Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Process Control 14 (2004) 389–398
JOURNAL OF PROCESS CONTROL
www.elsevier.com/locate/jprocont

The impact of compression on data-driven process analyses

Nina F. Thornhill [a,*], M.A.A. Shoukat Choudhury [b], Sirish L. Shah [b]

[a] Department of Electronic and Electrical Engineering, University College London, Torrington Place, London WC1E 7JE, UK
[b] Department of Chemical and Materials Engineering, University of Alberta, Edmonton, AB, Canada T6G 2G6

Received 18 December 2002; received in revised form 17 April 2003; accepted 17 June 2003

Abstract

Stored process data in the form of high fidelity time trends are a resource for data-driven process analyses such as statistical monitoring, minimum variance control loop benchmarking, fault detection, data reconciliation and development of inferential sensors. However, many commercial data historians compress the data before archiving it and a question therefore arises of how useful the compressed data are for the intended purposes.

This article examines the impact of compression on data-driven methods and presents an automated algorithm by which the presence of piecewise linear compression may be inferred during the pre-processing phase of a data-driven analysis.

The results show that compression interferes with many types of data-driven analyses and the paper strongly recommends caution in the use of compressed process data archives.

Keywords: Chemical process; Data compression; Data reconciliation; Fault detection; Controller performance monitoring; Non-linearity; Process monitoring; Process trend; Spectrum; Statistical process control

1. Introduction

The motivations for data compression include reduction of the costs of storage of historical data and reduction of cost of transmission of process data through a telecommunications link. For instance, in pharmaceutical manufacturing the regulatory authorities demand long term storage of manufacturing records and the cost of storage media would then be a consideration. In off-shore oil production the cost is in the satellite linkage to an on-shore headquarters. The trend towards remote monitoring of their installed systems by technology vendors also requires data transmission through a telecommunications link.

Data compression, however, has hidden costs if the data become unsuitable for their intended purposes. The operation of restoring the original signal from the archived data is called reconstruction. Once the data have been compressed they lose information and the reconstructed trends are deficient in various ways compared to the originals. End uses of the reconstructed data may be very different [1] and include:

- Calculation of daily statistics such as daily means, daily standard deviations.
- Averaging for data reconciliation and mass balancing.
- Archiving of data trends for subsequent high fidelity reconstruction.
- Data smoothing by removal of high frequency noise.
- Feature extraction and recovery of events.

For example, the transmitted data from an off-shore production platform are used to determine daily totals of oil flow into the pipeline for taxation purposes while remote monitoring of a model predictive controller at a refinery may need high fidelity data for identification of a dynamic process model.

The contribution of this paper is to give new insights into the impact of data compression on data-driven plant performance analysis and to give a recommendation about how much compression can be tolerated. The findings that compression causes trouble may seem obvious in retrospect but there appears to have been no systematic study in the literature to date. An automated

* Corresponding author. Tel.: +44-20-7679-3983; fax: +44-20-7388-9325.
E-mail address: n.thornhill@ee.ucl.ac.uk (N.F. Thornhill).

0959-1524/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jprocont.2003.06.003

means of detecting the severity of compression is also presented. Application of the algorithm during the data pre-processing phase of a plant audit means less time wasted in evaluation of unsuitable data. It also avoids the loss of credibility of the methods and their practitioners that might arise if wrong conclusions were to be drawn from bad data.

The next two sections of the paper motivate the study by means of an example, outline normal practices in industrial data compression and introduce three industrial data sets. Section 4 presents measures by which the impact of data compression methods may be evaluated. Section 5 gives results and discussion. An automated means for detection of compression is demonstrated in Section 6 and its application to industrial data considered. The paper ends with a conclusions section.

2. Motivating example

Compression using piecewise linear trending is in widespread use in industrial data historians. For instance, AspenTech described an adaptive method based upon the box-car/backward slope (BCBS) method [2] while OSI state that their PI data historian uses a type of swinging door compression algorithm involving a compression deviation blanket with a width equal to twice the compression deviation specification [3].

Fig. 1 shows a data set from a data historian typical of those from which engineers and consultants wish to extract useful information (courtesy of Celanese Canada Inc.). The straight line segments characteristic of industrial data compression can be seen in many of the time trends. It will be shown in Section 6 that compression factors of up to 94 were in use. The original uncompressed data were lost forever when they were compressed and archived and it is now not possible to determine what features have been lost. Later sections will show that most of these data trends are too compressed and that data-driven process analysis would, if attempted, give a misleading indication of the results that the original data would have given.

3. Methods

3.1. Overview of data compression

There is extensive literature on compression methods for images, speech and text [4]. Compression techniques for electrocardiogram (ECG) signals are also at an advanced stage [5–7]. The motives in ECG are like those for process data compression in regard to transmission of the ECG signals by telephone. Some developments in data compression have arisen from that field, for example wavelet compression has moved from ECG to process applications.

