International Journal of Production Research, 2013
Vol. 51, No. 15, 4704–4717, http://dx.doi.org/10.1080/00207543.2013.793423

Taylor & Francis
Taylor & Francis Group

A run-to-run control algorithm based on timely and delayed mixed-resolution information

Kaibo Wang* and Jing Lin

Department of Industrial Engineering, Tsinghua University, Beijing 100084, China

(*Received 22 February 2012; final version received 30 March 2013*)

In certain run-to-run (R2R) processes, timely accurate measurements are difficult to obtain due to slow laboratory measurement operations. Instead, only low-resolution categorical observations are observed online for important quality variables; continuous measurements for the same variables are provided after a specific amount of delay. Currently, most conventional R2R controllers cannot be applied if no continuous observations are available. It is therefore important to develop online algorithms for R2R process control based on mixed-resolution information that is partially timely and partially delayed. In this study, we take the lapping process in semiconductor manufacturing as an example and propose parameter estimation models with these mixed-resolution data for processes with the first-order autoregressive, AR(1), disturbance series. We also derive control strategies to generate recipes between production runs for better process control. The computational results of a performance evaluation show that the control performance of the proposed method is competitive compared to existing methods that are based on accurate measurements.

Keywords: run-to-run process control; statistical process adjustment; statistical process control

1. Introduction

Low-resolution categorical observations are frequently observed in manufacturing processes. An important reason is that in some processes, practical constraints, such as cost and instruments, make it impossible to collect timely quantitative measurements. Alternatively, qualitative observations are collected for quality assurance. Wang and Tsung (2007) studied a deep reactive ion etching (DRIE) process. In this process, products could be measured only by scanning electron microscopy (SEM), which results in a bottleneck in large-volume manufacturing. Instead, visual inspection, which produces quality-related readings of wafer trenches on a positive/normal/negative scale, can be used for quality control. Spanos and Chen (1997) studied a plasma etching process, in which the samples are classified as 'very rough', 'rough', 'smooth', 'very smooth', etc., based on the roughness of the etched sidewalls. Lu, Jeng, and Wang (2009) also emphasised that the analysis of a new data type, including categorical data, is an important research topic in nanotechnology.

Sometimes, although only categorical data is collected online, high-resolution continuous measurements can be obtained after a certain delay. In the DRIE process above, for example, precise values of the angle of the wafer trenches can also be measured by an SEM offline after a sufficient delay time. As another example, in a footwear manufacturing process, gluing is one important step that glues two piece of leather together. Factors such as heating time and pressure may affect the binding force. To adjust this process, the practitioners usually take samples from the process, and try to rip two pieces apart manually and classify the samples as weakly, strongly, or very strongly bound. Instead, the samples could be sent to the laboratory to measure the binding force using a tension gauge, which can give an accurate reading. In this case, the timely categorical and delayed accurate information could be used together. Thus, observations with two types of resolution (i.e. categorical data versus continuous data) would be attained in this case; these observations are called multi-resolution observations in this paper. Therefore, when multi-resolution data is available in a process, controlling the process status and the product quality based on this new type of information becomes necessary and important.

In this study, the lapping process in semiconductor manufacturing is taken as an example. A general wafer preparation process consists of slicing, lapping, chemical vapour deposition (CVD), and polishing. A lapping run can be described as the following steps: (a) wafer loading; (b) machine setup; (c) lapping; (d) wafer unloading; and (e) testing. Lapping is critical to forming quality characteristics for downstream processing, because it is the first step of the

*Corresponding author. Email: kbwang@tsinghua.edu.cn

mechanical treatment on the wafer surface after slicing. Thickness of lapped wafers is an important geometric quality parameter in the lapping step, which is largely dominated by lapping time and incoming thickness. To control the amount of removal and to achieve an ideal thickness, lapping time is usually adjusted between runs. More about this process is also introduced in Li, Wang, and Yeh (2013) and Lin and Wang (2012)

