuals from the estimated model (4) are still autocorrelated.

Therefore, when raw data contains outliers, a robust estimation of the model such as Huber's M-estimator (1981) is necessary. However, as shown by Martin (1979), the M-estimators have zero breakdown points when AO departs from perfectly observed autoregressive models. Denby and Martin (1979) suggested the use of generalized M-estimator to bound the influence of outlying points by means of some weight functions. Even though the GM-estimator has the positive breakdown point bounded above by $1/(p+2)$ (Maronna, 1976), it can easily fail when the order of the AR(p) model is high.

To improve the robustness of the model estimates, we propose to estimate autocorrelation coefficients by a multivariate location and scatter estimator separately. Then, the Yule–Walker equations (Box, Jenkins, & Reinsel, 1994) are applied to solve the AR(p) model parameters based on all the kth autocorrelation coefficients. For the observations $\{y_t\}_{t=1}^N$, its kth autocorrelation coefficient can be estimated by transforming the original univariate series $\{y_t\}_{t=1}^N$ into a bivariate series $\{Y_t = (y_t, y_{t-k})\}_{t=k+1}^N$ and applying one of the multivariate robust estimation methods instead of the univariate M- or GM-estimators. For example, the minimum covariance determinant (MCD) estimator developed by Rousseeuw (1984) and Rousseeuw and Driessen (1999) is a good alternative. The following Theorem shows that the proposed robust estimator for the AR(p) models is more robust than the M- and GM-estimators.

Theorem 1. *Consider observations $\{y_t\}_{t=1}^N$ from an AR(p) process data. If each kth ($k = 1, 2, \ldots, p$) order autocorrelation coefficient is independently estimated by a robust estimator which has a breakdown point of h ($0 < h < 1/2$), then the AR(p) model estimator obtained from the Yule–Walker equations has a breakdown point of $h/2$ when N tends to infinity.*

Proof. See Appendix A. □

Theorem 1 constructs an AR(p) model estimator whose breakdown point depends only on the breakdown point of a multivariate covariance matrix estimator. For example, if we apply the MCD method, whose highest breakdown point is 0.5, then the highest breakdown point of the AR(p) model estimator can be as high as 0.25 no matter how large p is. The MCD method is one of the most efficient estimators in computation. This enables our method to be applicable on-line.

Once all the autocorrelation coefficients are obtained, then the Yule–Walker equations are solved to obtain the AR(p) model parameters and the MT filter-cleaner can be applied to

filter future observations. In the following section, we briefly introduce the original MT filter-cleaner and then develop the revised MT filter-cleaner.

4. The original and the revised MT filter-cleaner

4.1. The MT filter-cleaner

Suppose the underlying process data model can be well approximated by an AR(p) model. Given an AO model (3), Martin and Thomson (1982) proposed the following filter-cleaner algorithm in state space form:

$$X_t = \Phi X_{t-1} + U_t,$$

where

$$X_t^{\mathrm{T}} = [x_t, x_{t-1}, \ldots, x_{t-p+1}],$$
$$U_t^{\mathrm{T}} = [\epsilon_t, 0, \ldots, 0],$$
$$\Phi = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & & 0 \\ \vdots & 0 & \cdots & \vdots & \vdots \\ \vdots & & \cdots & \vdots & \vdots \\ 0 & & \cdots & 1 & 0 \end{bmatrix},$$

The filter-cleaner computes robust estimates of the vector X_t according to a matrix M_t:

$$\hat{X}_t = \Phi \hat{X}_{t-1} + \tilde{m}_t s_t \Psi\left(\frac{y_t - \hat{y}_t^{t-1}}{s_t}\right),$$

where $\tilde{m} = m_t/s_t^2$, and m_t is the first column of the $p \times p$ matrix M_t. The matrix M_t is computed recursively as

$$M_{t+1} = \Phi P_t \Phi^{\mathrm{T}} + Q,$$
$$P_t = M_t - w\left(\frac{y_t - \hat{y}_t^{t-1}}{s_t}\right)\frac{m_t m_t^{\mathrm{T}}}{s_t^2},$$

where Q is a matrix with all zero entries except $Q_{11} = \sigma_\epsilon^2$. The time-varying scale is defined as

$$s_t^2 = m_{11,t} = \text{1-1} \in M_t,$$

where $m_{11,t}$ is the one-one element of M_t. The symbol $\hat{y}_t^{t-1}$denotes a robust one-step ahead prediction of y_t based on $Y_{t-1} = (y_1, \ldots, y_{t-1})$, and is given by

$$\hat{y}_t^{t-1} = (\Phi \hat{X}_{t-1})_1,$$

and $\hat{y}_t^{t-1}$is the first element of $\Phi \hat{X}_{t-1}$.

With the AO model (3), where x_t and v_t independent, a best predictor of y_t is also a best predictor of x_t, and so the robust one-step ahead predictor $\hat{x}_t^{t-1}$ of x_t satisfies $\hat{x}_t^{t-1} = \hat{y}_t^{t-1}$. Finally, the cleaned data at time t is given by the first element of $\hat{X}_t$,

$$\hat{x}_t = (\hat{X}_t)_1.$$

The psi-function, Ψ, and weight-function, w, are essential to obtain robustness. If Ψ and w are always equal to 1, then the MT filter-cleaner is the commonly used Kalman filter. In order to obtain robustness, both functions should be bounded and continuous. One commonly used value of w is defined by

$$w(\tau) = \frac{\Psi(\tau)}{\tau}.$$

In practice, the Ψ function is generally chosen as the three-sigma edit rule

$$\Psi(\tau) = \begin{cases} \tau, & |\tau| < K \\ 0, & |\tau| \geq K \end{cases}$$

or

$$\Psi(\tau) = \begin{cases} \tau, & |\tau| < K \\ \mathrm{sign}(\tau)K, & |\tau| \geq K \end{cases} \tag{5}$$

where $K = 3$. It should be noticed that the MT filter-cleaner assumes that the process data model is known. Here we propose the revised MT filter-cleaner which is applicable when the process data model is unknown, and its decorrelation model is adaptively estimated through an AR(p) model over a moving window. After the decorrelation model is obtained by the method suggested in Section 3, the rest of the data cleaning procedure is the same as in the MT filter-cleaner.