Compression techniques can be divided into two main functional groups, direct methods and transform methods. Those in industrial use are the direct methods (also known as piecewise linear trending methods) because these can be applied in real-time to spot data. Mah et al. [8] and Watson et al. [9] have given comparative reviews of various compression methods. Mah et al. [8]

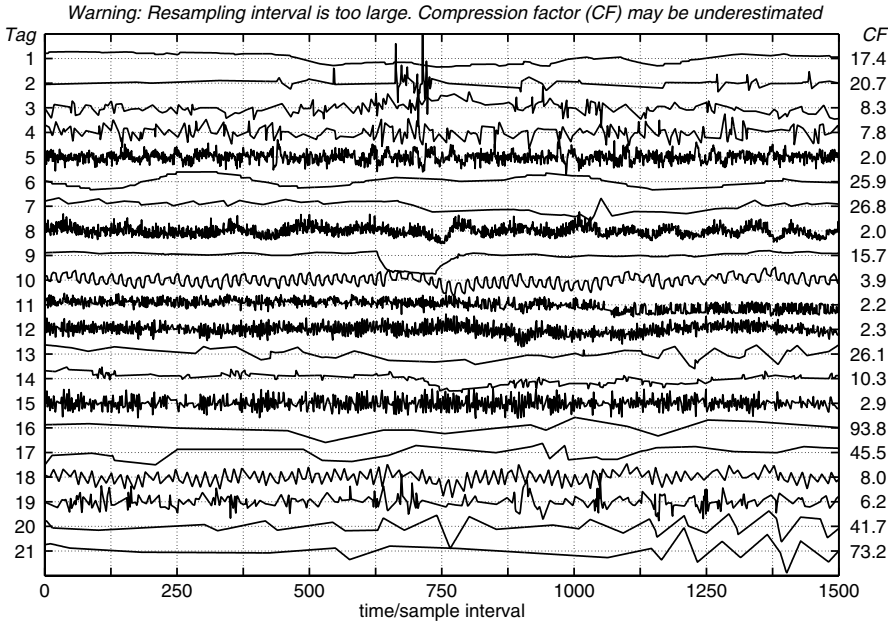

Fig. 1. An industrial data set with compression in some tags. Time trends are mean centered and normalised. CF is an estimate of the compression factor (see Section 6.3).

compared piecewise linear trending methods and introduced a new method (PLOT). The work of Watson et al. [9] studied piecewise linear trending and also wavelet and Fourier compression. They also introduced to process applications a method using vector quantization and discussed its benefits.

A direct method makes the archiving decision in real time as the data are captured from the process. The BCBS method [10] and the swinging door method [11] use heuristic rules to decide whether to archive a spot value and the rules are tuned to achieve the capture of exceptions and linear trends. They reconstruct a data trend as a series of linear segments connecting the archived spot values of the data.

A transform method performs an integral transform of the original data set and the compression is performed in the transformed domain. Wavelet compression falls into this category [12]. Such methods are not real-time. They require historical data since the transform is computed from an ensemble of data.

Many authors have explored wavelet compression. Nesic et al. [13] demonstrated its superior performance in process data from paper making machines. Other authors have explored various wavelet functions selected on a case-by-case basis [14–16]. Bakshi and Stephanopoulos [15] and Misra et al. [16] applied time-varying wavelet packets to achieve on-line feature extraction and noise removal from non-stationary signals. Misra et al. [17] described the use of adaptive compression thresholds to control the reconstruction error.

Vedam et al. [18] used a multiscale representation with coarse and fine resolution linear B-splines which comprise two piecewise linear segments. The multiscale formulation gave spline compression localisation features similar to wavelet compression.

3.2. Implementation of piecewise linear compression

The aim of this paper is to examine the impact of data compression on activities such as minimum variance control loop benchmarking, fault detection, data reconciliation and development of inferential sensors. It concerns industrial process data and therefore focuses upon piecewise linear trending. The swinging door method was selected for detailed study as representative of industrial practice. Similar results were observed with the BCBS and PLOT methods.

The swinging door method was implemented as described by [11]. Fig. 2 shows the principle. The first black point y_a, which has already been archived, is taken to be the start of a trend. Upper and lower pivot points marked $\times$ are calculated at $y_a \pm \Delta$.

As new spot values arrive, lines are drawn from the pivot points to form a triangular envelope that includes all the spot values since y_a. The sides of the triangle are the "doors". For instance, in Fig. 2 all points up to y_d can be enveloped in a triangle. However, the next point, y_e cannot be included in a triangle because, as shown by the dashed line, the upper and lower doors have opened wider than parallel. This signifies that a new trend started at y_d. Point y_d is archived and the procedure starts again at y_d.

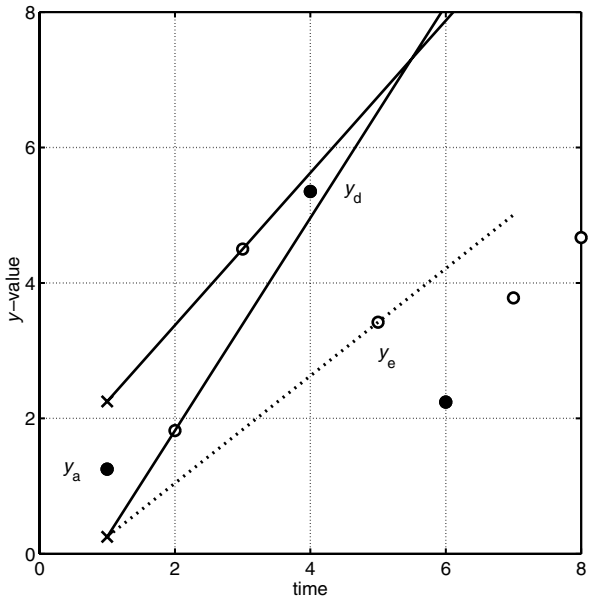

Fig. 2. Illustration of swinging door compression. Black circles represent archived spot values, values with open circles are not archived. At time step 5 (point y_e) the lower door (dotted line) opens up wider than parallel showing that a new trend started at y_d.