In a case in which wafer thickness before and after lapping are both known online, classical run-to-run controllers, such as an exponentially weighted moving average (EWMA) controller and its extensions (Ingolfsson and Sachs 1993; Tseng, Tsung, and Liu 2007), a double EWMA controller (Chen and Guo 2001) and other controllers (Fan et al. 2002; Fan 2005; He, Wang, and Jiang 2009; Jin and Tsung 2009; Jen, Jiang, and Wang 2011), can be applied to generate recipes and to guide setting adjustments between runs. However, in the lapping process, accurate thickness values must be measured in a special inspection room with the aid of an expensive testing machine, which is both costly and time-consuming. In industrial practice, a batch of wafers is moved together to the inspection room for intermediate testing until the whole batch finishes lapping. Since long breaks are not allowed between production runs, it is impossible to obtain accurate data immediately after each lapping run. Alternatively, a less expensive but less accurate machine is equipped to help classify lapped wafers into different categories, namely, very thin, thin, normal, and thick to very thick, immediately after the lapping operation. After some time, during intermediate testing, precise thickness values will be collected for these lapped wafers. Thus, mixed-resolution data can be obtained in this lapping process for quality control purposes.

To maintain a better thickness quality, a controller is necessary to help generate the optimal lapping time for each run. However, only categorical observations are available directly after each run, and a traditional EWMA controller is not applicable because it requires numerical data. Meanwhile, if delayed quantitative measurements are used for process adjustments, then the performance and robustness of an EWMA controller is not good (Good and Qin 2002, 2006; Chamness et al. 2001). Therefore, a new type of controller that can generate control actions based on timely categorical observations and delayed continuous observations is needed in such a process.

To develop a controller that works with mixed-resolution observations, there are two major critical challenges to address. First, a process model should be built and estimated using mixed-resolution data, and second, optimal control actions must be generated based on mixed-resolution data.

Different model-building methods using only categorical observations can be found in the existing literature. A cumulative logistic model is one of the most popular categorical data models. Spanos and Chen (1997) utilised this model in the study of an etching process and estimated unknown parameters through the maximum likelihood (ML) method. The nonlinear optimisation problem related to ML functions has been discussed in detail by Agresti (1990). McCullagh (1980) replaced the logit link function in the cumulative logistic model with other link functions, such as the probit, and thus successfully extended this model to a generalised linear model (GLM). Liu and Agresti (2005) investigated the choice between the logit and probit functions. Bayesian-based methods for parameter estimation can also be found in the literature. Chipman and Hamada (1996) proposed a Bayesian approach to estimating the parameters in the GLM using Gibbs sampling with the assumption that the categorical observations are uncorrelated. Girard and Parent (2001) extended this Bayesian GLM to cases with autocorrelated observations. Lawrence et al. (2008) studied parameter estimation issues with multivariate categorical outputs. All of the aforementioned estimation strategies assume that historical observations are already collected before model fitting, and they estimate parameters in an offline manner given all of the information. However, in a lapping process or in other R2R processes, products are produced batch by batch; data arrive gradually in a stream. Therefore, it is meaningful to develop an online parameter estimation method that incorporates categorical observations. An adjustment ML method and a Bayesian online estimation method were recently proposed by Lin and Wang (2011) to address this problem. Lin and Wang (2012) used a Bayesian framework to tackle the same problem. These investigators also studied the lapping process in wafer production and applied an adjustment strategy to this process, using parameters that were estimated with their online method.

Recently, research on process control using low-resolution information has received considerable attention. Spanos and Chen (1997) first demonstrated the feasibility of implementing process monitoring and control with qualitative characteristics. Wang and Tsung (2007) introduced a feedback controller in a semiconductor manufacturing process. Shang, Wang, and Tsung (2009) improved this controller by considering misclassification errors in which misclassification possibilities were used to compensate for process adjustment bias. Wang and Tsung (2010) studied recursive parameter estimation with categorical observations that were available and proposed a Bayesian Categorical Controller. The authors assumed that all of the cut-points were known and only studied the estimation of a linear process model. Lin and Wang (2011) constructed a new approach to estimating both the intercept and the cutoff parameters online and to adjusting the process each run. Nonetheless, all of the controllers above absorb only categorical data without considering how to combine the delayed accurate information to achieve more efficient process control and quality improvement. Therefore, a new algorithm that can estimate model parameters online and generate recipes for run-to-run control using mixed-resolution observations must be developed.

