

Effect of Phase I Estimation on Phase II Control Chart Performance with Profile Data

YAJUAN CHEN

Pfizer, Andover, MA 01810

JEFFERY B. BIRCH and WILLIAM H. WOODALL

Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0439

This paper illustrates how Phase I estimators in statistical process control (SPC) can affect the performance of Phase II control charts. The deleterious impact of poor Phase I estimators on the performance of Phase II control charts is illustrated in the context of profile monitoring. Two types of Phase I estimators are discussed. One approach uses functional cluster analysis to initially distinguish between estimated profiles from an in-control process and those from an out-of-control process. The second approach does not use clustering to make the distinction. The Phase II control charts are established based on the two resulting types of estimates and compared across varying sizes of sustained shifts in Phase II. A simulated example and a Monte Carlo study show that the performance of the Phase II control charts can be severely distorted when constructed with poor Phase I estimators. The use of clustering leads to much better Phase II performance. We also illustrate that the performance of Phase II control charts based on the poor Phase I estimators not only have more false alarms than expected but can also take much longer than expected to detect potential changes to the process.

Keywords: Clustering, Control Chart, Mixed Models, Statistical Process Control.

1. Introduction

Profile monitoring is a form of statistical process control (SPC) where the quality of a product or process is characterized by a functional relationship, referred to as a “profile”, between a response variable and one or more explanatory variables. Profile monitoring, and SPC in general, is conducted over two Phases, referred to as Phase I and Phase II. In Phase I profile monitoring, a historical data set (HDS) is utilized to first estimate the individual profiles using appropriate regression methods. Then the estimated profiles are analyzed to determine which estimated profiles can be considered from the in-control process and which, if any, can be considered from

an out-of-control process. Finally, using those estimated profiles determined to be from the in-control process, appropriate estimates are obtained of in-control parameters and used to establish control limits for the Phase II control charts. Here, an out-of-control process is one where at some point there is a change in the functional relationship between the response and the explanatory variables. For example, such a change might be represented by a sustained shift in a parameter in the regression model representing the profiles. After detecting possible profiles from the out-of-control process and obtaining the control limits for Phase II control charts based on the profiles considered to represent the in-control process, Phase II profile monitoring consists of monitoring the future profiles to determine the on-going stability of the process.

In Phase I profile monitoring, profiles from the out-of-control process can be detected by using the Hotelling's T^2 statistic. For example, Kang and Albin¹, Kim *et al.*² and Mahmoud and Woodall³ utilized the Hotelling's T^2 statistic to detect the profiles from the out-of-control process based on the estimated regression parameters using linear regression models. Profiles estimated by nonlinear and nonparametric regression methods were studied by Jin and Shi⁴, Walker and Wright⁵, Gupta *et al.*⁶, Ding *et al.*⁷, Williams *et al.*⁸, Williams *et al.*⁹, Chicken *et al.*¹⁰ and Hung *et al.*¹¹. In addition, Jensen *et al.*¹², Jensen and Birch¹³, Qiu *et al.*¹⁴ and Abdel-Salam *et al.*¹⁵ provided mixed model approaches to monitor the profiles in order to account for the correlation structure typically existing within each profile.

Based on the mixed models approach, the Hotelling's T^2 statistic used to detect the presence of profiles from an out-of-control process can be obtained by comparing each individual estimated profile, referred to as the "profile specific (PS) curve" to the estimated population average (PA) curve. Jensen *et al.*¹² also showed that the Hotelling's T^2 statistic used to detect profiles from the out-of-control process in the parametric mixed model can be equivalently obtained by using the estimated best linear unbiased predictors (eblups) of each profile. The method of Jensen *et al.*¹² will be used throughout this paper and referred to as the "non-cluster-based method".

Unfortunately, during Phase I analysis, the Hotelling's T^2 method may fail to properly classify profiles into the two categories of in-control and out-of-control. That is, a profile from an in-control process may be misclassified as being from an out-of-control process, an error referred to a false negative. The data from such a profile may be mistakenly discarded and not

used in the subsequent calculations to establish Phase II control limits. Also, a profile from the out-of-control process may be misclassified as being from the in-control process, an error referred to as a false positive. The data from such a profile are mistakenly included in calculating the Phase II control limits. Chen *et al.*¹⁶ showed that if profiles from the out-of-control process are classified as being from the in-control process, the resulting estimated PA profile parameters will be biased, perhaps severely so. They provided an alternative profile monitoring method in Phase I, referred to as “cluster-based profile monitoring” to obtain the estimates that are robust to profiles from the out-of-control process contained in the HDS. Recall that a primary goal of Phase I analysis is to obtain the estimates as close to the true parameters for the in-control PA profile as possible and then use these estimates to set the proper control limits for Phase II control charts. Severely biased estimates of the PA profile in Phase I can give incorrect Phase II control limits and adversely affect the performance of Phase II control charts.

