A Phase I Cluster-Based Method for Analyzing Nonparametric Profiles

YAJUAN CHEN

Pfizer, Andover, MA 01845

JEFFERY B. BIRCH and WILLIAM H. WOODALL

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

A cluster-based method was used by Chen *et al.*[24] to analyze parametric profiles in Phase I of the profile monitoring process. They showed performance advantages in using their cluster-based method of analyzing parametric profiles over a non-cluster-based method with respect to more accurate estimates of the parameters and improved classification performance criteria. However, it is known that, in many cases, profiles can be better represented using a nonparametric method. In this study, we use the cluster-based method to analyze profiles that cannot be easily represented by a parametric function. The similarity matrix used during the clustering phase is based on the fits of the individual profiles with p-spline regression. The clustering phase will determine an initial main cluster set which contains greater than half of the total profiles in the historical data set. The profiles with in-control T^2 statistics are sequentially added to the initial main cluster set and upon completion of the algorithm, the profiles in the main cluster set are classified as the in-control profiles and the profiles not in the main cluster set are classified as out-of-control profiles. A Monte Carlo study demonstrates that the cluster-based method results in superior performance over a non-cluster-based method with respect to better classification and higher power in detecting out-of-control profiles. Also, our Monte Carlo study shows that the cluster-based method has better performance than a non-cluster-based method whether the model is correctly specified or not. We illustrate the use of our method with data from the automotive industry.

Keywords: Mixed Models; Nonparametric; Profile Monitoring; Robust; T^2 Statistic.

1. Introduction

Profile monitoring is a well-known approach in statistical process control (SPC) where the quality of a product or a process is characterized by a functional relationship between a response variable and one or more explanatory variables. This functional relationship here is referred to as a "profile". Profile monitoring, and SPC in general, is conducted over two phases, labeled as Phase I and Phase II. Literature reviews and an introduction to profile monitoring can be found in Woodall *et al.*[1] , Woodall [2] and Noorossana *et al.*[3]. In SPC, a single profile is commonly used to characterize the quality of a product or process. However, Noorossana *et al.*[4] also showed that in some cases, multiple profiles can be used simultaneously to characterize in a better way the quality of a product or process. For ease of illustration, a single profile will be used in our study.

In Phase I profile monitoring, one goal is to separate the in-control process data from the out-of-control process data in the historical data set (HDS). The performance of the Phase I analysis can be measured in terms of how well one can correctly identify the out-of-control process data in the HDS. 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. We consider a sustained shift in the functional relationship in our paper. The profiles from the out-of-control process are usually removed from the HDS and the remaining profiles are used to compute the statistics needed for establishing the control limits used in Phase II analysis.

To separate the in-control process data from the out-of-control process data, the first step is to fit each profile using some appropriate modeling technique. In some applications, the profile can be represented adequately by some parametric function. For example, Croarkin[5], Stover and Brill[6], Kang and Albin[7], Kim *et al.*[8], Mahmoud and Woodall[9], Wang and Tsung[10], Gupta *et al.*[11] and Zhang *et al.*[12] have all used parametric profiles in their work. In many other cases, profiles may not be well-modeled by parametric functions. Nonparametric profile applications were studied by Jin and Shi[13], Lada *et al.*[14], Walker and Wright[15], Ding *et al.*[16], Gupta *et al.*[11], Williams *et al.*[17], Williams *et al.*[18], Zou *et al.*[19] and Abdel-Salam *et al.*[20]. To analyze parametric profiles, the presence of profiles from the out-of-control process can be detected by the Hotelling's T^2 statistic based on the estimated regression parameters. Kang and Albin[7], Kim *et al.*[8] and Mahmoud and Woodall[9] used the Hotelling's T^2 statistic based on the estimated regression parameters to detect the out-of-control profiles. The Hotelling's T^2 statistic has also been

used to detect out-of-control nonparametric profiles. For example, Abdel-Salam *et al.*[20] proposed using estimated best linear predictors (eblups) to calculate the Hotelling's T^2 statistics.

In the profile monitoring context, data observed within each profile are generally correlated as a result of being obtained as repeated measurements on the same experimental unit. In order to account for the correlation structure within each profile, Jensen *et al.*[21], Jensen and Birch[22], Qiu *et al.*[23] and Chen *et al.*[24] proposed the use of mixed models to model the profiles. Based on the linear mixed model, Jensen *et al.*[21] and Jensen and Birch[22] proposed detecting profiles from an out-of-control process in Phase I by comparing each estimated profile specific (PS) curve to the estimated population average (PA) curve using the T^2 statistic. However, Chen *et al.*[24] found that the ability of this method to distinguish profiles from the in-control and out-of-control processes will be distorted if there is a moderate or large shift among the profiles. Instead of using the estimated PA parameter vector based on all profiles from the entire data set to calculate the T^2 statistic, Chen *et al.*[24] proposed a cluster-based method to cluster the profiles before estimating the PA parameter vector, and demonstrated the performance advantages of using their cluster-based method over the method provided by Jensen *et al.*[21] and other robust methods.

Chen *et al.*[24] however, only considered the cluster-based method applied to the monitoring of parametric profiles. In many cases, profiles cannot be well modeled by parametric functions. For example, for automobile engine data, used in Chen *et al.*[24], the relationship between the torque produced by the engine and engine's speed in revolutions per minute (RPM) is used to characterize the quality of the engine. A plot of the raw data for 20 engines contained in Table A-1 is shown in Figure 1.1. Parametric profile monitoring methods, using a quadratic curve for each engine, have been applied to this data by Amiri *et al.*[25] and Abdel-Salam *et al.*[20] and all engines were found to conform to the same in-control process. Using their cluster-based method applied to the parametric fits to the engine profiles (based on fitting quadratic curves to each engine), Chen *et al.*[24] detected engine 11 as being from a process different from the other engines and thus suggesting potential mechanical issues with engine 11. Further work by Abdel-Salam *et al.*[20] has shown that the relationship between torque and speed in RPM can be better represented by a nonparametric function. They used a nonparametric mixed model, based on p-splines, to detect engines 11 and 20 to be from the out-of-control process. Further analysis of this example data will be illustrated in Section 4.

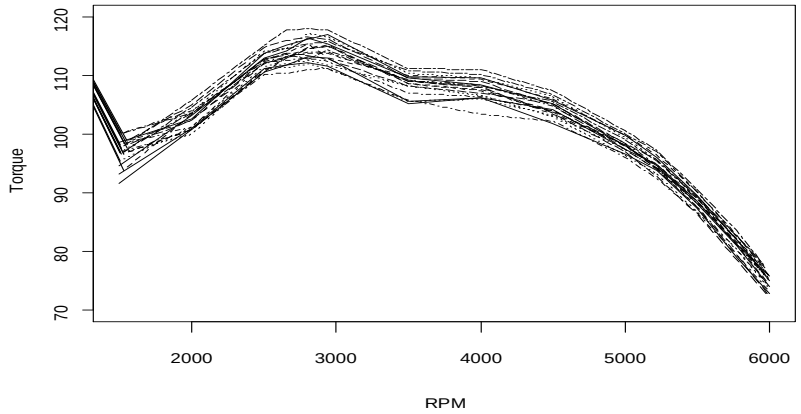

Figure 1.1: The plot of observed data for 20 automobile engines

2. Nonparametric Mixed Model in Profile Monitoring

The mixed model has been used to monitor correlated data within each profile in the HDS. Based on the mixed model, the i^{th} profile can be represented nonparametrically as

$$\boldsymbol{y}_i = f\left(\boldsymbol{x}_i\right) + \xi_i\left(\boldsymbol{x}_i\right) + \boldsymbol{\varepsilon}_i \,, i = 1,2,3,...,m, \tag{2.1}$$

where m is the number of profiles in the HDS, $\boldsymbol{y}_i$ is the $n_i \times 1$ response vector associated with the i^{th} profile and n_i is the number of observations within the i^{th} profile. Also, $f\left(\boldsymbol{x}_i\right)$ represents the mean response or PA function, common to all profiles, $\xi_i\left(\boldsymbol{x}_i\right)$ represents the random effects for the i^{th} profile where it is assumed that $\xi_i\left(\boldsymbol{x}_i\right) \sim N\left(0, \quad \sigma_\xi^2 I\right)$. In addition, $\boldsymbol{\varepsilon}_i$ is the $n_i \times 1$ vector of random errors for the i^{th} profile with $\boldsymbol{\varepsilon_i} \square\, MN\left(\boldsymbol{0}, \boldsymbol{R}_i\right)$, where $\boldsymbol{R}_i$ is the $n_i \times n_i$ covariance matrix. More details regarding the mixed model can be found in Schabenberger and Pierce[26], Seber and Wild[27], Ruppert *et al.*[28] and Demidenko[29].