The compression factor is not specified directly in swinging door compression. Instead, the parameter to be set is the deviation threshold Δ in engineering units. Therefore in conducting the compression tests described in this paper it was necessary to first conduct calibration trials to find the deviation thresholds corresponding to $CF = 10$ for each data set.

The trends were reconstructed from the archived spot values by linear interpolation between archived points at the original sampling instants. The compression factor (CF) for swinging door compression is defined as the ratio between the storage requirement of the original data set and that of the archived data. If the original data set had 1000 observations and 1000 time tags a direct method with $CF = 10$ would yield 100 observations, 100 time tags and 99 linear segments.

3.3. Process data for compression comparisons

Three contrasting examples were chosen for the evaluation of the impact of compression, courtesy of BP. They are uncompressed liquid flow trends from continuous processes operating at steady state. Each data set comprised nearly 3 h of 10 s samples representing deviations of flow in a process stream from the

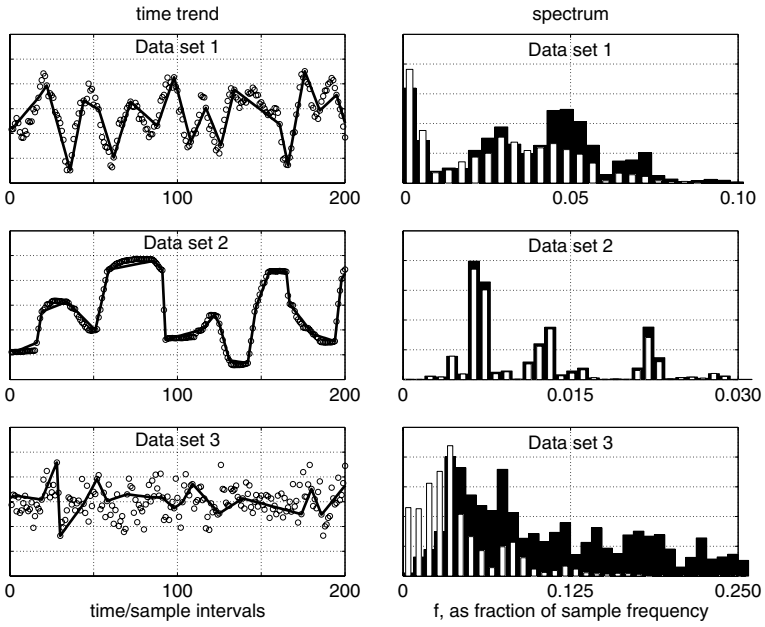

Fig. 3. *Left*: Time trends and reconstruction with compression factor 10. *Right*: Original spectra (black) and spectra of reconstructed signals (white).

mean value. Fig. 3 shows portions of the time trends (the open circles) while the black bars in Fig. 3 show their power spectra on linear vertical and horizontal axes [19]. Numerical scales on the y-axis were omitted by request.

Data set 1 shows a persistent oscillation characterized on average by about 20 samples per cycle. Fig. 3 shows that the spectrum of this signal has a broad peak at a frequency of 0.05 times the sampling frequency (i.e. 20 samples per cycle). The challenge for high fidelity compression and reconstruction is to retain the spectral peak in the frequency domain and the oscillatory features in the time domain.

Data set 2 has a tendency to stay at a value for a time and then to move rapidly to a new level. It is from a control loop which has a limit cycle caused by a sticking valve. The signal is predictable for long periods and its spectrum shows very low frequency features because the period of oscillation is long while a series of harmonics highlights the non-sinusoidal nature of the waveform. The low frequency features and harmonics should be preserved during compression and reconstruction.

Data set 3 has little predictability and has spectral features at all frequencies. This signal is dominated by random noise and is from a well tuned loop operating close to minimum variance.

4. Measures of performance

4.1. Statistical properties

Archived process data may be used for steady-state assessments such as plant production rates. Other uses include data reconciliation and mass balancing, for instance for the detection of leaks. Therefore if compressed archived data are to be used for these purposes the mean value of the reconstructed data should be the same as the mean of the original. The measure used is the percentage difference between the mean values (PDM) scaled by the standard deviation of the original data. The scaling allows the relative significance of any change in mean value to be assessed:

$$\mathrm{PDM} = 100\frac{\mathrm{mean}(y) - \mathrm{mean}(\hat{y})}{\sigma_y}$$

Process variability has an impact on profit [20,21] and plant audits usually begin with a determination of the standard deviations or the variances of the time trends. Therefore it is also necessary to determine the impact of compression on the observed variance. The measures used are the ratios between the variance of the reconstructed data ($\sigma_{\hat{y}}^2$) and the variance of the original data (σ_y^2) (RVC), and between σ_y^2 and the variance of the reconstruction error σ_e^2 where $e_i = y_i - \hat{y}_i$ (RVE). The measures are

$$\mathrm{RVC} = \sigma_{\hat{y}}^2/\sigma_y^2 \quad \text{and} = \sigma_e^2/\sigma_y^2$$

The second of these is similar to the NMSD measure used by Mah et al. [8] and Watson et al. [9] except that NMSD was expressed as a percentage.