4.2. The revised MT filter-cleaner

For a process data sequence, at time t, the revised MT filter-cleaner procedure consists of the following steps:

1. Choose a dataset $\{y_t\}_{i=t-N+1}^{N}$ with window size N.
2. Select the order p of the autoregressive process data.
3. Estimate the AR(p) decorrelation model based on the dataset $\{y_t\}_{i=t-N+1}^{N}$.
   - 3.1. Estimate the mean μ and variance γ_0 of $\{y_t\}_{i=t-N+1}^{N}$ based on a univariate robust estimator.
   - 3.2. Form the new multivariate datasets $\{Y_i^k = (y_i, y_{i-k})\}_{i-t-N+k+1}^{N}$ ($k = 1, 2, \ldots, p$). For the kth multivariate dataset $\{Y_i^k\}_{i=t-N+k+1}^{N}$, the covariance matrix $\begin{bmatrix} \gamma_{11}^k & \gamma_{12}^k \\ \gamma_{21}^k & \gamma_{22}^k \end{bmatrix}$ of this dataset $\{Y_i^k\}_{i=t-N+k+1}^{N}$ can be estimated by any robust multivariate location and scatter estimators (e.g., Rousseeuw's MCD method). Then, the kth autocorrelation coefficient $\rho_k = \gamma_{12}^k/\sqrt{\gamma_{11}^k \gamma_{22}^k}$.
   - 3.3. Solve the Yule–Walker equations:

$$\rho_j = \phi_1 \rho_{j-1} + \phi_2 \rho_{j-2} + \ldots + \phi_p \rho_{j-p},$$
$$j = 1, \ldots, p.$$

Denote

$$\phi = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{bmatrix}, \quad \rho = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_p \end{bmatrix},$$

$$P = \begin{bmatrix} 1 & \rho_1 & \cdots & \rho_p \\ \rho_1 & 1 & \cdots & \rho_{p-1} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_p & \rho_{p-1} & \cdots & 1 \end{bmatrix}.$$

Then $\phi = P^{-1}\rho$, and the process data model is estimated as

$$x_t = \frac{\mu}{1 - \phi_1 - \phi_2 - \cdots - \phi_p} + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \epsilon_t.$$

4. Filter and clean the data: Construct the state-space form for the process data over the user specified window and use the MT filter-cleaner to filter/clean the current data point y_t.
5. Go to step 1.

The order p of the autoregressive model in step 2 can be obtained from the Akaike information criterion (AIC) or the Bayesian information criterion (BIC).

However, from our experience, this selection is not critical in this procedure. Denote by $\hat{s}_\epsilon(p)$ the estimated noise standard deviation when the order of the autoregressive model is p. In practice, p can be simply chosen such that $\hat{s}_\epsilon(p+1)$ is not much smaller than $\hat{s}_\epsilon(p)$, so a lower order model often suffices.

5. Comparison of the revised MT filter-cleaner with the hampel filter

The Hampel identifier (1) can be implemented on-line in a moving window and the detected outliers can be replaced by the median of the data in the moving window. This has so far been considered one of the most robust and efficient on-line data filtering and cleaning tools (Albuquer & Biegler, 1996). However, as mentioned earlier, many outlier identifiers such as the Hampel identifier assume that the underlying process data is i.i.d.. When the process data is highly autocorrelated, such identifiers may fail to capture most outliers due to strong autocorrelation. In this section, we shall use the Hampel identifier as a benchmark to demonstrate the effectiveness of the revised MT filter-cleaner in the presence of autocorrelated data obtained from a process in a transient or dynamic state.

Our comparison will be conducted from two aspects. First, we compare the outlier detection efficiency of the two identifiers for time-invariant processes, i.e., process data model does not change with time. Then, a numerical example of a time-varying processes is used to illustrate the performance of the identifiers. Naturally, these comparisons are limited and could never cover all process scenarios. Further study of other linear and nonlinear process data is under investigation.

5.1. Time invariant process data

We now first analyze the outlier detection efficiency of both the Hampel identifier and the revised MT filter-cleaner. A natural measure of outlier detection efficiency is the outlier detection rate when outliers are present. This strongly depends on outlier sizes and distributions. Albuquer and Biegler (1996) introduced a notation of "relative efficiency" to measure outlier detection efficiency of outlier identifiers. The relative efficiency is defined as the ratio of the error variances between the theoretical optimal estimator and the actual estimator, i.e.,

$$E = \frac{\sigma^2_{\text{opt}}}{\sigma^2_{\text{act}}},$$

where σ^2_{opt} is the error variance of the theoretical optimal estimator and σ^2_{act} is the error variance attained by the actual estimator. The relative efficiency does not depend on conditions of outliers. Generally, it is conceived that, the higher relative efficiency an outlier detection method, the higher outlier detection rate. However, no work has been done to show the relationships between the two measures.