According to the Equation (2.1), the j^{th} response from the i^{th} profile from the nonparametric mixed model can be written as:

$$y_{ij} = f\left(x_{ij}\right) + \xi_i\left(x_{ij}\right) + \varepsilon_{ij} \,, i = 1,2,3,...,m \,, j = 1,2,3,...,n_i.$$

Abdel-Salam *et al.*[20] approximated both $f\left(x_{ij}\right)$ and $\xi_i\left(x_{ij}\right)$ by using p-spline regression with a truncated polynomial basis of order p (other basis functions can be utilized as well). Details of using p-spline regression can also be found in Ruppert *et al.*[28]. Using p-spline regression with the truncated polynomial basis, $f\left(x_{ij}\right)$ and $\xi_i\left(x_{ij}\right)$ can be approximated as

$$f\left(x_{ij}\right) \approx \beta_0 + \sum_{l=1}^{p} \beta_l x_{ij}^{\ l} + \sum_{k=1}^{K_1} u_{pk}\left(x_{ij} - \kappa_k\right)_+^p \qquad \sum_{k=1}^{K_1} u_k^2 \le c \tag{2.2}$$

and

$$\xi_i\left(x_{ij}\right) \approx b_{i0} + \sum_{l=1}^{p} b_{il} x_{ij}^{\ l} + + \sum_{k=1}^{K_2} t_{ik}\left(x_{ij} - \kappa_k\right)_+^p \quad i = 1,2,3,...,m\,,\ j = 1,2,3,...,n_i \qquad \sum_{k=1}^{K_2} t_{ik}^2 \le c_i, \tag{2.3}$$

respectively, where p is the order of the polynomial and $\kappa_1, \kappa_2, ..., \kappa_K$ are the knots. The values K_1 and K_2 are the number of knots chosen for $f\left(x_{ij}\right)$ and $\xi_i\left(x_{ij}\right)$, respectively. Additionally, $\left(x_{ij} - \kappa_k\right)_+^p$ is defined as 0 for $x_i \le \kappa_k$ and $\left(x_{ij} - \kappa_k\right)^p$ otherwise. The values c and c_i for $i = 1,2,3,...,m$ are fixed constants, used to control the smoothness of the nonparametric components, and bounded by 0 and ∞. Given the relationship between the p-spline regression approximation and the linear mixed model (see Ruppert *et al.*[28]), the approximation for the i^{th} profile in Equation (2.1) can be described succinctly in the linear mixed model framework as

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u} + \boldsymbol{X}_i\boldsymbol{b}_i + \boldsymbol{E}_i\boldsymbol{t}_i + \boldsymbol{\varepsilon}_i, i = 1,2,...,m, \tag{2.4}$$

where

$$\boldsymbol{X}_i = \begin{bmatrix} 1 & x_{i1} & \cdots & x_{i1}^{\ p} \\ 1 & x_{i2} & \cdots & x_{i2}^{\ p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{in} & \cdots & x_{in}^{\ p} \end{bmatrix},\ \boldsymbol{Z}_i = \begin{bmatrix} \left(x_{i1} - \kappa_1\right)_+^p & \cdots & \left(x_{i1} - \kappa_{K_1}\right)_+^p \\ \vdots & \ddots & \vdots \\ \left(x_{in} - \kappa_1\right)_+^p & \cdots & \left(x_{in} - \kappa_{K_1}\right)_+^p \end{bmatrix} \text{and } \boldsymbol{E}_i = \begin{bmatrix} \left(x_{i1} - \kappa_1\right)_+^p & \cdots & \left(x_{i1} - \kappa_{K_2}\right)_+^p \\ \vdots & \ddots & \vdots \\ \left(x_{in} - \kappa_1\right)_+^p & \cdots & \left(x_{in} - \kappa_{K_2}\right)_+^p \end{bmatrix}.$$

The fixed parameter vectors are $\boldsymbol{\beta}^T = \left[\beta_0, \beta_{1,}..., \beta_p\right]$ and $\boldsymbol{b}_i^T = \left[b_{i0}, b_{i1,}..., b_{ip}\right]$. The penalized parameter vectors are $\boldsymbol{\mu}^T = \left[\mu_0, \mu_{1,}..., \mu_{K_1}\right]$ and $\boldsymbol{t}_i^T = \left[t_{i0}, t_{i1,}..., t_{iK_2}\right]$ with $\boldsymbol{\mu}^T\boldsymbol{\mu} \le c$ and $\boldsymbol{t}_i^T\boldsymbol{t}_i \le c_i, i = 1,\ldots, m.$ According to Equation (2.4), the i^{th} estimated profile can be written as

$$\hat{\boldsymbol{y}}_i = \boldsymbol{X}_i\hat{\boldsymbol{\beta}} + \boldsymbol{Z}_i\hat{\boldsymbol{u}} + \boldsymbol{X}_i\hat{\boldsymbol{b}}_i + \boldsymbol{E}_i\hat{\boldsymbol{t}}_i, i = 1, 2, ..., m, \tag{2.5}$$

and the corresponding estimated PA profile is

$$\hat{\boldsymbol{y}}_{PA} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{u}}. \tag{2.6}$$

If we define $\hat{\boldsymbol{\varphi}}_i = \begin{bmatrix}\hat{\boldsymbol{b}}_i \\ \hat{\boldsymbol{t}}_i\end{bmatrix}$ for $i = 1, 2, ...m,$ then, according to Abdel-Salam *et al.*[20], the T^2 statistic for the i^{th} profile can be obtained based on the eblups $\hat{\boldsymbol{\varphi}}_i$ and can be calculated as

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

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

The successive difference variance-covariance estimator is used here since Sullivan and Woodall[30] showed that use of the successive difference estimate $\hat{\boldsymbol{V}}_D$ is effective in detecting sustained step changes in the process that may occur in Phase I data. Other robust variance-covariance estimators (see Chenouri *et al.*[31]) can be used instead of $\hat{\boldsymbol{V}}_D$ if a change in the process is suspected to be due to something other than a sustained shift. Unusual profiles can be determined by comparing T_i^2 with a value from chi-squared distribution. The i^{th} estimated PS curve will be declared as outlying if $T_i^2 \ge \boldsymbol{\chi}^2_{(df,\alpha)}$ for $i = 1, 2, ..., m,$ where α represents the significance level, typically obtained by using a Bonferroni adjustment, and df represents the degree of freedom which is equal to $p+K_2$, where K_2 is the number of knots and p is the order of the truncated polynomial basis in the p-spline regression.

The T^2 statistic based on the eblups from the mixed model works well in the situation where the size of the sustained shift for the out-of-control process is small. However, Chen *et al.*[24] showed that this method can be distorted if the shift size from the out-of-control process is moderate to large. They

proposed to remedy this situation by using a cluster-based method in monitoring parametric profiles. Also, they demonstrated the performance advantages of using their cluster-based method to detect the out-of-control profiles over the non-cluster-based method. In addition, Chen *et al.*[24] demonstrated that the probability of signal (POS), a commonly used statistic for evaluating the performance of a method in detecting a change in the process during Phase I analysis, can be very misleading. They proposed supplementing the POS in Phase I analysis with the metrics formed from the classification table (Table 2.1) after completing the Phase I analysis

Table 2.1: Classification table for Phase I analysis

Classified set / Actual set	Out-of-control process	In-control process
Out-of-control process	A	B
In-control process	C	D

The performance metrics based on this table are the fraction correctly classified (FCC), sensitivity, specificity, false positive rate (FPR) and false negative rate (FNR). Using Table 2.1, the FCC can be defined as $(A+D)/(A+B+C+D)$; the sensitivity and the specificity can be calculated as $A/(A+B)$ and $D/(C+D)$ respectively; while the FPR and the FNR are computed, respectively, as $C/(A+C)$ and $B/(B+D)$. All these metrics are bounded by 0 and 1. A method will perform well in Phase I analysis by achieving large values for FCC, sensitivity and specificity and small values for FPR and FNR. Fraker *et al.*[32] pointed out that similar metrics are used in biosurveillance for applications in which outbreak time periods are to be distinguished from non-outbreak time periods.

In our study, the cluster-based method will be used in analyzing nonparametric profiles. We assume there is a correlation structure within each profile and use a mixed model. Details of the proposed method will be given in Section 3 and the performance of the proposed method will be compared to the method of Abdel-Salam *et al.*[20], which is referred to herein as the non-cluster-based method.