If the two measures sum to 1 then the reconstruction error is the orthogonal complement of the compressed signal (i.e. the sequence $y_i - \hat{y}_i$ is uncorrelated with the sequence $\hat{y}_i$). The significance of this observation is considered in Section 5.

4.2. Non-linearity measure

Non-linearity assessment is starting to be used as a diagnostic tool for troubleshooting of hardware faults that may be present in the control loops [22,25] and to make decisions about the type of model needed in inferential sensing [23,24]. Therefore it is necessary to determine how the use of reconstructed data would influence non-linearity assessment.

A distinctive characteristic of a non-linear time series is the presence of phase coupling such that the phase of one frequency component is determined by the phases of others. Phase coupling leads to higher order spectral features which can be detected in the bicoherence of a signal. The non-linearity test applied here used bicoherence to assess non-linearity. The squared bicoherence is

$$\mathrm{bic}^2(f_1,f_2)=\frac{|B(f_1,f_2)|^2}{E(|Y(f_1)Y(f_2)|^2)E(|Y(f_1+f_2)|^2)}$$

where $B(f_1,f_2)$ is the bispectrum at frequencies (f_1,f_2) given by

$$B(f_1,f_2)=E(Y(f_1)Y(f_2)Y^*(f_1+f_2))$$

In the above expressions $Y(f)$ is the Fourier transform of y at frequency f, $Y^*(f)$ is its complex conjugate and E is the expectation operator. A key feature of the bispectrum is that it has a non-zero value if there is significant phase coupling in the signal y between frequency components at f_1 and f_2. The bicoherence gives the same information but is normalised as a value between 0 and 1.

As described in [22], the non-linearity assessment examines the mean value of the bicoherence over all frequencies and its maximum value, both quantities being tested against statistical thresholds that represent their expected values when no non-linearity is present.

4.3. Harris index measures

The widely used Harris index [26] is a minimum variance benchmark of control loop performance. Significant industrial implementations are being reported [27–29]. It is known that the use of data compression influences the Harris index [30] and an issue for practitioners is to know whether compressed archived data can be used for the purposes of a minimum variance benchmark calculation.

The Harris indexes for the three data sets were calculated using the method described in Desborough and Harris [26] with an estimated time delay of five samples. The index is determined from the residuals between the measured controller error denoted by y and a b-step ahead prediction, $\tilde{y}$.

$$r(i)=y(i)-\tilde{y}(i)$$

The model for $\tilde{y}$ used 30 autoregressive terms (i.e. $m=30$) as discussed in [30] and in this case the prediction horizon was $b=5$ since the time delay was estimated to be 5 sample intervals.

$$\tilde{y}(i+b)=a_0+a_1y(i)+a_2y(i-1)+\cdots+a_my(i-m+1)$$

The minimum variance benchmark is

$$1-\frac{\sigma_r^2}{\mathrm{mse}(y_i^2)}$$

where σ_r^2 is the variance of the residuals r and $\mathrm{mse}(y_i^2)$ is the mean square value of the controller error. An index of 0 represents minimum variance control while an index of 1 represents poor control in which $y\approx\tilde{y}$ and r is negligible. It means the controller is failing to deal with predictable components such as steady offsets or a predictable oscillatory disturbance.

The concern is that reconstructed data are more predictable and thus have a worse (larger) Harris index than the original because compression removes noise and the piecewise linear segments have high local predictability. Thus there is a danger that unnecessary maintenance effort may be spent on repair of control loops wrongly identified as performing poorly.

5. Results and discussion

5.1. Visual observations

The left hand panels in Fig. 3 show close-up portions of the original data (open circles) and reconstructions (solid line) with compression factor of 10 for data sets 1–3. Each complete data set had 1024 samples. Features of note are

- Swinging door compression did not follow all the oscillations in data set 1 because with CF = 10 the average duration of each linear segment was longer than half of the oscillation period.
- High fidelity compression was possible with data set 2 but with data set 3 much of the randomness was lost from the reconstructed trends.

The right hand panel in Fig. 3 shows reconstruction in the frequency domain. The power spectra of the original signal are in black with the spectra of the reconstructed signals with CF = 10 overlaid in white. When the two are not the same then there is a reconstruction error.

- The spectral feature in data set 1 at 0.05 samples per cycle was not fully captured by the reconstructed data set.
- Data sets 1 and 3 had errors at low frequency and a non-zero spectral error at $f=0$. Therefore the signal

reconstructed after compression had a different mean value than the original.
- The low frequency harmonics of data set 2 were reproduced well but the high frequencies of data set 3 were not captured.