To measure the power of outlier detection methods, we introduce statistical definition of type I and II errors. For a fixed level of outlier size and contamination rate, we define detection and misidentification rates as the proportion of correct outlier identification from the total number of outliers and proportion of misidentification from the total number of "good" data points, respectively. To be precise, the Hampel identifier declares an outlier when inequality (1) holds, while the revised MT filter-cleaner declares an outlier when $\Psi(\tau) = \text{sign}(\tau)K$ in (5). Threshold parameters $g(N, aN)$ in the Hampel method and K in the revised MT method are chosen so that the type I errors of both methods are 0.01, i.e., the misidentification rate is 0.01 when no outliers are present.

We investigate an ARIMA process and an open-loop system for an output error (OE) process. Detection and misidentification rates are obtained from Monte Carlo simulations. The simulation was run with 10,000 test data points for each process data model. The window length of both identifiers are selected as 100 to obtained the reliable process data model. For simplicity, an AR(1) model was used in the revised MT-method since the selection of model order p is not critical in our procedure.

5.1.1. Data from an ARMA and ARIMA models

For illustration, the two outlier identifiers are applied to process data generated by stationary ARMA(1, 1) models. Although the Hampel filter is developed to estimate the mean and variance of an i.i.d. process, when data sample is large enough, these estimators are robust if the outlier contamination rate is less than 50%, i.e., it can correctly capture the mean and variance of the ARMA process and the estimated error follows exactly the same ARMA(1, 1) model as the underlying process data

$$(1-\phi q^{-1})x_t = (1-\theta q^{-1})a_t, \tag{6}$$

where a_t is a stochastic noise with mean 0 and variance σ_a^2. A closed-form expression for the autocorvariance coefficient function follows (Pandit & Wu, 1990))

$$\gamma_k = \begin{cases} \dfrac{(\phi-\theta)(1-\phi\theta)}{1-\phi^2}\phi^{k-1}\sigma_a^2, & k \geq 1 \\ \dfrac{1-2\phi\theta+\theta^2}{1-\phi^2}\sigma_a^2, & k = 0 \end{cases} \tag{7}$$

and the relative efficiency of the Hampel filter can be computed as

$$E_{\text{Hampel}} = \frac{\sigma_{\text{opt}}^2}{\sigma_{\text{act}}^2} = \frac{\sigma_a^2}{\sigma_e^2} = \frac{1-\phi^2}{1+\theta^2-2\phi\theta}.$$

On the other hand, when the data sample is large enough and the outlier contamination rate is not high, if an AR(1) model is used in the revised-MT estimator to estimate the real ARMA(1,1) model (6), then estimated errors e_t can be expressed as

$$e_t = (1-\phi^* q^{-1})x_t.$$

Therefore, e_t follows an ARMA(2, 1) process

$$(1-\phi^* q^{-1})e_t = \frac{1-\theta q^{-1}}{1-\phi q^{-1}}a_t,$$

where ϕ^* is determined by minimizing the estimated variance, i.e.,

$$E(e_t, e_t) = \min_{\phi^*} \frac{1+(\phi-\theta-\phi^*)[2\phi^*\theta\phi-(\theta+\phi^*)] + (\phi^*\theta)^2 - \phi(\theta+\phi^*)}{1-\phi^2}\sigma_a^2$$

$$= \frac{(1-2\phi\theta+\theta^2)^2 - (\theta-\phi)^2(1-\phi\theta)^2}{(1-\theta^2-2\phi\theta)(1-\phi^2)}\sigma_a^2.$$

The relative efficiency of the revised MT filter-cleaner is

$$E_{\text{Re-MT}} = \frac{\sigma_{\text{opt}}^2}{\sigma_{\text{act}}^2} = \frac{(1-\theta^2-2\phi\theta)(1-\phi^2)}{(1-2\phi\theta+\theta^2)^2 - (\theta-\phi)^2(1-\phi\theta)^2}.$$

Fig. 3 shows the relative efficiency of the two methods for this ideal case, i.e., when the sample size is infinite and outlier contamination rate is not high. It is obvious that the relative efficiency based on the revised MT filter-cleaner is always greater than that based on the Hampel filter. More importantly, when the autocorrelation is high, the relative efficiency of the Hampel filter is very small. Therefore, it is

