Jurnal Teknologi

Full paper

Quality Control in Cocoa Powder Production Process: A Robust MSPC Approach

S. L. Lee[a*], M. A. Djauhari[a]

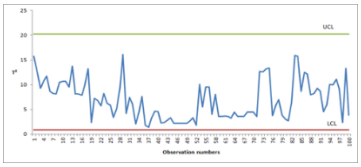
[a]*Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, johor, Malaysia*

*Corresponding author: leesiawli@gmail.com

Article history

Received : 21 January 2013
Received in revised form :
7 May 2013
Accepted : 25 June 2013

Graphical abstract



Abstract

To monitor a multivariate process mean, Hotelling's T^2 control chart is often used. However, the presence of multiple outliers may go undetected due to the masking effect or swamping effect. In this study, we propose a robust Hotelling's T^2 control charts where the mean vector and the covariance matrix are estimated by using fast minimum covariance determinant (FMCD) which gives a high breakdown point estimates. This study found that the latter approach performs far better than the former in terms of the ability in detecting an out-of-control situation during the start-up stage. We present and discuss our experience in monitoring the process mean of cocoa powder production process in a Malaysian company located in Johor Bahru.

Keywords: Control chart; statistical process control; multivariate normal process; robust estimation; fast minimum covariance determinant

Abstrak

Untuk memantau min proses multivariat, carta kawalan T2 Hotelling sering digunakan.Walau bagaimanapun, kehadiran pelbagai titik terpencil mungkin tidak dapat dikesan oleh kesan pelekat. Dalam kajian ini, kami mencadangkan carta kawalan T2 Hotelling di mana vektor min dan matriks kovarians adalah dianggarkan dengan menggunakan minimum kovarians penentu (FMCD) yang memberikan anggaran titik kerosakan yang tinggi. Kajian ini mendapati bahawa pendekatan kedua jauh lebih baik daripada yang pertama dari segi keupayaan untuk mengesan keadaan di luar kawalansemasa peringkat permulaan. Kami hadir dan membincangkan pengalaman kami dalam memantauproses pengeluaran serbuk koko di sebuah syarikat Malaysia di Johor Bahru.

Kata kunci: Carta kawalan; kawalan proses statistik; proses normal multivariat; anggaran teguh; penentu kovarians minimum

1.0 INTRODUCTION

The concept of quality has long been understood and defined by many people as the way a physical product compared to some defined ideal. If the product close to the ideal, its quality was considered good; otherwise the quality was poor. However, nowadays the concept of quality has expanded to mean as "*a measure of superior or a state of being free from defectiveness and significant variations, brought about by the rigorous and consistent adherence to measurable and verifiable standards to achieve uniformity of output that satisfies specific customer or user requirements*". (Business Dictionary)

Due to increase of customer demands on products, monitoring the process variables are become complex and multivariate in nature. Monitoring these process variables separately will be misled because without taking their correlation into consideration [1]. Therefore, some great reviews discussed about Multivariate Statistical Process Control (MSPC) with the use of multivariate control chart in considering the correlation among the process variables can be found in Mason and Young [2], Johnson and Wichern [3], and Montgomery [4].

According to Alt [5], monitoring process variables consists of two distinct phase – Phase I and Phase II. Although both phases are devoted in identify the out-of-control signals, but each phase has unique goal. In Phase I, based on a historical data set, a data subset (so-called reference sample) is selected which is clean from outliers. The reference sample is then used to estimate all parameters. In Phase II, the results of Phase I are then used to monitor the process by detecting departures from the statistical parameters that have been estimated. Jensen *et. al.* [6] have remark that a successful of monitoring process in Phase II depends on a successful analysis during Phase I.

As we mention in the previous paragraph, the main problem in Phase I is to estimate the parameters of in control process. Any

changes in the process parameters can give high impact on a critical quality attribute and therefore, the choice of the parameters estimation is important. There are two parameters that we need to monitor, namely mean vector and covariance matrix. However, in real data sets, it often exists that some observations are different from majority (so-called outliers) which strongly influence the parameter estimation [7]. In consequence, estimating the parameters from the reference sample is not effective and leads to poor properties in detecting out-of-control signals.