The observations from the spectra reinforce and illuminate the observations from the time domain plots. The frequency domain plots also give insight into why data set 2 is more compressible than data set 3. Data set 2 has very few spectral features and they are at low frequency (i.e. of long duration) while data set 3 has features over the whole frequency range. Data set 2 is therefore a much simpler signal with fewer different types of behavior to capture.

5.2. Statistical properties

Fig. 4 shows the behavior of the mean value and variance measures as a function of compression factor. Noteworthy observations are

- The mean of the signal reconstructed from the archive differs from the mean of the original.
- The variance of the reconstructed data is smaller than the variance of the original signal.
- The variance measures at a given compression factor do not sum to 1.

It is concluded that data compression gives misleading information about basic statistical properties of the data. Compression alters both the mean and variance. The changes in the means are only a small percentage of the standard deviation. It is noted, however, that the purpose of data reconciliation is often to find small shifts in the mean value that may be indicative of problems such as leaks. The shift in mean due to data compression may therefore be wrongly interpreted as evidence of a leak. Decisions of the type used in statistical process control [31] may also be wrong if the warning and alarm limits have been based upon a statistical distribution determined from compressed archived data.

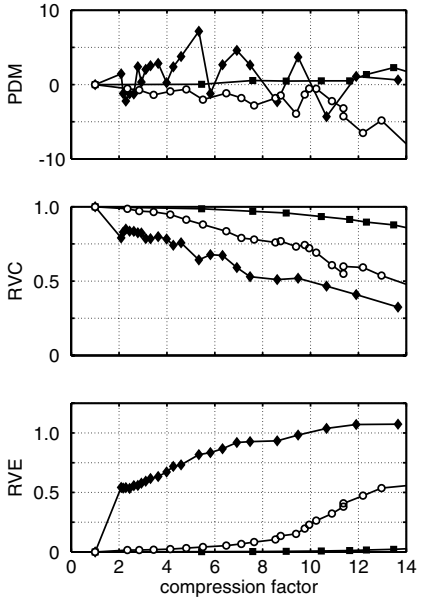

Fig. 4. Statistical measures as a function of CF for data set 1 (circles), data set 2 (squares) and data set 3 (diamonds).

The sum of the measures of error variance (RVE) and compressed signal variance (RVC) was not 1. Thus there exists a correlation between the part of the signal deleted during compression and the compressed signal itself. The implication for data-driven methods such as inferential sensing is that some informative features have been thrown away or that some unwanted features have been retained.

5.3. Non-linearity assessment

Table 1 shows results from non-linearity assessment of the three data sets. It shows that compression induces non-linearity in the signal because two of the three data sets (data 1 and data 3) were linear in their uncompressed state (CF = 1) but became non-linear after compression and reconstruction when the compression factor exceeded 3. Compression is a non-linear operation and the principle of superposition does not apply, i.e.

$$g(x_1(t)) + g(x_2(t)) \neq g(x_1(t) + x_2(t))$$

and

$$g(a \times x_1(t)) \neq a \times g(x_1(t))$$

where $x_1(t)$ and $x_2(t)$ are time domain signals, $g(x(t))$ is a compressed time trend and a is a scalar factor. In the case of swinging door compression, if the signal were twice as large then the compressed signal would not merely be twice as large at the retained spot values but it also would have more piecewise linear segments because more spot values would hit the condition for archiving.

The use of compressed archived data to assess non-linearity, for instance in an audit of control valves, may be misleading. Time may be wasted in inspection and testing of valves that are in fact operating normally.

Table 1
Non-linearity at various compression factors (CF)

CF	Data 1	Data 2	Data 3
1	No	Yes	No
2	No	Yes	No
3	No	Yes	No
4	Yes	Yes	Yes

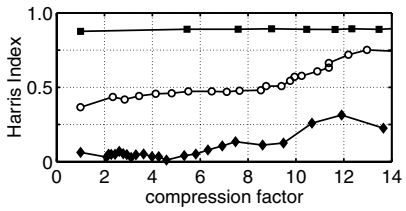

Fig. 5. Harris index as a function of compression factor for data set 1 (circles), data set 2 (squares) and data set 3 (diamonds). In this formulation, 0 represents minimum variance performance and 1 means poor performance.

5.4. Harris index

Fig. 5 shows the Harris index results, where 0 represents good performance close to minimum variance and 1 represents poor performance. The index increases with compression for all the data sets.

It is concluded in the case of data sets 1 and 3 that compression increases the predictability of the signal and thus affects the Harris index. Data set 2 was inherently predictable (see Section 3.3). The Harris index for data set 2 therefore indicated poor performance even in the uncompressed case and did not change as much on compression as for the other two data sets.

5.5. Summary of performance

The following comments focus the results and discussion of the previous section onto the industrial requirements.

