Quality Technology & Quantitative Management

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ttqm20

An EWMA chart for high dimensional process with multi-class out-of-control information via random forest learning

Mingze Sun, Lei Qian, Amitava Mukherjee & Dongdong Xiang

To cite this article: Mingze Sun, Lei Qian, Amitava Mukherjee & Dongdong Xiang (07 Aug 2023): An EWMA chart for high dimensional process with multi-class out-of-control information via random forest learning, Quality Technology & Quantitative Management, DOI: 10.1080/16843703.2023.2244213

To link to this article: https://doi.org/10.1080/16843703.2023.2244213

QUALITY TECHNOLOGY & QUANTITATIVE MANAGEMENT
https://doi.org/10.1080/16843703.2023.2244213

Taylor & Francis
Taylor & Francis Group

An EWMA chart for high dimensional process with multi-class out-of-control information via random forest learning

Mingze Sun[a,b], Lei Qian[a,c], Amitava Mukherjee[d] and Dongdong Xiang[a]

[a]KLATASDS-MOE, School of Statistics, East China Normal University, Shanghai, China; [b]Tsinghua Shenzhen International Graduate School, Tsinghua-Berkeley Shenzhen Institute, China; [c]Center for Data Science, Peking University, Beijing, China; [d]Production, Operations and Decision Sciences Area, XLRI-Xavier School of Management, Jamshedpur, India

ABSTRACT

Modern manufacturing and quality monitoring involve multi-class out-of-control (OOC) information from the training sample. It is essential to use such information during online monitoring of data streams from complex processes. In this paper, a monitoring framework is designed by combining the random forest technique with the exponentially weighted moving average method for monitoring complex processes with multi-class OOC information. To be specific, a process surveillance technique in the form of a control chart is proposed based on the probability that the online data is classified as an in-control (IC) sample, and the control chart triggers an alarm when the probability is lower than the control limit. Our numerical findings based on the Monte–Carlo simulation show that the proposed control chart performs more effectively than its competitors under various distributions and data types, especially for high-dimensional cases when multi-class OOC information is known in advance. Moreover, the proposed method is illustrated with an application using the data related to the hard disk manufacturing processes.

ARTICLE HISTORY
Received 16 October 2022
Accepted 21 July 2023

KEYWORDS
Complex process; MEWMA; random forest; statistical process control

1 Introduction

Because of the rapid growth of the manufacturing sectors in the Industry 4.0 era, the necessity for monitoring high-precision product quality and related production processes is gradually increasing. Over the years, surveillance of item quality in complex manufacturing systems has become an increasingly challenging issue to ensure stable and excellent output quality. A shift in the location or scale parameters in the process distribution of the quality characteristics can be detected using statistical process monitoring (SPM) tools. An efficient SPM tool can identify a process shift rather quickly using statistical methods. In modern industrial production, it is necessary to monitor multiple indicators or quality characteristics of the products. Several researchers Hotelling (1947), Woodall and Ncube (1985) and Lowry et al. (1992) proposed various multivariate Shewhart-type, multivariate cumulative sum (MCUSUM)-type, and multivariate exponentially weighted moving average control charts (MEWMA) to address multivariate SPM (MSPM) problems. Zou and Qiu (2009) designed a LASSO-based multivariate exponentially weighted moving average variable selection control chart. Mehmood et al. (2020) used bivariate ranked set schemes to develop an MCUSUM control chart, which can quickly monitor small variations in the process mean vector. Sabahno et al., 2020 designed an adaptive MSPM scheme that monitors both the mean and the

CONTACT Dongdong Xiang ddxiang@sfs.ecnu.edu.cn KLATASDS-MOE, School of Statistics, East China Normal University, Shanghai, China

covariance matrix of a multivariate normal distribution. F. Xie et al. (2022) proposed the one-sided adaptive truncated exponentially weighted moving average scheme. Bersimis et al., 2007 gave a comprehensive overview of MSPM schemes. Interested readers may also see the book by Qiu, 2013. In recent years, MSPM schemes have been widely used in industrial manufacturing, medical services, and other fields. This paper proposes an MSPM scheme using the random forest learner to monitor the complex high-dimensional processes where a process shift may lead to a mixture distribution of differently clustered data. In the context of process monitoring, for more details on applications of random forest classifications in profile classification, we refer to Alshraideh et al. (2020).

In practice, quality-related datasets in large-scale manufacturing are often very complex. This article is motivated by a complex process monitoring in the Hard Disk Drive Monitoring System (HDDMS) Zhang et al. (2015). The hard disk drive is an electromechanical data storage device that stores and repossesses digital data, but it often crashes, resulting in the loss of part or even all of the stored information. Moreover, recovery of the hard drive is a complex and expensive process. Therefore, it is essential to monitor the hard disk performance to determine its real-time condition. To this end, we consider the dataset containing four quality attributes of the hard disk performance, including some continuous and count variables. For example, reading error rate, seek error rate, and power recorded on an hourly basis are continuous variables, while the current pending sector count is a categorical variable. The dataset to be monitored is of a mixed type which is not purely continuous or purely categorical. Clearly, for this type of data, an assumption of multivariate normality is not meaningful. Also, it is difficult to fit any standard multivariate distributions to such data. Another characteristic of this process is that it can produce many datasets in a short time. This process can generate more than 1 million in-control (IC) profiles in a few weeks, along with a large number of out-of-control (OOC) profiles. The historical data contain various OOC profiles, which can be divided into a few classes. Using this historical information efficiently for process monitoring and improvement is always a challenging problem for engineers.

There are many other examples from different domains, such as the semiconductor manufacturing industry. A dataset related to semiconductor quality is available in the UC Irvine Machine Learning Repository, consisting of 1567 observations and 591 variables Mukherjee and Marozzi, (2021). The dataset contains several variables that are almost constant, and there are many missing values. In some cases, we have zero-inflated data. Obviously, the data do not follow the multivariate normal distribution. Fitting some suitable well-known multivariate distribution to this dataset is immensely challenging, primarily because of its large dimension and diverse nature of variables. Consequently, traditional control charts based on the multivariate normal or other continuous distribution may not be ideal. Therefore, it is necessary to establish a control chart that does not use any stringent distributional assumption. Similar applications include monitoring the commercial-scale cell culture expansion bioreactor in the biopharmaceutical field Tulsyan et al. (2018) and the high-dimensional complex process of users' search terms on social media like Wikipedia Weese et al. (2016).

In most multivariate and high-dimensional SPM problems similar to the ones discussed in previous paragraphs, the actual process distribution remains unknown. Parametric MEWMA and other control charts based on the normality assumptions cannot effectively monitor such processes. However, using a large volume of available historical data, the above charting schemes may be redesigned by adjusting the control limits via suitable methods. For example, many nonparametric SPM schemes are proposed using certain rank-based statistical methods. Interesting papers regarding the above issue have been published by Bakir (2004), Bakir & Reynolds (1979), Graham et al. (2011), Qiu & Li (2011), Abbasi et al. (2013), Zhou et al. (2016), Bush et al. (2010), Huwang et al. (2019), Chakraborti (2004), Maboudou-Tchao et al. (2022a), Dastoorian & Wells (2021), and Tran et al. (2022).

Although these SPM schemes are useful when the process distribution is unknown, most of these methods do not consider the complete information from the historical datasets. Usually, during

a Phase-I analysis, OOC samples are identified and removed from the historical data. The remaining sample observations form a reference sample and are used as the gold standard. The OOC signals obtained during Phase-I analysis with training data are not usually used to identify the possible nature of the shifts in the OOC samples. This practice results in the loss of some essential OOC information. The historical OOC data may contain multiple shift classes, so the monitored data can be considered as mixed distributed. Even with much historical data, it is difficult to accurately estimate the data distribution, which also affects the effectiveness and robustness of the above control charts.

The increase in the data complexity reduces the usability and effectiveness of the conventional SPM schemes. Most traditional charting schemes use only IC samples of historical data for benchmarking and ignore information in the OOC samples. The signals may be carefully investigated as some may be false alarms. However, in the semiconductor wafer's quality monitoring and other similar contexts, we often get many OOC signals that are not false alarms and are correctly classified as OOC samples. Statistical learning tools may allow us to use historical data, including evidence in historical OOC samples, and facilitate online monitoring of complex processes with better information. Some researchers designed control charts based on statistical learning. S. Chen & Yu (2019) developed a deep recurrent neural network (RNN) model to detect mean shifts in autocorrelated processes. Wang et al. (2019) used a differential evolution algorithm to determine the optimal parameter selection of the support vector machine (SVM) classifiers and built a single-side control graph based on SVMs to monitor multiple quality characteristics. X. Xie & Qiu (2022) used certain existing machine learning control charts together with a recursive data de-correlation procedure. Lee et al. (2022) proposed a control chart using the support vector machine under gamma distribution. Maboudou-Tchao et al. (2022b) compared penalized methods and support vector methods for Shewhart-type and accumulative-type control charts. To obtain more information about other control charts based on traditional machine learning, please refer to Huang et al. (2022), Ding et al. (2023), and Chan et al. (2023).

Zhang et al. (2015) creatively proposed to combine SVMs with EWMA sequence and make full use of historical IC data and OOC data information to construct the SVMEWMA control chart. It used the improved SVMs classifier to develop the monitoring scheme and achieved good performances in monitoring complex processes. At present, the control charts designed based on statistical learning mainly use traditional machine learning methods. While technique like SVMs is primarily applied to the dichotomy problem, which has certain limitations for monitoring problems with various types of possible OOC conditions. The traditional classifiers usually need to carry out feature decomposition when processing high-dimensional data, which significantly increases the computational complexity. Simultaneously, with the increase in data dimensions and classification numbers, the monitoring effect of control charts based on traditional classifiers may also decrease significantly (Zhang et al. 2015).

Based on the discussion above, it is desirable to develop a new monitoring scheme that applies even to complex, high-dimensional data with unknown distribution under possible multiple categories of shifts using the complete information from the historical data. This paper introduces the random forest ensemble classifier and EWMA sequence to construct the RFEWMA control chart. The random forest model does not need feature selection when processing high-dimensional data with multiple classes. Consequently, such a classifier is highly efficient for complex processes. The RFEWMA control chart can effectively solve complex process monitoring using the random forest model, which is simple and highly precise for different data types. By calculating the voting situation of the decision trees in the random forest, the monitoring of the high-dimensional and multi-classified data is realized. The control chart is proved to have good performance through numerical study. Although there have been some studies using the random forest to monitor profile signals, the OOC information in the current study is only binary and the dimension is small (Alshraideh et al. (2020)) [35]. We emphasize the managerial implication of the proposed method in improving total quality management in the Industry 4.0 era, where high-dimensional processes

are more common than ever and require more sophisticated tools and techniques. Most classical process monitoring practices ignore possible out-of-control (OOC) situations leading to multi-class OOC conditions. Proper knowledge of the OOC class makes corrective actions easier, and the proposed techniques may play a vital role. Our control chart combines the random forest model with the EMWA sequence to realize efficient monitoring of high-dimensional and multi-classification data.

The remainder of this paper is organized as follows: the random forest theory is briefly reviewed in Section 2. The statistical framework of the problem is discussed, and the implementation design of the proposed control chart is described in detail in Section 3. The numerical performance of the proposed scheme is thoroughly investigated in Section 4. In Section 5, the usability of the proposed scheme is illustrated using an example based on real data. Finally, Section 6 concludes with some discussions on future research directions.

2 A review of random forest

In this section, the theoretical framework of random forest classification is briefly described. According to the description of the theory, the solution to the corresponding problem can be given.

The nomenclature table

To better comprehend the notations and symbols, we provide a nomenclature table in this section. See Table 1 for details.

2.2 The theoretical framework of random forest

The random forest technique (Ho (1998)) is essentially an ensemble learning method for regression, classification, and other tasks that build an assembly of decision trees using a training sample that leads to the appropriate classifications of data or offers suitable predictors of one or more characteristics of the individual trees. Ensemble learning can accomplish the learning task by constructing and combining multiple learners, and the performance of a single learner with a weak function can be significantly improved through an ensemble. Current ensemble learning can be divided into two categories, namely, strong dependence relationship between individuals and weak strong dependence relationship between individuals. The former is represented by Boosting, while the latter is bootstrap aggregating or Bagging. A random forest constructs a Bagging ensemble based on decision trees and introduces random attribute selection during the training process Breiman (2001). The random forest model generates decision trees by randomizing features (columns) and data (rows). There is no correlation between each decision tree. When there is a new input sample, each decision tree in the forest classifies (votes), respectively, and finally

Table 1. The nomenclature table.

Symbol	Symbol interpretation
P	characteristic dimension
k	class number
T	number of decision trees
c	category
h	decision tree
F_0	IC distribution
F_j	OOC distribution
E_i	MEWMA statistics
L	the control limit
λ	the smoothing parameter
M	pieces of the input data to construct the new MEWMA sequence

summarises the result of the decision trees, the category with the most votes is the prediction result of the random forest model.