3. Cluster-Based Method in Monitoring Nonparametric Profiles

Recall that in Equation (2.1), the j^{th} response from the i^{th} nonparametric profile can be written as

$$y_{ij} = f\left(x_{ij}\right) + \xi_i\left(x_{ij}\right) + \varepsilon_{ij}\ , i = 1,2,3,...,m,\ \ j = 1,2,3,...,n_i,$$

where $f\left(x_{ij}\right)$ and $\xi_i\left(x_{ij}\right)$ are nonparametric functions and can be estimated by using a nonparametric mixed model, as for example, model (2.4). With the proposed cluster-based method one will first fit a curve individually to each profile using the p-spline regression method (using, for example, the truncated polynomial basis). Each estimated profile curve will then be represented by a vector of estimated parameters. The corresponding variance-covariance matrix of these *m* vectors, $\boldsymbol{V}$, can be estimated by some appropriate method. The second step is to obtain the similarity matrix, $\boldsymbol{S}$, based on the estimated parameter vectors and variance-covariance matrix. An appropriate clustering method is then used to cluster the profiles based on $\boldsymbol{S}$.

To obtain a set of similar profiles, hierarchical clustering with a proper linkage is performed until a main cluster set containing at least half the profiles is formed. The profiles in the main cluster set are used to obtain $\hat{f}\left(\boldsymbol{x}\right)$, via the mixed model, an estimate of the PA function $f\left(\boldsymbol{x}\right)$. Then $\hat{f}\left(\boldsymbol{x}\right)$ is used with the previously estimated variance-covariance matrix, $\hat{\boldsymbol{V}}$, to calculate the T^2 statistics for the profiles not contained in the initial main cluster set. The profiles which have in-control T^2 statistics (that is, T^2 is less than the control limit of the T^2 chart) are then added to the main cluster set to form a new main cluster set. One repeats the above procedure of updating the main cluster set by adding the profiles not contained in the main cluster set. The iteration stops with either the smallest T^2 statistic for the remaining profiles outside of the main cluster set is beyond the control limit or all the profiles in the HDS have been moved to the main cluster set. Upon completion of the algorithm, those profiles contained in the main cluster set are labeled as profiles from the "in-control process" and those not included in the main cluster set are labeled as profiles from an "out-of-control process". A similar iterative procedure was also used in a multivariate control chart setting by Shiau and Sun[33]. The proposed algorithm is now outlined in detail.

Step 1. Fit the i^{th} profile by using the p-spline regression method. Note that the p-spline with the first order truncated polynomial basis is used as an example in our paper. We have

$$y_{ij} = f_i\left(x_{ij}\right) + \varepsilon_{ij} = \beta_{0i} + \beta_{1i} + \sum_{k=1}^{K} \mu_{ki}\left(x_{ij} - \kappa_k\right)_+ + \varepsilon_{ij},\ i = 1,2,3,....,m,\ j = 1,2,3,...,n_i,\ \kappa = 1,2,3,...,K,$$

$$\sum_{k=1}^{K} \mu_{ki}^2 \le c_i, \quad 0 < c_i < \infty.$$

The i^{th} estimated profile can be represented by a vector of estimated parameter $\hat{\boldsymbol{\varphi}}_i$, where $\hat{\boldsymbol{\varphi}}_i = \left(\hat{\beta}_{0i} \quad \hat{\beta}_{1i} \quad \hat{\mu}_{1i} \quad ... \quad \hat{\mu}_{Ki}\right)$. An appropriate estimated variance-covariance matrix, $\hat{\boldsymbol{V}}$, then can be estimated based on the $\hat{\boldsymbol{\varphi}}_i$ vectors. For example, the successive difference estimator is used here since it is assumed that the out-of-control profiles are due to a sustained shift. Consequently, $\boldsymbol{V}$, is estimated by

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

Step 2. Using $\hat{\boldsymbol{V}}$, obtained in step 1, compute a $m \times m$ similarity matrix $\boldsymbol{S}$, where the i, j entry is defined as

$$s_{ij} = \left(\hat{\boldsymbol{\varphi}}_i - \hat{\boldsymbol{\varphi}}_j\right)^T \hat{\boldsymbol{V}}^{-1} \left(\hat{\boldsymbol{\varphi}}_i - \hat{\boldsymbol{\varphi}}_j\right).$$

Step 3. Perform a hierarchical cluster analysis using an appropriate linkage function on the given similarity matrix to obtain the main cluster set of profiles. The initial main cluster set is defined as the first cluster set that contains greater than half of the profiles. We denote the set of indices for the profiles in the main cluster set by C. Stop the clustering process as soon as at least greater than half of the profiles are contained in the initial main cluster set C.

Step 4. Use the p-spline mixed model approach to estimate the parameters for the PA profile based on the profiles in C , The estimated PA profile can be represented as

$$y_{PA,j} = \hat{f}(x_j) \approx \hat{\beta}_0 + \hat{\beta}_1 + \sum_{k=1}^{K} \hat{\mu}_k \left(x_j - \kappa_k\right)_+, j = 1,2,...,n,$$

$$\sum_{k=1}^{K} \mu_k^2 \le c \qquad 0 < c < \infty.$$

We define $\overline{\boldsymbol{\varphi}} = \left(\hat{\beta}_0 \quad \hat{\beta}_1 \quad \hat{\mu}_1 \quad \hat{\mu}_2 \quad ... \quad \hat{\mu}_K\right)$. Then for all profiles not in the main cluster set C, compute

$$T_i^2 = \left(\hat{\boldsymbol{\varphi}}_i - \bar{\boldsymbol{\varphi}}\right)^T \hat{\boldsymbol{V}}^{-1} \left(\hat{\boldsymbol{\varphi}}_i - \bar{\boldsymbol{\varphi}}\right),$$

where " i " denotes the i^{th} profile not contained in C and add the profiles which have $T_i^2 < \chi^2_{[1-\alpha/m],df}$ to C. Here, $\chi^2_{[1-\alpha/m],df}$ is the [1- α/m] quantile of a chi-squared distribution. Here α is the level of the test and [α/m] is the Bonferroni adjustment for multiple comparisons. The degrees of freedom used for the chi-square distribution is $K+p$, where K is the number of knots and p is the order of the truncated polynomial basis in the p-spline regression.

Step 5. Repeat step 4 until no profile can be added to the main cluster set C or all profiles have been added to the main cluster set C. All these profiles in the main cluster set C will be classified as profiles from the in-control process and will be used to estimate the PA profile with the nonparametric mixed model for Phase II analysis. The estimated PA profile can then be obtain using Equation (2.6) based on all the profiles in C. Individual profiles in C can be estimated using Equation (2.5). The estimated PA profile may then be used to set control limits for the process in Phase II.

Nearly identical results were obtained in all examples and simulations using either "ward" or "complete" linkage. Other linkage functions may work equally as well. Complete linkage was used in all results presented in subsequent sections. Also, in this algorithm, the successive difference variance-covariance estimator is recommended if there is concern about a sustained shift. Other robust variance-covariance estimators such as minimum volume ellipsoid (MVE), minimum covariance determinant (MCD), reweighted MVE or the reweighted MCD can be used as well. The user needs to decide which variance-covariance matrix estimator is best suited for his/her situation.

4. The Automobile Engine Application

In the automotive application from Amiri *et al.*[26] the goal is to study the relationship between the torque produced by an engine and the engine speed in revolutions per minture (RPM). The application considered here is for the engine type TU3, which are assembled for a French automobile, the Peugeot. In this example, there are 20 engines in the HDS and for each egnine, the speed values were set equal to 1500, 2000, 2500, 2660, 2800, 2940, 3500, 4000, 4500, 5000, 5225, 5500, 5775, and 6000 RPM and the engine's corresponding torque values were measured. An engine with mechanical or other issues will

yield a relationship that is different from other engines. The raw data set (see Table A-1 in the Appendix), where individual data points for each engine are connected by straight-line segments, is shown in Figure 1.1. This data set has been analyzed using profile analysis methods by Abdel-Salam *et al.*[20] and Chen *et al.*[24]. In our study, the cluster-based method with a nonparametric mixed model, using the p-spline method with the first-order truncated basis, will be used to fit the relationship and to detect any unusual engines.

Step 1. Using p-spline regression, we fit to the data for each engine with the model

$$f_i\left(x_{ij}\right) = \beta_{0i} + \beta_{1i} + \sum_{k=1}^{K} \mu_{ki}\left(x_{ij} - \kappa_k\right)_+ + \varepsilon_{ij},\ i = 1,2,3,....,20,\ \ j = 1,2,3,...,14,\ \kappa = 1,2,3,...,\mathrm{K},$$

$$\sum_{k=1}^{K} \mu_{ki}^2 \le c_i \qquad 0 < c_i < \infty.$$

We define $\boldsymbol{\varphi}_i = \left(\beta_{0i} \quad \beta_{1i} \quad \mu_{1i} \quad ... \quad \mu_{Ki}\right)$ and the corresponding $\hat{\boldsymbol{\varphi}}_i$ for the i^{th} engine can be represented as $\hat{\boldsymbol{\varphi}}_i = \left(\hat{\beta}_{0i} \quad \hat{\beta}_{1i} \quad \hat{\mu}_{1i} \quad ... \quad \hat{\mu}_{Ki}\right)$. In this case, since there are 14 observations for each engine, choosing K=4 equally spaced knots seems reasonable. Table 4.1 lists $\hat{\boldsymbol{\varphi}}_i$ for each engine.