We mentioned above that the EWMA controller fails to work well with delayed accurate data. In fact, the performance and stability properties of EWMA controllers with such measurement delays have been discussed by several researchers. Good and Qin (2006) studied the stability region for both the single-input single-output (SISO) and multiple-input multiple-output (MIMO) EWMA controllers while handling different metrology delays. The robustness of double EWMA controllers was extensively investigated in Good and Qin (2002). A comparison between several run-to-run algorithms, including the EWMA with a measurement delay, has been conducted by Chamness et al. (2001). All of these investigations show that both the stability properties and the performance of EWMA controllers are worse when there is a delay in the measurement. Thus, researchers began to investigate and improve EWMA-type run-to-run controllers under a metrology delay. Jin and Tsung (2009) developed the Smith–EWMA run-to-run controller, introducing the Smith predictor, which was created specifically for time delay systems in control theory for EWMA controllers. A performance comparison with the EWMA and recursive-least-square controllers under the first-order autoregressive (AR(1)) and integrated moving average (IMA(1)) (1) disturbance conditions based on simulation is conducted; this comparison shows that the Smith–EWMA controller has better stability properties and also has a more satisfactory performance for process control. However, these algorithms still rely on continuous measurement for process adjustment.

In this research, a strategy for online parameter estimation and process adjustment based on mixed-resolution observations is proposed. The remainder of this paper is organised as follows. Section 2 illustrates the formulation of the model to be studied. Section 3 presents the method for online parameter estimation and process control using multi-resolution data. Section 4 studies and analyses the performance of the proposed method. Finally, Section 5 concludes this paper and discusses topics that are related to mixed-resolution data, which deserves further research.

2. Process modelling

In this section, we still use the lapping process for illustration. In many R2R processes, linear models can be used to characterise processes with continuous inputs and outputs (see, e.g., Del Castillo and Hurwitz 1997; Wang and Tsung 2007, 2008; Shang, Wang, and Tsung 2009; Wang and Tsung 2010). Othman et al. (2006) showed that a linear model between the removal rate and the controller factors is adequate for the lapping process. In this research, we also studied the lapping process via experimental design and found that it is adequate to represent the lapping process by the following equation:

$$y_t = a + bu_{t-1} + cx_t + d_t, \tag{1}$$

where y_t is the output thickness that is obtained at time t; u_{t-1} is the lapping time set at time $t-1$, which is a controllable factor in lapping; x_t is the incoming wafer thickness generated by the slicing stage, which can be observed but not changed; d_t represents the process disturbance; and a, b and c are the linear coefficients. Here, we assume that d_t obeys a first-order autoregressive model, i.e. an AR(1) model. Thus, d_t can be written as follows:

$$d_t = \rho d_{t-1} + \varepsilon_t, \tag{2}$$

where $\varepsilon_t \sim N(0, \sigma^2)$ and ρ is the autogressive coefficient. AR(1) is a general model that can be used to represent an autocorrelated series, and autocorrelation is also seen in the literature and many real applications. For example, Fan et al. (2002) developed a triple-EWMA controller for a process having an AR(1) disturbance series; Vanli et al. (2007) also employed an autoregressive disturbance series when doing model selection for run-to-run control; and the real lithography process in semiconductor manufacturing the authors studied has an autoregressive disturbance model. Our engineering experience also suggests that the AR(1) disturbance series is adequate for some real processes. Therefore, the AR(1) disturbance model is used in this work. If a different disturbance model is used, some of the following derivations presented in this work should be changed accordingly, but the general framework of incorporating mixed-resolution information for R2R control can still be applied. Without loss of generality, in this equation, d_0 is assumed to be equal to zero; therefore, we have $d_1 = \varepsilon_1 \sim N(0, \sigma^2)$.