In the random forest model, N pieces of training data x_i and their corresponding label y_i are given. For each decision tree, N training samples are randomly drawn from the training sample as the training set of the tree. We assume that the characteristic dimension of each sample is P, specify a constant $p < P$, randomly select P features from P ones, and select the optimal feature from the P features every time the decision tree is divided, the random forest model is obtained by training finally. For a $(k+1)$-classification task, it is assumed that there are a total of T decision trees in the random forest model. For each input x, the corresponding label y should be selected from the category $\{c_0, c_1, \ldots, c_k\}$ by the h_t $(t = 1, 2, \ldots, T)$ decision tree, and here the method of plurality voting is used to determine the final result. We express the predicted output of h_t on sample $\boldsymbol{x}$ as a $(k+1)$-dimensional vector $\left(\boldsymbol{h_t^0}(\boldsymbol{x}); \boldsymbol{h_t^1}(\boldsymbol{x}); \ldots; \boldsymbol{h_t^k}(\boldsymbol{x})\right)$, where $h_t^j(\boldsymbol{x})$ is the output of h_t on c_j category, and $h_t^j(\boldsymbol{x})$ belongs to $\{0, 1\}$. If h_t predicts sample $\boldsymbol{x}$ as category c_j, the value of $h_t^j(\boldsymbol{x})$ is 1, otherwise is 0. Then the predictive classification of $\boldsymbol{x}$ is:

$$H(x) = c_{\underset{j}{\operatorname{argmax}} \sum_{t=1}^{T} h_t^j(\boldsymbol{x})}. \tag{1}$$

When the random forest model gives the classification result, we can also calculate the probability that $\boldsymbol{x}$ is divided into the c_j class:

$$P\left(\boldsymbol{x} \in c_j\right) = \frac{\sum_{t=1}^{T} h_t^j(\boldsymbol{x})}{T}. \tag{2}$$

This also provides the basis for constructing the monitoring statistics later.

2.3 Properties of random forest

Unlike the neural network and other conventional methods, which require a large amount of computation, the random forest model involves a lesser computational load for the decision tree and improves prediction accuracy. Random forest is not sensitive to multivariate collinearity, and its results are robust to both missing data and unbalanced data, so it can reasonably predict the effects of up to thousands of explanatory variables.

The efficiency of the random forest model is mainly determined by calculating the out-of -bag error (OOB error). Since only part of the samples is used in the training of each decision tree, the remaining samples can be used as a validation set to estimate the generalization performance of the model Breiman (1996) and Wolpert & Macready, (1999). Let D be the sample set for training, D_t is the sample set, in which the sample is used for training in the decision tree h_t, and let $H^{oob}(\boldsymbol{x})$ represents the out-of-bag prediction of a sample $\boldsymbol{x}$, that is, for each input $\boldsymbol{x}$, consider the predicted results of the decision tree training without $\boldsymbol{x}$:

$$H^{oob}(\boldsymbol{x}) = c_{\underset{j}{\operatorname{argmax}} \sum_{t=1}^{T} h_t^j(\boldsymbol{x}) \times I(\boldsymbol{x} \notin D_t)}, \tag{3}$$

where $I(\cdot)$ denotes the sign function. Then the out-of-bag estimation of the generalization error of the random forest model is:

$$\varepsilon^{oob} = 1/|D| \times \sum_{(\boldsymbol{x}, y) \in D} I\left(H^{oob}(\boldsymbol{x}) \neq y\right). \tag{4}$$

The selection of parameters of the random forest model is mainly to minimize the ε^{oob} (out-of-bag error), which will be discussed in the following Section.

3 A monitoring framework for complex processes

This section formalizes the problem, introduces our control chart framework for complex processes in detail, and explains how to determine the control limit L when the process distribution is unknown. Finally, we discuss the choice of some design parameters for the control chart.

3.1 Problem description

In the monitoring of a complex process, we can often observe a large volume of historical observations, say, $\{\boldsymbol{x_1},\boldsymbol{x_2},\ldots,\boldsymbol{x_n}\}$ during Phase-I. Each of the sample observations may include information on multiple characteristics and could be high-dimensional. Here, some of the characteristics may be measured on a continuous scale, and the rest could be categorical. In this context, we assume that each observation $\boldsymbol{x_i}$ $(i = 1, 2 \ldots, n)$ is p-dimensional $(p \geq 1)$. In a large number of historical observations, IC and OOC observations may be mixed. It is necessary to separate OOC observations from the IC samples through appropriate Phase-I analysis. Classical statistical Phase-I analysis tools are not suitable when some variables are continuous and some discrete in a multivariate and high-dimensional set-up. However, the problem may be addressed using machine learning tools, such as classification technology. Note that, the OOC observations may be divided into k different classes $(k \geq 1)$. Let us denote F_0 as the IC distribution, and F_j as the distribution of the j^{th} class under OOC. We assume that the online surveillance starts from $(n+1)^{\text{th}}$ and a possible process shift occurs at the $(n+m+1)^{\text{th}}$ stage. Precisely, we consider a change-point model for the online monitoring problem:

$$\boldsymbol{x_i} \sim \begin{cases} F_0\left(\boldsymbol{x};\, \mu_{\boldsymbol{0}}\right), & \boldsymbol{x} = n+1, \ldots, n+m, \\ F_j\left(\boldsymbol{x};\, \mu_{\boldsymbol{j}}\right), & \boldsymbol{x} = n+m+1, \ldots \end{cases} \tag{5}$$

where $F_*(\cdot)$ is the unknown process distribution. $\mu_{\boldsymbol{0}}$ and $\mu_{\boldsymbol{j}}$ are the parameters of interest during the IC and the j^{th} OOC states, respectively, and $\mu_{\boldsymbol{0}} \neq \mu_{\boldsymbol{j}}$, for some $j = 1, 2, \ldots, k$.

There is no requirement to know the functional forms of the IC and OOC process distributions in our monitoring framework, so $F_*(\cdot)$ is unknown. Meanwhile, in the actual monitoring process, the online data median of the OOC classes, $\mu_{\boldsymbol{j}}$'s, are also unknown. We expect to detect and classify the shift quickly in the online data with the help of the historical OOC data acquired from Phase-I.

3.2 The monitoring framework

Our monitoring framework consists of Phase-I and Phase-II procedures. In Phase-I, a large amount of raw data can be clustered by some traditional Phase-I methods. Then, we use the random forest model to construct the RFEWMA control chart to monitor the online data in Phase-II. OOC observations are separated from the raw data in the traditional Phase-I monitoring framework, and IC parameters are estimated using the IC samples. As discussed earlier, different causes may produce different shifts. The OOC data often contain more than one type of shift, and we intend to capture different kinds of OOC shifts in our monitoring framework. Therefore, after separating the IC data and OOC samples through traditional classification methods, we repeatedly use clustering and other ways to divide OOC data into k different subcategories, and each separate subcategory contains a different shift type. Subsequently, we use the information about IC samples and partitions of OOC samples into k diverse classes during Phase-II monitoring. In this paper, the research emphasis for Phase-II of online data monitoring, and methods in Phase-I here are not described in detail.

In Phase-II, we assume that sufficient historical IC data and OOC samples are already available a-priori. For multiple shift classes in the data, it is difficult to effectively monitor such processes by estimating the data distribution. So we apply the random forest model introduced above. The final

classification result of the random forest is determined by counting the voting result of each decision tree, which provides great convenience for the design of our monitoring statistics. For each input data, the decision trees in the random forest determine the data classification either as the IC class or any one of the k OOC class and vote, respectively. Let the voting result of the t^{th} decision tree be h_t, then the probability that the input data $\boldsymbol{x}$ belongs to the j^{th} type is shown in formula 2.

The specific monitoring process in Phase-II is explained in detail. Suppose we have obtained a set of IC samples and OOC observations of k different shift types in Phase-I. The above data are used to establish a random forest model. To make full use of historical data information and make the control chart more sensitive to the small shift, we use the MEWMA-type scheme. Due to a large amount of data, we have made some adjustments to the sequence construction to prevent the MEWMA sequence from tending to a fixed value. We assume that the input data is $\boldsymbol{x_i}$ $(i = 1, \ldots)$, we only use the input data and the first M pieces of the input data to construct the new MEWMA sequence. We define the sequence:

$$\boldsymbol{S_{m,l}} = \lambda \boldsymbol{x_{m+l-1}} + (1-\lambda)\boldsymbol{S_{m,l-1}}, \quad m \geq 1, 1 \leq l \leq M \tag{6}$$

$$\boldsymbol{E_i} = \begin{cases} \boldsymbol{S_{1,i}}, & 1 \leq i \leq M \\ \boldsymbol{S_{i-M,M}}, & i > M \end{cases} \tag{7}$$

where the initial vector $\boldsymbol{S_{m,0}}$ $(j = 1, \ldots)$ is defined as the mean vector estimated from the historical IC samples, and $\{\boldsymbol{E_i}\}$ values form a sequence of the MEWMA statistics. Using this MEWMA sequence data, we developed a random forest model to calculate the probability that $\boldsymbol{E_i}$ belongs to the IC distribution. The higher the probability, the more we tend to consider the system in a controlled state, and the lower the probability, the more we tend to consider it in an uncontrolled state. We recommend triggering an alarm when the probability falls below a certain threshold L, which serves as the control limit of our RFEWMA control chart.

$$P(\boldsymbol{E_i} \in \text{IC Data}) \leq L \tag{8}$$

The control limit L is mainly determined by Monte Carlo simulation, and the corresponding method is explained in Subsection 3.2.

The RFEWMA control chart requires us to consider two additional parameters, namely λ and M. The value of λ should be chosen appropriately depending on the specific monitoring problem, while N. Chen et al. (2016) recommend that M should range between 15 and 30 but not exceed a certain value. In our study, we found $M = 20$ to be optimal for the random forest and MEWMA chart. These details are elaborated in Subsection 3.3.

3.3 Determination of the control limit L

Our proposed RFEWMA scheme does not depend on the distribution of the process data, and the process distribution is also unknown a-priori. Therefore, the traditional method of calculating the control limit using the distribution function of the pivot is not applicable. In the current study, we recommend using a search algorithm-based Monte Carlo simulation to determine the control limit. Assume that we have collected many IC data and OOC data in Phase-I. First, a random forest model is trained with a set of IC data and OOC data. Then, for fixed L, a piece of data x_i is randomly chosen from this IC dataset at each time, the MEWMA sequence stated previously is applied to the random forest model until the corresponding probabilistic output $P(\boldsymbol{E_i} \in \text{IC Data})$ is smaller than L. We repeatedly choose samples from the IC dataset and calculate the corresponding ARL_0. The average ARL_0 calculated above is the corresponding ARL_0 under the fixed L. If the attained ARL_0 under a certain control limit L is close to the target ARL_0, then the control limit is obtained. While there is a large difference from the target ARL_0,

we use the dichotomy to determine the specific control limit L. It can be checked that there is a monotonic relationship between the control limit L and ARL_0. That is, as the control limit L decreases, ARL_0 gradually increases. We denote the upper and lower bounds as h_l and h_u, and we set their initial values to 1 and 0. Let $L = (h_l + h_u)/2$, and calculate the corresponding ARL_0. If attained $ARL_0 <$ target ARL_0, then let $h_l = (h_l + h_u)/2$ and h_u remains unchanged; otherwise, if attained $ARL_0 >$ target ARL_0, let $h_u = (h_l + h_u)/2$ and h_l remains the same. Repeat the above iteration process and stop iteration when attained ARL_0 under a certain L is the same as the target ARL_0. The specific control limit L of the control charts with different ARL_0 can be determined by the above method.

3.4 Choice of the design parameters

We first discuss the choice of the smoothing parameter λ of the MEWMA sequence. Note that $\lambda \in [0, 1]$ and $\lambda = 1$ corresponds to a Shewhart-type scheme. As a rule of thumb, the smaller value of λ should be selected for monitoring small shifts, and the relatively higher larger value of λ should be selected for monitoring large shifts. The larger λ allows the MEWMA sequence to consider more information from the latest sample observations. To choose the value of λ properly, one needs to specify a target shift size, which unfortunately is often not known a-priori. The numerical study reveals that the value of λ between 0.05 and 0.2 ensures the robustness of the proposed scheme in monitoring various complex processes. Details of the Monte-Carlo study are deferred to Section 4.

Here, we also need to discuss another parameter, M, which specifies the number of previous samples used in the MEWMA statistic. In practical applications, the selection of M should consider the size of the sample data and the requirements of the application. In our RFEWMA chart, the purpose of constructing this MEWMA sequence is to make full use of the online historical data to improve the monitoring performance of the control chart. This means that when we construct the monitoring statistics, we need to calculate the probability that these MEWMA sequences come from the corresponding IC random forest model. For this purpose, we need MEWMA sequence data based on IC data and OOC data to train the random forest models. Unfortunately, if we have enough historical data on which to construct MEWMA sequences, the observed values may eventually tend to be constant, which makes it difficult to train appropriate random forest models based on these small differences. The trained models are unable to fit the true distribution of data in the application. Therefore, we need to strike a balance between monitoring effects and model training. So we should choose a moderate M. N. Chen et al. (2016) suggested setting M between 15 and 30, but not too large. Based on our empirical results, we recommend setting M to 20, and a detailed discussion on this can be found in Section 4.5.