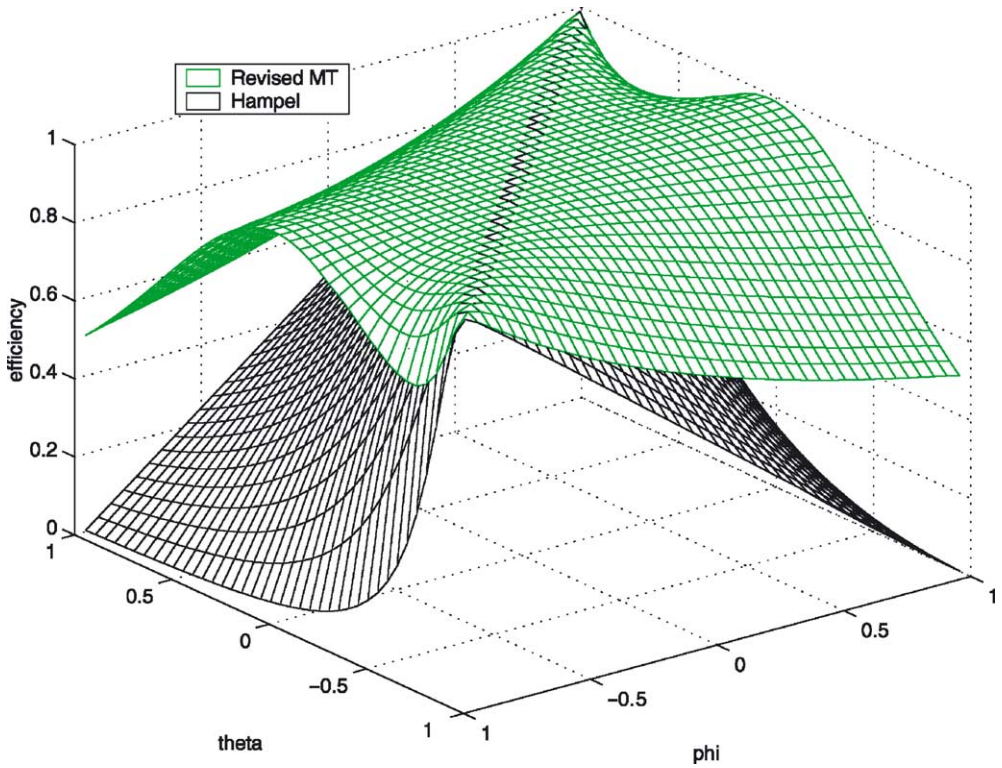

Fig. 3. Relative efficiency comparison between the revised MT filter-cleaner and the Hampel filter.

Table 1
On-line outlier detection rates for data from ARMA(1, 1) process

ϕ	θ	ρ_1	Outlier size	Revised MT			Hampel		
				Relative efficiency (%)	Mis-ID rate (%)	Detection rate (%)	Relative efficiency (%)	Mis-ID rate (%)	Detection rate (%)
0.0	0.0	0.00	4.0	1.00	0.55	82.83	1.00	0.63	84.23
0.0	0.0	0.00	5.0	1.00	0.43	95.41	1.00	0.44	96.41
0.0	−0.5	0.40	4.0	0.95	0.54	78.24	0.80	0.86	73.85
0.0	−0.5	0.40	5.0	0.95	0.41	94.81	0.80	0.85	92.61
0.0	−0.9	0.50	4.0	0.73	0.58	65.87	0.55	0.81	53.89
0.0	−0.9	0.50	5.0	0.73	0.45	86.03	0.55	0.66	75.25
0.5	0.0	0.50	4.0	1.00	0.49	82.44	0.50	0.71	73.65
0.5	0.0	0.50	5.0	1.00	0.40	95.01	0.50	0.80	92.02
0.5	−0.5	0.71	4.0	0.86	0.54	74.85	0.43	0.72	41.52
0.5	−0.5	0.71	5.0	0.86	0.46	90.22	0.43	0.75	62.87
0.9	0.0	0.90	4.0	1.00	0.59	79.84	0.19	0.47	12.57
0.9	0.0	0.90	5.0	1.00	0.37	93.01	0.19	0.56	25.15

expected that the outlier detection rate of the Hampel filter is also low when the autocorrelation is high, and the revised MT filter-cleaner should offer improvement over the Hampel method.

When the two methods are implemented on-line with a limited window, it is generally hard to derive their relative efficiencies. To compare the performance of the two methods with finite samples, their outlier detection rates are evaluated via a Monte Carlo simulation. Table 1. shows the first order autocorrelation coefficient, the theoretical relative efficiency (when sample size is infinitely large), the misidentification and detection rates based on the simulation study under different ARMA(1, 1) process data when outlier contamination rate is 5%. Outliers of 4 or 5 standard deviations of white noise a_t were generated and added to the process in order to investigate the impact of outlier size. As shown in Table 2, the Hampel method shows a significant deterioration when process autocorrelation becomes high. For example, when $\phi = 0$ and $\theta = 0$ (i.e., the process is i.i.d.), the Hampel method identifies about 84% of outliers of size $4\sigma_a$. When $\phi = 0.9$ and $\theta = 0$ (i.e., the first-order correlation coefficient is 0.9), it can only identify 13% of the outliers. Larger outlier size helps the outlier detection, but not significantly. On the other hand, the revised-MT method has about the same detection rate (83%) when the process is i.i.d., and has 80% detection rate when the autocorrelation is high. The reason is that the revised-MT method has theoretically the same relative efficiency whether the process data is i.i.d. or not.

At the same time, the revised-MT method always has a lower misidentification rate than the Hampel method, i.e., the former is able to capture the system dynamics more precisely than the later. It is important to note that the misidentification rate is less than 1% which is the target type I error. This is because, with outliers involved, the actual error variance is higher than the error variance when outliers are not present. Consequently, wider limits are obtained and fewer misidentification results. This becomes more obvious for large outliers (e.g., $5\sigma_a$). The detection power increases for large outliers as they are more easily noticed. When we tested for a higher (e.g., 10%) outlier contamination rate, similar results were observed.

Take the last process as an example. Fig. 4 shows a snapshot of outlier detection and data cleaning comparison for the revised MT filter-cleaner versus the Hampel filter, where the underlying model is

$$x_t = 0.9x_{t-1} + a_t$$

and the actual outliers are added at every 20th sample instance (i.e., at $0, 20, 40, \ldots$) points. Here if a data point is out of the upper and lower limits, the point is considered as an outlier and replaced by its expected values. Comparing the two methods, the revised MT filter-cleaner is able to detect five additional outlier (as shown by arrows) that the Hampel filter fails to detect.