Phase II performance based on Phase I estimators has been studied in the past by others under different scenarios than considered in this paper. For example, Mahmoud¹⁷ investigated the performance of three Phase II simple linear profile approaches, those of Kang and Albin¹⁸, Kim *et al.*², and Mahmoud *et al.*¹⁹, with estimated parameters. Aly *et al.*²⁰, on the other hand, compared the in-control performance of three Phase II simple linear profile monitoring approaches based on the standard deviation of the average run length (SDARL).

2. Review of Methods in Phase I Profile Monitoring

In Phase I profile monitoring, we assume that the i^{th} profile can be represented by using the linear mixed model (LMM) (Laird and Ware²¹) given by

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, 2, \dots, m. \quad (2.1)$$

where m is the number of profiles, \mathbf{y}_i is the $n_i \times 1$ response vector associated with the i^{th} profile, n_i is the sample size for the i^{th} profile, \mathbf{X}_i and \mathbf{Z}_i are $n_i \times p$ and $n_i \times q$, respectively, matrices of explanatory variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of PA parameters, \mathbf{b}_i is a $q \times 1$ vector of random effects for the i^{th} profile with $\mathbf{b}_i \sim MN(\mathbf{0}, \mathbf{G})$ and \mathbf{G} is a $q \times q$ covariance matrix. In addition, $\boldsymbol{\varepsilon}_i$ is the $n_i \times 1$ vector of random errors for the i^{th} profile with $\boldsymbol{\varepsilon}_i \sim MN(\mathbf{0}, \mathbf{R}_i)$, where \mathbf{R}_i is the $n_i \times n_i$

covariance matrix. For more details of the LMM, see Schabenberger and Pierce²², Ruppert *et al.*²³, Seber and Wild²⁴ and Demidenko²⁵.

Using the LMM, the i^{th} estimated PS curve is given by

$$\widehat{PS}_i = \mathbf{X}_i \hat{\boldsymbol{\beta}}_{LMM} + \mathbf{Z}_i \hat{\mathbf{b}}_i, \quad (2.2)$$

where $\hat{\boldsymbol{\beta}}_{LMM}$ is the estimator for $\boldsymbol{\beta}$ and $\hat{\mathbf{b}}_i$ is the eblup vector for the random effect \mathbf{b}_i . If we let

$\hat{\boldsymbol{\beta}}_i = \begin{pmatrix} \hat{\boldsymbol{\beta}}_{LMM} \\ \hat{\mathbf{b}}_i \end{pmatrix}$ and $\tilde{\mathbf{X}}_i = [\mathbf{X}_i \quad \mathbf{Z}_i]$, then equation (2.2) can also be expressed as

$$\widehat{PS}_i = \tilde{\mathbf{X}}_i \hat{\boldsymbol{\beta}}_i. \quad (2.3)$$

The estimated PA curve is

$$\widehat{PA} = \mathbf{X} \hat{\boldsymbol{\beta}}_{LMM}, \quad (2.4)$$

where \mathbf{X} is the $\tilde{n} \times p$ model matrix composed of all the distinct rows contained in each \mathbf{X}_i and \tilde{n} is the number of such distinct rows.

The first step of Phase I profile monitoring is to classify the profiles as either from the in-control or out-of-control process. Using the LMM, Jensen *et al.*¹² proposed using the distance of the estimated vector for the i^{th} profile, $\hat{\boldsymbol{\beta}}_i$, from the center of the group of estimated vectors, $\hat{\boldsymbol{\beta}}_{LMM}$. They introduced a formula for the T^2 statistic for the i^{th} profile, denoted as T_i^2 , based on comparing $\hat{\boldsymbol{\beta}}_i$ to the sample mean of the $\hat{\boldsymbol{\beta}}_i$ vectors, $\hat{\boldsymbol{\beta}}_{LMM}$, which is defined as

$$T_i^2 = \left(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_{LMM} \right)^T \hat{\mathbf{V}}^{-1} \left(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_{LMM} \right), \quad (2.5)$$

where $\hat{\mathbf{V}}$ is the estimated variance-covariance matrix of $\hat{\boldsymbol{\beta}}_i$. The successive difference estimator $\hat{\mathbf{V}}_D$, is preferred here. Sullivan and Woodall²² showed that use of $\hat{\mathbf{V}}_D$ is effective in detecting sustained step changes in the process that may occur in Phase I data, the situation we consider. The successive difference estimator of \mathbf{V} is

$$\hat{\mathbf{V}}_D = \frac{1}{2(m-1)} \sum_{i=1}^{m-1} \left(\hat{\boldsymbol{\beta}}_{i+1} - \hat{\boldsymbol{\beta}}_i \right) \left(\hat{\boldsymbol{\beta}}_{i+1} - \hat{\boldsymbol{\beta}}_i \right)^T. \quad (2.6)$$

Jensen *et al.*¹² showed that the distribution of T^2 follows asymptotically a chi-squared distribution with p degrees of freedom for large samples, where p is the number of estimated parameters. Since $\sum_{i=1}^m \hat{\mathbf{b}}_i = \mathbf{0}$, it follows that (Jensen *et al.*¹²) the Equations (2.5) and (2.6) can be written equivalently as

$$T_i^2 = \hat{\mathbf{b}}_i^T \hat{\mathbf{V}}_D^{-1} \hat{\mathbf{b}}_i,$$

and

$$\hat{\mathbf{V}}_D = \frac{1}{2(m-1)} \sum_{i=1}^{m-1} (\hat{\mathbf{b}}_{i+1} - \hat{\mathbf{b}}_i)(\hat{\mathbf{b}}_{i+1} - \hat{\mathbf{b}}_i)^T.$$