It is also important to correctly specify the parameters of the random forest model. The random forest parameters can increase the model's prediction ability and reduce the time required to train the model. We focus on two essential parameters: the maximum number of features that a random forest allows a single decision tree to use (*mtry*) and the number of decision trees in the random forest (*mtree*). An increase in *mtry* generally gives the model more options to consider on each node. However, it reduces the diversity of a single tree and slows down the algorithm by increasing *mtry* Wu et al. (2021). We recommend setting *mtry* close to the data dimension when monitoring small shifts and setting *mtry* as the positive square root of the data dimension when monitoring large shifts. Note that having more subtrees can give the model better performance, but at the same time make the code slower and increase the possibility of overfitting. For our control chart, however, increasing *mtree* has little influence on the monitoring effect of the control chart, so based on our empirical findings, we recommend setting the number of *mtree* as $50 - 300$. See Section 4.5 for a detailed discussion.

4 Numerical studies

In this section, we discuss various numerical results related to the proposed SPM schemes. We compare the efficacy of the proposed RFEWMA scheme with the traditional SPM schemes for different data types and distributions and show that the proposed RFEWMA scheme can address complex process monitoring problems. The numerical comparison is arranged as follows. In Section 3.4, we consider monitoring of continuous processes and sample data simulated from continuous distributions for both situations: a single OOC class ($k = 1$) and multiple OOC classes ($k > 1$) in historical OOC samples. Section 4.1 investigates monitoring the categorical and mixed data of single classification and multi-classification in historical OOC samples. In Section 4.2, we consider the case where the nature of the shift is different from historical shifts. In Section 4.4, we study the possibility of unbalanced training data. We discuss the parameter λ in the MEWMA sequence and the choice of critical parameters in the random forest model in Section 4.5. In Section 4.6, we explore the sample size required to determine the control limit L. We use Python 3.6 to implement the simulation. Throughout the simulation, we mainly detect the mean shift and use $\lambda = 0.1$ in Section 3.4–4.4. Because of the high dimension of the data being considered, both ARL_0 and ARL_1 values conditional on the given IC dataset and OOC dataset are obtained from 5000 replications of simulations, and the whole process is repeated 100 times. The ARL_0 is fixed at 200.

4.1 Monitoring of continuous data

This section focuses on the continuous processes, covering both cases where historical OOC data have only one class and multiple classes. We show the better performance of the proposed RFEWMA scheme compared to the SVMEWMA scheme Zhang et al. (2015), and the MEWMA scheme Lowry et al. (1992) based on Hotelling's T^2 statistic. We choose the multivariate normal and multivariate t (denoted by $t_{p,\varepsilon}$) distributions to represent symmetric thin-tailed and heavy-tailed processes and the asymmetric multivariate chi-square distribution, denoted by $X^2_{p,\varepsilon}$. If not specified, we assume that the mean of the IC process is a null vector and the covariance matrix is the identity matrix in all three cases. The parameter ε of the multivariate t distribution and the multivariate chi-square distribution is fixed at 5. We use the relative mean index (RMI) as a metric to compare the SPM schemes. Small RMI means that the control chart has a quick and robust performance in detecting mean changes (Han & Tsung (2006)).

$$RMI = \frac{1}{n} \times \sum_{i=1}^{n} \frac{ARL_{1i}(T) - \min(ARL_{1i})}{\min(ARL_{1i})} \tag{9}$$

where $ARL_{1i}(T)$ presents the ARL_1 value of the T chart under the i^{th} condition, and $\min(ARL1_i)$ presents the minimum ARL_1 value among the considered charts under the i^{th} condition.

For only one class of shift ($k = 1$) in the training OOC data, we assume that the shift occurs in the first dimension of the mean vector. We randomly generate 100,000 IC data and 100,000 OOC data, respectively, to construct the RFEWMA chart and the SVMEWMA chart ($N_0 = N_1 = 100000$). We assume that the type of shift for the online data is the same as the historical shift. In the process of simulating online data monitoring, the first 100 online data are IC data, and then the data is turned from in-control to out-of-control. Besides, for the three different continuous distributions, we consider the data dimensions $p = 6$ and $p = 10$. We calculate ARL_1 and the corresponding standard deviation of the RFEWMA chart, the SVMEWMA chart, and the MEWMA chart, respectively, and the specific simulation results are shown in Table 2. First, we simulate the situation of the multivariate normal distribution, and it is obvious that the overall performance of RFEWMA is better than that of MEWMA and the monitoring effect of RFEWMA and SVMEWMA is similar. It can be found that the effect of RFEWMA and SVMEWMA is better than that of MEWMA when the shift size is small, and the gap between the three methods is

Table 2. Comparison of ARL_1 values for multivariate normal data, multivariate t data, and multivariate chi-square data. The SDRL value is shown in parentheses.

		Normal			t ($df = 5$)			$X2$ ($df = 5$)		
p	δ	RFEWMA	SVMEWMA	MEWMA	RFEWMA	SVMEWMA	MEWMA	RFEWMA	SVMEWMA	MEWMA
6	0.5	18.5(14.8)	17.5(11.8)	39.2(33.1)	28.6(23.1)	28.3(22.9)	81.0(61.6)	26.9(22.8)	25.6(21.2)	54.5(47.9)
	1.0	8.1(4.1)	8.3(4.0)	13.1(6.8)	10.0(5.6)	10.2(6.1)	28.4(20.3)	6.1(3.3)	6.0(3.0)	11.1(6.4)
	1.5	6.0(2.3)	5.6(2.2)	8.1(3.1)	6.9(3.0)	7.0(3.2)	13.3(6.0)	5.1(1.6)	3.8(1.5)	4.9(2.0)
	2.0	5.6(1.8)	4.5(1.5)	6.0(2.0)	5.9(2.3)	5.4(2.0)	9.5(3.5)	3.1(0.8)	3.1(1.0)	3.2(1.0)
10	0.5	18.9(14.4)	17.5(13.3)	49.2(42.5)	25.7(20.8)	26.0(21.5)	102.9(69.4)	26.3(22.2)	24.0(19.9)	67.8(55.6)
	1.0	8.2(4.0)	7.8(3.7)	14.6(7.9)	10.6(5.9)	10.8(6.0)	43.2(35.4)	5.9(2.9)	5.8(3.0)	13.1(8.2)
	1.5	5.9(2.3)	5.4(2.2)	8.8(3.5)	7.0(3.1)	7.0(3.2)	18.0(9.8)	4.8(1.7)	3.8(1.5)	5.5(2.3)
	2.0	4.8(1.6)	.3(1.5)	6.7(2.2)	5.7(2.2)	5.4(2.1)	11.4(4.4)	3.2(0.9)	2.8(0.8)	3.7(1.2)
RMI		0.0891	0.0031	0.8135	0.0198	0.0081	1.7689	0.1161	0.0000	0.7691

narrowed to a certain extent when the shift size is large. For monitoring of different dimensions, the effect of RFEWMA is less affected by the improvement of data dimensions, while the effect of MEWMA for higher dimensions is reduced. This indicates that our control chart is robust for monitoring the multivariate normal distribution data.

We now investigate the case under non-normal distributions. First, we consider the case of the multivariate t distribution with the parameter ε equal to 5. Compared with the normal distribution, the efficacy of the MEWMA decreases significantly, especially in the case of small shift size. However, the RFEWMA chart still maintains a decent performance for the case of the multivariate t distribution. With increasing data dimension, the decrease of MEWMA monitoring effect is obvious, while RFEWMA still has a robust monitoring effect. Finally, we consider a multivariate skewed chi-square distribution, and the parameter of the multivariate chi-square distribution is still chosen as 5. According to the findings of the multivariate t distribution, the overall efficacy of RFEWMA is better than that of MEWMA, especially for small shift size, and the effectiveness of RFEWMA is almost not affected by the increase in data dimension. When there is only one type of shift ($k = 1$) in the historical OOC data, the efficacy of the SVMEWMA is slightly better than that of the RFEWMA scheme.

We now consider a more general situation. We assume that there are shifts of multiple types ($k > 1$) in the historical OOC data, and that the class of the real shift is one of the historical shifts. We consider the same three multivariate distributions, namely, normal, t, and chi-square, as before. The parameter settings for the training IC set are also the same as the case of $k = 1$. We randomly generate 100,000 IC data and k-classes OOC data. The number of each class of OOC data is 25000; that is, $(N_0 = 100000, N_{11} = N_{12} = \ldots = N_{1k} = 25000)$. The above data is used to design and compare the RFEWMA, SVMEWMA and MEWMA schemes. To verify that our control chart is suitable for high-dimensional and multi-class situations, we choose $k = 2, 4, 6, 8, 10$, and $p = 6, 10, 20, 30, 50, 100$. Without loss of generality, we all assume that the magnitude of shifts is 1 and the shift takes place on only one dimension at a time. See Table 3 for specific simulation results.

We observe from Table 3 that the efficacy of the RFEWMA chart is better than that of the SVMEWMA chart and the MEWMA chart in the case of multi-classification. When the underlying density is multivariate normal, the difference between the three methods is negligible when the data dimension p is less than 10. With the increase in the classification number k and the data dimension p, the efficacy of the MEWMA scheme gradually decreases, but the effectiveness of the RFEWMA scheme is hardly affected and is slightly better than SVMEWMA. For the multivariate t distribution, the efficacy of the three methods is reduced compared to that for the multivariate normal distribution, but the monitoring effect of the MEWMA schemes decreases significantly when p exceeds 6. For higher dimensions and more OOC classes, the MEWMA scheme is practically unable to effectively monitor the shift. The effectiveness of the SVMEWMA scheme also decreases, but the

Table 3. Comparison of ARL_1 values when multiple OOC classes in historical OOC data. The SDRL value is shown in parentheses.

