Feature extraction in control chart patterns with missing data

R Haghighati[1] **and A Hassan**[1]

[1]Department of Materials, Manufacturing & Industrial Engineering, Faculty of Mechanical Engineering, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, Malaysia

Abstract. Data preprocessing and feature extraction are critical steps in control chart pattern (CCP) recognition for reducing dimensionality and irrelevant information. To ensure good quality of input representation, it is important to handle missing values on control charts before feature extraction. Excluding missing values and imputing them with plausible values are two common missing data handling approaches in the literature. In this paper imputation capability of exponentially weighted moving average (EWMA) was investigated. Incomplete process data for three missingness mechanisms namely, missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) were investigated using computer simulation. Missing data at severity levels i.e. 0, 5, 10, 15, 20, 25 and 50 % were evaluated. The investigation covers feature mean, standard deviation, skewness, kurtosis and quartiles extracted from imputed patterns. The imputation performance was measured by comparing the deviation between full patterns and patterns with missing values in term of mean square error (MSE). The results show that EWMA imputation was highly reliable to recover missing values as evident form low feature deviations, MSE values; 0.04 (random), 0.04 (trend-up), 0.3 (shift-up) and 0.5 (cycle) respectively. The results suggest that EWMA imputation technique is superior than the mean and median imputations.

1. Introduction

Control chart pattern (CCP) recognition has been widely used to identify assignable causes, detect process disturbances and equipment malfunction through studying abnormal patterns in control charts. To automate the recognition, artificial neural networks (ANN) has been used. In ANN, either raw data or features drawn from the data provide the input and potential process patterns defined expected output [1]. Replacing raw data with features has several advantages such as providing better interpretation of the prediction model; improving the performance and efficiency of a learning process by removing irrelevant and redundant data, and reducing dimensionality [2]-[4].

Facing missing samples or missing individual observations within samples is very common in process monitoring [5]. Human error and misunderstanding, equipment malfunctioning and faulty data transmission can cause data corruption and missing during the whole process of data collection, storage, and preparation [6]. Sensors commonly experience faults and communication errors. Some data are visible or hidden by noise or an item in the sample could be defective and unable to be measured. As a result, data may not be reported for a region causing missingness [5].

Control charting is all about interpretation by statistical analysis. The presence of missing data, however, brings an inevitable uncertainty to statistical analysis [7]. Control chart is created by plotting mean (or other

statistic) of samples taken from the process versus sample time or sample number. Then, each plotted point is compared with the control limits. If it falls within the control limits, then it shows that the variability is due to common causes and that the process is in control. Otherwise it provides strong evidence that a special cause is present, and immediate action should be taken to find and remove this special cause.

Missing values can be estimated using data imputation techniques to fill the gaps and to obtain complete data set [8]. Mahmoud *et al*. [9] discussed the effect of four imputation methods namely mean imputation, regression, stochastic regression and the expectation maximization algorithm for estimating Phase I historical data set in control charts and then estimated the unknown parameters in the Hotelling's T^2 chart statistic. They reported that the stochastic regression method has the best overall performance among all the competing methods.

Unlike the phase I that deals with historical samples, phase II control chart is maintained by online data. In a study on estimating the parameters from a "treated Phase I data set" on the performance of the control chart in Phase II, missing data handling techniques, such as complete case, mean substitution, regression, stochastic regression, and the expectation–maximization algorithm methods were evaluated [10]. Simulation results showed that imputation is preferred over case deletion methods. In particular, the regression-based imputation gained the best overall performance.

However, Waterhouse [11] argued that imputing with sample mean or regression methods may destroy the relationships between variables and reduce variability. They also advised against deleting incomplete records and suggested multiple imputation when a large amount of imputation is required.

Missing data literature in process monitoring seek the answer to the question that which missing data handling technique can ensure that control chart behave in the same way either by the incomplete or complete data? Until now, influence of missing data was only focused on primary function of control charts i.e. testing the hypothesis that the process is still in statistical control. However, control charts can be used for process diagnosis too. Unstable processes may also produce CCPs such as cyclic, linear trend, sudden shift, mixtures, stratification and systematic when plotted on a Shewhart X-bar chart. Features extracted from incomplete CCPs may have deviated feature values, so that those features cannot characterize the originated CCP effectively, they have lower discrimination power and when they function as input for classifiers, classification performance is degraded.