Step 2. Using $\hat{\boldsymbol{\varphi}}_i$ computed in step 1, we obtained the similarity matrix $\boldsymbol{S}$.

Step 3. Perform the cluster analysis on the similarity matrix $\boldsymbol{S}$ using complete-linkage. The cluster process is represented by the dendrogram in Figure 4.1. Since there are 20 engines in total, the initial main cluster must consist of at least 11 engines. The cluster history is listed in Table 4.4. One can see that the initial main cluster set contains 9 profiles at step 17 and that 6 more profiles are added to this initial main cluster set in cluster step 18, resulting in 15 profiles in the main cluster. Since this is the first step that the main cluster set contains greater than half of the profiles, the cluster step of the algorithm stops at this point. The cluster history (Table 4.4) shows that the proposed algorithm ended up with 15 engines in the initial main cluster set, consisting of engines 1-10 and 13-17.

Step 4. The corresponding estimated PA profile is then obtained by fitting the p-spline mixed model to the data for the 15 engines in the main cluster. The estimated PA profile is

$$y_{PA,j} = \hat{f}(x_j) \approx \hat{\beta}_0 + \hat{\beta}_1 + \sum_{k=1}^{K=4} \hat{\mu}_k\left(x_j - \kappa_k\right)_+,\ j = 1,2,...,n.$$

If we define $\bar{\boldsymbol{\varphi}} = \left(\hat{\beta}_0 \quad \hat{\beta}_1 \quad \hat{\mu}_1 \quad \hat{\mu}_2 \quad \hat{\mu}_3 \quad \hat{\mu}_4\right)$, then $\bar{\boldsymbol{\varphi}}$ based on 15 engines is

$$\bar{\boldsymbol{\varphi}} = \left(71.831 \quad 0.0160 \quad -0.0176 \quad -0.0040 \quad -0.0071 \quad -0.0151\right).$$

Using $\bar{\boldsymbol{\varphi}}$ and $\hat{\boldsymbol{\varphi}}_i$ in Equation (3.1), the T^2 statistics for the engines not included in the initial main cluster set are calculated and listed below. The cutoff value for the T^2 statistic here is $\chi^2_{1-\frac{\alpha}{m},df} = 18.38$, where $\alpha = 0.05$ and $df = K + p = 5$. According to the observed T^2 statistics and the cutoff value, all engines in the minor sets are added to the initial main cluster set except the 11th engine and 20th engine.

Table 4.1: Estimated $\hat{\boldsymbol{\varphi}}_i$, $i = 1, 2..20.$ for each engine

Index of Engines *i*	$\hat{\beta}_{0i}$	$\hat{\beta}_{1i}$	$\hat{\mu}_{1i}$	$\hat{\mu}_{2i}$	$\hat{\mu}_{3i}$	$\hat{\mu}_{4i}$
1	73.3139	0.016	-0.0141	-0.0085	-0.0065	-0.0157
2	71.7374	0.0157	-0.0119	-0.0107	-0.0056	-0.0139
3	73.6999	0.0146	-0.0173	-0.0019	-0.0074	-0.0139
4	75.8218	0.0134	-0.0153	-0.0037	-0.0063	-0.0152
5	74.4416	0.0149	-0.0127	-0.009	-0.0078	-0.0143
6	79.9753	0.0128	-0.0133	-0.0044	-0.0084	-0.0147
7	66.3589	0.0187	-0.0193	-0.0057	-0.0074	-0.0147
8	71.8467	0.0157	-0.0119	-0.0107	-0.0056	-0.0139
9	70.2356	0.0174	-0.0169	-0.0069	-0.0075	-0.0136
10	80.1105	0.0128	-0.0126	-0.0057	-0.0079	-0.0125
11	71.6214	0.0172	-0.0207	-0.0031	-0.008	-0.0154
12	68.9162	0.0188	-0.0214	-0.0034	-0.0075	-0.0150
13	66.0206	0.0179	-0.0211	-0.002	-0.0068	-0.0143
14	65.7138	0.0185	-0.0203	-0.0042	-0.0063	-0.0166
15	70.4448	0.0162	-0.0184	-0.0031	-0.0083	-0.0132
16	75.8862	0.0141	-0.0166	-0.0023	-0.0081	-0.0151
17	71.938	0.0163	-0.0227	0.0024	-0.0084	-0.0143
18	70.6005	0.0172	-0.018	-0.0043	-0.0085	-0.0143
19	62.7847	0.0191	-0.0243	-0.0001	-0.0062	-0.0181
20	74.8780	0.0149	-0.0154	-0.004	-0.0088	-0.015

Table 4.2: T^2 statistic for engines in the minor set

Index of Engines	11	12	18	19	20
T_i^2 Statistic	18.450	**14.564**	**8.762**	**17.544**	20.387

Step 5. Repeat step 4 and use the mixed model to update the estimated PA profile by using the engines 1-10 and 12-19. The updated $\bar{\varphi}$ and the T^2 statistics are obtained as

$$\bar{\varphi} = \begin{pmatrix} 70.999 & 0.0165 & -0.0185 & -0.0035 & -0.0072 & -0.0153 \end{pmatrix}$$

Table 4.3: T^2 statistic for engines in the minor set

Index of Engines	11	20
T_i^2 Statistic	19.644	18.559

Since the T^2 statistics above show that no profile can be added, the algorithm stops with the in-control engines identified to be 1-10 and 12-19.

Table 4.4: Cluster history based on eblups for 20 engines

Step	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	1	2	3	4	5	6	7	2	8	9	10	11	12	13	14	15	16	17	18	19
3	1	2	3	4	5	6	7	2	8	9	10	11	12	13	14	6	15	16	17	18
4	1	2	3	4	5	6	7	2	7	8	9	10	11	12	13	6	14	15	16	17
5	1	2	3	4	5	6	7	2	7	8	9	10	11	12	3	6	13	14	15	16
6	1	2	3	4	5	6	7	2	7	8	9	10	11	11	3	6	12	13	14	15
7	1	2	3	4	5	6	7	2	7	8	9	10	11	11	3	6	12	13	14	13
8	1	1	2	3	4	5	6	1	6	7	8	9	10	10	2	5	11	12	13	12
9	1	1	2	3	4	5	6	1	6	3	7	8	9	9	2	5	10	11	12	11
10	1	1	2	3	4	4	5	1	5	3	6	7	8	8	2	4	9	10	11	10
11	1	1	2	3	4	4	1	1	1	3	5	6	7	7	2	4	8	9	10	9
12	1	1	2	3	4	4	1	1	1	3	5	6	7	7	2	4	2	8	9	8
13	1	1	2	3	4	4	1	1	1	3	5	5	6	6	2	4	2	7	8	7
14	1	1	2	3	4	4	1	1	1	3	5	5	6	6	2	4	2	7	7	7
15	1	1	2	3	4	4	1	1	1	3	5	5	2	2	2	4	2	6	6	6
16	1	1	2	1	3	3	1	1	1	1	4	4	2	2	2	3	2	5	5	5
17	1	1	2	1	1	1	1	1	1	1	3	3	2	2	2	1	2	4	4	4
18	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**2**	**1**	**1**	**1**	**1**	**1**	**3**	**3**	**3**
19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2
20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

The cluster dendrogram in Figure 4.1 shows that clustering phase ends up with a main cluster set of engines and two minor sets, one minor set contains engines 11-12 and the other minor set contains the engines 18-20. After the sequential addition of the remaining engines to the initial main cluster set, the

cluster-based method identified engines 11 and 20 as from the out-of-control process and engines 12, 18-19 as from the in-control process. The result from the cluster-based method is consistent with the result from Abdel-Salam *et al*[20].

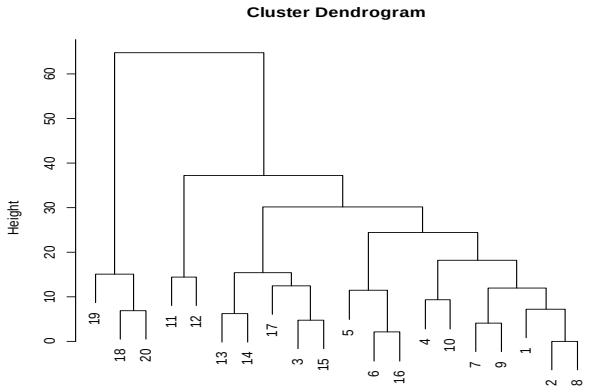

Figure 4.1: Cluster dendrogram for clustering of 20 engines by nonparametric approach

5. A Monte Carlo Study

If the researcher choses a parametric model that is adequate and estimates the parameters for this model appropriately, then the researcher's model is correctly specified and maximum information regarding the curves can be extracted from the data. On the other hand, if the researcher's model is not adequate then this model misspecification inhibits maximum information extraction from the data and incorrect decisions from such an analysis are likely to result. Use of a nonparametric model and a nonparametric method, such as a spline regression technique, are warranted when no adequate parametric model can be determined. In this case, improved fits to the data can often be obtained over use of an inadequate parametric model and better fits often lead to fewer errors in the decision making process. Our Monte Carlo study demonstrates how model misspecification affects the performance of the cluster-based method and the non-cluster-based method.