k	p	historical OC mean	Normal			t ($df = 5$)			$X2$ ($df = 5$)		
			RFEWMA	SVMEWMA	MEWMA	RFEWMA	SVMEWMA	MEWMA	RFEWMA	SVMEWMA	MEWMA
2	6	(1,0,0,0,0,0)	9.5(4.6)	9.4(4.5)	13.3(6.9)	**12.2(6.9)**	12.7(7.0)	28.3(20.3)	**6.6(3.4)**	7.0(3.5)	10.7(6.2)
		(0,1,0,0,0,0)	9.3(4.5)	9.1(4.4)	13.3(7.0)	12.8(7.2)	12.7(7.1)	28.1(20.6)	6.9(3.4)	6.8(3.4)	10.8(6.3)
4	6	(1,0,0,0,0,0)	9.7(4.8)	9.7(4.9)	12.9(6.4)	**15.0(8.6)**	15.4(9.2)	29.5(23.4)	7.9(3.9)	7.8(3.9)	10.2(6.1)
		(0,1,0,0,0,0)	10.7(5.7)	10.6(5.1)	13.1(7.0)	17.1(9.9)	16.6(9.7)	29.4(21.4)	**7.3(3.7)**	7.5(3.8)	10.2(6.1)
		(0,0,1,0,0,0)	**9.8(4.6)**	10.1(4.9)	13.0(6.5)	14.5(8.3)	14.5(8.9)	29.6(21.9)	**7.0(3.5)**	7.2(3.6)	10.2(6.0)
		(0,0,0,1,0,0)	10.4(5.0)	10.4(5.0)	13.3(6.8)	15.9(9.7)	15.4(9.3)	30.7(23.6)	7.7(3.8)	7.6(3.8)	10.0(6.0)
4	10	1 in 1st com	10.9(5.3)	10.7(5.0)	15.3(9.0)	**14.5(8.5)**	15.3(9.0)	41.8(34.8)	**7.2(3.6)**	7.7(3.9)	13.3(8.0)
		1 in 2nd com	10.7(5.5)	10.5(5.2)	15.9(8.7)	15.6(9.2)	15.1(8.9)	41.5(34.0)	7.6(3.9)	7.6(3.9)	13.1(8.5)
		1 in 3rd com	**10.0(5.0)**	10.4(4.9)	15.8(8.4)	**15.6(9.0)**	15.6(10.0)	42.4(33.7)	**7.9(4.0)**	8.0(4.1)	13.3(7.7)
		1 in 4th com	11.1(5.6)	10.5(5.0)	15.5(8.4)	**14.0(8.0)**	16.1(9.5)	41.6(32.3)	**7.7(3.8)**	7.9(4.1)	13.3(8.6)
6	10	1 in 1st com	10.9(5.0)	10.9(5.4)	15.1(8.2)	**17.4(10.1)**	19.1(11.9)	45.5(37.6)	**8.3(4.0)**	8.4(4.1)	13.5(8.5)
		1 in 2nd com	11.5(5.7)	11.5(5.7)	15.2(8.9)	**16.3(9.7)**	16.9(10.3)	45.3(37.1)	**7.9(3.9)**	8.3(4.0)	13.2(8.0)
		1 in 3rd com	**11.0(5.6)**	11.1(5.3)	15.3(8.4)	**16.9(10.1)**	18.9(11.4)	43.7(36.5)	**7.9(4.0)**	8.3(4.1)	13.5(8.4)
		1 in 4th com	10.6(5.3)	10.6(5.1)	15.3(8.6)	**17.6(10.7)**	18.0(10.9)	45.2(37.3)	**7.8(3.9)**	8.2(4.1)	13.6(8.6)
		1 in 5th com	10.9(5.3)	10.7(5.1)	15.2(8.4)	**16.2(9.6)**	16.6(10.5)	43.8(35.4)	**8.1(4.0)**	8.2(4.4)	13.0(7.7)
		1 in 6th com	**11.2(5.7)**	11.5(5.7)	14.8(7.7)	**16.2(9.8)**	16.8(10.1)	43.4(37.7)	**7.8(4.1)**	8.2(4.0)	13.2(8.1)
6	20	1 in 1st com	**11.4(5.8)**	10.9(5.6)	19.5(11.6)	**16.1(9.6)**	16.5(9.9)	80.7(61.2)	**7.7(37)**	8.5(4.5)	19.4(13.4)
		1 in 2nd com	**11.5(5.9)**	10.9(5.4)	19.7(12.3)	**16.7(9.8)**	18.1(11.0)	81.4(61.2)	**8.0(3.9)**	8.0(4.2)	18.0(12.3)
		1 in 3rd com	**10.7(5.0)**	11.1(5.6)	19.6(12.8)	**16.5(10.4)**	17.8(10.7)	81.5(61.6)	**8.9(4.4)**	9.0(4.5)	18.8(13.0)
		1 in 4th com	**10.5(5.3)**	11.2(5.7)	18.5(11.0)	**16.0(9.1)**	17.7(11.2)	85.2(63.1)	**8.0(4.0)**	8.5(4.2)	19.2(14.1)
		1 in 5th com	**11.0(5.4)**	11.1(5.6)	20.0(13.1)	**16.1(9.1)**	16.8(10.6)	78.0(61.3)	**7.6(3.7)**	8.2(4.1)	18.7(13.6)
		1 in 6th com	10.8(5.2)	10.7(5.2)	18.9(11.3)	**15.9(9.5)**	16.5(10.2)	86.7(63.5)	**8.7(4.2)**	8.5(4.4)	19.5(14.1)
6	30	1 in 1st com	12.5(6.3)	12.3(6.1)	21.5(13.6)	**15.1(9.3)**	17.8(10.9)	92.5(66.4)	**7.8(4.0)**	8.2(4.3)	22.7(16.9)
		1 in 2nd com	**11.2(5.3)**	11.5(5.8)	21.3(17.7)	**17.1(10.1)**	17.5(10.5)	91.7(65.1)	**8.0(3.9)**	8.2(4.1)	21.9(16.6)
		1 in 3rd com	11.8(5.6)	11.3(5.6)	22.0(13.6)	**18.3(12.1)**	17.6(10.6)	93.7(66.5)	**7.4(3.7)**	8.2(4.1)	22.1(17.1)
		1 in 4th com	**10.5(5.2)**	11.2(5.4)	21.7(13.8)	**15.9(10.0)**	17.8(11.1)	90.9(66.6)	**8.4(4.1)**	8.8(4.6)	20.7(15.4)
		1 in 5th com	**11.0(5.3)**	11.4(5.6)	21.5(13.7)	**17.2(10.0)**	18.0(11.1)	90.8(66.2)	**7.7(3.8)**	8.4(4.3)	21.5(15.9)
		1 in 6th com	11.9(5.8)	11.6(5.7)	22.0(14.5)	**15.8(10.0)**	19.3(13.1)	92.3(67.7)	**7.8(4.0)**	8.4(4.4)	23.0(17.8)
8	30	1 in 1st com	**11.4(5.8)**	11.5(5.9)	22.9(15.2)	**17.5(10.3)**	19.0(12.2)	77.6(61.0)	**8.4(4.1)**	9.3(4.9)	22.1(16.2)
		1 in 2nd com	**11.3(5.7)**	11.5(5.8)	22.4(15.2)	**17.0(10.5)**	18.6(11.3)	82.9(63.5)	**8.2(4.1)**	9.1(4.6)	22.5(18.2)
		1 in 3rd com	12.1(6.1)	11.8(5.7)	22.0(14.1)	20.1(13.2)	18.7(11.6)	82.1(62.1)	**8.4(4.1)**	9.0(4.5)	23.0(17.8)
		1 in 4th com	**11.5(5.9)**	11.8(5.9)	21.5(13.6)	**16.1(9.6)**	16.6(10.3)	80.7(63.3)	**8.6(4.3)**	9.0(4.6)	23.1(18.4)
		1 in 5th com	12.4(6.0)	11.9(5.7)	21.8(13.0)	**15.8(9.3)**	17.1(10.5)	81.4(60.8)	**8.6(4.1)**	8.9(4.6)	22.2(16.7)
		1 in 6th com	**12.0(6.2)**	12.6(6.1)	22.4(14.7)	**14.9(8.8)**	15.8(9.9)	81.9(63.1)	**8.4(4.2)**	9.1(4.7)	23.2(17.8)
		1 in 7th com	12.0(6.3)	11.6(6.1)	22.2(14.0)	**16.4(9.9)**	17.4(10.5)	82.7(62.3)	**8.6(4.3)**	9.3(4.6)	23.7(17.5)
		1 in 8th com	11.8(5.8)	11.6(5.9)	22.1(14.3)	**15.6(9.2)**	16.4(10.5)	76.6(61.7)	**8.5(4.1)**	9.1(4.5)	22.5(18.2)
8	50	1 in 1st com	13.4(6.4)	12.7(6.5)	27.3(20.3)	**19.5(12.7)**	20.4(13.4)	102.9(69.5)	**8.0(3.8)**	8.6(4.2)	30.5(24.5)
		1 in 2nd com	13.5(7.0)	13.5(7.0)	27.8(20.0)	**18.5(11.8)**	20.4(13.0)	100.6(69.9)	9.1(4.5)	9.1(4.4)	29.6(23.9)

(*Continued*)

Table 3. (Continued).

k	p	historical OC mean	Normal			t ($df = 5$)			$X2$ ($df = 5$)		
			RFEWMA	SVMEWMA	MEWMA	RFEWMA	SVMEWMA	MEWMA	RFEWMA	SVMEWMA	MEWMA
		1 in 3rd com	**11.8(5.7)**	12.5(6.1)	27.6(19.8)	**18.6(11.5)**	20.1(13.0)	107.1(70.6)	**8.1(3.8)**	9.0(4.7)	29.0(23.5)
		1 in 4th com	**11.9(5.6)**	12.8(6.5)	26.3(18.2)	**15.1(9.1)**	17.5(11.40)	99.4(68.9)	**8.2(4.1)**	9.4(4.8)	29.9(24.1)
		1 in 5th com	**11.9(5.9)**	11.9(6.0)	26.9(19.3)	**13.3(7.9)**	17.4(10.6)	100.8(70.5)	**8.6(4.3)**	8.9(4.7)	30.5(26.0)
		1 in 6th com	12.4(6.1)	12.4(6.0)	28.1(21.2)	**20.7(14.1)**	21.4(14.1)	104.5(71.7)	**8.8(4.4)**	9.3(4.9)	30.3(24.6)
		1 in 7th com	**11.6(5.5)**	12.5(6.3)	27.1(18.7)	**13.4(7.7)**	16.6(10.3)	101.1(69.7)	**8.4(4.0)**	9.0(4.6)	29.8(23.8)
		1 in 8th com	11.3(5.5)	11.2(5.7)	27.6(19.8)	**17.8(11.3)**	21.0(14.2)	103.4(71.9)	**8.4(3.8)**	8.9(4.3)	28.8(24.0)
10	50	1 in 1st com	13.3(7.1)	12.1(6.1)	28.2(20.0)	**18.1(11.3)**	19.8(12.7)	105.3(70.6)	**8.6(4.3)**	9.1(4.6)	29.4(24.4)
		1 in 2nd com	14.0(7.2)	12.2(6.2)	27.3(19.6)	**17.7(10.9)**	18.8(12.3)	106.3(69.7)	**8.4(4.2)**	9.0(4.6)	29.9(24.1)
		1 in 3rd com	**11.4(5.6)**	12.4(6.1)	27.4(19.7)	**20.7(13.4)**	19.1(12.4)	100.8(70.1)	**8.3(4.2)**	9.3(4.8)	28.4(22.5)
		1 in 4th com	**12.1(6.0)**	12.3(5.9)	26.7(18.6)	**16.8(10.5)**	19.4(13.7)	102.0(70.1)	**8.8(4.2)**	8.9(4.5)	28.4(24.9)
		1 in 5th com	13.5(7.1)	12.9(6.3)	26.9(19.7)	21.0(13.3)	19.1(12.6)	109.7(70.1)	**8.6(4.1)**	9.3(4.9)	28.9(23.3)
		1 in 6th com	13.1(6.1)	13.1(6.4)	28.5(20.2)	**15.5(9.3)**	18.6(11.3)	103.6(71.0)	**9.0(4.6)**	9.5(4.7)	29.1(23.8)
		1 in 7th com	13.8(7.4)	12.0(6.1)	28.0(19.7)	**16.0(9.9)**	19.3(13.0)	102.7(71.5)	**8.1(4.0)**	9.0(4.5)	29.2(23.7)
		1 in 8th com	12.2(5.9)	12.2(6.1)	26.2(18.6)	**16.6(10.3)**	18.4(11.5)	105.3(69.5)	**8.7(4.5)**	9.9(5.3)	28.7(22.8)
		1 in 9th com	**12.5(6.4)**	13.0(6.6)	29.1(20.6)	**17.9(10.9)**	21.1(13.9)	106.2(69.6)	**8.8(4.4)**	9.7(4.8)	27.8(22.6)
		1 in 10th com	**11.7(5.7)**	12.7(6.3)	28.2(20.5)	**19.8(12.5)**	24.2(17.5)	105.9(71.1)	**8.9(4.3)**	9.0(4.8)	28.7(22.9)
10	100	1 in 1st com	**11.7(5.6)**	13.4(7.1)	36.1(28.3)	23.8(16.9)	20.6(13.5)	113.5(72.2)	**8.7(4.1)**	10.0(5.1)	44.4(37.7)
		1 in 2nd com	**12.9(7.3)**	13.3(7.3)	35.9(28.5)	**18.4(11.6)**	21.5(15.3)	116.5(72.1)	**8.7(4.1)**	10.1(5.2)	41.0(36.8)
		1 in 3rd com	**11.2(5.6)**	12.2(6.2)	35.2(27.2)	**17.2(11.0)**	22.6(14.9)	117.2(72.3)	**8.8(4.4)**	10.4(5.4)	42.5(37.7)
		1 in 4th com	13.9(7.0)	13.7(7.9)	35.9(27.7)	**20.0(12.9)**	20.0(13.6)	115.2(71.9)	**8.5(4.1)**	9.7(5.4)	41.7(35.8)
		1 in 5th com	**12.2(6.3)**	13.1(7.2)	36.9(29.7)	**17.1(10.8)**	20.4(14.2)	119.6(72.1)	**8.2(4.1)**	9.5(5.1)	42.1(36.7)
		1 in 6th com	**11.9(6.0)**	12.7(6.5)	35.8(28.0)	**19.7(12.3)**	23.6(15.9)	118.2(72.7)	**8.4(4.0)**	9.2(4.9)	41.2(35.5)
		1 in 7th com	**13.1(6.8)**	13.7(7.2)	37.5(29.1)	**17.5(10.7)**	19.8(13.2)	115.3(71.2)	**8.7(4.6)**	10.3(5.6)	40.7(36.7)
		1 in 8th com	**11.8(5.8)**	12.2(6.4)	36.0(28.8)	**15.2(8.8)**	20.3(13.2)	119.2(72.2)	**8.4(4.3)**	9.7(5.4)	43.1(38.3)
		1 in 9th com	**11.6(5.6)**	13.2(7.3)	37.2(29.1)	**18.8(11.5)**	22.6(16.0)	114.4(71.2)	**8.9(4.6)**	10.2(5.3)	40.9(37.2)
		1 in 10th com	**11.1(5.6)**	12.6(6.5)	37.8(30.0)	**14.6(8.4)**	20.0(13.0)	115.1(72.5)	**8.7(4.2)**	10.1(5.5)	44.1(38.3)
RMI			**0.0155**	0.0258	1.0437	**0.0087**	0.0989	3.9603	**0.0010**	0.0713	1.9526

RFEWMA scheme has little impact on the increase of p and k, and it still maintains a good efficacy even when $p = 100$. For the multivariate chi-square distribution, the effectiveness of the RFEWMA scheme is even better than that of the multivariate normal distribution, and efficacy remains robust for the increase of p and k. The above results show that the RFEWMA scheme can effectively monitor high-dimensional data with multiple possible classes of shifts. Although the SVMEWMA and MEWMA schemes are slightly better in some cases, the RFEWMA scheme is significantly better than its competitors when we consider the RMI metric.

So far, we investigate the effectiveness of the RFEWMA scheme when the actual shift of online data is the same as that of the historical OOC shift patterns. However, the RFEWMA scheme could also be helpful when the process encounters a different type of shift in real-time, not observed earlier in the training dataset. This aspect is discussed in Section 4.2. Since the random forest technique is an excellent classifier of complex and high-dimensional data in the presence of multiple classes, the RFEWMA scheme is robust for monitoring different process distributions and high-dimensional data.