The T^2 statistic provided by Jensen *et al.*¹² works well when the HDS contains very few or no profiles from the out-of-control process. However, if there are even a modest number of profiles in the HDS from the out-of-control process, this method can be adversely affected and will result in biased estimates of the PA profile (Chen *et al.*¹⁶). In order to provide a T^2 statistic that is robust to the profiles from the out-of-control process, Chen *et al.*¹⁶ introduced the cluster-based profile monitoring method for Phase I profile monitoring analysis. Instead of computing the T^2 statistic based on all the profiles in the HDS, their method is first to cluster the estimated profiles to obtain an initial main cluster of estimated profiles with similar shapes. The profiles in the initial main cluster set are used to obtain the initial PA parameter estimates which are used to compute the T^2 statistic for each profile not contained in the initial main cluster set. Those profiles with “small” T^2 statistics are added to the initial main cluster set to obtain a new set of profiles. By iteratively updating the profiles not in the main cluster in this manner until no profiles can be added to the main cluster set, the estimated PA parameters are obtained by using the profiles in the final main cluster set. The details of algorithm can be found in Chen *et al.*¹⁶.

During Phase I analysis, if the HDS contains profiles from the out-of-control process resulting from a sustained shift, Chen *et al.*¹⁶ illustrated the following two important features of their cluster-based profile monitoring method: 1) their method is more likely to correctly distinguish profiles from the in-control process from those from the out-of-control process than the method of Jensen *et al.*¹² and 2), their method results in PA estimated parameters with less

bias than those estimates resulting from the method of Jensen *et al.*¹². The impact of these features on the Phase II monitoring performance is illustrated in Sections 3 and 4.

3. Phase II Control Charts based on Phase I Estimates

It is intuitive that an estimator with favorable Phase I properties, such as small bias and small variance, will result in better Phase II control chart performance than an estimator with less favorable Phase I properties. This effect is further illustrated using the following example of the Phase I profile monitoring analysis (Chen *et al.*¹⁶).

In this example, using the LMM, the first nine profiles were generated from the in-control process as

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_2 + b_{1i})x_{ij} + (\beta_3 + b_{2i})x_{ij}^2 + \varepsilon_{ij}, i = 1, 2, \dots, m_1, j = 1, 2, \dots, n \quad (3.1)$$

and the last three profiles were generated as

$$y_{ij} = (\beta'_0 + b_{0i}) + (\beta'_1 + b_{1i})x_{ij} + (\beta'_2 + b_{2i})x_{ij}^2 + \varepsilon_{ij}, i = m_1 + 1, \dots, m, j = 1, 2, \dots, n \quad (3.2)$$

where $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \beta_2) = (12.5, -7, 2)$ for the in-control process and $\boldsymbol{\beta}'^T = (\beta'_0, \beta'_1, \beta'_2) = (21.875, -14.5, 3.5)$ for the out-of-control process with $n=10$, $m_1 = 9, m = 12$, $\sigma_0^2 = \sigma_1^2 = \sigma_2^2 = 0.5$, and $\sigma^2 = 4$. The regressor vector, \mathbf{x} , contains 10 equally spaced values from 1 to 10. Thus, profiles 1 through 9 represent profiles from the in-control process and profiles 10, 11, and 12 represent profiles from the out-of-control process. The plot of the observed data is given in Figure 1 where straight line segments are used to connect the ten observations for each profile. We note that it is not at all clear from looking at Figure 1 that profiles 10, 11, and 12 are from the out-of-control process.

The non-cluster-based method leads one to conclude that the 6th profile is from the out-of-control process, and then yielded the following estimates for the PA parameters

$$\hat{\boldsymbol{\beta}}_{PA}^T = (16.27, -9.71, 2.18)$$

The estimated variance-covariance matrix was

$$\hat{\mathbf{V}} = \begin{bmatrix} 3.660 & -2.470 & -0.405 \\ -2.470 & 2.035 & -0.132 \\ -0.405 & -0.132 & 0.493 \end{bmatrix}.$$

The cluster-based method, on the other hand, correctly detected the 10th, 11th and 12th profiles as from the out-of-control process and yielded as the estimates for the PA parameters and variance-covariance matrix

$$\hat{\boldsymbol{\beta}}_{PA}^T = (12.55, -7.24, 1.78),$$

and

$$\hat{\mathbf{V}} = \begin{bmatrix} 0.203 & 0.120 & 0.270 \\ 0.120 & 0.236 & 0.090 \\ 0.270 & 0.090 & 0.389 \end{bmatrix},$$

respectively.