To handle this problem, we need a high breakdown point robust parameter estimator. There are many different robust estimation methods available in the literatures, but the most popular method is Fast Minimum Covariance Determinant (FMCD) which is introduced by Rousseeuw and Van Driessen [8]. FMCD is a hybrid algorithm based on an iterative scheme and the Minimum Covariance Determinant (MCD) estimators. It is widely used because of the computation efficiency and time saving.

In what follows, we compare control charts constructed by using the classical approach and the robust approach. The rest of the paper is organized as follows. The next section recalls the Hotelling's T^2 statistic based on the classical sample mean vector and sample covariance matrix. Section 3 presents the robust Hotelling's T^2 statistic where the parameters are estimated by using FMCD. In section 4, we present an industrial example. At the end of this paper, we will discuss the conclusion.

2.0 CLASSICAL APPROACH

Suppose that we have a data matrix X of size $(m \times p)$ which consisting of m independent observations from the p–variate normal distribution $N_p(\mu, \sum)$. The p variables are to be monitored simultaneously based on that data matrix. Based on the historical data set, the unknown parameters μ and Σ are estimated and replaced with sample mean vector $\bar{X}$ and sample covariance matrix S, respectively,

$$\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i .$$

and

$$S = \frac{1}{m-1}\sum_{i=1}^{m}\left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)' .$$

Based on this approach, Hotelling's T_i^2 statistics is

$$T_i^2 = \left(X_i - \bar{X}\right)' S^{-1}\left(X_i - \bar{X}\right) \quad \text{for } i = 1,2,\ldots,m \tag{1}$$

The exact distribution of T_i^2 is proportional to a beta distribution,

$$T_i^2 \sim \frac{(m-1)^2}{m} BETA_{\left(\frac{p}{2},\frac{m-p-1}{2}\right)} \quad \text{for } i = 1,2,\ldots,m \tag{2}$$

See Chou *et al.* [9] for the details of the proof of equation (2). Therefore, the corresponding upper control limit (UCL) and lower control limit (LCL) are

$$LCL = \frac{(m-1)^2}{m} BETA_{\left(\alpha,\frac{p}{2},\frac{m-p-1}{2}\right)} \tag{3}$$

$$UCL = \frac{(m-1)^2}{m} BETA_{\left(1-\alpha,\frac{p}{2},\frac{m-p-1}{2}\right)} \tag{4}$$

3.0 ROBUST APPROACH

Consider a random vector data set of p–variate normal observations. The FMCD algorithm is as follows (see [8] for more details):

- H_{old} is an arbitrary subset containing $h = \frac{m+p+1}{2}$ data points.
- Compute the mean vector $\bar{X}_{H_{old}}$ and covariance matrix $S_{H_{old}}$ of all observations belonging to H_{old}. Then compute

$$T^2_{H_{old}}(i) = \left(X_i - \bar{X}_{H_{old}}\right)' S^{-1}_{H_{old}}\left(X_i - \bar{X}_{H_{old}}\right) \quad \text{for } i = 1,2,\ldots,m$$

- Sort these $T^2_{H_{old}}(i)$ value in increasing order,

$$T^2_{H_{old}}(\pi(1)) \le T^2_{H_{old}}(\pi(2)) \le \ldots \le T^2_{H_{old}}(\pi(m)) .$$

where π is a permutation on $\{1,2,\ldots,m\}$.

- Define $H_{new} = \{X_{\pi(1)}, X_{\pi(2)}, \ldots, X_{\pi(h)}\}$. Calculate $\bar{X}_{H_{new}}$, $S_{H_{new}}$ and $T^2_{H_{new}}(i)$.
- If $\det(S_{H_{new}}) = 0$ or $\det(S_{H_{new}}) = \det(S_{H_{old}})$, the process is stopped. Otherwise, the above process is continued until the k-th iteration. Thus,

$$\det(S_{H_1}) \ge \det(S_{H_2}) \ge \ldots \ge \det(S_{H_k}) = \det(S_{H_{k+1}}) .$$