To verify the validity of our proposed scheme, we consider the case where the covariance matrix in the IC process is not the identity matrix. We use σ_{ij} to represent the element of the i^{th} row and j^{th} column of the covariance matrix and assume that it is equal to $(0.5)^{|i-j|}$. We set $k = 6$ and $p = 10$. Specific simulation results are shown in Table 4. From Table 4, it can be seen that the RMI values of our scheme are smaller than those of the SVMEWMA scheme and the MEWMA scheme under these three distributions, which indicates that our scheme still performs well when the covariance matrix is not the identity matrix. Without loss of generality, we assume that the covariance matrix is the identity matrix in the subsequent simulation.

4.2 Monitoring of categorical and mixed-type data

In this subsection, we mainly discuss the monitoring of categorical and mixed-type data. We choose the SVMEWMA and MEWMA schemes as the competitors for the performance assessment of the proposed RFEWMA scheme. When simulating the monitoring problem of categorical data, we first assume that the historical data dimension is 2 ($p = 2$), one variable has two levels, and the other has three levels. Their marginal probabilities are {0.4,0.6} and {0.3,0.4,0.3}. The shift of the historical OOC data occurs in the first dimension, and its marginal probabilities become $\{0.4 + \delta, 0.6 - \delta\}$. The number of the historical IC data and the historical OOC data are both 100,000. The simulation results of the three control schemes are shown in Table 5. It can be seen from the table that the monitoring effect of the RFEWMA scheme and the SVMEWMA scheme is better than that of MEWMA, and RFEWMA has certain advantages in monitoring small shifts.

We then simulate the situation where there are multiple classes of shifts in the historical OOC data. In this case, we assume that the historical data dimension is 10 ($p = 10$) and OOC data is divided into 4 classes ($k = 4$).The marginal probability of each variable is separately {0.4, 0.6}, {0.5, 0.5}, {0.7, 0.3}, {0.8, 0.2}, {0.3, 0.4, 0.3}, {0.2, 0.3, 0.5}, {0.2, 0.3, 0.3, 0.2}, {0.1, 0.2, 0.3, 0.4}, {0.2, 0.2, 0.2, 0.2, 0.2} and {0.1, 0.2, 0.4, 0.2, 0.1}.We assume that the shift of historical data only occurs in one dimension, and the classes are $\{0.4 + \delta, 0.6 - \delta\}, \{0.2 + \delta, 0.3, 0.5 - \delta\}, \{0.1 + \delta, 0.2, 0.3, 0.4 - \delta\}, \{0.1 + \delta, 0.2, 0.4 - \delta, 0.2, 0.1\}$, and the real shift type is one of the above conditions. We generate the above historical data, in which the number of IC data is 10,000 and the number of each class of OOC data is 25,000. The results are tabulated in Table 6.

It can be found from Table 6 that the RMI value of the RFEWMA scheme is 0.043, which is smaller than that of the other two schemes. Although the efficacy of the RFEWMA scheme is slightly inferior in a few cases, overall, the RFEWMA scheme is the most robust for monitoring the attribute data with multiple OOC classes, especially in monitoring the small shifts, such as $\delta \leq 0.02$. At the same time, with the increase in shift size, the efficacy of the RFEWMA scheme increases most rapidly. The above results show that the RFEWMA scheme is also good at monitoring the attribute data with multiple OOC classes.

Table 4. Comparison of ARL_1 values when the covariance matrix is not the identity matrix. The SDRL value is shown in parentheses.

k	p	historical OC mean	Normal			t $(df = 5)$			$X2$ $(df = 5)$		
			RFEWMA	SVMEWMA	MEWMA	RFEWMA	SVMEWMA	MEWMA	RFEWMA	SVMEWMA	MEWMA
6	10	1 in 1st com	**8.5(3.9)**	8.9(4.0)	12.2(5.8)	**12.7(6.7)**	13.2(6.9)	29.5(22.2)	**8.0(4.0)**	8.1(4.1)	9.2(5.2)
		1 in 2nd com	**8.5(3.6)**	8.8(3.7)	10.8(4.8)	**11.6(5.6)**	11.7(6.2)	23.8(15.9)	**7.3(3.5)**	7.3(3.7)	11.1(6.1)
		1 in 3rd com	**8.8(3.6)**	9.2(3.7)	10.8(4.7)	**12.5(6.0)**	12.6(6.3)	24.5(15.9)	**7.8(3.8)**	7.8(3.8)	10.9(5.9)
		1 in 4th com	**8.5(3.4)**	8.8(3.7)	10.7(4.8)	**11.2(5.4)**	12.0(5.9)	23.2(14.9)	**7.7(3.8)**	8.1(4.1)	11.1(6.3)
		1 in 5th com	**8.6(3.7)**	9.2(3.8)	11.0(4.8)	**11.6(5.4)**	12.0(6.1)	23.2(16.6)	**7.8(3.9)**	7.3(3.6)	10.9(6.0)
		1 in 6th com	**8.0(3.4)**	8.8(3.8)	10.9(4.8)	**11.5(5.3)**	11.8(6.0)	22.6(14.6)	**7.5(3.7)**	7.7(3.7)	11.0(6.0)
RMI			**0**	0.0639	0.3161	**0**	0.0313	1.0619	**0.0114**	0.0152	0.4116

Table 5. Comparison of ARL_1 values for categorical data. The SDRL value is shown in parentheses.

δ	RFEWMA	SVMEWMA	MEWMA
0.01	**122.4(71.2)**	130.4(71.4)	131.0(71.1)
0.02	**116.5(71.0)**	125.1(72.5)	127.3(73.0)
0.03	113.9(71.6)	108.7(69.6)	121.1(74.2)
0.04	104.3(70.1)	102.6(70.1)	139.1(71.5)
0.05	106.0(70.1)	91.7(76.6)	140.9(71.2)
0.10	58.4(49.1)	55.2(47.8)	142.1(70.2)
RMI	0.0464	0.0232	0.4573

Table 6. Comparison of ARL_1 values for categorical data when there are multiple OOC classes in historical OOC data. The SDRL value is shown in parentheses.

	historical OC	RFEWMA	SVMEWMA	MEWMA
0.01	OC 1	126.7(72.3)	149.6(67.7)	125.7(72.5)
	OC 2	**127.3(71.7)**	149.4(67.9)	129.4(72.7)
	OC 3	**122.9(71.8)**	141.1(69.7)	131.7(71.2)
	OC 4	**124.7(73.0)**	148.7(66.9)	126.9(72.5)
0.02	**OC 1**	**125.4(71.5)**	137.4(71.0)	125.4(73.7)
	OC 2	**128.6(72.6)**	140.1(71.2)	129.7(72.3)
	OC 3	**118.8(73.2)**	139.1(69.7)	132.5(71.2)
	OC 4	**127.5(72.5)**	133.6(72.5)	128.3(72.9)
0.03	OC 1	139.7(69.9)	119.3(73.2)	118.5(72.6)
	OC 2	130.8(71.2)	135.9(72.9)	119.6(73.6)
	OC 3	137.9(71.2)	128.8(73.3)	129.5(72.8)
	OC 4	133.3(69.5)	127.9(72.0)	123.2(72.8)
0.04	OC 1	**127.7(71.6)**	133.2(70.4)	132.1(71.5)
	OC 2	134.6(71.7)	131.3(72.0)	139.9(69.3)
	OC 3	**110.1(71.0)**	130.0(70.7)	145.6(69.6)
	OC 4	**125.8(71.4)**	128.9(72.0)	143.0(70.2)
0.05	OC 1	131.4(70.5)	128.7(73.1)	116.2(73.1)
	OC 2	127.0(70.4)	141.9(71.3)	116.0(74.9)
	OC 3	**121.6(70.6)**	143.4(70.2)	130.4(75.0)
	OC 4	136.6(68.5)	130.5(72.0)	121.3(73.8)
0.1	OC 1	**108.6(69.8)**	129.5(70.7)	134.1(72.7)
	OC 2	**109.8(70.9)**	109.7(70.8)	147.2(69.3)
	OC 3	56.5(48.2)	46.6(39.2)	167.8(60.4)
	OC 4	**83.6(63.7)**	144.1(71.0)	154.0(67.3)
RMI		**0.0426**	0.1267	0.2039

There are hardly a few effective methods for monitoring mixed-type data. We still choose the SVMEWMA and MEWMA schemes for comparing the effectiveness of our proposed method. In this case, we assume that the historical data dimension is 10 ($p = 10$). Among them $(x_1, \ldots, x_8)$ are continuous data, $x_9 = I_{\{x_1+x_2>2\}} - I_{\{x_1+x_2<2\}}$ and $x_{10} = I_{\{x_6>1\}} - I_{\{x_6<-1\}}$ are discrete data. We consider three cases where continuous data are subject to multivariate normal distribution, the multivariate t distribution, and the multivariate chi-square distribution, respectively. The parameter ε of the multivariate t distribution and the multivariate chi-square distribution is fixed at 5. We assume that the mean of the IC process is the null vector and the covariance matrix is the identity matrix. The classes of historical OOC data are 4 ($k = 4$), and their shifts are, respectively: mean of x_1; x_2 or x_6 changes with magnitude equal to 1 (a shift in both continuous and categorical data); mean of x_3; x_4; x_5; x_7 or x_8 changes with magnitude equal to 1 (a shift in continuous data); a threshold of categorical variables change from 2 to 1 ($x_9 = I_{\{x_1+x_2>1\}} - I_{\{x_1+x_2<1\}}$); a threshold of categorical variables change from 1 to 0.5 ($x_{10} = I_{\{x_6>0.5\}} - I_{\{x_6<-1\}}$). The numerical results based on simulation are shown in Table 7. We observe that the RFEWMA scheme has good efficacy for mixed data of various types. The results obtained in the current subsection and the previous subsection indicate that the RFEWMA scheme is robust with high efficacy for processes with different distributions, including continuous, discrete, and mixed-type, especially in the case of

Table 7. Comparison of ARL_1 values for mixed-type data when there are multiple OOC classes in historical OOC data. The SDRL value is shown in parentheses.

		Normal			t $(df = 5)$			$X2$ $(df = 5)$		
Shift location	Magnitude	RFEWMA	SVMEWMA	MEWMA	RFEWMA	SVMEWMA	MEWMA	RFEWMA	SVMEWMA	MEWMA
x1	1	**9.9(5.1)**	10.0(5.1)	142.4(71.1)	**14.2(8.2)**	14.9(8.8)	137.1(72.0)	7.3(3.5)	7.2(3.8)	12.3(7.4)
x3	1	**10.0(5.1)**	10.7(5.3)	23.9(18.6)	**13.0(7.5)**	14.3(7.8)	41.4(35.4)	7.5(3.8)	7.2(3.6)	12.3(7.7)
x9	Threshold becomes 1.0	18.0(13.7)	17.6(12.4)	185.3(46.7)	29.4(24.2)	23.0(18.1)	172.5(57.1)	**121.1(72.3)**	125.4(72.8)	119.0(74.0)
x10	Threshold becomes 0.5	45.7(37.3)	25.4(19.7)	86.7(64.6)	59.1(49.6)	50.9(43.0)	110.8(71.6)	**111.9(70.6)**	119.0(71.9)	122.7(74.6)
RMI		0.2055	0.0200	6.6789	0.1098	0.0373	4.6291	**0.0183**	0.0293	0.3783

high-dimensional and multi-classification data. Consequently, the proposed scheme provides a good solution for monitoring many complex processes.

4.3 When historical shifts and real shifts are slightly dissimilar

The real-time process shifts considered in the previous subsections belong to one of the classes of shift patterns identified in the historical data. However, it is difficult to guarantee this assumption in practice. This section mainly simulates different types of shifts in real-time that are not members of the k OOC classes realized in Phase-I. As before, we compare the proposed scheme with the SVMEWMA and MEWMA schemes under a similar IC framework with the three continuous distributions used in Subsection 3.4. We take the case of $k = 4$ and $p = 10$ as an example, that is, suppose there are four classes of historical OOC data, and the size of each class is 25,000. The shifts of historical OOC data occurred in the first four dimensions, and the shift magnitudes are 1. We simulate the case that the shift magnitude in Phase-II is $0.1 - 1.0$, and the shift occurs at the first dimension. We also consider the possibility that the changes occur on two dimensions. Specific simulation results are shown in Table 8.

Table 8 shows that there is little difference between the RFEWMA and SVMEWMA schemes in the case of the multivariate normal distribution. However, in the case of non-normal distributions, the performance of the RFEWMA scheme is better than that of the other two schemes, especially in the case of small real-time shifts. When the change in Phase-II is similar to that observed in some classes of training samples, the effectiveness of the RFEWMA scheme does not decrease considerably. The results are alike for other types of shifts. This further proves that our proposed SPM scheme has good performance for complex process monitoring.

4.4 When the historical data is unbalanced

Our proposed scheme is designed using the historical data which has been collected and suitably classified. The previous simulation designs assume that all the k OOC classes are of equal size, but this is impractical in real situations. Different classes may contain different sizes of data. This section assumes that the historical data are unbalanced, consists of 100,000 IC sample observations and a varying number of OOC observations from other classes. We take the same three multivariate continuous distributions as before to compare the efficacies of the three SPM schemes. Similar results can be obtained for different types of data but are omitted for brevity. The simulation results are shown in Table 9.