Although few recent works have addressed missing data in control charts, we were unable to locate published investigation on effect of missing data in univariate CCPs. The objective of this paper is to study the effect of four imputation approaches in feature retrievability in CCPs with incomplete data. An imputation technique based on EWMA is proposed and its performance in feature retrievability is compared with common missing data handling techniques.

The paper proceeds as follows. In Section 2 we outline data simulation followed by feature selection and EWMA imputation. In Section 3, major results and discussions are elaborated. Section 4 provides conclusion and recommendations.

2. Materials and method

2.1 Data simulation

One of the important criteria for choosing suitable approach of handling missing data is to investigate how data samples have gone missing which is called missingness mechanism in the literature [12]. Missingness mechanism contains very informative assumptions regarding missing data for example determine whether missingness can be ignored. Little and Rubin [12] introduced three major mechanisms. If the cause of missingness is independent of data, missingness is called missing completely at random (MCAR). For instance, MCAR explains a situation when container of a liquid sample breaks by accident and thus the sample cannot be measured. On the other hand, in missing at random (MAR) mechanism, missingness

depends on data which is observed yet independent of the unobserved data.

The example of MAR is power outage that leads to occasional failure of a sensor. In this example, the cause of the incomplete data is not the actual missing variables but some other external influence. Finally, third mechanism termed missing not at random (MNAR) because the pattern of missing data is non-random and depends on the missing variable. If a sensor cannot acquire information outside a certain range, data are missing due to MNAR factors. No general method of handling missing data has been identified for the third scenario in which data that could provide valuable information is lost [12]. Statistical definition of described missingness mechanisms are shown in table 1, by defining whether probability of missingness (M) is related to probability of observed values (Y_o) and missing values (Y_m) in each respective missingness mechanism. Missing mechanism is ignorable, when data are MCAR or MAR, which means analysis of data can be done irrespective of reasons data were missing.

This study focused on four types of x-bar chart patterns, namely, in-control, trend-up, shift-up and cyclic patterns. In-control process represents process which is in the state of statistical control. Cyclic pattern was included to show stationary mean pattern while trend-up and shift-up patterns represented non-stationary mean patterns [13].

Table 1. Statistical definition of four missingness mechanisms.

Missingness Mechanism	Abv.	Statistical Definition
Missing at Random	MAR	$P(M \mid Y_o, Y_m) = P(M \mid Y_o)$
Missing Completely at Random	MCAR	$P(M \mid Y_o, Y_m) = P(M)$
Missing Not at Random	MNAR	$P(M \mid Y_o, Y_m)$ = cannot be computed

MATLAB 7.12.0 (R2011a) was used to simulate patterns with MCAR, MAR and MNAR mechanisms. A dataset of size 30 dependent variables (Y) was created using explanatory variables x_1, x_2, x_3 with added random components according to [14]

$$Y(t) = x_1(t) + 2x_2(t) + 3x_3(t) \quad (1)$$

The dataset (Y) was standardized (Z_t) and then process patterns were simulated according to commonly investigated methods in the CCPR. The parameters for simulating individual process data are given in table 2. These parameters are commonly used by other researchers [13].

Table 2. Parameters for simulating individual process data.

Pattern Types	Parameters (symbol)	Value
Trend-up	Gradient (s)	0.015 to 0.025
Shift-up	Magnitude (h)	0.7 to 2.5
Cyclic	Amplitude (c1, c2)	0.5 to 2.5
	Period (T)	10
Stable Process	Baseline Noise (b)	1/3
	Standardized	N (0,1)

Process patterns were simulated based on the following models (Equation 2, 3, 4 and 5), in which μ and σ stands for mean and standard deviation of process when it is in the state of statistical control. Z_t is the dependent variable based on three explanatory independent variables which was standardized and generated normal variate at time t.

$$X_{\text{in-control}} = \mu + (b\, Z_t\, \sigma_x) \quad (2)$$

$$X_{\text{trend-up}} = \mu + s\,(t-t_0)\,\sigma x + (b\, Z_t\, \sigma x) \quad (3)$$

$$X_{\text{shift-up}} = \mu + h\,\sigma_x + (b\, Z_t\, \sigma_x) \quad (4)$$

$$X_{\text{cyclic}} = \mu + (b\, Z_t\, \sigma_x) + c_1\sigma_x \text{Cos}\,[2\pi\,(t-t_0)/T] + c_2\sigma_x \text{Sin}\,[2\pi\,(t-t_0)/T] \quad (5)$$

To simulate MCAR mechanism, certain ranges namely, 0, 5, 10, 15, 20, 25, and 50 % of data were randomly deleted. Data was sorted according to one of the x explanatory variables (e.g. x_1), and the upper values were deleted in several rates to represent MAR data. To represent MNAR, data was sorted according to the actual pattern values and the extreme values were deleted at each respective range[14]. Incomplete raw data or contaminated statistical features extracted from incomplete raw data can degrade recognizers' performance.