- Let $\bar{X}_R$ and S_R are the sample mean and covariance matrix given by that process. Hotelling's $T^2_{R,i}$ statistic is defined as

$$T^2_{R,i} = \left(X_i - \bar{X}_R\right)' S_R^{-1}\left(X_i - \bar{X}_R\right) \quad \text{for } i = 1,2,\ldots,m \tag{5}$$

By using robust estimate, the distributional property of $T^2_{R,i}$ is still open to be explored. In Jensen *et al.* [6], the distribution of the $T^2_{R,i}$ converges in distribution to Chi-square distribution for $i = 1,2,\ldots,m$ as $m \to \infty$. In this paper, χ^2_p distribution approximation is used. According to Chi-square distribution, the cut-off values are as follows:

$$UCL = \chi^2_{\alpha,p} \tag{6}$$

$$LCL = \chi^2_{1-\alpha,p} \tag{7}$$

A friendly user toolbox in LIBRA [10] contains implementations of robust methods and functions for location and scale estimation. In this study, MATLAB will be used to get the results. Besides that, in this paper we plan to use $h = 0.75m$. When a large proportion of contamination is presumed, intermediate value for $h = 0.75m$ is recommended to obtain a higher-sample efficiency [10].

4.0 INDUSTRIAL EXAMPLE

In this section, we compare the classical approach and robust approach during Phase I operation for the production process of cocoa powder in a food industry, Industry A Sdn Bhd. The name of Industry A is kept confidential upon request made by the management of the industry. There are p = 7 quality variables, namely x_1 = intrinsic color L, x_2 = intrinsic color a, x_3 = intrinsic color b, x_4 = fineness, x_5 = pH, x_6 = fat content, and x_7 = moisture. The number of individual observations is m = 147. Observations from the first 100 sample are utilized as the set

of historical data during Phase I and remain 47 samples are reserved for future observations during Phase II.

In order to develop the reference samples, outlier purging process is undergoing. Here, the T^2 statistic is applied to detect out-of-control signals. Firstly, we showed the difference in Phase I between classical estimation and robust estimation. The mean vector and covariance matrix based on classical estimation are

$$\bar{X} = \begin{bmatrix} 6.9434 \\ 16.8654 \\ 15.5955 \\ 18.2913 \\ 10.5271 \\ 2.6642 \\ 99.8685 \end{bmatrix}$$

$$S = \begin{bmatrix} 0.00401 & & & & & & \\ 0.02155 & 0.79767 & & & & & \\ 0.01420 & 0.14372 & 0.18578 & & & & \\ 0.04480 & 0.23166 & 0.41506 & 1.28588 & & & \\ 0.00002 & 0.06607 & 0.00897 & 0.00407 & 0.06876 & & \\ 0.00622 & -0.03157 & 0.01770 & 0.04481 & 0.00080 & 0.04838 & \\ 0.00105 & 0.00755 & 0.00727 & 0.02074 & 0.00265 & 0.00320 & 0.00182 \end{bmatrix}$$

From the table of Beta distribution with degree of freedom p = 7 and probability of false alarm 0.0027, we get $UCL = 20.24245$ and $LCL = 0.84065$. Figure 1 visualizes Phase I based on classical approach and we can clearly see that no observation lies outside the control limits.

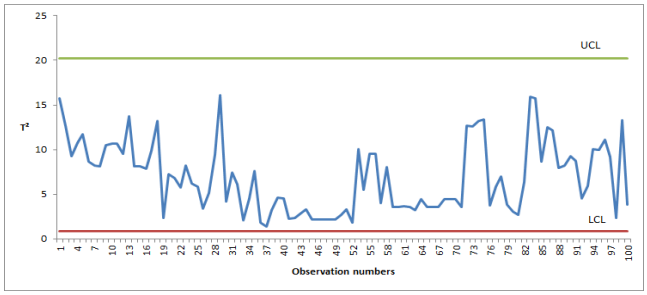


Figure 1 Phase I based on classical Hotelling's T^2 statistic

Next we continue to further analyze Phase I using robust Hotelling's T^2 statistic to see whether masking and swamping effects occur. The mean vector and covariance matrix based on robust estimation by using FMCD are:

$$\bar{X}_R = \begin{bmatrix} 6.9308 \\ 17.1304 \\ 15.6907 \\ 18.4248 \\ 10.5586 \\ 2.6243 \\ 99.8847 \end{bmatrix}$$