Our Monte Carlo study is used to evaluate the average performance of the cluster-based method and the non-cluster-based method when analyzing profiles estimated with a nonparametric method such as p-splines. The performance will be measured by the following metrics: FCC, sensitivity, specificity,

FPR, FNR and the POS. In this Monte Carlo study, the in-control PA profile is generated by using a combination of two functions: a second order polynomial in one variable, $\boldsymbol{x}$, and a nonlinear function of $\boldsymbol{x}$ of the form $\gamma\left(10\left(Sin\left(\frac{\pi(x_j-1)}{2.25}\right)\right)\right)\gamma\left(10\left(Sin\left(\frac{\pi(x_j-1)}{2.25}\right)\right)\right)$. The nonlinear component represents the departure of the PA curve from the second order polynomial, the assumed model. The value of γ, ranging from 0 to 4, represents the amount of departure of the actual model from the user's model.

Each profile specific curve is generated as a random curve about the PA profile using the linear mixed model. Consequently, the j^{th} response in the i^{th} in-control profile is from the model

$$y_{ij}=\beta_{0i}+\beta_{1i}x_{ij}+\beta_{2i}x_{ij}{}^2+\gamma\left(10\left(Sin\left(\frac{\pi(x_{ij}-1)}{2.25}\right)\right)\right)+\varepsilon_{ij}, i=1,2,...,m_1,\ j=1,2,...,n_i, \quad (5.1)$$

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, and n_i is the number of observations within i^{th} each profile and for ease of illustration, $n_i=n=20$ for all profiles. In addition,

$$\beta_{0i}=\beta_2\bar{x}^2+b_{0i},$$

$$\beta_{1i}=\beta_1-2\beta_2\bar{x}+b_{1i},$$

$$\beta_{2i}=\beta_2\bar{x}^2+b_{2i}$$

$$x_{ij}=j,\ i=1,2,...,m_1.,$$

where the random effects satisfy

$$\begin{bmatrix}b_{0i}\\b_{1i}\\b_{2i}\end{bmatrix}\sim MN\left[\mathbf{0}\ ,\begin{pmatrix}\sigma_0^2&0&0\\0&\sigma_1^2&0\\0&0&\sigma_2^2\end{pmatrix}\right],$$

and

$$\varepsilon\sim N\left[\mathbf{0},\quad \sigma^2 I\right].$$

The out-of-control profiles are also generated from Equation (5.1) but with a sustained shift contained in profiles m_1 through m as

$$\beta_{0i} = (\beta_2 + shift)\bar{x}^2 + b_{0i},$$

$$\beta_{1i} = \beta_1 - 2(\beta_2 + shift)\bar{x} + b_{1i},$$

$$\beta_{2i} = (\beta_2 + shift)\bar{x}^2 + b_{2i}.$$

$$i = m_1, m_1 + 1, ..., m.$$

One can show that the PA profile for the in-control process and out-of-control process are

$$y_{PA,j} = \beta_1 x_j + \beta_2 \left(x_j - \bar{x}\right)^2 + \gamma\left(10\left(Sin\left(\frac{\pi(x_j - 1)}{2.25}\right)\right)\right),\ j = 1, 2, ..., n,$$

and

$$y_{PA,j} = \beta_1 x_j + (\beta_2 + shift)\left(x_j - \bar{x}\right)^2 + \gamma\left(10\left(Sin\left(\frac{\pi(x_j - 1)}{2.25}\right)\right)\right),\quad j = 1, 2, ..., n,$$

respectively. Here, $\sigma_0^2 = \sigma_1^2 = \sigma_2^2 = 0.5$, $\sigma^2 = 4$ and $\beta_1 = 3, \beta_2 = 2$. It is assumed that we have $m_1 = 20$ in-control profiles, and $m = 30$ profiles in the HDS. Thus, 10 of the 30 profiles are from the out-of-control process. In our Monte Carlo study, the shift values are set at 0.05, 0.1, 0.15, 0.2, 0.25, and 0.3. For each value of the shift factor, the performance measures FCC, sensitivity, specificity, FNR, FPR and the POS were averaged over 5,000 replications.

The parameter γ in above equations is called the misspecification parameter and varies from 0 to 4. When γ is 0, the PA curve is exactly the quadratic function, the specified model, and when γ is 4, the PA curve departs considerably from the quadratic model and represents severe model misspecification. Values of γ between 0 and 4 represent a continuous departure from the quadratic model. A plot of the PA profiles for $\gamma = 0, 1, 2, 3,$ and 4 is given in Figure 5.1.

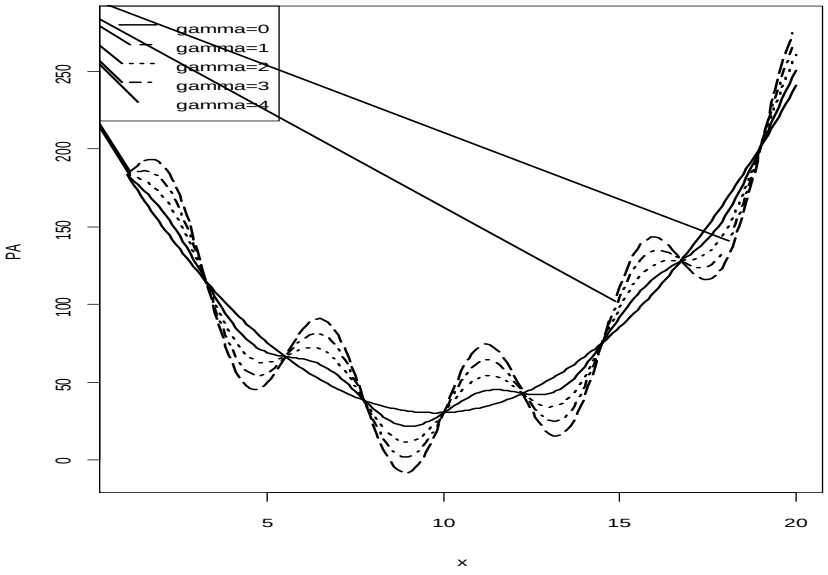

Figure 5.1: Plot of PA profile with different γ values.

Figure 5.2 to Figure 5.4 display the plots of the FCC, FPR and the POS based on the nonparametric fit to profiles with the cluster-based method and the non-cluster-based method when no, moderate, and high ($\gamma = 0, \gamma = 2$ and $\gamma = 4$) model misspecification respectively.

Recall that the FCC measures the method's ability to correctly identify the in-control and the out-of-control profiles; a better method will have the larger values of the FCC. Figure 5.2 shows that for different γ values, the FCC values from two methods are very close when the shift value is small. However, when the shift value is moderate or large, the cluster-based-method works uniformly better than the non-cluster-based method. Also, Figure 5.2 shows that the FCC values for the two methods are increasing as the shift value increases, as expected. The FPR measures the proportion of falsely classified in-control profiles and the smaller values represent the better performance. In Figure 5.3, the average FPR for the cluster-based method is uniformly smaller than the ones from the non-cluster-based method when the shift is moderate or large regardless of γ values, and while shift is small, the FPR values for the two methods are very close.

The POS is another performance metric commonly used in SPC. Figure 5.4 displays the POS of the two methods with different shift values. The conclusion is consistent with the conclusions from the plots of the FCC and the FPR, i.e., the cluster-based method works better than the non-cluster-based method.