2.2 Feature selection

Eight statistical features namely mean, standard deviation, skewness, kurtosis and three quartiles were selected to investigate the sensitivity to missing data and missing mechanism in process patterns contaminated with 0, 1, 5, 10, 15, 20, 25 and 50% missing data. The values of these features were calculated for complete dataset and incomplete dataset with 'treated' missing data. Built-in functions in MATLAB was used to calculate seven statistical features in various combinations of four missing mechanism and seven missing rates for four types of CCPs.

2.3 EWMA imputation

Exponential smoothing has been one of the most common forecasting methods for more than six decades [15]. It is known to be a simple and transparent approach with solid theoretical foundation [16]. The principle idea of EWMA i.e. 'using weighted average of values in a period' was used to develop an imputation technique. Forecast of all data using exponential smoothing provided a back-up for data. Then, each missing instance is replaced by the forecasted value and this process continue to the next missing values until all the missing values in the pattern are replaced by EWMA estimates F_t (6).

$$F_t = \alpha\, A_{(t-1)} + (1-\alpha)\, F_{(t-1)} \quad (6)$$

where F_t and A_t stand for forecasted and actual data at time t and the smoothing factor, α ranges between 0 and 1. This study selected α=0.4 for in-control pattern and α=0.7 for abnormal CCPs including trend-up, shift-up and cyclic patterns. These values are selected by comparing several EWMA imputations with various alpha values to select which alpha value result in better estimation and the lowest mean square error (MSE) in each CCP.

Recent changes in the data have greater influence for values of α close to one. On the other hand, smoothing effect is greater and recent changes are less important if values of α closer to zero are selected. The results indicated that good estimation of missing data in stable pattern depends on smoothed historical data, while missing data in out of control patterns are better estimated by recent changes in the data. Another important parameter in good estimation is choosing a right initial estimate $F_{(1)}$. It was assumed that the initial estimate in all patterns is equal to the target value i.e. $F_1 = 0$.

Performance of EWMA imputation technique in estimating process features were compared with mean

and median imputations as well as case deletion (i.e. excluding missing data). In both mean and median imputations, the values of missing samples were replaced by the mean/median value of the observed data, respectively. Case deletion and imputation based on mean and median have been investigated extensively in the literature and they have been standard default options of major statistical software such as SPSS, SOLAS, S-PLUS, SAS, BMDP and BUGS, among others [8], [9], [14]. The performance of missing data handling techniques was measured by mean square error (MSE) in which estimation error of selected features from the actual complete dataset was compared in various techniques. The best technique for handling missing data is the one that result in the least estimation error (MSE) in various missing data properties.

3. Results and discussion

Performance of missing data handling techniques in stable pattern is shown in figure 1. In some previous researches [17], [18] mean and median imputations were grouped together with regards to their estimation power. However, we observed their retrievability vary by type of pattern, mechanism and amount of missing values.

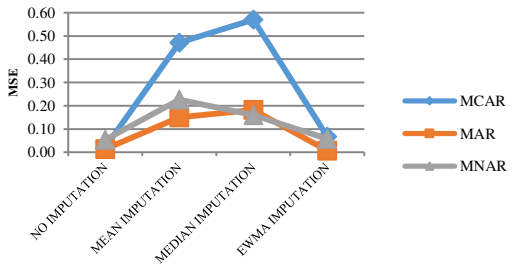
It was observed that skewness and kurtosis are the most challenging features to estimate in patterns with 5 to 50 % missing data. In trend-up pattern the treated feature kurtosis with all techniques but EWMA imputation resulted in poor estimation. EWMA imputation estimated the kurtosis almost correctly with low MSE value of 0.01.