The variable y_t in Equation (1) above should be measured on a continuous scale. We assume its metrology to be delayed for d steps, therefore y_t would remain unknown until step $t+d$. Therefore, y_t is treated as a latent variable at step t, and another categorical variable, Y_t, is assumed to be observable and linked with y_t by the following mapping function (Chipman and Hamada 1996; Girard and Parent 2001; Wang and Tsung 2007):

$$Y_t = j \Leftrightarrow \gamma_{j-1} < y_t \leq \gamma_j, \quad j = 1, \ldots, c,$$

where $\gamma = [\gamma_0, \gamma_1, \ldots, \gamma_{c-1}, \gamma_c]^T$ is a vector of cut-points against which samples are classified. For the case when y_t is an unbounded variable (i.e. no boundaries for the worst and best values of y_t), we assume that $\gamma_0 = -\infty$ and $\gamma_c = \infty$. Meanwhile, at step t, the wafer sample of step $t - d$ would be measured precisely; in other words, y_{t-d} can be obtained at that point. Thus, the mixed-resolution dataset observed at step t consists of an online categorical observation Y_t and an offline continuous observation y_{t-d}. That is, Y_t and y_{t-d} always become available at every time t for $t = 1, 2, 3, \ldots$.

In Equation (1), we assume that b and c are known and that a is unknown. This assumption results from the fact that a easily fluctuates with the temperature of the lapping pans and the machine setup when a new order arrives, while b and c are dominated by a physical mechanism and therefore are relatively stable. The wafer thickness before lapping, x_t, is to be measured accurately after slicing; thus, it is available in the lapping stage. The autogressive coefficient ρ in Equation (2) is also assumed to be unknown. The cut-points γ are neither known nor able to be measured directly. Therefore, γ must be estimated.

3. Online estimation and adjustment using categorical observations

At the end of each production run, the following tasks are performed sequentially: (a) measurement of samples to obtain categorical data and fetching of the delayed accurate data from the intermediate test in the inspection room; (b) updating of estimates of unknown parameters using categorical and accurate observations that are available at this run; and (c) recommendation of parameter settings for the next run. Task (a) is performed manually by operators. In the following sections, the treatment of tasks (b) and (c) is introduced.

3.1 *Online parameter estimation*

In this section, a recursive method to estimate the unknown parameters a, γ and ρ online is presented. This method is developed in a Bayesian framework, utilising Gibbs sampling to simulate the posterior distribution of the unknown parameters. For each run, whenever a new mixed-resolution observation dataset becomes available, the categorical and continuous observations in this dataset will be used to calculate the updated estimates of the parameters.

3.1.1 *Fully conditional distribution in Gibbs sampling*

In Bayesian theory, prior knowledge is assumed to be possessed. This knowledge serves as the prior distributions of unknown parameters with respect to estimations. It is reasonable to assume that $a \sim N(a_0, \sigma_a^2)$ and $\gamma \sim N(\gamma_0, \Sigma_{\gamma,0})$, where $\gamma_0 = (\gamma_{1,0}, \ldots, \gamma_{c-1,0})^T$, and that $\sum_{\gamma,0} = \sigma_{\gamma,0}^2 \cdot I$. Meanwhile, γ is restricted such that $\gamma_{\min} < \gamma_1 < \gamma_2 < \ldots < \gamma_{c-1} < \gamma_{\max}$. Here, $\gamma_{\min}$ and $\gamma_{\max}$ are constants. Because ρ always falls into the interval $(-1, 1)$, we assume that it obeys a truncated normal distribution, which is $\rho \sim N\left(\rho_0, \sigma_\rho^2\right) \cdot I(|\rho| < 1)$. We also assume that a, γ and ρ are independent of one another.

Based on the prior distributions and the new observations, the corresponding posterior distributions can be calculated with Bayes' rules directly or with simulation approaches. In this work, a classical MCMC method, Gibbs sampling, will be applied. For Gibbs sampling, it is critical to obtain the fully conditional distributions of the unknown parameters.

The conditional distributions to derive are as follows:

$$f(y_t|a, \gamma, \rho, Y_t, \ldots, Y_1, y_{t-d}, \ldots, y_1),$$

$$f(a|y_t, \gamma, \rho, Y_t, \ldots, Y_1, y_{t-d}, \ldots, y_1),$$

$$f\left(\gamma_j\middle|a, \gamma_{i\neq j}, \rho, y_t, Y_t, \ldots, Y_1, y_{t-d}, \ldots, y_1\right),$$

$$\text{and } f(\rho|y_t, a, \gamma, Y_t, \ldots, Y_1, y_{t-d}, \ldots, y_1).$$

For convenience, we omit the latter condition items and denote those above as $f(|\cdot)$ hereafter.