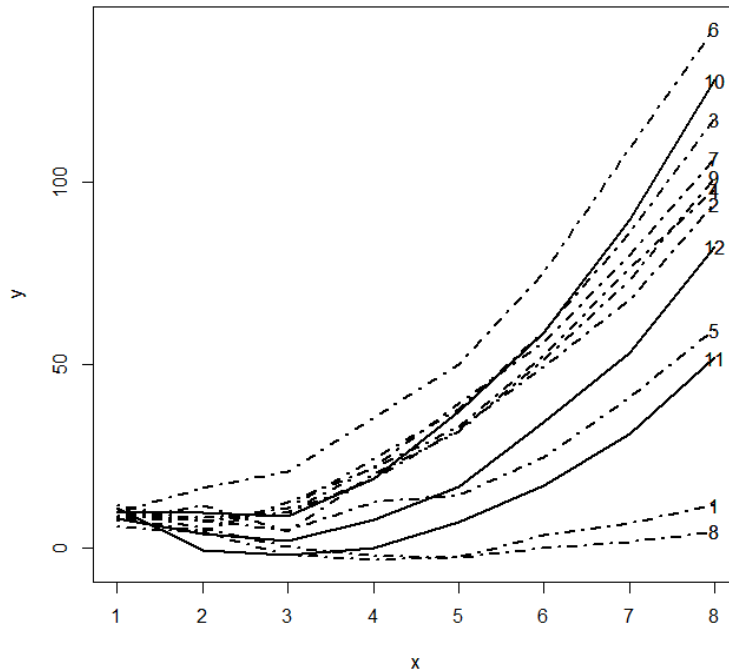


Figure 1: The plot of observed data. The dashed curves represent profiles from the in-control process; the solid curves represent profiles from the out-of-control process.

The Phase II control limits were obtained by simulating the in-control process with the Phase I estimates and an in-control ARL_0 set equal to 200. In order to choose the upper control limit (UCL) so that the in-control $ARL_0=200$, first define the in-control T^2 as

$$T^2 = (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{PA})^T \hat{\mathbf{V}}^{-1} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{PA}),$$

where $\hat{\boldsymbol{\beta}}$ is the estimated profile parameter vector for the in-control process profile simulated by the LMM with the estimated PA parameter vector $\hat{\boldsymbol{\beta}}_{PA}$ and estimated corresponding covariance matrix $\hat{\mathbf{V}}$ resulting from the Phase I analysis. The UCL is then obtained so that the average number (based on 5000 simulations) of simulated in-control profiles until a signal (a T^2 greater than the UCL) occurs is 200. A good start value of UCL could be based on the average of the maximum of every 200 simulated in-control T^2 statistics. Note, since the simulated in-control profiles are based on the Phase I estimates, and not on the true in-control parameters, it is likely for the Phase II ARL to not equal 200 even when the profiles result from an in-control process.

After calibrating the control limits with ARL_0 equaling 200 for the non-cluster-based T^2 control chart and the cluster-based T^2 control chart, the performance of each chart can be evaluated by comparing the out-of-control ARL, denoted as ARL_1 , for different Phase II shift values. The PA curve in this example may be written as

$$PA = \beta_0 + \beta_1 x + \beta_2 x^2 \quad (3.3)$$

where $\boldsymbol{\beta}^T = (12.5, -7, 2)$. Equation (3.3) can be rewritten as

$$PA = \beta_{C1} x + \beta_{C2} (x - 2.5)^2 \quad (3.4)$$

where $\boldsymbol{\beta}_C^T = (\beta_{C1}, \beta_{C2}) = (3, -2)$. Assuming that the out-of-control process has the PA profile

$$PA = \beta_{C1} x + (\beta_{C2} + shift)(x - 2.5)^2 \quad (3.5)$$

with $shift=(0,0.25,0.5,0.75,1,1.25,1.5)$, the in-control ARL_0 and the out-of-control ARL_1 based on the cluster-based T^2 control chart and the non-cluster-based T^2 control chart are given in Table 1 to compare the performance of these two control charts.

Table 1: ARL_{CB} and ARL_{NCB} with $ARL_0 \approx 200$ for Phase II with Varying Shift Values. ARL_{CB} represents the ARL for the cluster-based method and ARL_{NCB} represents the ARL for the non-cluster-based method (bolded cells represent the better values).

Phase II Shift	ARL_{CB}	ARL_{NCB}
0	193.8	64.9
0.25	72.13	120.3
0.5	17.4	261.4
0.75	3.5	30.36
1	2.0	7.9
1.25	1.3	2.6
1.5	1.1	1.6

Table 1 shows that when Phase II shift size is greater than 0 the ARL_1 based on the cluster-based T^2 control chart are uniformly smaller than the ARL_1 based on the non-cluster-based T^2 control chart, especially when the Phase II shift size is small. For example, when the Phase II shift size equals 0.25, the ARL_1 based on the non-cluster-based T^2 control chart is 73.2 while the ARL_1 based on the cluster-based T^2 control chart is 120.3. When the Phase II shift size equals 1, the cluster-based T^2 control chart has an ARL_1 value of 2.0. The non-cluster-based T^2 control chart, on the other hand, takes about 8 samples, on the average, to detect the change of process. Also, the simulation results show that when Phase II shift equals 0, the ARL for the cluster method is 193.8 and for the non-cluster method the ARL is 64.9. This result indicates that when the Phase I estimates are far from the true parameters, the ARL is much smaller than 200, the desired value, and thus leads to more false alarms.