The mechanism by which the data have gone missing is one of the determinant factors for choosing correct imputation technique. Estimation was very poor when data were MCAR in compare to patterns with MAR and MNAR mechanisms, particularly when mean and median imputations were applied.

The results indicated regardless of the missing data handling technique, features are better estimated if data are missing at random (MAR) in compare to scenarios in which data missing by other mechanisms. This result confirms previous works by Scheffer [14]. Her study on two statistics i.e. mean and standard deviation changes in normal patterns which are contaminated with missing data, showed that at 50% missing rate, the mean value deviated 20% when missing mechanism was MNAR, but mean was only deviated 6% with MAR mechanism. Also, it was observed that the standard deviation deviated by 7% under MAR, but it deviated 38% under MNAR when 50% of values in a normal dataset are missing [14].

It was observed that in most of the scenarios of missing data and with all feature values, mean and median imputation techniques resulted in subordinated performance. The reason is that mean and median imputation techniques corrupt variability of data by replacing all the missing values with the same central value regardless of considering the place where the missingness occurred. But with EWMA imputation, samples weights have a geometrically decreasing order in which higher weights belongs to the most recent samples while little weight is distributed among the most distant samples [19].

1-a



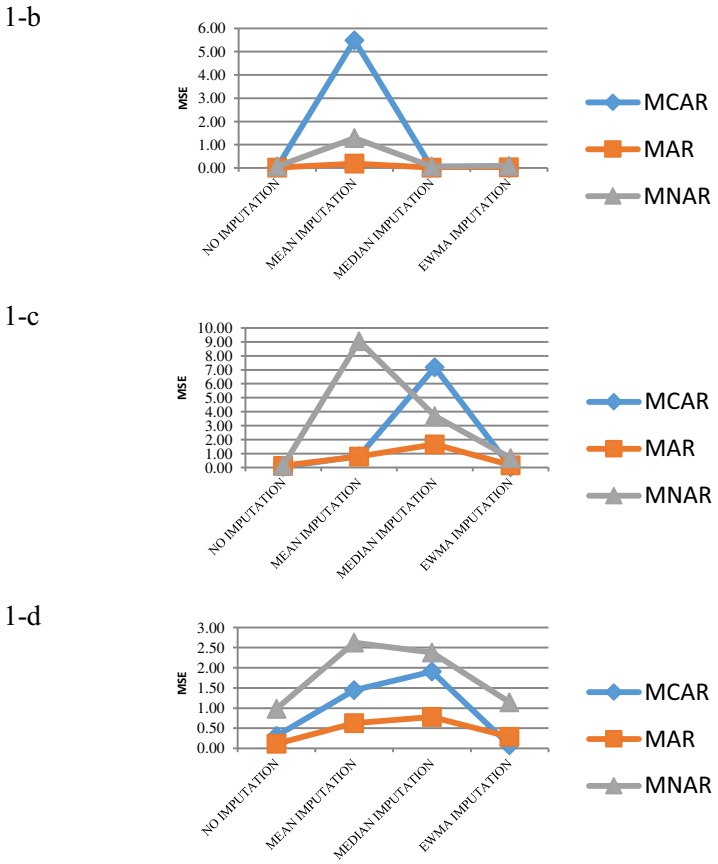


Figure 1. Comparison of imputation performance for the in-control(1-a), trend-up(1-b), shift up(1-c) and cyclic (1-d) pattern by mean squared error (MSE).

Thus, EWMA imputation maintained the dynamic behavior and resulted in better estimations. Moreover, EWMA provided unique estimation for each missing value that differs across the various missing values.

Scheffer [14] reported that both mean imputation and no imputation should be avoided for normal pattern. This study confirms her results regarding normal pattern. However, in abnormal patterns avoiding imputations and estimating feature values directly from available data, sometimes better maintain approximate values of statistical features, particularly when the missing rate is high in the abnormal pattern.

Level of missing data also affects the performance of imputation. This result is in line with literature that reported more than 15% missing observations in dataset severely impact any kind of interpretation [20]. This study provided an extension to analysis of normal pattern in Scheffer's work [14]. We extended her work to non-normal patterns namely, trend-up, shift-up and cyclic patterns. Moreover, we compared skewness, kurtosis and three quartiles in addition to mean and standard deviation and we proposed and evaluated performance of EWMA as a promising imputation technique.