Table 2
On-line outlier detection rates for data from an ARIMA(1, 1, 1) process

ϕ	θ	Outlier size	Revised MT		Hampel	
			Mis-ID rate (%)	Detection rate (%)	Mis-ID rate (%)	Detection rate (%)
0.0	0.5	4.0	0.94	43.16	0.40	9.39
0.0	0.5	5.0	0.18	69.93	0.21	14.99
0.0	−0.9	4.0	1.53	20.18	0.71	1.10
0.0	−0.9	5.0	1.87	39.26	0.91	1.80
0.5	0.0	4.0	1.12	19.28	0.89	1.10
0.5	0.0	5.0	0.86	38.96	0.71	1.20

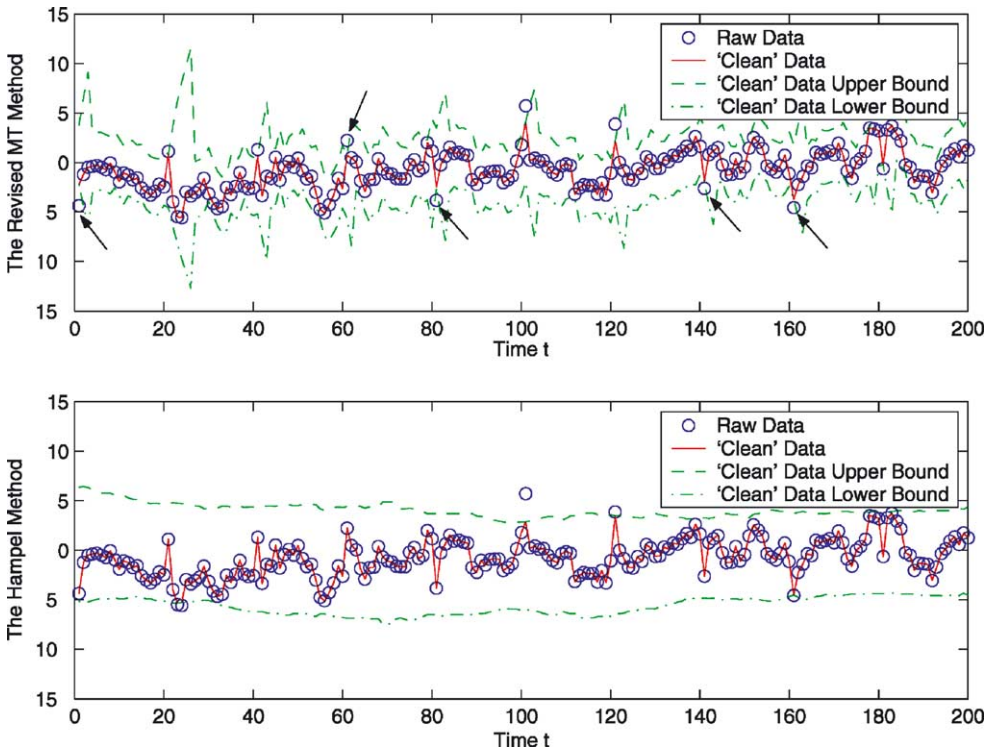

Fig. 4. Outlier detection and cleaning comparison: arrows in the top graph indicate five additional outliers that revised MT method is able to detect. The Hampel method considers these five points as normal data.

It should be pointed out that the relative efficiency of a filter-cleaner is not sufficient to determine the power of the filter-cleaner against outliers. Both detection rate and misidentification rate are strongly related to outlier sizes. Generally the larger the outlier, the easier it is to detect the outlier and the estimated model is more precise.

Table 2 shows simulation results for nonstationary ARIMA(1, 1, 1) process data. The Hampel filter now becomes so bad that it fails to detect outlier in most of cases. The nonstationarity also deteriorates the revised-MT method, but still keeps its detection rate at a reasonable level. It also can be observed that the misidentification rates in some cases are greater than 0.01 for the revised MT filter-cleaner. That means that the nonstationary property of a data process model may affect the outlier detection performance of the revised MT filter-cleaner in some unknown ways. This suggests that further modifications of the revised MT method may be necessary, for example, by taking differences before fitting AR(p) models. This procedure is currently under investigation in a separate project.

5.1.2. Data from an output error model

We now discuss data filtering and cleaning for an OE model. Consider a case where processes are subjected to an output error model as shown in Fig. 5. Assume that the process is described by the following OE model:

$$x_t = \frac{0.037q^{-1} + 0.07172q^{-2} + 0.00785q^{-3}}{1 - 1.3422q^{-1} + 0.4455q^{-2} - 0.045q^{-3}}u_t + a_t,$$

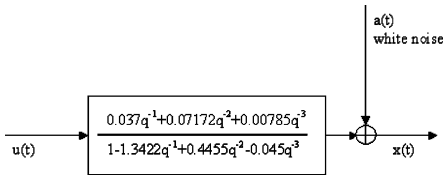

Fig. 5. Details of the output error model for generating process data.

where q^{-1} is the backshift operator, u_t is the input, and a_t is a stochastic disturbance with mean 0 and variance 1. The input u_t is a random binary sequence (RBS) between -4 and 4. With the steady-state process gain of 2, the size of the random binary sequence input is large enough so that the contribution from input changes are not be masked by the noise.

A simulation is carried out to compare the outlier detection performance of the two methods. Table 3 presents

Table 3
On-line outlier detection for data from an output error model process

Outlier size	Outlier percentage	Revised MT		Hampel	
		Mis-ID rate (%)	Detection rate (%)	Mis-ID rate (%)	Detection rate (%)
6	5	0.64	54.6	0.94	39.31
6	10	0.42	42.69	0.57	34.15
8	5	0.53	84.92	0.84	70.07
8	10	0.32	73.14	0.58	65.82