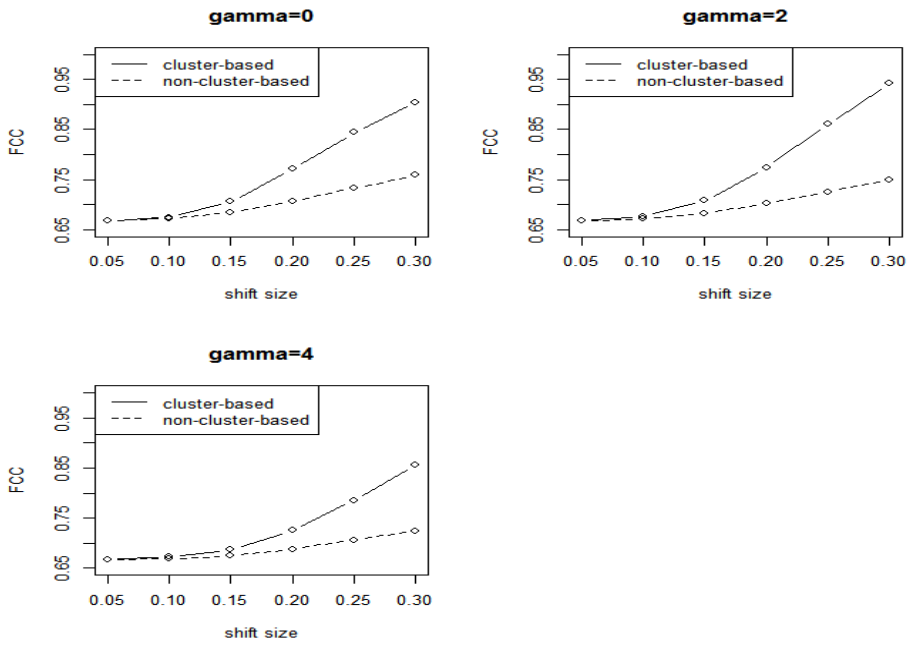

Figure 5.2: Comparing FCC **(**fractional correctly classified) for different γ values

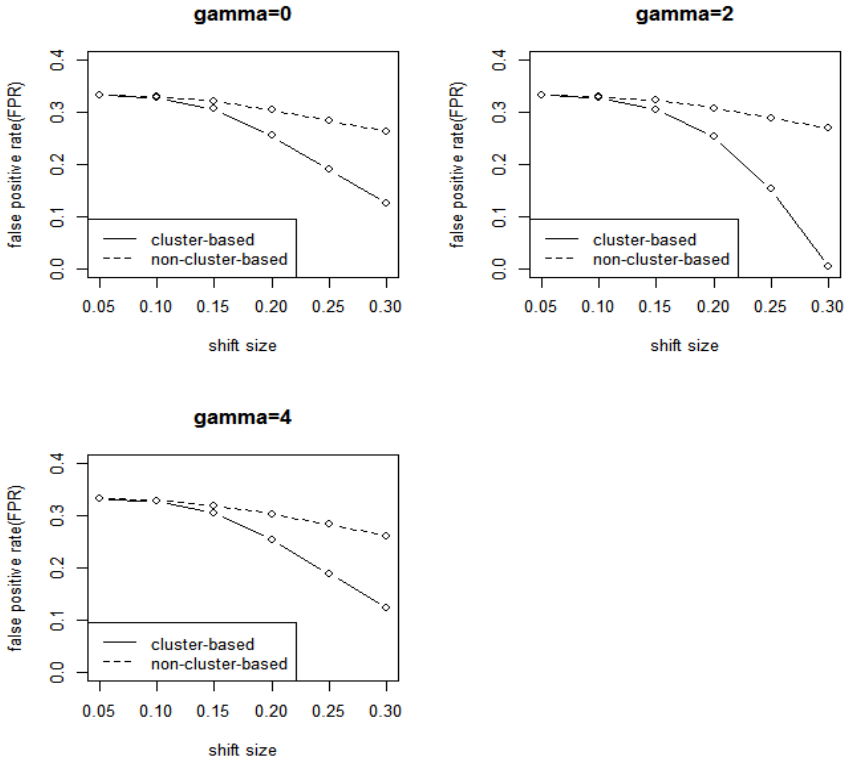

Figure 5.3: Comparing FPR (false positive rates) for different γ values

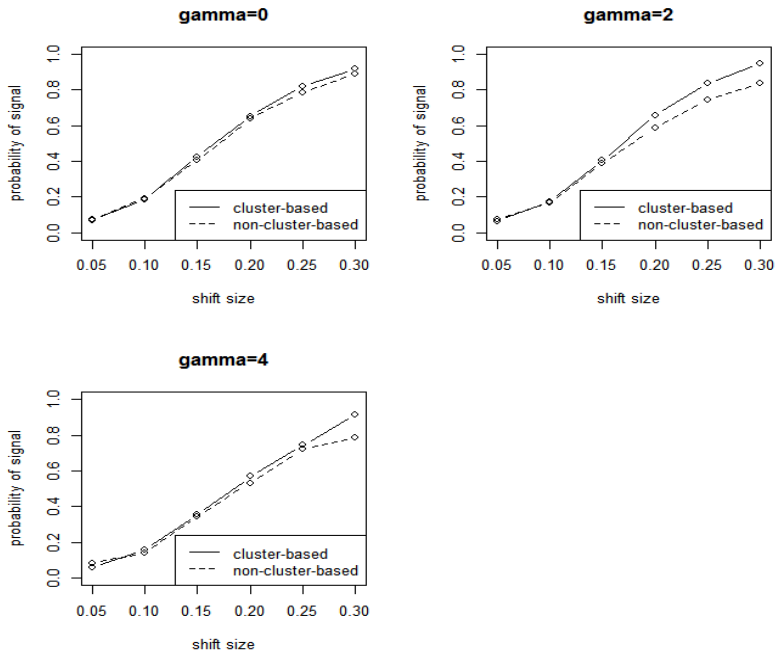

Figure 5.4: Comparing POS **(**probability of signal) for different γ values

In our Monte Carlo study, a nonparametric model was used to fit each profile. However, in applications, users may use an incorrect parametric model. Figure 5.5 compares the FCC of using the parametric and nonparametric model to fit the profiles when $\gamma = 0$ and $\gamma = 4$.

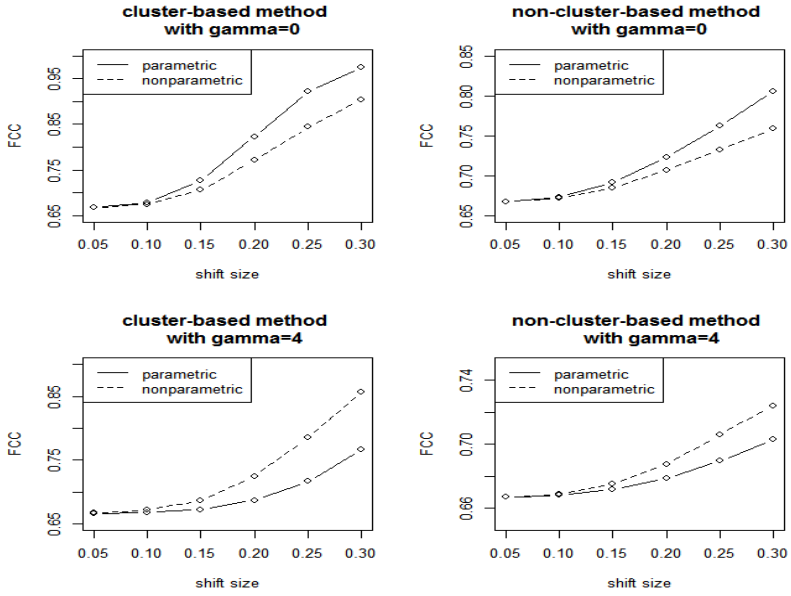

Figure 5.5: Comparing the FCC (fractional correctly classified) for different scenarios

Figure 5.5 shows that when $\gamma = 0$, the parametric model works better than the nonparametric model. Also, according to the results in Appendix Table A-2, one can see that when $\gamma = 0$, the parametric model with the cluster-based method works uniformly better than other three methods when the shift is moderate or large. Further, when $\gamma = 4$, the nonparametric method works better than the parametric method and the cluster-based method works uniformly better than the non-cluster-based method, whether the parametric or nonparametric methods are used. All simulation results can be found in the tables in the Appendix.

In addition, one can see that the nonparametric method is more robust compared to the parametric method. For example, when $\gamma = 0$, the specificity from the parametric method is 30% larger than the corresponding specificity from the nonparametric method, while the specificity from the nonparametric method is 90% larger than the specificity from the parametric method when $\gamma = 4$.

6. Conclusions and Outlook of Future Work

The goal of this research is to use a cluster-based method in monitoring nonparametric profiles and to demonstrate an improvement over other methods. The Monte Carlo study demonstrates that, in the presence of a sustained shift, the proposed cluster-based method results in superior performance over a non-cluster-based method regardless of whether the model is correctly specified or not. Further, if the model cannot be correctly specified, the researcher should incorporate a nonparametric model and a corresponding appropriate nonparametric procedure, such as p-splines, in conjunction with the cluster-based method to improve the profile monitoring results over using a non-cluster-based method. Specifically, our Monte Carlo study indicates that the cluster-based method performs uniformly better than the non-cluster-based method when there is a moderate or large sustained shift. The cluster-based method not only has larger POS values, but also has better performance in classifying the in-control and out-of-control profiles.