After Y_t is observed, we begin by investigating the fully conditional distribution of y_t. It is a special case when $t = 1$ because $d_1 = \varepsilon_1 \sim N(0, \sigma^2)$; thus, we calculate it first. It can be given as follows:

$$f(y_1|\cdot) \propto N\left(a + bu_0 + cx_1, \sigma^2\right) \cdot I\left(\gamma_{Y_1-1} < y_1 < \gamma_{Y_1}\right). \tag{3}$$

When $t \geq 2$, due to the autoregressive properties between the sample series, we have $d_t = \rho(\hat{y}_{t-1} - a - bu_{t-2} - cx_{t-1}) + \varepsilon_t$, where $\hat{y}_{t-1}$ is the estimate of y_{t-1} that is attained in the run $t-1$. Hence,

$$\begin{aligned} y_t &= a + bu_{t-1} + cx_t + \rho(\hat{y}_{t-1} - a - bu_{t-2} - cx_{t-1}) + \varepsilon_t \\ &= \rho\hat{y}_{t-1} + (1-\rho)a + b(u_{t-1} - \rho u_{t-2}) + c(x_t - \rho x_{t-1}) + \varepsilon_t \end{aligned}$$

Therefore, the conditional distribution of y_t becomes related to the observations in the foregoing steps and can be written as follows:

$$f(y_t|\cdot) \propto N\left(\rho\hat{y}_{t-1} + (1-\rho)a + b(u_{t-1} - \rho u_{t-2}) + c(x_t - \rho x_{t-1}), \sigma^2\right) \cdot I\left(\gamma_{Y_t-1} < y_t < \gamma_{Y_t}\right), \tag{4}$$

Based on Bayes' theorem, the fully conditional distribution of a can be written as follows:

$$\begin{aligned} f(a|\cdot) \propto N\left(\mu_{a,t-1}, \sigma^2_{a,t-1}\right) &\cdot N\left(\frac{y_t - \rho\hat{y}_{t-1} - b(u_{t-1} - \rho u_{t-2}) - c(x_t - \rho x_{t-1})}{(1-\rho)}, \frac{\sigma^2}{1-\rho}\right) \\ &\cdot N\left(\frac{y_{t-d} - \rho y_{t-1-d} - b(u_{t-1-d} - \rho u_{t-2-d}) - c(x_{t-d} - \rho x_{t-1-d})}{(1-\rho)}, \frac{\sigma^2}{1-\rho}\right), \end{aligned} \tag{5}$$

which is still a normal distribution. Equation **(5)** could be simplified to be $f(a|\cdot) \propto N\left(\mu_{a,t}, \sigma^2_{a,t}\right)$, which satisfies the following two recursive equations:

$$\begin{aligned} \mu_{a,t} = \Bigg(&\frac{\mu_{a,t-1}}{\sigma^2_{a,t-1}} + \frac{y_t - \rho\hat{y}_{t-1} - b(u_{t-1} - \rho u_{t-2}) - c(x_t - \rho x_{t-1})}{\sigma^2/(1-\rho)} \\ &+ \frac{y_{t-d} - \rho y_{t-1-d} - b(u_{t-1-d} - \rho u_{t-2-d}) - c(x_{t-d} - \rho x_{t-1-d})}{\sigma^2/(1-\rho)}\Bigg) \Bigg/ \left(\frac{1}{\sigma^2_{a,t-1}} + \frac{2(1-\rho)^2}{\sigma^2}\right), \end{aligned} \tag{6}$$

and

$$\sigma^2_{a,t} = 1 \Bigg/ \left(\frac{1}{\sigma^2_{a,t-1}} + \frac{2(1-\rho)^2}{\sigma^2}\right), \tag{7}$$

where $\mu_{a,0} = a_0$, and $\sigma^2_{a,0} = \sigma^2_a$.