In addition, Table 1 shows that the ARL values based on the non-cluster-based T^2 control chart are greater than 200 when Phase II shift size is 0.5. This result indicates that the non-cluster-based T^2 control chart requires many more observations than expected to detect the change of process at this point. This ARL-bias feature occurs because the non-cluster-based method in Phase I misclassified the profiles from both the in-control and the out-of-control processes, and the resulting estimates of the PA profile parameters are severely biased. Recall that the profiles from the out-of-control process in the HDS were generated with the PA curve

$$PA = 21.875 - 14.5x + 3.5x^2,$$

which can be written as

$$PA = 3x + (2 + 1.5)(x - 2.5)^2 \quad (3.6)$$

Note equation (3.6) is equivalent to equation (3.5) with a shift size equal to 1.5. The non-cluster-based estimator (Jensen *et al.*¹²) involves 3 profiles from the out-of-control process with Phase I shift size equal to 1.5 and, as a result, its estimated PA curve is pulled to the direction of the PA curve with shift value equal to 1.5. The corresponding Phase II control chart is also distorted, resulting in many more samples, on the average, than expected to detect changes in the process and many more false alarms when the process is in-control.

4. A Monte Carlo Study Comparing Phase II Control Charts

In this section, the average performance of Phase II control charts for the two methods is compared based on the average performance of Phase I estimators. The average Phase I estimates were obtained from the Monte Carlo study provided by Chen *et al.*¹⁶. In this Phase I Monte Carlo study, they assumed that the in-control profiles are randomly generated from the linear mixed model

$$y_{ij} = \beta_{0i} + \beta_{1i}x_{ij} + \beta_{2i}x_{ij}^2 + \varepsilon_{ij}, i = 1, 2, \dots, m_1, j = 1, 2, \dots, n, \quad (4.1)$$

where y_{ij} is the j^{th} observation for the i^{th} profile, m_1 is the number of profiles from the in-control process, n is the number of observations within each profile, and where

$$\begin{aligned} \beta_{0i} &= \beta_2 \bar{x}^2 + b_{0i}, \\ \beta_{1i} &= \beta_1 - 2\beta_2 \bar{x} + b_{1i}, \\ \beta_{2i} &= \beta_2 + b_{2i}, \end{aligned}$$

with $\bar{x} = \frac{\sum_{i=1}^m \sum_{j=1}^n x_{ij}}{mn}$. Here, $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \beta_2)$ represents the fixed effect parameters and $\mathbf{b}_i^T = (b_{0i}, b_{1i}, b_{2i})$ represents the random effects. Note the corresponding PA parameter vector can also be written as $\boldsymbol{\beta}_{PA}^T = (\beta_2 \bar{x}^2, \beta_1 - 2\beta_2 \bar{x}, \beta_2)$. Consequently, the PA profile, evaluated at an arbitrary vector of $\mathbf{x}_0 = (x_{01}, x_{02}, \dots, x_{0n_0})$ can be written as

$$PA_j = \beta_2 \bar{x}^2 + (\beta_1 - 2\beta_2 \bar{x})x_{0j} + \beta_2 x_{0j}^2, \quad j = 1, 2, \dots, n_0. \quad (4.2)$$

It is easy to show that the PA profile can be simplified to give

$$PA_j = \beta_1 x_{0j} + \beta_2 (x_{0j} - \bar{x})^2, \quad j = 1, 2, \dots, n_0. \quad (4.3)$$

The profiles in the out-of-control process were also generated from a model of the same form, but with a sustained shift. Thus we have

$$\begin{aligned} y_{ij} &= \beta_{0i} + \beta_{1i} x_{ij} + \beta_{2i} x_{ij}^2 + \varepsilon_{ij}, \quad i = m_1 + 1, m_1 + 2, \dots, m, \quad j = 1, 2, \dots, n, \\ \beta_{0i} &= (\beta_0 + shift) \bar{x}^2 + b_{0i}, \\ \beta_{1i} &= \beta_1 - 2(\beta_2 + shift) \bar{x} + b_{1i}, \\ \beta_{2i} &= (\beta_2 + shift) + b_{2i}, \end{aligned} \quad (4.4)$$

where m is the number of profiles. The shift is assumed to be sustained for the final $m - m_1$ profiles. Thus, the corresponding PA profile is

$$\begin{aligned} PA_j &= (\beta_0 + shift) \bar{x}^2 + [\beta_1 - 2(\beta_2 + shift) \bar{x}] x_{0j} + (\beta_2 + shift) x_{0j}^2, \\ j &= 1, 2, \dots, n_0. \end{aligned} \quad (4.5)$$

Also, equation (4.5) can be simplified to give

$$PA_j = \beta_1 x_{0j} + (\beta_2 + shift) (x_{0j} - \bar{x})^2, \quad j = 1, 2, \dots, n_0. \quad (4.6)$$