Similar to the continuously balanced case results, the overall monitoring effects of the RFEWMA and SVMEWMA schemes are better than that of the MEWMA scheme. In the case of the multivariate normal distribution, the monitoring effect of the SVMEWMA scheme is slightly better than that of the RFEWMA scheme, but in the case of non-normal distributions, the monitoring effect of the RFEWMA scheme is better and more robust. Correlative conclusions can also be drawn from the RMI values, which indicates that the proposed scheme maintains a high efficacy for unbalanced data, which is also in line with the characteristics of the random forest model.

4.5 Determination of some important parameters

In this subsection, we mainly discuss selecting some critical parameters in the model during the construction of the RFEWMA scheme. For different data distributions and data types, there may be different possible choices of optimal parameters. The multivariate normal distribution and the multivariate t distribution are taken as examples. We minimize ARL_1 as the standard to determine the optimal parameters under the condition that the shift of OOC data only occurs in the first

Table 8. Comparison of ARL_1 values when a similar shift pattern has appeared. The SDRL value is shown in parentheses.

	Normal			t $(df = 5)$			$X2$ $(df = 5)$		
Type of online OC data	RFEWMA	SVMEWMA	MEWMA	RFEWMA	SVMEWMA	MEWMA	RFEWMA	SVMEWMA	MEWMA
0.1 in 1st com	98.8(68.8)	94.8(69.7)	116.1(71.3)	118.7(71.8)	122.6(70.3)	129.6(71.2)	123.0(72.2)	118.3(72.4)	123.5(71.4)
0.2 in 1st com	77.1(62.8)	72.9(60.9)	101.3(69.1)	**98.6(67.3)**	107.3(70.2)	122.8(71.5)	**107.1(72.7)**	110.1(72.7)	116.0(72.7)
0.3 in 1st com	**53.2(44.1)**	55.0(46.7)	86.2(64.0)	**75.2(61.5)**	82.9(65.3)	120.6(72.3)	93.8(67.0)	88.9(67.3)	109.0(70.2)
0.4 in 1st com	**35.6(30.3)**	36.7(32.2)	65.7(54.0)	**61.5(50.6)**	66.3(54.8)	113.0(71.8)	**65.9(55.8)**	67.4(55.8)	89.0(67.3)
0.5 in 1st com	26.8(21.4)	24.8(20.0)	47.2(39.8)	**44.9(37.3)**	49.0(40.9)	106.9(70.0)	**40.6(35.2)**	42.4(36.2)	70.3(56.5)
0.6 in 1st com	20.8(15.0)	19.4(13.0)	36.9(29.0)	**33.8(27.2)**	37.5(30.6)	89.2(64.9)	**25.3(20.2)**	25.9(21.4)	50.0(44.3)
0.7 in 1st com	16.3(10.5)	16.0(9.6)	27.2(20.0)	**25.7(19.2)**	28.2(21.7)	78.4(60.0)	**17.1(11.7)**	17.2(12.4)	34.7(28.9)
0.8 in 1st com	**13.5(7.4)**	13.6(7.5)	20.9(13.5)	**20.8(14.1)**	22.3(15.4)	69.9(53.7)	**11.5(7.0)**	12.1(1.3)	24.9(20.5)
0.9 in 1st com	**11.4(6.0)**	11.8(6.4)	17.2(10.6)	**18.0(11.1)**	18.1(11.8)	54.1(45.3)	**9.3(4.8)**	9.8(5.4)	17.3(11.9)
1.0 in 1st com	9.9(5.1)	9.8(4.9)	15.2(8.1)	**15.2(8.2)**	15.4(9.2)	45.5(35.3)	**7.4(3.7)**	7.6(3.9)	13.4(8.5)
0.5 in 1st,2nd com	16.8(10.7)	16.7(10.4)	26.0(18.8)	**32.0(26.0)**	33.2(27.5)	79.7(61.6)	25.8(21.7)	25.4(21.0)	37.3(32.6)
1.0 in 1st,2nd com	8.0(3.4)	7.4(3.1)	9.9(4.2)	12.0(6.1)	11.6(5.3)	21.2(12.4)	**5.5(2.4)**	5.7(2.4)	7.1(3.4)
RMI	0.0307	0.0089	0.5907	**0.0029**	0.0607	1.2939	**0.0092**	0.0245	0.5864

Table 9. Comparison of ARL_1 values when the historical data are unbalanced. The SDRL value is shown in parentheses.

				Normal			t ($df = 5$)			$X2$ ($df = 5$)		
k	p	historical OC mean	OC number	*RFEWMA*	SVMEWMA	MEWMA	RFEWMA	SVMEWMA	MEWMA	RFEWMA	SVMEWMA	MEWMA
2	6	(1,0,0,0,0,0)	25000	9.6(4.6)	9.3(4.6)	13.2(7.2)	**13.8(7.3)**	13.9(7.8)	31.8(23.8)	7.1(3.6)	6.9(3.4)	10.5(6.1)
		(0,1,0,0,0,0)	50000	9.1(4.4)	9.0(4.6)	13.4(6.7)	12.5(6.6)	12.4(6.8)	29.9(22.1)	**6.3(3.1)**	6.6(3.5)	10.2(5.9)
4	6	(1,0,0,0,0,0)	10000	11.6(5.6)	11.4(5.6)	13.3(6.8)	20.0(13.0)	18.3(11.4)	29.3(21.0)	**8.7(4.2)**	8.7(4.3)	10.8(6.4)
		(0,1,0,0,0,0)	20000	10.1(5.0)	10.0(5.0)	13.2(6.9)	**17.5(10.4)**	18.2(11.2)	29.3(21.6)	**8.0(3.9)**	8.1(4.0)	10.8(6.5)
		(0,0,1,0,0,0)	30000	**9.5(4.8)**	9.7(4.8)	13.2(6.8)	**14.3(8.1)**	15.1(9.0)	28.5(21.3)	**7.4(3.6)**	7.7(3.9)	11.0(6.7)
		(0,0,0,1,0,0)	40000	**9.8(4.8)**	10.3(5.0)	13.1(6.5)	**13.6(8.0)**	13.8(8.3)	29.1(21.8)	**6.9(3.3)**	6.9(3.5)	10.4(5.9)
6	6	1 in 1st com	10000	**11.5(5.7)**	11.6(5.8)	13.4(6.7)	**16.6(10.3)**	18.5(11.3)	28.6(20.6)	**8.4(4.3)**	8.9(4.4)	10.3(5.9)
		1 in 2nd com	20000	10.7(5.4)	10.5(5.3)	13.5(6.8)	16.9(10.2)	14.9(9.6)	29.5(21.9)	7.9(4.2)	7.7(3.9)	10.4(6.2)
		1 in 3rd com	30000	10.7(5.2)	10.2(5.1)	13.7(6.7)	15.6(9.3)	15.5(8.8)	29.6(20.6)	**8.1(4.0)**	8.3(4.2)	10.4(6.1)
		1 in 4th com	5000	**13.4(6.7)**	13.5(6.7)	13.6(6.7)	**23.4(16.9)**	25.3(17.4)	31.4(23.5)	**9.3(4.5)**	9.6(4.7)	10.4(6.1)
		1 in 5th com	15000	11.5(5.9)	11.0(5.4)	13.2(7.0)	19.5(13.0)	18.7(12.3)	30.7(23.0)	7.9(3.8)	7.6(3.7)	10.4(6.2)
		1 in 6th com	25000	9.8(7.0)	9.8(4.7)	13.4(6.6)	17.0(10.6)	17.0(10.4)	29.7(22.0)	7.8(3.8)	7.8(3.8)	10.2(5.8)
RMI				0.0154	0.0074	0.2915	**0.0237**	0.0261	0.8722	**0.0079**	0.0181	0.3651

dimension. We assume that the process dimension is 10 ($p = 10$) and that there is only one type of shift in the OOC data ($k = 1$). Both IC data and OOC data used to train the random forest model are 100,000, and ARL_0 is fixed at 200.

We first discuss the determination of parameter λ in the MEWMA sequence. In the case of $\lambda = 0.05, 0.1$ and 0.2, we simulate the change in the ARL_1 value corresponding to different shift sizes, as shown in Figure 1(a, b). It can be seen from the figure that the results of the multivariate normal distribution and the multivariate t distribution are similar: in the case of small shifts, the ARL_1 value is the smallest when $\lambda = 0.05$, and with the increase in the shift sizes, the difference between the ARL_1 values corresponding to three kinds of λ gradually decreases. For large shifts, the efficacy of the proposed scheme is slightly better when $\lambda = 0.2$. To sum up, the traditional concept of using a smaller λ for small shifts and a larger λ for large shifts is also valid for our proposed scheme.

We study the choice of the maximum number of features used by a single decision tree (*mtry*) and the number of decision trees (*mtree*) in the random forest model. First, we set $\lambda = 0.1$ and $mtree = 300$ in the MEWMA model to simulate the situations with $mtry = 1, 3, 5, 7, 9$. From Figure 2(a, b), we see that in the case of a small shift, such as $\delta = 0.5$, the results of optimal *mtry* in the two distributions are different. For the multivariate normal distribution, $mtry = 9$ should be selected, while for the multivariate *t* distribution, *mtry* should be set to 7. In the case of large shifts, the results are similar in both distributions. When $mtry = 3$, a point of inflection appears in the graph. To ensure the running speed of the algorithm, we recommend setting the value of *mtry* as the positive square root of the data dimension in the case of large shifts. When determining the value of *mtree*, we also choose $\lambda = 0.1$ in the MEWMA model and $mtry = 3$ to simulate the situation of $mtree = 50, 100, 200, 300, 400, 500$. From Figure 2(c, d), it is observed that the smaller the number of decision trees, the smaller the value of ARL_1. In large shifts, the change in the number of decision trees has little effect on the monitoring effect. Here, we recommend that the value of *mtree* should be between 50 and 300.

Finally, we discuss the determination of parameter M in the MEWMA sequence, taking the multivariate *t* distribution as an example. We use the minimization of ARL_1 as the criterion to determine the optimal parameters when OOC data occurs in different dimensions. We assume that the process dimension is 6 ($p = 6$) and there are six types of OOC data ($k = 6$). Both IC and OOC data used to train the random forest model are 100,000, and ARL_0 is fixed at 200. We first set $\lambda = 0.1$, $mtree = 300$, and the value of *mtry* is set to the positive square root of the data dimension in the MEWMA model. As shown in Table 10, the ARL_1 value is generally minimized when $M = 20$. Therefore, we recommend setting M to 20.

4.6 Required sample size to determine the control limit

In previous subsections, we assumed that sufficient historical IC data and OOC sample observations were available in Phase-I. However, it may be necessary to determine the specific total sample size in practical applications. In this subsection, we take only one shift class in OOC data as an example ($k = 1$) to study the total sample size required to determine the robust control limit L under the multivariate normal and multivariate *t* distributions. We only consider the case where the number of historical IC data m_0 is equal to the number of historical OOC data m_1. The data dimension is selected as 10 ($p = 10$), and the shift of OOC data occurs in the first dimension, and the magnitude is 1. The summary of the numerical results based on simulation is presented in Table 11. We use 100,000 IC data and 100,000 OOC data to train the model and construct the RFEWMA scheme. We use these data to calculate the scheme L under the two distributions. The scheme L of 0.312 and 0.135 is calculated, respectively, under the condition that ARL_0 is fixed at 200. We simulate different m_0 conditions and train new models. An independent set of IC data is regenerated, and we calculate the new ARL_0 values with the control limit calculated before when $m_0 = 100000$. We

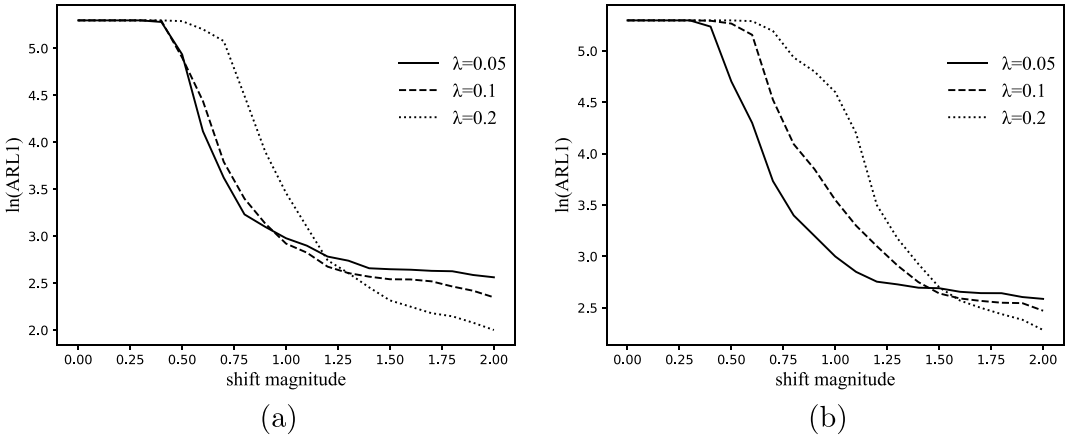

Figure 1. The ARL_1 values of different λ in (a) the multivariate normal distribution, and (b) the multivariate t distributions.