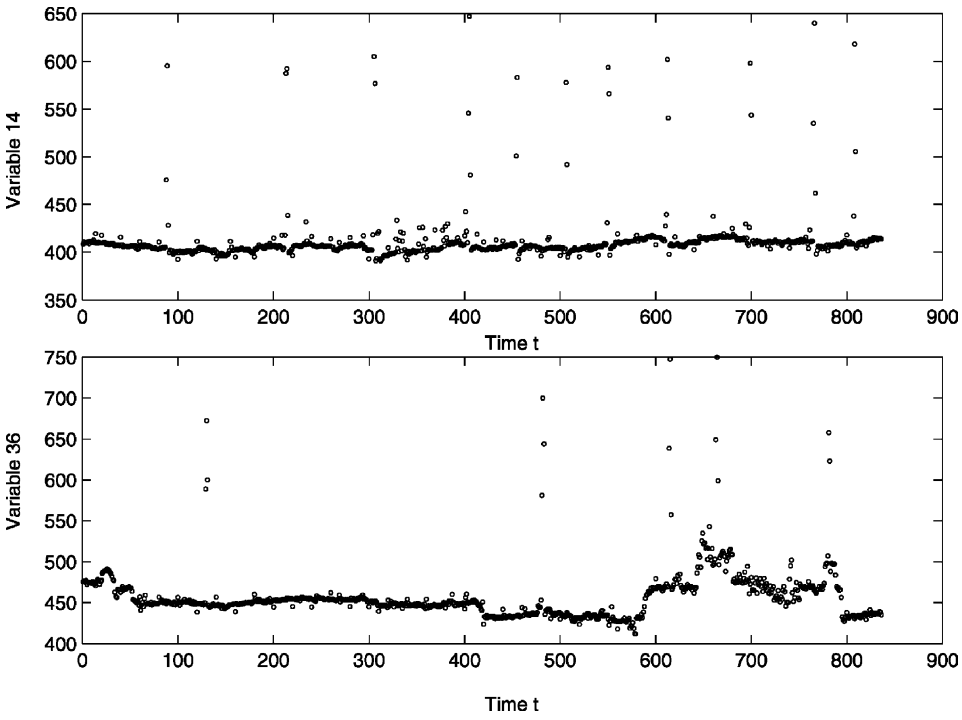

Fig. 6. Time series trajectory of raw data from Syncrude Canada Ltd.

the simulation results of misidentification rates and outlier detection rates for different outlier size and percentage. As before, the outlier detection rates of the revised MT method are much greater than those of the Hampel method.

5.2. Time varying process data

In practice, process data may change over time, e.g., baseline changes. For data from time-varying processes in the presence of outliers are present, process tracking becomes

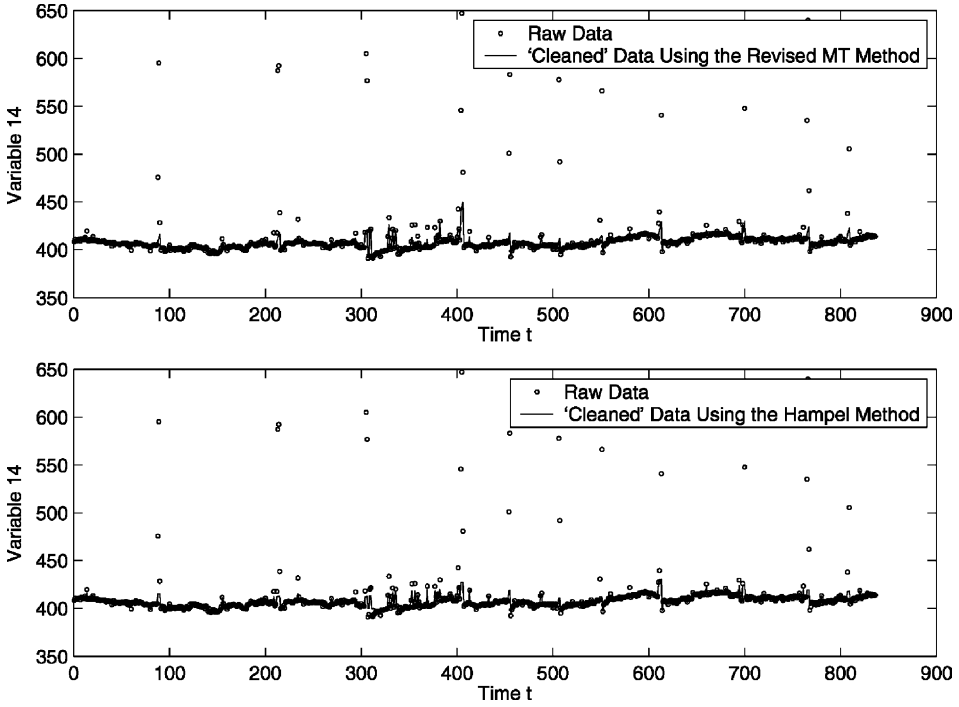

Fig. 7. Time series trajectory of raw and "clean" data using the revised MT method and the Hampel method for variable 14.