In equation (4.4), it is assumed that

$$\begin{pmatrix} b_{0i} \\ b_{1i} \\ b_{2i} \end{pmatrix} \sim MN \left(\mathbf{0}, \begin{bmatrix} \sigma_0^2 & 0 & 0 \\ 0 & \sigma_1^2 & 0 \\ 0 & 0 & \sigma_2^2 \end{bmatrix} \right), \quad \varepsilon_i \sim MN(\mathbf{0}, \sigma^2 \mathbf{I}_{n \times n}).$$

Here, $\sigma_0^2 = \sigma_1^2 = \sigma_2^2 = 0.5$, $\sigma^2 = 1$, $\beta_1 = 3$, $\beta_2 = 2$ and $x_{ij} = j$, $i = 1, 2, \dots, m$, $j = 1, 2, \dots, n$.

Consequently, the PA parameter vector for the in-control process is $\beta_{PA}^T = (60.5, -19, 2)$.

Additionally, $m_1 = 20$, $m = 30$ and $n = 10$. Thus, one third of the profiles are from the out-of-control process. The Phase II control limits were obtained based on the Phase I average (based on 5,000 samples) PA parameter vector and average variance-covariance matrix estimates to achieve the simulated in-control ARL, ARL_0 , equal to 200. The performance of Phase II control

charts are obtained by using either a small (shift=0.05), moderate (shift=0.175) or large (shift=0.3) Phase I shift value. The profiles for Phase II were generated via the same model as described above for Phase I with twelve Phase II shift values ranging from 0 (that is, the process is in-control) to 0.3.

Table 2 lists the average ARLs of the Phase II T^2 control charts, listed for appropriate values of the shift in Phase II, obtained by using the estimates from both methods when the out-of-control process has a shift=0.05 in Phase I. When the Phase I shift = 0.05, the simulation results from Chen *et al.*¹⁶ showed that the average estimated PA parameter vector based on 5000 samples with the cluster-based method was $\hat{\boldsymbol{\beta}}^T = (61.00, -19.18, 2.02)$ and the estimated variance-covariance matrix was

$$\hat{\mathbf{V}} = \begin{bmatrix} 0.387 & 0.007 & -0.003 \\ 0.007 & 0.407 & 0.002 \\ -0.003 & 0.002 & 0.441 \end{bmatrix}.$$

To achieve an $ARL_0=200$, the UCL for the cluster-based T^2 control chart, determined via simulation using the Phase I estimates, is 50.8.

The average estimated PA parameters based on 5000 samples, with the non-cluster-based method from Chen *et al.*¹⁶, was $\hat{\boldsymbol{\beta}}^T = (61.01, -19.18, 2.02)$ and corresponding estimated variance-covariance matrix was

$$\hat{\mathbf{V}} = \begin{bmatrix} 0.476 & -0.001 & -0.002 \\ -0.001 & 0.467 & 0.003 \\ -0.002 & 0.003 & 0.501 \end{bmatrix}.$$

With $ARL_0=200$, the UCL for the non-cluster-based T^2 control chart is 41.4. Clearly, the estimates resulting from the two methods are nearly equal because the size of the Phase I shift is very small. The ARLs for different Phase II shifts are presented in Table 2.

Table 2: ARL_{CB} and ARL_{NCB} with Phase I shift=0.05, $ARL_0 \approx 200$. ARL_{CB} represents the ARL for the cluster-based method and ARL_{NCB} represents the ARL for the non-cluster-based method (bolded cells represent the better values).

Phase II Shift	ARL_{CB}	ARL_{NCB}
0	186.2	186.1
0.05	77.6	78.3
0.075	22.1	22.4
0.1	8.2	8.2
0.125	3.7	3.7
0.15	2.1	2.1
0.175	1.4	1.4
0.2	1.2	1.2
0.225	1.1	1.1
0.25	1.0	1.0
0.275	1.0	1.0
0.3	1.0	1.0

When the Phase I shift=0.15, the simulation results from Chen *et al.*¹⁶ showed that the average estimated PA parameters using the cluster-based method was $\hat{\beta}^T = (61.71, -19.44, 2.04)$ and the estimated variance-covariance matrix was

$$\hat{V} = \begin{bmatrix} 0.966 & -0.190 & 0.014 \\ -0.190 & 0.507 & -0.004 \\ 0.014 & -0.004 & 0.476 \end{bmatrix}.$$

With $ARL_0=200$, the UCL for the cluster-based T^2 control chart is 27.39.

The average estimated PA parameter vector based on the non-cluster-based method from Chen *et al.*¹⁶ was $\hat{\beta}^T = (62.01, -19.55, 2.05)$ and corresponding estimated variance-covariance matrix was

$$\hat{V} = \begin{bmatrix} 1.425 & -0.347 & 0.029 \\ -0.347 & 0.591 & -0.008 \\ 0.029 & -0.008 & 0.503 \end{bmatrix}.$$

With $ARL_0=200$, the UCL for the non-cluster-based T^2 control chart is 21.58. The ARLs for different Phase II shifts are presented in Table 3.