Similarly, the fully conditional distribution of γ_j can be obtained as follows:

$$f(\gamma_j|\cdot) \propto \begin{cases} N(\mu_{\gamma_j,t-1}, \sigma^2_{\gamma_j,t-1}) \cdot I\left(\max\limits_{i \leq t-d}\{y_i|Y_i = j\} < \gamma_j < \min\limits_{i \leq t-d}\{y_t, y_i|Y_i = j+1\}\right), & j = Y_t - 1 \\ N(\mu_{\gamma_j,t-1}, \sigma^2_{\gamma_j,t-1}) \cdot I\left(\max\limits_{i \leq t-d}\{y_t, y_i|Y_i = j\} < \gamma_j < \min\limits_{i \leq t-d}\{y_i|Y_i = j+1\}\right), & j = Y_t \\ N(\mu_{\gamma_j,t-1}, \sigma^2_{\gamma_j,t-1}) \cdot I\left(\max\limits_{i \leq t-d}\{y_i|Y_i = j\} < \gamma_j < \min\limits_{i \leq t-d}\{y_i|Y_i = j+1\}\right), & o.w. \end{cases} \tag{8}$$

Lastly, we will derive the conditional distribution for the autoregressive coefficient ρ, which can be written as follows:

$$f(\rho|\cdot) \propto N\left(\mu_{\rho,t-1}, \sigma^2_{\rho,t-1}\right) \cdot N\left(\frac{y_t - a - bu_{t-1} - cx_t}{\hat{y}_{t-1} - a - bu_{t-2} - cx_{t-1}}, \left(\frac{\sigma}{\hat{y}_{t-1} - a - bu_{t-2} - cx_{t-1}}\right)^2\right)$$
$$\cdot N\left(\frac{y_{t-d} - a - bu_{t-1-d} - cx_{t-d}}{y_{t-1-d} - a - bu_{t-2-d} - cx_{t-1-d}}, \left(\frac{\sigma}{y_{t-1-d} - a - bu_{t-2-d} - cx_{t-1-d}}\right)^2\right) \cdot I(|\rho| < 1).$$

It is not difficult to verify that the above equation is a truncated normal distribution and therefore can be rewritten as $f(\rho|\cdot) \propto N\left(\mu_{\rho,t}, \sigma^2_{\rho,t}\right) \cdot I(|\rho| < 1)$, which leads to the following:

$$\mu_{\rho,t} = \left(\frac{\mu_{\rho,t-1}}{\sigma^2_{\rho,t-1}} + \frac{(y_t - a - bu_{t-1} - cx_t) \cdot (\hat{y}_{t-1} - a - bu_{t-2} - cx_{t-1})}{\sigma^2}\right.$$
$$\left. + \frac{(y_{t-d} - a - bu_{t-1-d} - cx_{t-d}) \cdot (y_{t-1-d} - a - bu_{t-2-d} - cx_{t-1-d})}{\sigma^2}\right)$$
$$\Big/ \left(\frac{1}{\sigma^2_{\rho,t-1}} + \frac{(\hat{y}_{t-1} - a - bu_{t-2} - cx_{t-1})^2}{\sigma^2} + \frac{(y_{t-1-d} - a - bu_{t-2-d} - cx_{t-1-d})^2}{\sigma^2}\right), \tag{9}$$

$$\sigma^2_{\rho,t} = 1 \Big/ \left(\frac{1}{\sigma^2_{\rho,t-1}} + \frac{(\hat{y}_{t-1} - a - bu_{t-2} - cx_{t-1})^2}{\sigma^2} + \frac{(y_{t-1-d} - a - bu_{t-2-d} - cx_{t-1-d})^2}{\sigma^2}\right), \tag{10}$$

where $\mu_{\rho,0} = \rho_0$, $\sigma^2_{\rho,0} = \sigma^2_\rho$.

Now we have finished calculating the fully conditional distributions of all of the unknown parameters. These calculations serve as preparation for Gibbs sampling.

3.1.2 *Online parameter estimation procedure*

In the following, a Bayesian online procedure for estimating the unknown parameters and for generating control actions via Gibbs sampling is proposed.