$$S_R = \begin{bmatrix} 0.00373 & & & & & & \\ -0.00764 & 0.28910 & & & & & \\ -0.01184 & 0.08651 & 0.16518 & & & & \\ -0.04502 & 0.31641 & 0.40592 & 1.28521 & & & \\ 0.00268 & 0.03482 & -0.01182 & 0.01152 & 0.06209 & & \\ 0.00512 & 0.01029 & -0.00443 & -0.02442 & 0.01123 & 0.04778 & \\ -0.00032 & 0.00047 & 0.00009 & 0.00669 & 0.00169 & -0.00079 & 0.00041 \end{bmatrix}$$

From the table of Chi-square distribution, with p = 7 and probability of false alarm 0.0027, $UCL = 16.01276$ and $LCL = 1.68987$. Figure 2 shows the Phase I control chart based on robust Hotelling's T^2 statistic. Sample number 1, 2, 3, 4, 5, 6, 7, 12, 13, 14, 15, 16, 17, 18, 29, 35, 72, 73, 74, 75, 82, 83, 84 and 99 have the largest T^2 value and lie outside the control limits.

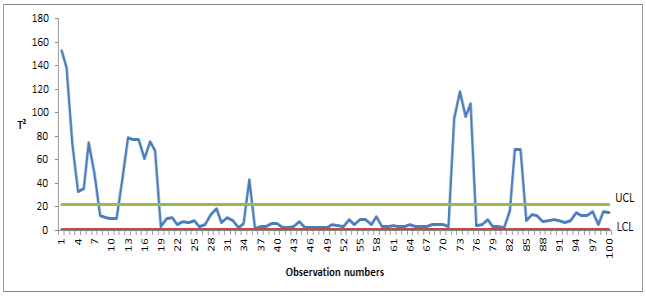


Figure 2 Phase I based on robust Hotelling's T^2 statistic

The figure presents the information that cannot be provided by classical Hotelling's T^2 chart. Therefore 24 outliers need to remove for further analysis. By removing all 24 outlying observations and recalculating the parameter estimates with m = 76 and p = 7, the new mean vector and sample covariance matrix were obtained. The results are presented as below:

$$\bar{X}_R = \begin{bmatrix} 6.9292 \\ 17.1222 \\ 15.7170 \\ 18.4859 \\ 10.5673 \\ 2.6218 \\ 99.8856 \end{bmatrix}$$

$$S_R = \begin{bmatrix} 0.00376 & & & & & & \\ -0.00802 & 0.29486 & & & & & \\ -0.01157 & 0.09687 & 0.15047 & & & & \\ -0.04489 & 0.34380 & 0.37403 & 1.22714 & & & \\ 0.00330 & 0.04129 & -0.01584 & 0.00334 & 0.05535 & & \\ 0.00498 & 0.01294 & -0.00434 & -0.02363 & 0.01086 & 0.04750 & \\ -0.00030 & 0.00052 & -0.00056 & 0.00542 & 0.00182 & -0.00068 & 0.00040 \end{bmatrix}$$

Reconstructing the control chart, we observe that in Figure 3 there is none of the observation lie outside the control limits. If we compare that result with that given by classical approach, (see Figure 1) where no outlier is detected, the use of classical approach will be misleading.