Table 3: ARL_C and ARL_{NCB} with Phase I shift=0.15, $ARL_0 \approx 200$. ARL_{CB} represents the ARL for the cluster-based method and ARL_{NCB} represents the ARL for the non-cluster-based method (bolded cells represent the better values).

Phase II Shift	ARL_{CB}	ARL_{NCB}
0	104.5	97.5
0.05	590.5	927.2
0.075	146.7	392.8
0.1	36.6	99.0
0.125	11.8	27.7
0.15	5.0	9.8
0.175	2.6	4.3
0.2	1.7	2.4
0.225	1.3	1.6
0.25	1.1	1.4
0.275	1.0	1.1
0.3	1.0	1.0

When the Phase I shift =0.3, the simulation results from Chen *et al.*¹⁶ showed that the average estimated PA parameter vector using the cluster-based method was $\hat{\beta}^T = (61.73, -19.08, 2.01)$ and the estimated variance-covariance matrix was

$$\hat{V} = \begin{bmatrix} 0.748 & -0.108 & 0.006 \\ -0.108 & 0.489 & -0.002 \\ 0.006 & -0.002 & 0.481 \end{bmatrix}.$$

With $ARL_0=200$, the UCL for the cluster-based T^2 control chart is 24.55.

The average estimated PA parameters based on the non-cluster-based method from Chen *et al.*¹⁶ was $\hat{\beta}^T = (63.52, -20.10, 2.10)$ and corresponding estimated variance-covariance matrix was

$$\hat{V} = \begin{bmatrix} 2.874 & -0.878 & 0.076 \\ -0.878 & 0.785 & -0.025 \\ 0.076 & -0.025 & 0.504 \end{bmatrix}.$$

With $ARL_0=200$, the UCL for the non-cluster-based T^2 control chart is 19.27. The ARLs for different Phase II shifts are presented in Table 4.

Table 4: ARL_CB and ARL_NCB with Phase I shift=0.3, $ARL_0 \approx 200$. ARL_CB represents the ARL for the cluster-based method and ARL_NCB represents the ARL for the non-cluster-based method (bolded cells represent the better values).

Phase II Shift	ARL_CB	ARL_NCB
0	189.4	50.3
0.05	33.9	371.4
0.075	11.1	706.1
0.1	4.8	891.8
0.125	2.5	718.4
0.15	1.6	375.3
0.175	1.2	146.1
0.2	1.1	51.5
0.225	1.0	18.5
0.25	1.0	7.5
0.275	1.0	3.7
0.3	1.0	2.2

Tables 2-4 show that the cluster-based T^2 control chart can detect a change in the process at all Phase II shifts with far fewer observations than the ARL_0 of 200. The non-cluster-based T^2 control chart continues to require a very large number of observations to detect a process shift, especially for Phase II shifts in the 0.05 to 0.2 range.

Tables 2-4 also show that the cluster-based T^2 control chart can detect the shift in the process with very few observations even when the shift value is very small. For example, Table 4 shows that, on average, this chart only requires about 34 samples to signal when the Phase II shift is 0.05, and it signals almost immediately when the Phase II shift is equal to or greater than 0.15. Also, Tables 2-4 show that the non-cluster-based T^2 control chart incorrectly detects that an observed profile is from an out-of-control process (a “false negative”) far sooner than expected than the cluster-based T^2 control chart when the process is in-control.

5. Conclusion

The goal of this paper is to illustrate how the performance of the Phase I analysis in profile monitoring can affect the performance of the Phase II control chart. From the example and the Monte Carlo study, one can conclude that the better the Phase I estimates are, the better the Phase II results will be as indicated by an in-control ARL nearer the desired value and smaller ARL_1 values for smaller Phase II sustained shift sizes. The poorer the Phase I estimates are, the poorer the Phase II results will be, as indicated by very large ARL_1 values even for moderate Phase II sustained shift sizes.

In the example, one can see that the Phase I estimates from both the cluster-based method and the non-cluster based method are not very close to the true parameters due to the small amount of data and the impact of the data from the out-of-control process. Both methods provided poor Phase I estimates and ended up with poor performance of the Phase II control charts. For example, both Phase II control charts have many more false alarms than expected. However, the non-cluster based control chart not only has more false alarms, it takes much longer than expected to detect changes in the process. The Monte Carlo study demonstrated the average effect of Phase I estimates on the performance of Phase II control charts across a variety of sustained shift values. From Tables 2-4 in Section 5, one can conclude that the cluster-based T^2 control chart works uniformly better than the non-cluster-based T^2 control chart for the cases considered due to its better Phase I estimates. Also, we can see that poor Phase I estimates can severely compromise the performance of Phase II control charts.

Perhaps a better method of conducting the Monte Carlo study would be to perform the Phase II analysis separately based on each of the 5,000 individual estimated PA parameter vectors and their corresponding variance-covariance matrices rather than the average of the estimated PA parameter vectors and the average of the variance-covariance matrices. This former approach is what we used for the example and would mimic the way in which the Phase I analysis leads to the Phase II analysis in real applications. The problem we encountered when implementing the former approach is the tremendous computing resources needed to complete the analysis. The method of using the average values required far less computational effort and produced results that clearly favor the use of the cluster-based method.