In our study, the cluster-based method is illustrated for the case where the response variable is continuous and normally distributed. However, these assumptions will not always be reasonable. For example, the response variable could be counts, a binary variable or a continuous variable from a distribution other than the normal distribution. In these cases, the profile or relationship between the

response variable and explanatory variables can be represented by using the generalized linear mixed model. For future work, the cluster-based methods could be applied in those situations where the response variable comes from the exponential family and the relationship between the response variable and explanatory variables can be represented by using the generalized linear model.

Our algorithm was programmed using R and the program is available from the authors upon request. The algorithm is surprisingly fast. For example, the case study analysis required only a second using a moderately equipped PC.

References

1. Woodall WH, Spitzner DJ, Montgomery DC, Gupta, S. Using Control Charts to Monitor Process and Product Quality Profiles. *Journal of Quality Technology* 2006; 36: 09-320.

2. Woodall WH. Current Research on Profile Monitoring. *Produção* 2007; 17: 420-425.

3. Noorossana R, Saghaei A, Amiri A. Statistical Analysis of Profile Monitoring, Wiley Series in Probability and Statistics, John Wiley & Sons Inc., 2011.

4. Noorossana R, Eyvazian M, Amiri, A, Mahmoud, MA. Statistical Monitoring of Multivariate Multiple Linear Regression Profiles in Phase I with Calibration Application. *Quality and Reliability Engineering International* 2010; 26: 291-303.

5. Croarkin C. Measurement Assurance for Dimensional Measurements on Integrated-Circuit Photomasks. *NBS Technical Note 1164, U.S. Department of Commerce, Washington, D.C., USA*, 1982.

6. Stover FS, Brill RV. Statistical Quality Control Applied to Ion Chromatography Calibrations. *Journal of Chromatography A* 1998; 804: 37-43.

7. Kang L, Albin SL. On-Line Monitoring When the Process Yields a Linear Profile. *Journal of Quality Technology* 2000; 32: 418-426.

8. Kim K, Mahmoud MA, Woodall WH. On the Monitoring of Linear Profiles. *Journal of Quality Technology* 2003; 35: 317-328.

9. Mahmoud MA, Woodall WH. Phase I Analysis of Linear Profiles with Calibration Applications. *Technometrics* 2004; 46: 380-391.

10. Wang KB, Tsung F. Using Profile Monitoring Techniques for a Data-Rich Environment with Huge Sample Size. *Quality and Reliability Engineering International* 2005; 21: 677-688.

11. Gupta S, Montgomery DC, Woodall WH. Performance Evaluation of Two Methods for Online Monitoring of Linear Calibration Profiles. *International Journal of Production Research* 2006; 44: 1927-1942.

12. Zhang J, Li Z, Wang Z. Control Chart Based on Likelihood Ratio for Monitoring Linear Profiles. *Computational Statistics & Data Analysis* 2009; 53: 1440-1448.

13. Jin JH, Shi JJ. Automatic Feature Extraction of Waveform Signals for in-Process Diagnostic Performance Improvement. *Journal of Intelligent Manufacturing* 2001; 12: 257-268.

14. Lada EK, Jye-Chyi L, Wilson JR. A Wavelet-Based Procedure for Process Fault Detection. *IEEE Transactions on Semiconductor Manufacturing* 2002; 15: 79-90.

15. Walker E, Wright SP. Comparing Curves Using Additive Models. *Journal of Quality Technology* 2002 ; 34: 118-129.

16. Ding Y, Zeng L, Zhou S. Phase I Analysis for Monitoring Nonlinear Profiles in Manufacturing Processes. *Journal of Quality Technology* 2006; 38: 199-216.

17. Williams JD, Birch JB, Woodall WH, Ferry NM. Statistical Monitoring of Heteroscedastic Dose-Response Profiles from High-Throughput Screening. *Journal of Agricultural Biological and Environmental Statistics* 2007; 12: 216-235.

18. Williams JD, Woodall WH, Birch JB. Statistical Monitoring of Nonlinear Product and Process Quality Profiles. *Quality and Reliability Engineering International* 2007; 23: 925-941.

19. Zou C, Tsung F, Wang Z. Monitoring Profiles Based on Nonparametric Regression Methods. *Technometrics* 2008; 50: 512-526.

20. Abdel-Salam ASG, Birch JB, Jensen WA. A Semiparametric Mixed Model Approach to Phase I Profile Monitoring. *Quality and Reliability Engineering International* 2013; 29: 555-569.

21. Jensen WA, Birch JB, Woodall WH. (2008). Monitoring Correlation within Linear Profiles Using Mixed Models. *Journal of Quality Technology* 2008; 40: 167-183.

22. Jensen WA, Birch JB. Profile Monitoring Via Nonlinear Mixed Models. *Journal of Quality Technology* 2009; 41: 18-34.

23. Qiu P, Zou C, Wang Z. Nonparametric Profile Monitoring by Mixed Effects Modeling. *Technometrics* 2010; 52: 265-277.

24. Chen Y, Birch JB, Woodall WH. Cluster-Based Profile Monitoring in Phase I Analysis. *Journal of Quality Technology* 2014, To Appear.

25. Amiri A, Jensen WA, Kazemzadeh, RB. A Case Study on Monitoring Polynomial Profiles in the Automotive Industry. *Quality and Reliability Engineering International* 2010; 26: 509-520.

26. Schabenberger O, Pierce FJ. *Contemporary Statistical Models for the Plant and Soil Sciences*, Boca Raton: CRC Press, 2002.

27. Seber GAF, Wild CJ. *Nonlinear Regression*. Hoboken, N.J.: Wiley-Interscience, 2003.

28. Ruppert D, Wand MP, Carroll RJ. *Semiparametric Regression*. Cambridge University Press: Cambridge, NY, 2003.

29. Demidenko E. *Mixed Models : Theory and Applications*. Hoboken, N.J.: Wiley-Interscience, 2004.

30. Sullivan JH, Woodall WH. A Comparison of Multivariate Control Charts for Individual Observations. *Journal of Quality Technology* 1996; 28: 398-408.

31. Chenouri SE, Steiner SH, Variyath AM. A Multivariate Robust Control Chart for Individual Observations. *Journal of Quality Technology* 2009; 41: 259-271.

32. Fraker SE, Woodall WH, Mousavi S. Performance Metrics for Surveillance Schemes. *Quality Engineering* 2008; 20: 451-464.

33. Shiau JJH, Sun JH. A New Strategy for Phase I Analysis in SPC. *Quality and Reliability Engineering International* 2010; 26: 475-486.

Appendix

Table A-1: *The Automotive Industry Data 20 Automobile Engines, Torque (T) vs. RPM*

RPM	T_E1	T_E2	T_E3	T_E4	T_E5	T_E6	T_E7	T_E8	T_E9	T_E10
1500	98.53	96.35	96.7	96.75	97.61	100.06	94.55	96.48	96.83	100.07
2000	102.65	100.74	100.05	100.87	102.46	103.6	103.22	100.87	103.78	103.91
2500	113.82	110.67	111.17	110.14	112.18	112.74	112.99	110.81	114.3	112.52
2660	115.26	113.06	111.51	110.48	112.99	113.56	114.18	113.2	114.62	113.25
2800	116.24	114.58	112.01	110.94	114.54	112.85	116.48	114.73	117.19	114.1
2940	117.06	114.98	111.23	111.17	115	114.49	115.33	115.13	116.61	114.1
3500	109.89	108.55	105.64	105.78	108.99	108.95	109.59	108.69	110.43	109.21
4000	109.65	107.41	106.02	103.37	107.95	108.24	108.47	107.55	109.61	108.34
4500	105.72	103.9	103.11	102.23	103.65	105.56	105.27	104.03	106.32	104.87
5000	99.74	97.99	97.4	96.06	96.94	98.92	97.9	98.12	99.44	98.35
5225	95.97	94.27	93.88	92.39	92.78	95.41	94.67	94.39	95.62	94.76
5500	89.47	88.45	88.17	86.54	86.41	89.19	88.23	88.56	89.46	88.93
5775	81.96	81.44	81.18	79.31	78.6	81.85	80.86	81.54	82	82.19
6000	74.9	75	75.03	73.13	71.97	75.09	73.93	75.09	75.83	100.07
RPM	**T_E11**	**T_E12**	**T_E13**	**T_E14**	**T_E15**	**T_E16**	**T_E17**	**T_E18**	**T_E19**	**T_E20**
1500	97.98	97.29	93.13	93.11	95.38	98.28	96.79	96.45	91.53	98.37
2000	104.98	105.86	101.02	103.43	101.25	101.29	103.64	104.52	100.72	102.4
2500	114.9	115.25	111.25	112.02	111.53	112.2	112.73	113.78	110.71	112.67
2660	116.06	117.83	111.83	113.2	112.11	112.57	113.92	114.59	111.72	113.76
2800	116.65	117.97	113.27	113.77	112.6	113.06	113.35	115.4	112.29	115.41
2940	116.18	117.77	113.04	113.77	111.76	112.37	112.78	115.86	111.61	113.01
3500	109.65	111.31	105.6	109.15	108.12	107.03	108.2	110.78	105.21	110.08
4000	109.06	110.97	106.15	108.05	106.62	106.37	107.06	110.21	106.22	109.51
4500	105.01	107.37	104.12	103.46	102.92	104.1	105.27	106.75	101.73	106.09
5000	97.43	100.53	97.45	98.26	96.35	98.01	98.47	99.94	96.59	99.84
5225	94.04	97.17	94.68	94.26	93.14	94.21	95.67	96.94	93.78	96.46
5500	87.51	90.47	88.59	89.09	86.75	87.53	89.41	90.24	87.29	90.16
5775	79.36	83.51	81.08	81.06	80.27	80.08	82.57	82.65	78.97	82.74
6000	72.34	76.34	75.77	74.14	73.47	73.9	76.31	76.76	72.8	75.82