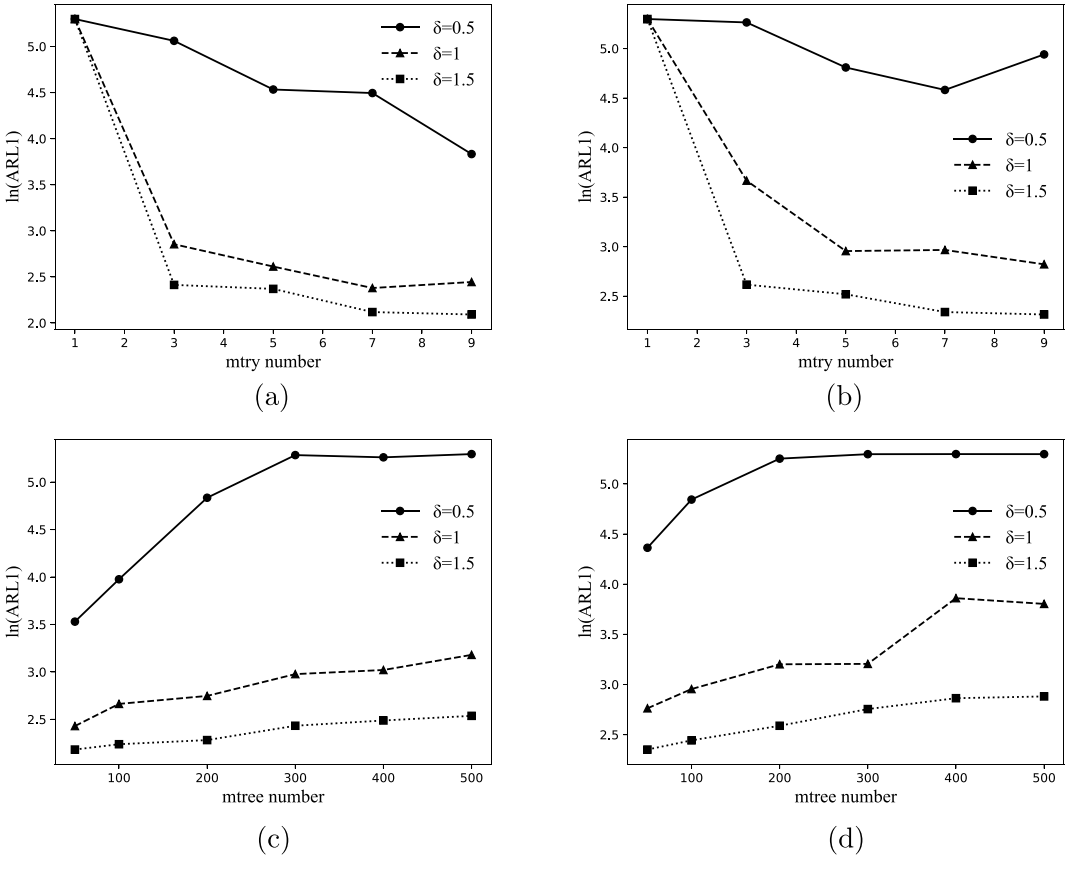

Figure 2. The ARL_1 values corresponding to different mtry values in (a) the multivariate normal distribution, and (b) the multivariate t distribution when $mtree = 300$; and the ARL_1 values corresponding to different mtree values in (c) the multivariate normal distribution, and (d) the multivariate t distribution when $mtry = 3$.

Table 10. Comparison of ARL_1 values when M is different. The SDRL value is shown in parentheses.

k	p	historical OC mean	t ($df = 5$)				
			M=10	M=15	M=20	M=25	M=30
6	6	1 in 1st com	24.8(20.7)	22.6(17.3)	**21.5(16.1)**	21.7(14.8)	24.4(14.2)
		1 in 2nd com	29.2(22.4)	27.0(18.8)	25.3(17.4)	**24.5(16.9)**	29.2(20.6)
		1 in 3rd com	16.0(10.4)	**15.3(10.5)**	15.4(11.0)	16.3(12.7)	17.1(12.0)
		1 in 4th com	82.1(66.8)	65.6(54.4)	**51.6(42.1)**	52.2(43.1)	56.2(43.0)
		1 in 5th com	41.2(40.4)	31.4(24.8)	**27.1(20.1)**	28.9(21.8)	30.5(22.9)
		1 in 6th com	36.4(30.8)	33.1(25.0)	**27.9(20.0)**	28.2(19.4)	31.9(21.3)
RMI			0.3012	0.1283	**0.0065**	0.0272	0.1337

Table 11. Simulation results about ARL_1 on sample size under the multivariate normal distribution and the multivariate t distribution with $\lambda = 0.1$. The SDRL value is shown in parentheses.

m_0	Normal	t ($df = 5$)
10000	187.4(177.4)	190.5(182.8)
30000	186.4(176.9)	201.5(193.6)
50000	218.8(206.5)	195.5(186.1)
70000	198.3(189.0)	198.8(189.5)
90000	201.4(192.0)	198.4(191.0)

can see that when m_0 is 10,000, there is a certain gap between the recalculated ARL_0 value and 200, but with the increase in the sample size, the gap has gradually narrowed down. For the multivariate normal distribution, the ARL_0 is close to 200 when m_0 is 70,000, while in the case of the multivariate t distribution, a good effect has been achieved when m_0 is 30,000. However, the determination of the sample size under more different data types and distributions needs more research in the future.

5 Illustrative example

We now apply our method to a real monitoring example mentioned earlier. In this example, we use the dataset provided by Zhang et al. (2015). The authors divided the raw data into IC data and OOC data. Considering that the system can record a large number of historical IC data and OOC data quickly, we apply the RFEWMA scheme to monitor the HDDMS process mentioned in Section 1. More than 3 million pieces of IC data and 120,000 pieces of OOC data are collected, and we explore the distribution of the data through visualization. We take the 5^{th}, 6^{th} and 10^{th} attributions of the data as the representative, and draw the scatter plots of 200 random samples, as shown in Figure 3(a)–(c) and the normal Q–Q plots as shown in Figure 3(d–f). We can see that each marginal has a different distribution, and it is difficult to approximate them by some well-known distribution, and the SPM scheme based on the distributional assumption may not solve the problem of monitoring the process well. Thus we can apply our method to this dataset.

In Phase-I, we need to classify the raw data. Since the IC data and the OOC data above have been classified beforehand, we only need to determine the specific classification number of the OOC data. Here we use the $k-means$ clustering method to cluster the OOC data based on the classification criterion of minimizing the sum of squared error (SSE) MacQueen (1967). The OOC data is divided into two classes ($k = 2$). The sample sizes of each class are 79,500 and 45,277. The result of clustering is shown in Figure 4.

We randomly select 100,000 pieces of IC sample observations and 250,000 pieces of OOC observations from each of the two OOC classes from the historical data set to construct the RFEWMA scheme. At the same time, we select another 100 pieces of IC and 200 pieces of OOC samples from the Phase-I data, which do not use for training the random forest model to generate MEWMA sequences for online testing. Here, we set the number of the decision trees in the random

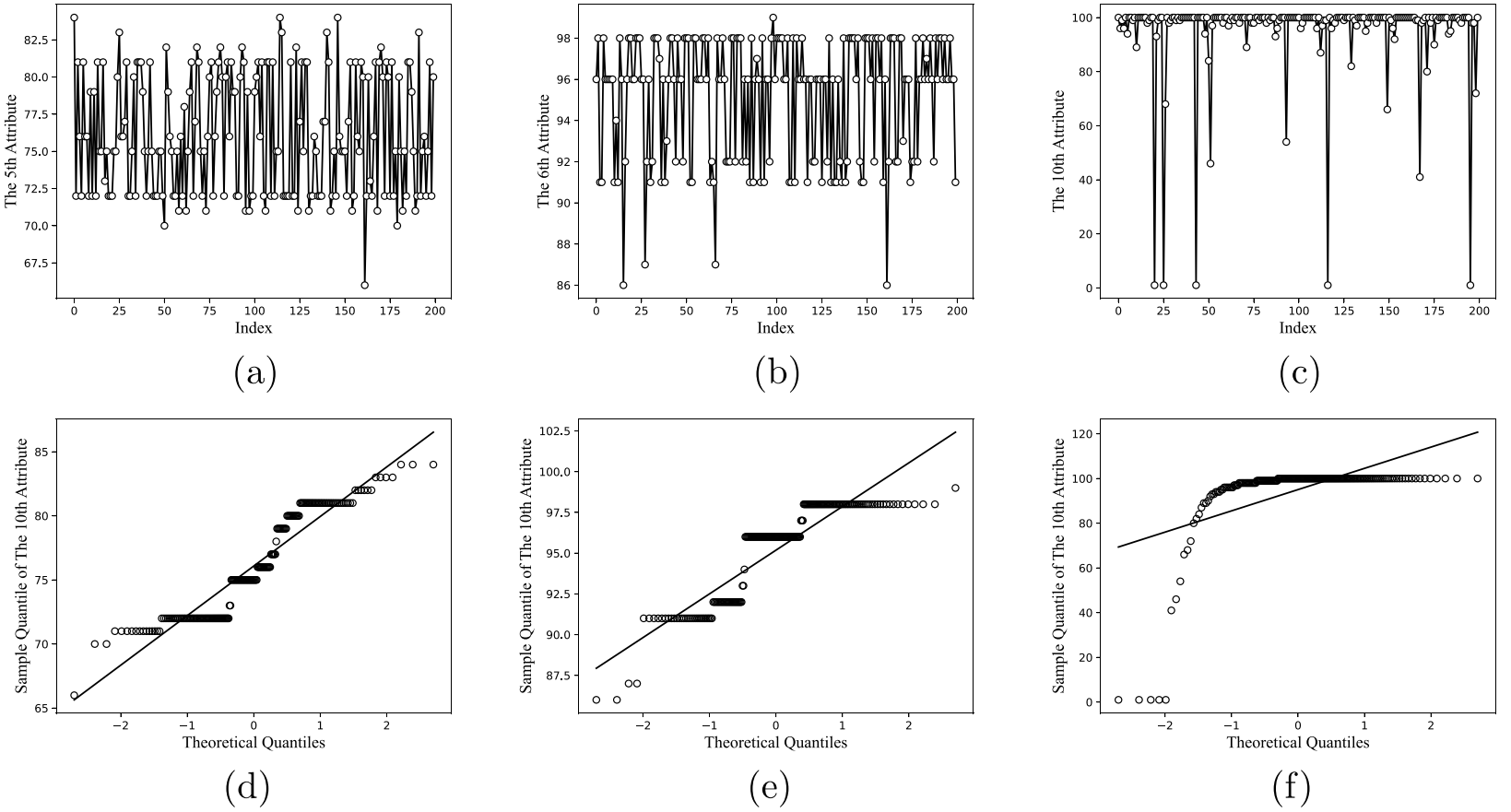

Figure 3. Plot of (a) the 5th, (b) 6th, and (c) 10th attribute of 200 pieces of randomly picked data; and (d) to (f): the normal QQ plots for these three attributes, respectively.

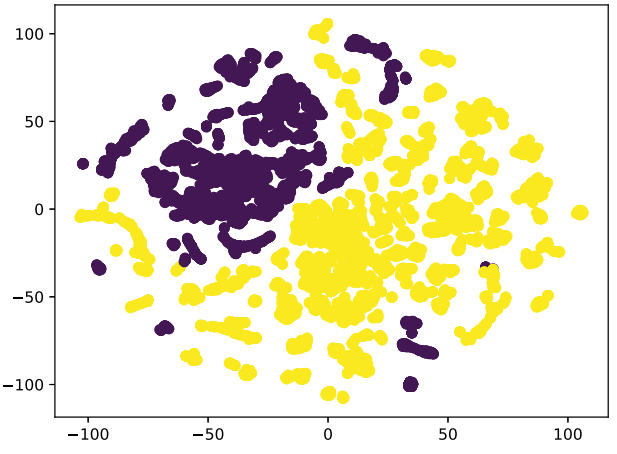

Figure 4. Two-dimensional scatter plot of 12,000 OOC samples.