- Data compression changes the statistical properties of the data. Fig. 4 suggests that even a small amount of compression has an effect on the mean and variance.
- Averaging, data reconciliation and mass balancing applications should calculate the required quantities directly from the original data because data archived with swinging door compression have a different mean value after reconstruction. This could have serious implications, for instance if the reconstructed data represented oil flow from an off-shore facility being monitored for taxation purposes.
- High fidelity reconstruction requires that the statistical properties of the reconstructed signal are similar to those of the original. Minimum variance and non-linearity assessment are two procedures that require high fidelity data. Swinging door compression alters these measures significantly.
- Data smoothing, feature extraction and reconstruction of events require the events and features of interest to be retained during compression. The non-orthogonality of piecewise linear trending means that the condition is not met because the reconstruction error is correlated with the reconstructed signal. Thus, for instance, the magnitude of a transient event may not be reconstructed accurately.

The performance measures for data set 3, which was a random signal, were influenced by compression even at small compression factors. For instance, the RVE and RVC measures changed significantly. No random value is any more significant than any other, but the compression algorithm makes some points more significant by choosing to archive them and therefore the reconstructed signal does not have the same randomness. The performance measures for data sets 1 and 2, however, did not change as much for small compression factors up to about 3. For instance, in Fig. 4 the results for a CF of 3 were very similar to those for the uncompressed case when CF was 1, while Table 1 shows that non-linearity was not induced for a CF of 3 or less. Therefore a heuristic rule is proposed so that at least some compressed archived data can be exploited:

Data having CF $\leqslant 3$ may be used with caution for data-driven process analyses.

It is noted, however, that certain types of process trends may allow for higher compression factors because their intended use is to record constant values such as set points, targets and high and low limits.

6. Automated detection of compression

6.1. Motivation

The previous discussion showed that compression induces changes to many of the quantities commonly used in data-driven process analyses. However, engineers are not always in a position to examine data closely enough to detect compression because plotting and examining the time trends is time consuming. Rather, data pre-processing activity usually focuses upon finding and replacing bad data such as missing values. If archived data are to be used for an automated analysis it is first necessary to test for the presence of compression.

If the number of spot values in the compressed archive and the original sampling rate are known then the compression factor may be determined by calculation as the ratio between the expected number of observations and the number of archived observations. However, such information is not always available when data sets are sent off-site to consultants or universities, and it may be necessary to estimate the compression factor from the reconstructed data only. An automated method for detection of piecewise linear compression is now presented and some guidelines given for its application to industrial data.

6.2. Compression detection procedure

Since the reconstructed data set is piecewise linear, its second derivative is zero everywhere apart from at the places where the linear segments join. Therefore the presence of the characteristic linear segments can be detected by counting of zero-valued second differences $\Delta(\Delta\hat{y})$ calculated from

$$\Delta(\Delta\hat{y})_i = \frac{(\hat{y}_{i+1} - \hat{y}_i)/h - (\hat{y}_i - \hat{y}_{i-1})/h}{h} = \frac{\hat{y}_{i+1} - 2\hat{y}_i + \hat{y}_{i-1}}{h^2}$$

where $\hat{y}$ is the reconstructed signal and h is the sampling interval. The index i ranges from 2 to $N-1$, where N is the number of samples. Suppose the original data set had N values and after compression there are m archived spot values and $m-1$ linear segments. If the reconstructed data are differenced twice there will be $n = N - m$ second differences whose values are zero. Therefore the compression factor can be determined from

$$\mathrm{CF}_{\mathrm{est}} = \frac{N}{m}$$

where $m = N - n$. For example, with 10 data points compressed to four archived values and three linear segments there are $10 - 4 = 6$ second differences whose values are zero.

The method can be extended to other piecewise reconstruction methods using polynomials. For instance, if cubic spline compression were in use [18] the fourth derivatives would be zero everywhere except at the knot points where the splines join. In that case the compression factor would be determined from the number of fourth differences having zero values.

Fig. 6 shows results for data sets 1, 2 and 3. The compression factor derived from counting the zero second differences was a good estimate of the true compression factor.

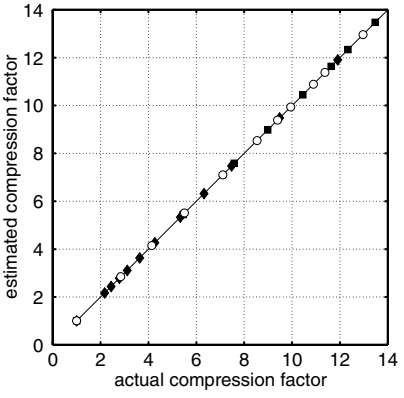

Fig. 6. Results from the compression estimation algorithm for data sets 1–3.

6.3. Implementation considerations

Enhancements are needed to the basic algorithm for industrial implementation for the following reasons:

- The sampling interval of the reconstructed signal may be larger than the original; e.g. the compression algorithm may have used 10 s samples but the reconstruction may use 1 min samples.
- The effects of finite precision arithmetic mean that some computed second differences may not be exactly zero.

Suggestions for handling these cases are given here and illustrated with the industrial data of Fig. 1.