We believe that clustering methods in Phase I will lead to better Phase II performance. Generally, we encourage more research on the effect of Phase I estimation error on Phase II performance when monitoring profiles. This suggestion is in agreement with the idea expressed in Jensen *et al.*¹² that parameter estimation in Phase I may greatly affect the performance of control charts during Phase II.

References

1. Kang L and Albin SL. On-Line Monitoring When the Process Yields a Linear Profile. *Journal of Quality Technology* 2000; 32: 418-426.
2. Kim K, Mahmoud MA and Woodall WH. On the Monitoring of Linear Profiles. *Journal of Quality Technology* 2003; 35: 317-328.
3. Mahmoud MA and Woodall WH. Phase I Analysis of Linear Profiles with Calibration Applications. *Technometrics* 2004; 46: 380-391.
4. Jin JH and Shi JJ. Automatic Feature Extraction of Waveform Signals for in-Process Diagnostic Performance Improvement. *Journal of Intelligent Manufacturing* 2001; 12: 257-268.
5. Walker E and Wright SP. Comparing Curves Using Additive Models. *Journal of Quality Technology* 2002 ; 34: 118-129.
6. Gupta S, Montgomery DC and Woodall WH. Performance Evaluation of Two Methods for Online Monitoring of Linear Calibration Profiles. *International Journal of Production Research* 2006; 44: 1927-1942.
7. Ding Y, Zeng L and Zhou S. Phase I Analysis for Monitoring Nonlinear Profiles in Manufacturing Processes. *Journal of Quality Technology* 2006; 38: 199-216.
8. Williams JD, Birch JB, Woodall WH and Ferry NM. Statistical Monitoring of Heteroscedastic Dose-Response Profiles from High-Throughput Screening. *Journal of Agricultural Biological and Environmental Statistics* 2007; 12: 216-235.
9. Williams JD, Woodall WH and Birch JB. Statistical Monitoring of Nonlinear Product and Process Quality Profiles. *Quality and Reliability Engineering International* 2007; 23: 925-941.

10. Chicken E, Pignatiello JJ Jr. and Simpson JR. Statistical Process Monitoring of Nonlinear Profiles Using Wavelets. *Journal of Quality Technology* 2009; 41: 198-212.
11. Hung YC, Tsai WC, Yang SF, Chuang SC and Tseng YK. Nonparametric Profile Monitoring in Multi-Dimensional Data Spaces. *Journal of Process Control* 2012; 22: 397-403.
12. Jensen WA, Birch JB and Woodall WH. Monitoring Correlation within Linear Profiles Using Mixed Models. *Journal of Quality Technology* 2008; 40: 167-183.
13. Jensen WA and Birch JB. Profile Monitoring Via Nonlinear Mixed Models. *Journal of Quality Technology* 2009; 41: 18-34.
14. Qiu P, Zou C and Wang Z. Nonparametric Profile Monitoring by Mixed Effects Modeling. *Technometrics* 2010; 52: 265-277.
15. Abdel-Salam ASG, Birch JB and Jensen WA. A Semiparametric Mixed Model Approach to Phase I Profile Monitoring. *Quality and Reliability Engineering International* 2013; 29: 555-569.
16. Chen Y, Birch JB and Woodall WH. Cluster-Based Profile Monitoring in Phase I Analysis. *Journal of Quality Technology* 2014; to appear.
17. Mahmoud, MA. The Performance of Phase II Simple Linear Profile Approaches when Parameters Are Estimated. *Communications in Statistics - Simulation and Computation* 2012; 41:1816-1833
18. Kang, L and Albin, SL. On-line Monitoring when the Process Yields a Linear Profile. *Journal of Quality Technology* 2000; 32:418–426.
19. Mahmoud, MA, Morgan, JP and Woodall, WH. The Monitoring of Simple Linear Regression Profiles with Two Observations per Sample. *Journal of Applied Statistics* 2010; 37(8):1249-1263.
20. Aly, AA, Mahmoud, MA and Woodall, WH. A Comparison of the Performance of Phase II Simple Linear Profile Control Charts when Parameters Are Estimated. *Communications in Statistics Simulation and Computation* 2014; to appear.
21. Laird NM and Ware JH. Random-effects Models for Longitudinal Data. *Biometrics* 1982; 38: 963–974.
22. Schabenberger O and Pierce FJ. *Contemporary Statistical Models for the Plant and Soil Sciences*, Boca Raton: CRC Press, 2002.

23. Ruppert D, Wand MP and Carroll RJ. *Semiparametric Regression*. Cambridge University Press: Cambridge, NY, 2003.
24. Seber GAF and Wild CJ. *Nonlinear Regression*. Hoboken, N.J.: Wiley-Interscience, 2003.
25. Demidenko E. *Mixed Models : Theory and Applications*. Hoboken, N.J.: Wiley-Interscience, 2004.
26. Sullivan JH and Woodall WH. A Comparison of Multivariate Control Charts for Individual Observations. *Journal of Quality Technology* 1996; 28: 398-408.