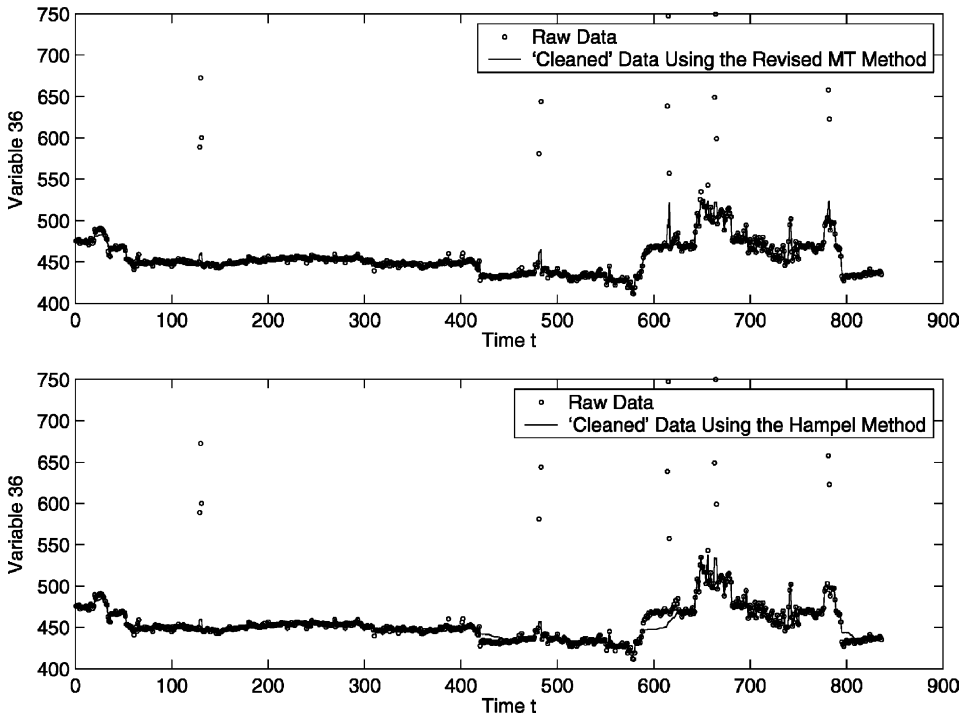

Fig. 8. Time series trajectory of raw and "clean" data using the revised MT method and the Hampel method for variable 36.

extremely difficult since it is hard to differentiate change points and outliers when changes or outliers first occurs. In this section, a real example of an industrial data set from Syncrude Canada Ltd. is used to illustrate the on-line capability of the revised MT method in capturing the dynamic change of the process data model in presence of outliers.

Two variables, V14 and V36, have been selected from the process data since outliers and change points are present in both of variables. Their time series are displayed in Fig. 6. It is clear that some observations are very different from their neighborhoods and are suspected to be outliers. We applied the two methods on-line with a window of size 100.

Figs. 7 and 8 show the on-line effects of the two filter-cleaners. Since the process data model for variable V14 does not change much over time, Fig. 7 shows that both the revised MT filter-cleaner and Hampel filter work well for this variable. On the other hand, the process data model of V36 is observed to change over time, especially its mean level. When the process data model changes, the Hampel filter often misidentifies some points (e.g., data points from 430 to 450, from 590 to 640 and from 795 to 810) to be outliers. In contrast, the revised MT filter-cleaner can quickly adjust its estimates of the process data and is thus able to capture the dynamic change of the process data model and remove outliers efficiently.

6. Conclusions

In this paper, a revised MT filter-cleaner is proposed for on-line outlier detection and data cleaning when the process data model is unknown. The main procedure involves on-line outlier-resistant process data model estimation and data cleaning. To the best knowledge of the authors there are very few outlier detection and replacement schemes that work on-line by estimating a model. To robustly estimate the process data on-line using an AR(p) model, the revised MT filter-cleaner transforms a univariate parameter estimation problem into a multivariate data covariance matrix estimation problem. After deriving correlation coefficients using a robust method, the parameters of the AR(p) model are computed from the Yule–Walker equations. Unlike the MT method, the proposed method estimates the parameters of the AR(p) model separately and results in a high breakdown point independent of the order of the AR(p) model. Once a process data model is obtained, the modified Kalman filter is used to filter and clean the data.

It is important to note that the revised MT cleaner can be applied to clean data from autocorrelated and even non-stationary process data. Because the process data model is estimated on-line, the revised MT method can adaptively capture the dynamics of the process data. Thus, the method is applicable to both time invariant and time varying process data. More importantly, the proposed filter-cleaner is simple and reliable since it has a high breakdown point and can be easily applied in practice because only one or two parameters (e.g., window size) need to be adjusted. Numerical examples show that the proposed method outperforms the Hampel method in most cases, especially when the autocorrelation is strong or the process data model changes over time.

In this paper, we have mainly studied process data that can be approximated by autoregressive models. Only additive type outliers have been considered here. Research in progress is concerned with the treatment of process data with nonlinear properties and other complex distributions.

Acknowledgements

The authors are grateful to Syncrude Canada Ltd. for granting their permission to use their data. Financial supports from NSERC, Matrikon INC. and the Alberta Science and Research Authority (ASRA) towards the Industrial Research Chair Program at the University of Alberta is also gratefully acknowledged. The third author's (Wei Jiang) work is partially supported by NSF-DMI grant #0200224.

Appendix A

Proof of Theorem 1. For the univariate dataset $\{y_t\}_{t=1}^N$ and fixed k ($k = 1, 2, \ldots, p$), the new multivariate dataset can be expressed as $\{Y_t = (y_t, y_{t-k})\}_{t=k+1}^N$. If there are m outliers in the univariate dataset $\{y_t\}_{t=1}^N$, the maximum number of outliers in the multivariate dataset $\{Y_t = (y_t, y_{t-k})\}_{t=k+1}^N$ is equal to

$$\begin{cases} 2m, & \text{if } m \le \left[\dfrac{N-k}{2}\right] \\ N-k, & \text{if } m > \left[\dfrac{N-k}{2}\right] \end{cases}$$

□

where $[\cdot]$ is the function to "round-down" to the nearest integer. Therefore, if there are n outliers in $\{Y_t = (y_t, y_{t-xk})\}_{t=k+1}^N$, the smallest number of outliers in $\{y_t\}_{t=1}^N$ is equal to

$$\begin{cases} \left[\dfrac{n}{2}\right], & \text{if } n < \left[\dfrac{N-k}{2}\right] \text{ and} n \text{ is even} \\ \left[\dfrac{n+1}{2}\right], & \text{if } n < \left[\dfrac{N-k}{2}\right] \text{ and } n \text{ is odd} \\ \left[\dfrac{N-k}{2}\right], & \text{if } n \ge \left[\dfrac{N-k}{2}\right] \text{ and } N-k \text{ is even} \\ \left[\dfrac{N-k+1}{2}\right], & \text{if } n \ge \left[\dfrac{N-k+1}{2}\right] \text{ and } N-k \text{ is odd.} \end{cases} \tag{8}$$