4. Conclusions

Incomplete data in terms of missing, faulty or delayed values in process monitoring influence statistical features and reduce recognition performance. An imputation technique based on EWMA statistic was proposed to handle missing values in CCPs which are contaminated by missing data. Its performance was

evaluated by estimation error (MSE) of seven feature values, namely mean, standard deviation, skewness, kurtosis and three quartiles. Features extracted from incomplete patterns from three missing mechanisms, namely MCAR, MAR and MNAR and seven missing percentage ranging from zero to 50. The performance of EWMA imputation was then compared with two of the commonly used imputation techniques namely mean and median imputation as well as the alternative of no imputation. Main findings are the followings:

EWMA imputation is a promising imputation technique particularly for incomplete patterns with missing-at-random (MAR) mechanism.

Restricting inference to complete data is the best strategy for handling missing data in incomplete patterns with MNAR mechanism.

Skewness and the kurtosis are the most challenging features to estimate particularly in when missingness mechanism is MNAR and MCAR. Estimation error increases sharply as the rate of missing data increases.

Finally, the major practical implication of this research was to measure performance of missing data handling techniques by comparing actual feature values with treated feature values. Majority of works in the literature, however, compared actual raw data with imputed raw data to compare performance of various imputation techniques. Comparing feature values rather than raw data has the advantage of surpassing the normal variation (noise) and emphasizing on whatever change the statistical characteristic of pattern plagued by missing data. Maintaining actual feature values in control chart patterns with incomplete data is crucial for good input representation and successful recognition.

Future research includes investigation on non-determinant approaches for missing data approximation such as interpolation, extrapolation and likelihood based probabilistic modeling. Further investigation on multivariate process patterns with missing data is highly suggested. Studying shape features is another natural extension to the present work. The authors are currently studying proposed technique in real-life scenarios and extending the model to online and partially developed patterns as further steps.

5. References

[1] W Hachicha and A Ghorbel 2012 Computers & Industrial Engineering 63, 1
[2] S Abdullah, N R Sabar, M Z Ahmad Nazri and M Ayob 2014 Computers & Industrial Engineering 67
[3] S K Gauri and S Chakraborty 2009 Computers & Industrial Engineering 56, 4
[4] I Masood and A Hassan 2010 European Journal of Scientific Research 39, 3
[5] S R Wilson 2009 in Control Charts with Missing Observations, Vol. PhD.
[6] M G Rahman and M Z Islam 2015 Knowledge and Information Systems
[7] S C Chuang, Y C Hung, W C Tsai and S F Yang 2013 Computers & Industrial Engineering 64, 1
[8] E L Silva-Ramírez, R Pino-Mejías and M López-Coello 2015 Applied Soft Computing 29, 0 **(2015)**
[9] M A Mahmoud, N A Saleh and D F Madbuly 2014 Quality and Reliability Engineering International 30, 4
[10] D F Madbuly, P E Maravelakis and M A Mahmoud 2013 Communications in Statistics-Simulation and Computation 42, 6
[11] M Waterhouse, I Smith, H Assareh and K Mengersen 2010 Int J Qual Health C 22, 5
[12] R J Little and D B Rubin 2002 Statistical analysis with missing data, John Wiley, New York
[13] A Hassan, M S N Baksh, A M Shaharoun and H Jamaluddin 2003 International Journal of Production Research 41, 7
[14] J Scheffer 2002 Research Letters in the Information and Mathematical Sciences 3
[15] M A S Monfared, R Ghandali and M Esmaeili 2014 Journal of Industrial Engineering International 10, 4
[16] R Hyndman, A B Koehler, J K Ord and R D Snyder 2008 Forecasting with exponential smoothing: the state space approach, Springer Science & Business Media
[17] M Grzenda, A Bustillo and P Zawistowski 2012 Journal of Intelligent Manufacturing 23, 5
[18] M R Malarvizhi and D A S Thanamani 2012 International Journal of Engineering Research and

Development 5
[19] D C Montgomery 2007 Introduction to statistical quality control, John Wiley & Sons
[20] E Acuna and C Rodriguez 2004 The treatment of missing values and its effect on classifier accuracy, Springer pp. Springer.

6. Acknowledgements

This research was supported by Research Management Center Universiti Teknologi Malaysia through GUP Grant Q.J130000.2624.12J68. This support is gratefully acknowledged.