Dealing with a larger sampling interval: It is recommended to reconstruct compressed data using the same sampling interval as the original because reconstruction with a longer sampling interval leads to an underestimate of the compression factor. For instance, if five 10 s samples out of 120 were archived then the true compression factor is 24 ($m = 5$, $N = 120$). When the data are reconstructed using 1 min samples the number of piecewise linear segments does not change but there are only 20 samples in the reconstructed data so the compression factor appears to be 4. An effect of reconstruction with a longer sampling interval is that the true end points of the piecewise linear segments may fall between samples. Thus x_i would be the end of one linear segment, x_{i+1} would be the start of the next with the true end point somewhere in between. The effect on the second differences is that there are two non-zero second differences where the linear segments join instead of the one that would be expected. The presence of these pairs of non-zero second differences can be used as a warning of a sampling interval issue. If such pairs are detected then the calculation of the compression factor has to acknowledge that each pair represents just one true archived point and the expression for the compression factor is modified to

$$\mathrm{CF}_{\mathrm{est}} = \frac{N}{m/2}$$

Such pairs were detected in the industrial data of Fig. 1 and therefore the modified expression was used in the calculation. The estimated compression factors are shown on the right of Fig. 1. For instance, tag 20 has a compression factor of 41.7. It had 1428 zero second differences, 72 non-zero second derivatives in 36 pairs and 36 linear segments.

If the characteristic pairs are noticed then a warning must be given that the compression factors are under estimates and to reconstruct at the correct rate. Fig. 1 showed such a warning.

Finite precision arithmetic: A procedure was developed to deal with the effects of finite precision arithmetic.

The numerical values of the second differences were converted to integers. The *ceil* function in the following expressions rounds up to the next integer:

$$P = \text{ceil}(\log_{10}|x|)$$

$$y = x/10^{P}$$

$$z = y \times 10^{R}$$

x is the original entry in the data base having R significant figures, $P-1$ of which are to the left of the decimal point (e.g. $P=5$ and $R=10$ in 1478.144165). y has the same digits as x but has a zero to the left of the decimal point (e.g. 0.1478144165) and z is an integer with the same digits (e.g. 1478144165). The second difference calculations were applied to the integers z.

Some computed second differences may not be exactly zero because of arithmetic rounding errors. With the integer transformation above it would be expected that the errors would be ±1. It was observed, however, that errors of up to ±500 were present. That is to say, the precision of the arithmetic used by the data historian in the reconstruction was less than 10 significant figures although the results were reported to 10 significant figures. The following sequence illustrates the pattern of second differences observed in z for a portion of a straight line trend in tag 7 of the industrial data of Fig. 1.

$$\left\{\begin{array}{cccccccccc} -476, & 477, & 0, & -477, & 477, & 0, & -1, & -476, & 477, \\ 0, & 0, & -477, & 477, & 1, & 477, & -1, & -476, & 477, & 0 \end{array}\right\}$$

Any second difference in z whose absolute value was below 500 was counted for calculation of the compression factor.

If the data historian complies with a published numerical Standard (e.g. IEEE 854-1987) then the threshold for second differences may be determined from the Standard. Otherwise the threshold must be determined by observation of the arithmetical precision achieved and the number of significant figures in use, as was done here. There is no fundamental significance to the numerical value of ±477 in the example presented above. The observed rounding errors appear to arise from an interplay between the original data values, the arithmetic precision and the details of the data base.

Final recommendation: It has been demonstrated earlier that data with CF > 3 are not suitable for data-driven analyses. Compression factors of up to 93.8 were present in the industrial data set of Fig. 1 and only five tags had compression factors of three or below. It was concluded that this archived data set would not be suitable for data-driven analysis. Moreover, the algorithm issued a warning that the compression factors were underestimated. For improved estimates of compression factor the data set should be reconstructed with the original sampling interval. The reconstruction was not attempted because it was already clear that the data were much too compressed for data-driven analyses.

7. Conclusions

Time and frequency domain plots were presented for data from continuous processes to show how well the trends were reconstructed after compression. Piecewise linear compression using the swinging door algorithm altered key statistical features of the data set such as the average value and standard deviation. Other data-driven analyses were also altered.

A procedure was presented for detection of compression during the pre-processing stage of a data-driven analysis, together with additional features required for its application to industrial data. An expression based upon counting of zero-valued second derivatives gave a lower bound for the compression factor. It is important for an accurate assessment of the compression factor to reconstruct the data at the original sampling interval. If the reconstruction interval is longer than the original then characteristic pairs of non-zero second derivatives are noticed and a warning must then be given to reconstruct at the correct rate.

On the basis of the findings in this paper the authors strongly recommend caution in the use of compression in process data archives. It is noted that pressure from customers, together with the cheaper costs of storage, are now making an impact and that newer data historians (e.g. AspenWatch) are starting to use uncompressed data. It is hoped that the work reported in this paper will provide end-users wishing to eliminate the use of data compression with some solid, quantified reasons for doing so.

Acknowledgements

Nina Thornhill gratefully acknowledges financial support from the Royal Academy of Engineering (Foresight Award). Financial support to Shoukat Choudury, in the form of a CIDA scholarship from the Canadian government, is gratefully acknowledged. The project has also been supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), Matrikon Inc., and the Alberta Science and Research Authority (ASRA) in the form of an NSERC-Matrikon-ASRA Industrial Research Chair Program at the University of Alberta.

References

[1] J.P. Kennedy, Data treatment and applications—future of the desktop, in: Proceedings of FOCAPO, CACHE, 1993.