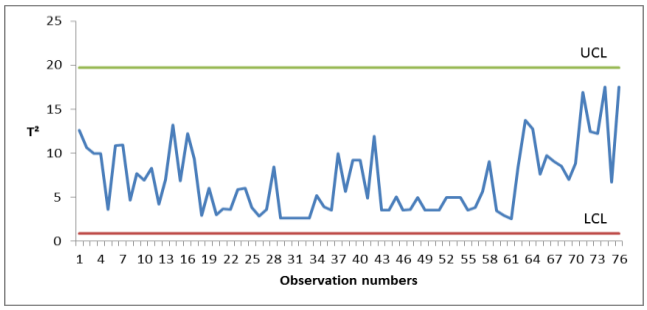
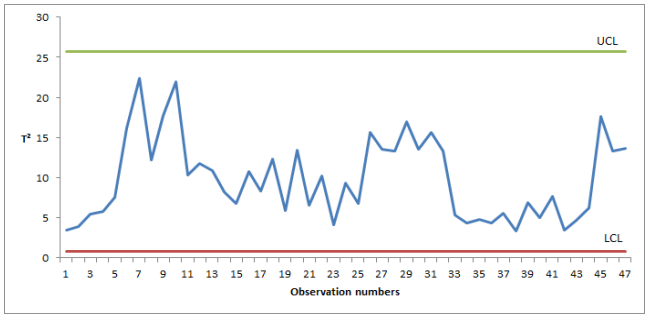
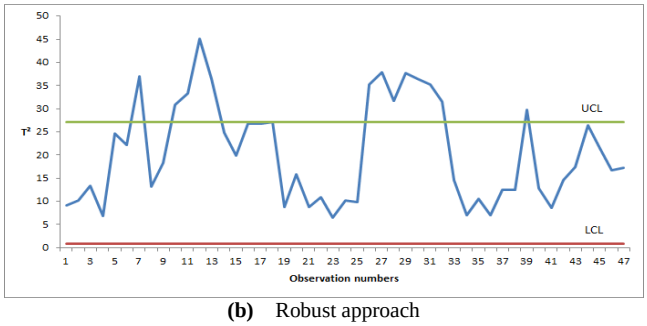


Figure 3 Robust Hotelling's T^2 control chart with reference sample

Now we present the advantage of the robust approach for Phase II process mean monitoring. Figure 4 represent the history of process mean using classical Phase I and robust Phase I. By comparing the performance of monitoring process mean in Figure 4, the robust Phase II monitoring is more sensitive compared to the classical Phase II monitoring.



(a) Classical approach



(b) Robust approach

Figure 4 Classical and robust Phase II monitoring

■5.0 CONCLUSION

In this paper, we compare the control charts by using the classical approach and robust approach. We successfully point out the inability of the classical Hotelling's T^2 control chart for detecting out-of-control signals due to special causes, e.g. masking effect and swamping effect. The robust T^2 charting technique has taken advantage of the robust estimators to address the problem of inevitable effects exist in cocoa powder production process.

Acknowledgement

This research is sponsored by the Ministry of Higher Education of Malaysia under FRSG vote number 02H18. The authors gratefully acknowledge Government of Malaysia for the sponsorships and Universiti Teknologi Malaysia for the opportunity to do this research. Special thanks go to Ms. Chan for the permission to use her company data set.

References

[1] Ryan, T. P. 1989. *Statistical Methods for Quality Improvement*. New York, NY: John Wiley and Sons.
[2] Mason, R. L. and Young, J. C. 2002. *Multivariate Statistical Process Control with Industrial Applications*. SIAM, Philadelphia, PA.
[3] Johnson, R. A. and Wichern, D. W. 2002. *Applied Multivariate Statistical Analysis*. 5th Ed. Upper Saddle River, NJ: Prentice Hall.
[4] Montgomery, D. C. 2005. *Introduction to Statistical Quality Control*. 5th Ed. New York, NY: John Wiley and Sons.
[5] Alt, F. B. 1982. Multivariate Quality Control: State of the Art. *ASQC Quality Congress Transactions*. 886–893.
[6] Jensen, W. A., Birch, J. B. and Woodall, W. H. 2005. High Breakdown Point Estimation Methods For Phase I Multivariate Control Charts, *Quality and Reliability Engineering International*. 23(5): 615–629.
[7] Hadi, A. S. 1992. Identifying Multiple Outliers in Multivariate Data. *Journal of Royal Statistical Society*. 54(3): 761–771.
[8] Rousseeuw, P. J. and Van Driessen, K. 1999. A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*. 41(3): 212–223.
[9] Chou, Y. M., Mason, R. L. and Young, J. C. 1999. Power Comparisons for a Hotelling's T^2 statistic, *Communications in Statistics, Part B-Simulation and Computation*. 28: 1031–1050.
[10] Verboven, S. and Hubert, M. 2005. LIBRA: A MATLAB Library for Robust Analysis. *Chemometrics and Intelligent Laboratory Systems*. 75:127–136.