Table A-2: Average performance metrics based on different methods with $\gamma = 0$

Parametric model with cluster-based method						
Shift	FCC	FPR	FNR	Sensitivity	Specificity	POS
0.05	0.6674	0.3324	0.3922	0.9981	0.0059	0.0864
0.1	0.6782	0.325	0.1016	0.9978	0.0391	0.2876
0.15	0.7268	0.2903	0.0154	0.9986	0.1832	0.6396
0.2	0.8234	0.2091	0.003	0.9993	0.4716	0.8790
0.25	0.9219	0.1044	0.0016	0.9994	0.7670	0.9750
0.3	0.9749	0.0359	0.0011	0.9995	0.9256	0.9956
Parametric model with non-cluster-based method						
Shift	FCC	FPR	FNR	Sensitivity	Specificity	POS
0.05	0.6670	0.3326	0.4429	0.9978	0.0055	0.0904
0.1	0.6731	0.328	0.2409	0.9955	0.0282	0.2812
0.15	0.6913	0.3145	0.1518	0.992	0.0899	0.5854
0.2	0.7227	0.2899	0.1176	0.9871	0.1940	0.8230
0.25	0.7627	0.256	0.0996	0.9821	0.3241	0.9336
0.3	0.8052	0.2163	0.089	0.9775	0.4604	0.9806
Nonparametric model with cluster-based method						
Shift	FCC	FPR	FNR	Sensitivity	Specificity	POS
0.05	0.6673	0.3326	0.3731	0.9985	0.0049	0.0722
0.1	0.6749	0.3274	0.1168	0.9981	0.0284	0.1876
0.15	0.7061	0.3056	0.0258	0.9984	0.1214	0.4258
0.2	0.7712	0.2548	0.0115	0.9982	0.3173	0.6506
0.25	0.8440	0.1891	0.004	0.9989	0.5342	0.8202
0.3	0.9048	0.1242	0.0036	0.9987	0.7169	0.9164
Nonparametric model with non-cluster-based method						
Shift	FCC	FPR	FNR	Sensitivity	Specificity	POS
0.05	0.6673	0.3327	0.3787	0.9988	0.0043	0.0693
0.1	0.6719	0.3292	0.2087	0.9972	0.0212	0.192
0.15	0.6848	0.3199	0.1306	0.9952	0.0644	0.4052
0.2	0.7066	0.3039	0.0885	0.9936	0.1326	0.6368
0.25	0.7327	0.2834	0.0759	0.9911	0.2159	0.7846
0.3	0.7589	0.2619	0.0672	0.9893	0.2981	0.8856

Table A-3: Average performance metrics based on different methods with $\gamma = 2$

Parametric model with cluster-based method						
Shift	FCC	FPR	FNR	Sensitivity	Specificity	POS
0.05	0.6668	0.333	0.4743	0.9987	0.0029	0.0712
0.1	0.6703	0.3305	0.1933	0.9983	0.0144	0.1622
0.15	0.6861	0.3198	0.0439	0.9986	0.061	0.3460
0.2	0.7275	0.2899	0.008	0.9993	0.1841	0.6180
0.25	0.801	0.2298	0.0017	0.9997	0.4036	0.8296
0.3	0.8795	0.0796	0.0015	0.9995	0.6396	0.9348
Parametric model with non-cluster-based method						
Shift	FCC	FPR	FNR	Sensitivity	Specificity	POS
0.05	0.6667	0.3328	0.5383	0.998	0.0035	0.0765
0.1	0.6697	0.3307	0.2642	0.9975	0.0142	0.1469
0.15	0.6782	0.3244	0.1792	0.9952	0.0442	0.3360
0.2	0.6925	0.3136	0.1466	0.992	0.0935	0.5522
0.25	0.7092	0.3004	0.1303	0.9889	0.1482	0.7071
0.3	0.7357	0.2792	0.1123	0.985	0.2371	0.8213
Nonparametric model with cluster-based method						
Shift	FCC	FPR	FNR	Sensitivity	Specificity	POS
0.05	0.6675	0.3324	0.3531	0.9985	0.0057	0.0653
0.1	0.6759	0.3268	0.0733	0.9988	0.0303	0.1745
0.15	0.7077	0.3045	0.0188	0.9988	0.1256	0.4030
0.2	0.7743	0.2525	0.0072	0.9988	0.3252	0.6571
0.25	0.8612	0.1527	0.0038	0.9989	0.5859	0.8375
0.3	0.9418	0.0034	0.0024	0.999	0.8273	0.9454
Nonparametric model with non-cluster-based method						
Shift	FCC	FPR	FNR	Sensitivity	Specificity	POS
0.05	0.6671	0.3327	0.3939	0.9987	0.0040	0.0774
0.1	0.6714	0.3296	0.1853	0.9979	0.0186	0.1672
0.15	0.6818	0.3221	0.1318	0.9959	0.0536	0.3882
0.2	0.7017	0.3075	0.0957	0.9938	0.1175	0.5840
0.25	0.7259	0.2889	0.0784	0.9917	0.1943	0.7446
0.3	0.7503	0.2691	0.0685	0.9900	0.2710	0.835

Table A-4: Average performance metrics based on different methods with $\gamma = 4$

Parametric model with cluster-based method						
Shift	FCC	FPR	FNR	Sensitivity	Specificity	POS
0.05	0.6666	0.3332	0.5195	0.9988	0.0022	0.0642
0.1	0.6680	0.332	0.3150	0.9985	0.0067	0.1076
0.15	0.6730	0.3287	0.1269	0.9984	0.0223	0.2266
0.2	0.6875	0.3189	0.0388	0.9987	0.065	0.3916
0.25	0.7166	0.298	0.0135	0.999	0.1519	0.6008
0.3	0.7661	0.2595	0.0043	0.9994	0.2996	0.7560
Parametric model with non-cluster-based method						
Shift	FCC	FPR	FNR	Sensitivity	Specificity	POS
0.05	0.6666	0.3331	0.5517	0.9987	0.0021	0.0678
0.1	0.6675	0.3321	0.4194	0.9975	0.0071	0.1161
0.15	0.6716	0.3294	0.2096	0.9973	0.0201	0.1858
0.2	0.6789	0.324	0.1579	0.9958	0.0453	0.3603
0.25	0.6895	0.3162	0.1349	0.9937	0.0812	0.5343
0.3	0.7031	0.3057	0.1214	0.9912	0.127	0.6255
Nonparametric model with cluster-based method						
Shift	FCC	FPR	FNR	Sensitivity	Specificity	POS
0.05	0.6673	0.3326	0.3668	0.9987	0.0044	0.0628
0.1	0.6723	0.3291	0.1625	0.998	0.0209	0.1571
0.15	0.6864	0.3194	0.0614	0.9979	0.0633	0.3538
0.2	0.725	0.2915	0.0181	0.9984	0.1784	0.5692
0.25	0.7855	0.243	0.0061	0.9989	0.3586	0.7441
0.3	0.8566	0.1763	0.0044	0.9987	0.5724	0.9156
Nonparametric model with non-cluster-based method						
Shift	FCC	FPR	FNR	Sensitivity	Specificity	POS
0.05	0.6671	0.3329	0.4187	0.9988	0.0034	0.0828
0.1	0.6686	0.3314	0.2938	0.998	0.0098	0.1424
0.15	0.6751	0.3269	0.1736	0.9967	0.0319	0.3442
0.2	0.6874	0.3181	0.1137	0.9954	0.0713	0.5324
0.25	0.7061	0.3041	0.0938	0.9932	0.1318	0.7212
0.3	0.7242	0.2902	0.0786	0.9919	0.1889	0.7888