It can be seen that the kth autocorrelation coefficient is obtained once the covariance matrix of $\{Y_t = (y_t, y_{t-k})\}_{t=k+1}^N$ is known. Hence, if the breakdown point of a multivariate location and covariance matrix estimator is h ($0 < h < 1/2$), the smallest number l of the contaminated data (outliers) that can cause the location and covariance matrix estimator to arbitrary values is expressed as

$$l = \begin{cases} [h(N-k)], & \text{if } h(N-k) - [h(N-k)] = 0, \\ [h(N-k)] + 1, & \text{if } h(N-k) - [h(N-k)] \ne 0. \end{cases}$$

From (8), the smallest percentage of contaminated data (outliers) that can cause the kth autocorrelation coefficient to arbitrary values is equal to

$$\begin{cases} \left[\dfrac{l}{2}\right] / N, & \text{if } l < [\dfrac{N-k}{2}] \text{ and } l \text{ is even} \\ \left[\dfrac{l+1}{2}\right] / N, & \text{if } l < \left[\dfrac{N-k}{2}\right] \text{ and } l \text{ is odd} \end{cases}. \tag{9}$$

If all the kth autocorrelation coefficients are known, the AR(p) model can be calculated by solving the Yule–Walker equations. Because all the kth autocorrelation coefficients can be estimated separately, the breakdown point of the AR(p) model estimation is equal to the maximum breakdown point of all the kth autocorrelation estimations. From (9), the breakdown point of the AR(p) model estimation via the Yule–Walker equations is $h/2$ when N tends to infinity.

References

Albuquer, J. S., & Biegler, L. T. (1996). Data reconciliation and gross-error detection for dynamic systems. *American Institute of Chemical Engineering Journal, 42*, 2841–2856.

Barnett, V., & Lewies, T. (1994). *Outliers in statistical data* (3rd ed.) New York: Wiley.

Bianco, A.M., Garcia B.M., Martinez, E.J, & Yohai, V.J. (1996). Robust procedures for regression models with ARIMA errors. In *COMPSTAT 96, Proceedings in Computational Statistics Part A* (pp. 27–38). Berlin: Physica-Verlag.

Bianco, A. M., Garcia, , Ben, M. G, Martinez, E. J., & Yohai, V. J. (2001). Outlier detection in regression models with ARIMA errors using robust estimations. *Journal of Forecasting, 20*, 565–579.

Box, G.E.P., Jenkins, G.M., & Reinsel, G.C. (1994). *Time series modeling for statistical process control* (3rd ed.). San Francisco, CA: Holden-Day.

Chang, I., Tiao, G. C., & Chen, C. (1988). Estimation of time series parameters in the presence of outliers. *Technometrics, 30*, 193–204.

Chen, C., & Liu, L. M. (1993). Forecasting time series with outliers. *Journal of Forecast, 12*, 13–35.

Davies, L., & Gather, U. (1993). The identification of multiple outliers. *Journal of the American Statistical Association, 88*, 782–792.

Denby, L., & Martin, R. D. (1979). Robust estimation of the first-order autoregressive parameter. *Journal of the American Statistical Association, 88*, 284–297.

Fox, A. J. (1972). Outliers in time series. *Journal of Royal Statistics Society B, 34*, 350–363.

Hampel, F. R. (1971). A general qualitative definition of robustness. *Annals of Mathematics Statistics, 42*, 1887–1896.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association, 69*, 382–393.

Huber, P.J. (1981). *Robust statistics*. New York: John Wiley & Sons.

Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. *The Annals of Statistics, 4*, 51–67.

Masreliez, C. J., & Martin, R. D. (1977). Robust Bayesian estimation for the linear model and robustifying the Kalman filter. *IEEE Transactions on Automatic Control, AC-22*(3), 361–371.

Martin, R.D. (1979). Robust estimation for time series autoregressions. In: Launer, R.L., Wilkinson, G.N. (Eds.). *Robustness in Statistics* (pp. 147–176). New York: Academic Press.

Martin, R. D., & Thomson, D. J. (1982). Robust-resistant spectrum estimation. *Proceeding of the IEEE, 70*, 1097–1115.

Nounou, M. N., & Bakshi, B. R. (1999). On-line multiscale filtering of random and gross errors without process models. *American Institute of Chemical Engineering Journal, 45*, 1041–1058.

Pandit, S.M., & Wu, S.M. (1990). *Time series and system analysis with applications*. Malabar, FL: Krieger Publishing Company.

Perarson, R. K. (2002). Outliers in process modeling and identification. *IEEE Transactions On Control Systems Technology, 10*, 55–63.

Rousseeuw, P. J. (1984). Least median of square regression. *Journal of the American Statistical Association, 79*, 871–880.

Rousseeuw, P.J., & Leroy, A.M. (1987). *Robust regression and outlier detection*. New York: Wiley.

Rousseeuw, P. J., & Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics, 41*, 212–223.

Tsay, R. S. (1988). Outliers, level shifts, and variance changes in time series. *Journal of Forecasting, 7*, 1–20.

Tsay, R. S. (1996). Time series model specification in the presence of outliers. *Journal of the American Statistical Association, 81*, 132–141.