[2] AspenTech, Analysis of data storage technologies for the management of real-time process manufacturing data, Retrieved July 19th 2003, from http://www.advanced-energy.com/Upload/symphony_wp_infoplus.pdf, 2001.

[3] OSI Software Inc., PI data storage component overview, Retrieved July 19th 2003, from http://www.osisoft.com/ 270.htm, 2002.
[4] A.B. Watson (Ed.), Digital Images and Human Vision, MIT Press, 1993.
[5] J.A. Crowe, N.M. Gibson, M.S. Woolfson, M.G. Somekh, Wavelet transform as a potential tool for ECG analysis and compression, Journal of Biomedical Engineering 14 (1992) 268–272.
[6] W. Philips, ECG data compression with time warped polynomials, IEE Transaction on Biomedical Engineering 40 (1993) 1095–1101.
[7] M. Karczewicz, M. Gabbouj, ECG data compression by spline approximation, Signal Processing 59 (1997) 43–59.
[8] R.S.H. Mah, A.C. Tamhane, S.H. Tung, A.N. Patel, Process trending with piecewise linear smoothing, Computers and Chemical Engineering 19 (1995) 129–137.
[9] M.J. Watson, A. Liakopoulos, D. Brzakovic, C. Georgakis, A practical assessment of process data compression techniques, Industrial and Engineering Chemistry Research 37 (1998) 267–274.
[10] J.C. Hale, H.L. Sellars, Historic data recording for process computers, Chemical Engineering Progress 77 (November) (1981) 38–43.
[11] E.H. Bristol, Swinging door trending: adaptive trend recording, in: ISA National Conference Proceedings, 1990, pp. 749–753.
[12] D.L. Donoho, M. Vetterli, R.A. DeVore, I. Daubechies, Data compression and harmonic analysis, IEEE Transactions on Information Theory 44 (1998) 2435–2476.
[13] Z. Nesic, M. Davies, G. Dumont, Paper machine data analysis and compression using wavelets, Tappi Journal 80 (1997) 191–203.
[14] V.K. Raghavan, J.R. Whiteley, Wavelet representation of sensor patterns for monitoring and control, in: AIChE Annual Meeting, St Louis, Paper 150b6, 1993.
[15] B. Bakshi, G. Stephanopoulos, Compression of chemical process data through functional approximation and feature extraction, AIChE Journal 42 (1996) 477–492.
[16] M. Misra, S. Kumar, S.J. Qin, D. Seeman, Error based criterion for on-line wavelet data compression, Journal of Process Control 11 (2001) 717–731.
[17] M. Misra, S. Kumar, S.J. Qin, D. Seeman, On-line data compression and error analysis using wavelet technology, AIChE Journal 46 (2000) 119–132.
[18] H. Vedam, V. Venkatasubramanian, B. Bhalodia, A B-spline based method for data compression, process monitoring and diagnosis, Computers and Chemical Engineering 22 (1998) S827–S830.
[19] P.D. Welch, The use of fast Fourier transforms for the estimation of power spectra, IEEE Transactions on Audio and Electroacoustics AU 15 (1967) 70–73.
[20] G.D. Martin, L.E. Turpin, R.P. Cline, Estimating control function benefits, Hydrocarbon Processing 70 (June) (1991) 68–73.
[21] J.P. Shunta, Achieving World Class Manufacturing Through Process Control, Prentice-Hall, NJ, 1995.
[22] M.A.A.S. Choudhury, S.L. Shah, N.F. Thornhill, Diagnosis of poor control loop performance using higher order statistics, Automatica, in press.
[23] H.E. Emara-Shabaik, J. Bomberger, D.E. Seborg, Cumulant/bispectrum model structure identification applied to a pH neutralization process, in: IEE Proceedings, UKACC International Conference on Control'96, UK, 1996, pp. 1046–1051.
[24] J.P. Barnard, C. Aldrich, M. Gerber, Identification of dynamic process systems with surrogate data methods, AIChE Journal 47 (2001) 2064–2075.
[25] M.A.A.S. Choudhury, S.L. Shah, N.F. Thornhill, Detection and diagnosis of system nonlinearities using higher order statistics, in: IFAC World Congress, 2002.
[26] L. Desborough, T. Harris, Performance assessment measures for univariate feedback control, Canadian Journal of Chemical Engineering 70 (1992) 1186–1197.
[27] M.A. Paulonis, J.W. Cox, A practical approach for large-scale controller performance assessment, diagnosis, and improvement, Journal of Process Control 13 (2003) 155–168.
[28] L. Desborough, R. Miller, Increasing customer value of industrial control performance monitoring—Honeywell's experience, in: AIChE Symposium Series No. 326, vol. 98, 2002, pp. 153–186.
[29] P. Fedenczuk, P. Fountain, R. Miller, Loop Scout, RPID and Profit Controller team up to produce significant benefits for BP, Retrieved July 19th 2003, from http://loopscout.com/loopscout/info/bpamoco.pdf, 1999.
[30] N.F. Thornhill, M. Oettinger, P. Fedenczuk, Refinery-wide control loop performance assessment, Journal of Process Control 9 (1999) 109–124.
[31] G.B. Wetherill, D.W. Brown, Statistical Process Control, Chapman and Hall, London, 1991.