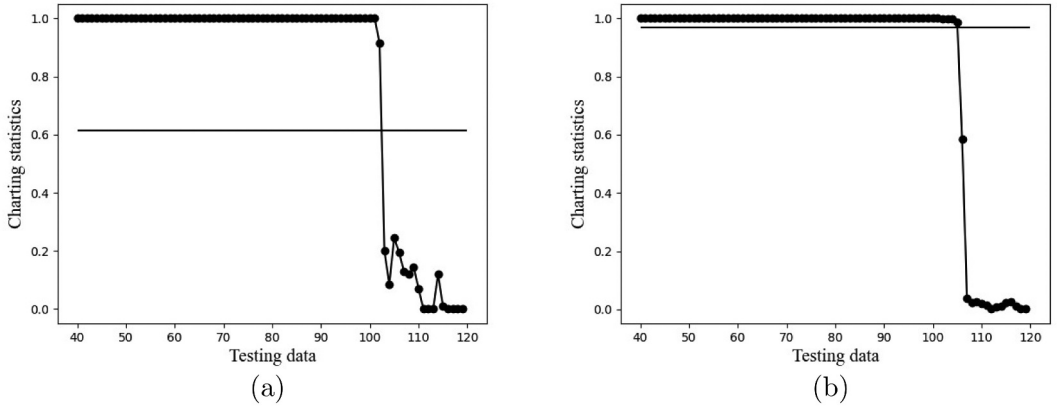

Figure 5. The RFEWMA chart (a) and the SVMEWMA chart (b) for monitoring the hard disk data. A change point occurs at the 100^{th} point.

forest as 200, the number of selected features as 3, and λ in the MEWMA sequence as 0.1. The control limit L is obtained by Monte Carlo simulation to ensure that the value of ARL_0 is 200 and is found to be 0.615. The final monitoring result is shown in Figure 5(a). The RFEWMA scheme gives an alarm at the 104^{th} point, that is, ARL_1 is 4, which proves that our proposed scheme can effectively solve the monitoring problem in this process. We use the same method to build the SVMEWMA scheme, and the final result is shown in Figure 5(b). The control limit L for the chart is 0.968 and the scheme gives an alarm at the 107^{th} point. The above comparison result also proves the high efficiency of our scheme.

6 Conclusion and future work

In this paper, the RFEWMA scheme based on the random forest model is proposed, making full use of historical data and the information therein to realize effective monitoring of complex processes of different natures. Through the numerical study of various situations and comparison with other SPM schemes, it is established that the proposed RFEWMA scheme has better monitoring performance, especially for the monitoring of high-dimensional and the data which may tend to belong to different classes under an OOC set-up. We apply our control scheme to the monitoring problem of

the real example and get the ARL_1 of 4, which proves that our control scheme can provide an effective method for the monitoring of such a process.

Although the content related to the RFEWMA control scheme has been discussed in detail, some problems need to be further studied. When monitoring the location shifts, our proposed SPM scheme mainly considers that the dimensions of the actual change are the same as those of the historical shifts. It is also essential to detect more location shift types and scale shifts. These issues not considered in this paper need to be studied in the future, and the RFEWMA scheme needs to be improved to realize the monitoring of various other types of shifts. In our study on the random forest parameters, only the independent effects of two main parameters are considered. Future works can deliberate the interaction effects of parameters and add the study to the effects of other parameters on the monitoring effect.

Acknowledgements

The authors sincerely acknowledge the efforts of the Editor, the Associate Editor and two anonymous referees that have resulted in significant improments of this paper. This work was supported by the National Key R&D Program of China [2021YFA1000101; 2021YFA1000102; 2022YFA1003801], National Natural Science Foundation of China [12071144; 71931004; 11771145], Basic Research Project of Shanghai Science and Technology Commission (22JC1400800).

Disclosure statement

Notes on contributors

Mingze Sun is a master's student at Tsinghua University. He received a B.S. degree under the supervision of Professor Dongdong Xiang from East China Normal University, Shanghai, China, in 2022.

Lei Qian is a postgraduate candidate in Peking University. He received B.S. in statistics under the supervision of Professor Dongdong Xiang from East China Normal University, Shanghai, China, in 2023.

Amitava Mukherjee, Professor of Production, Operations and Decision Sciences Area, XLRI-Xavier School of Management, Jamshedpur, India. His main research interest is Statistical Process Control.

Dongdong Xiang, Professor of School of statistics at East China Normal University. His main research interests are Statistical Process Control, Large-scale Multiple Tests and Machine Learning.
No potential conflict of interest was reported by the author(s).

References

Abbasi, S. A., Miller, A., & Riaz, M. (2013). Nonparametric progressive mean control chart for monitoring process target. *Quality and Reliability Engineering International*, *29*(7), 1069–1080. https://doi.org/10.1002/qre.1458

Alshraideh, H., Del Castillo, E., & Del Val, A. G. (2020). Process control via random forest classification of profile signals: An application to a tapping process. *Journal of Manufacturing Processes*, *58*, 736–748. https://doi.org/10.1016/j.jmapro.2020.08.043

Bakir, S. T. (2004). A distribution-free Shewhart quality control chart based on signed-ranks. *Quality Engineering*, *16*(4), 613–623. https://doi.org/10.1081/QEN-120038022

Bakir, S. T., & Reynolds, M. R. (1979). Nonparametric procedure for process control based on within-group ranking. *Technometrics*, *21*(2), 175–183. https://doi.org/10.1080/00401706.1979.10489747

Bersimis, S., Psarakis, S., & Panaretos, J. (2007). Multivariate statistical process control charts: An overview. *Social Science Electronic Publishing*, *23*(5), 517–543. https://doi.org/10.1002/qre.829

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/BF00058655

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bush, H. M., Chongfuangprinya, P., Chen, V. C., Sukchotrat, T., & Kim, S. B. (2010). Nonparametric multivariate control charts based on a linkage ranking algorithm. *Quality and Reliability Engineering International*, *26*(7), 663–675. https://doi.org/10.1002/qre.1129

Chakraborti, S. (2004). Nonparametric (distribution-free) quality control charts. *Encyclopedia of Statistical Sciences*, 1–27. https://doi.org/10.1002/0471667196.ess7150

Chan, K. M., Chong, Z. L., & Mukherjee, A. (2023). Exponentially weighted moving average Lepage-type schemes based on the lower-order percentile of the run-length metrics and their use in monitoring time-occupancy in Google applications. *Quality Technology & Quantitative Management*, *20*(5), 1–24. https://doi.org/10.1080/16843703.2022.2132452

Chen, S., & Yu, J. (2019). Deep recurrent neural network-based residual control chart for autocorrelated processes. *Quality and Reliability Engineering International*, *4*(8), 2687–2708. https://doi.org/10.1002/qre.2551

Chen, N., Zi, X., & Zou, C. (2016). A distribution-free multivariate control chart. *Technometrics*, *58*(4), 448–459. https://doi.org/10.1080/00401706.2015.1049750

Dastoorian, R., & Wells, L. J. (2021). A hybrid off-line/on-line quality control approach for real-time monitoring of high-density datasets. *Journal of Intelligent Manufacturing*, *34*(2), 669–682. https://doi.org/10.1007/s10845-021-01818-8

Ding, N., He, Z., He, S., & Song, L. (2023). Real-time profile monitoring schemes considering covariates using Gaussian process via sensor data. *Quality Technology & Quantitative Management*, 1–19. https://doi.org/10.1080/16843703.2023.2165284

Graham, M. A., Chakraborti, S., & Human, S. W. (2011). A nonparametric EWMA sign chart for location based on individual measurements. *Quality Engineering*, *23*(3), 227–241. https://doi.org/10.1080/08982112.2011.575745

Han, D., & Tsung, F. (2006). A reference-free cuscore chart for dynamic mean change detection and a unified framework for charting performance comparison. *Publications of the American Statistical Association*, *101*(473), 368–386. https://doi.org/10.1198/016214505000000556

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(8), 832–844. https://doi.org/10.1109/34.709601

Hotelling, H. (1947). *Multivariate quality control-illustrated by the air testing of sample bombsights. Techniques of Statistical Analysis* (pp. 111–184). McGraw Hill.

Huang, C. H., Huang, D. H., & Lin, Y. K. (2022). A novel approach to predict network reliability for multistate networks by a deep neural network. *Quality Technology & Quantitative Management*, *19*(3), 362–378. https://doi.org/10.1080/16843703.2021.1992072

Huwang, L., Lin, L. W., & Yu, C. T. (2019). A spatial rank–based multivariate EWMA chart for monitoring process shape matrices. *Quality and Reliability Engineering International*, *35*(6), 1716–1734. https://doi.org/10.1002/qre.2471

Lee, P. H., Torng, C. C., Lin, C. H., & Chou, C. Y. (2022). Control chart pattern recognition using spectral clustering technique and support vector machine under gamma distribution. *Computers & Industrial Engineering*, *171*, 108437. https://doi.org/10.1016/j.cie.2022.108437

Lowry, C. A., Woodall, W. H., Champ, C. W., & Rigdon, S. E. (1992). A multivariate exponentially weighted moving average control chart. *Technometrics*, *34*(1), 46–53. https://doi.org/10.2307/1269551

Maboudou-Tchao, E., Harrison, C. W., & Sen, S. (2022a). A comparison study of penalized likelihood via regularization and support vector-based control charts. *Quality Technology & Quantitative Management*, *20*(2), 147–167. https://doi.org/10.1080/16843703.2022.2096198

Maboudou-Tchao, E., Harrison, C. W., & Sen, S. (2022b). A comparison study of penalized likelihood via regularization and support vector-based control charts *Quality Technology* &. *Quality Technology & Quantitative Management*, *20*(2), 147–167. https://doi.org/10.1080/16843703.2022.2096198

MacQueen, J. B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281–297). Statistics, University of California Press.

Mehmood, R., Lee, M. H., Ali, I., Riaz, M., & Hussain, S. (2020). Multivariate cumulative sum control chart and measure of process capability based on bivariate ranked set schemes. *Computers & Industrial Engineering*, *150*, 106891. https://doi.org/10.1016/j.cie.2020.106891

Mukherjee, A., & Marozzi, M. (2021). Nonparametric phase-ii control charts for monitoring high-dimensional processes with unknown parameters. *Journal of Quality Technology*, *54*(1), 44–64. https://doi.org/10.1080/00224065.2020.1805378

Qiu, P. (2013). *Introduction to statistical process control*. Chapman and Hall/CRC. https://doi.org/10.1201/b15016

Qiu, P., & Li, Z. (2011). Supplementary material: On nonparametric statistical process control of univariate processes. *Technometrics*, *53*(4), 390–405. https://doi.org/10.1198/TECH.2011.10005

Sabahno, H., Amiri, A., & Castagliola, P. (2020). A new adaptive control chart for the simultaneous monitoring of the mean and variability of multivariate normal processes. *Computers & Industrial Engineering*, *151*, 106524. https://doi.org/10.1016/j.cie.2020.106524

Tran, P. H., Ahmadi Nadi, A., Nguyen, T. H., Tran, K. D., & Tran, K. P. (2022). Application of machine learning in statistical process control charts: A survey and perspective. *Control Charts and Machine Learning for Anomaly Detection in Manufacturing*, 7–42. https://doi.org/10.1007/978-3-030-83819-5_2

Tulsyan, A., Garvin, C., & Ündey, C. (2018). Advances in industrial biopharmaceutical batch process monitoring: Machine-learning methods for small data problems. *Biotechnology & Bioengineering*, *115*(8), 1915–1924. https://doi.org/10.1002/bit.26605

Wang, F. K., Bizuneh, B., & Cheng, X. B. (2019). One-sided control chart based on support vector machines with differential evolution algorithm. *Quality and Reliability Engineering International*, *35*(6), 1634–1645. https://doi.org/10.1002/qre.2465

Weese, M., Martinez, W., Megahed, F. M., & Jones-Farmer, L. A. (2016). Statistical learning methods applied to process monitoring: An overview and perspective. *Journal of Quality*, *48*(1), 4–24. https://doi.org/10.1080/00224065.2016.11918148

Wolpert, D. H., & Macready, W. G. (1999). An efficient method to estimate bagging's generalization error. An efficient method to estimate bagging's generalization error. *Machine Learning*, *35*(1), 41–55. https://doi.org/10.1023/A:1007519102914

Woodall, W. H., & Ncube, M. M. (1985). Multivariate CUSUM quality control procedures. *Technometrics*, *27*(3), 285–292. https://doi.org/10.1080/00401706.1985.10488053

Wu, R., Wang, J., Zhang, D., & Wang, S. (2021). Identifying different types of urban land use dynamics using Point-of-interest (POI) and random forest algorithm: The case of Huizhou, China. *Cities*, *114*, 103202. https://doi.org/10.1016/j.cities.2021.103202

Xie, F., Castagliola, P., Li, Z., Sun, J., & Hu, X. (2022). One-sided adaptive truncated exponentially weighted moving average $\bar{X}$ Schemes for detecting process mean shifts. *Quality Technology & Quantitative Management*, *19*(5), 533–561. https://doi.org/10.1080/16843703.2022.2033404

Xie, X., & Qiu, P. (2022). *Machine learning control charts for monitoring serially correlated data*. Springer International Publishing. https://doi.org/10.1007/978-3-030-83819-5_6

Zhang, C., Tsung, F., & Zou, C. (2015). A general framework for monitoring complex processes with both in-control and out-of-control information. *Computers & Industrial Engineering*, *85*, 157–168. https://doi.org/10.1016/j.cie.2015.03.007

Zhou, M., Zhou, Q., & Geng, W. (2016). A new nonparametric control chart for monitoring variability. *Quality & Reliability Engineering International*, *32*(7), 2471–2479. https://doi.org/10.1002/qre.1949

Zou, C., & Qiu, P. (2009). Multivariate statistical process control using LASSO. *Journal of the American Statistical Association*, *104*(488), 1586–1596. https://doi.org/10.1198/jasa.2009.tm08128