When a new mixed-resolution observation dataset $\{Y_t, y_{t-d}\}$ is collected, the Gibbs sampling procedure begins to sample y_t, a, $\gamma_1, \ldots, \gamma_{c-1}$ and ρ repeatedly a sufficiently large number of times. The updated parameter values in the posterior distributions can be obtained by calculating the sample mean and variance of a, $\gamma_1, \ldots, \gamma_{c-1}$ and ρ with the initial samples removed. The sampling process for each run of the online estimation is outlined as follows:

Step 1: Sample one y_t from Equation (3) or (4);

Step 2: Using the posterior distribution of a, $N(\mu_{a,t-1}, \sigma^2_{a,t-1})$ estimated in the previous run as the prior and y_t sampled from Step 1, calculate the conditional distribution of a for run t using Equations (6) and (7);

Step 3: Sample one a from its conditional distribution obtained from Step 2;

Step 4: Using the posterior distribution of γ_j, which is

$$N\left(\mu_{\gamma_j,t-1}, \sigma^2_{\gamma_j,t-1}\right) \cdot I\left(\max_{i \le t-d}\{y_i | Y_i = j\} < \gamma_j < \min_{i \le t-d}\{y_i | Y_i = j+1\}\right)$$

estimated in the previous run as the prior, y_t sampled from Step 1 and a from Step 3, calculate the conditional distribution of γ_j for run t in Equation (8);

Step 5: Sample $\gamma_1, \ldots, \gamma_{c-1}$ in order from their respective conditional distributions, one element at a time.

Step 6: Using the posterior distribution of ρ, $N\left(\mu_{\rho,t-1}, \sigma^2_{\rho,t-1}\right) \cdot I(|\rho| < 1)$, which was estimated in the previous run as the prior, y_t sampled from Step 1, a from Step 3 and γ from Step 5, calculate the conditional distribution of ρ for run t using Equations (9) and (10);

Step 7: Sample one ρ from its conditional distribution obtained from Step 6;
Step 8: Using the newly sampled a, γ and ρ, update the conditional distribution of y_t, and return to Step 1;
Step 9: Repeat Steps 1–8 a sufficiently large number of times;
Step 10: Calculate the sample mean and variance of a, γ_j and ρ, which will serve as the updated parameter values in the posterior distribution $N\left(\mu_{a,t}, \sigma^2_{a,t}\right)$, $N\left(\mu_{\gamma_j,t}, \sigma^2_{\gamma_j,t}\right) \cdot I\left(\max_{i\le t-d}\{y_i|Y_i=j\}<\gamma_j<\min_{i\le t-d}\{y_i|Y_i=j+1\}\right)$ and $N\left(\mu_{\rho,t}, \sigma^2_{\rho,t}\right) \cdot I(|\rho|<1)$. Next, continue to produce the next run and perform an inspection to collect new observations. Next, repeat Steps 1–10.

3.2 *Run-to-run process adjustment*

To control the process, maintaining it on target, and to compensate for any initial bias, a recipe for each run is generated under a specific criterion to minimise the process variability. Denote the target of the process (1) as T. We define the following quadratic loss function conditioning on all of the historical information as the objective of the process adjustment at Step t:

$$L = E\left[(y_{t+1} - T)^2 | F_t\right],$$

where $F_t = \{Y_t, \ldots, Y_1, y_{t-d}, \ldots, y_1, u_{t-1}, \ldots, u_0, x_t, \ldots, x_1\}$.

Equation (1) shows that $y_{t+1} = a + bu_t + cx_{t+1} + d_{t+1}$, and Equation (2) shows that $d_{t+1} = \rho d_t + \varepsilon_{t+1}$; thus, it follows that

$$L = E\left[(\rho y_t + (1-\rho)a + b(u_t - \rho u_{t-1}) + c(x_{t+1} - \rho x_t) + \varepsilon_{t+1} - T)^2 | F_t\right].$$

Considering $E(\varepsilon^2_{t+1}) = \sigma^2$, we have the following:

$$L = (\rho y_t + (1-\rho)a - b\rho u_{t-1} + c(x_{t+1} - \rho x_t) - T)^2 + b^2 u_t^2$$

$$+2(\rho y_t + (1-\rho)a - b\rho u_{t-1} + c(x_{t+1} - \rho x_t) - T)bu_t + \sigma^2.$$

Taking the partial derivative of the above equation with respect to u_t as zero and replacing the unknown parameter with its estimate leads to the optimal control action

$$u_t = \frac{T - \rho^{(t)}\hat{y}_t - (1-\rho^{(t)})a^{(t)} + b\rho^{(t)}u_{t-1} - c(x_{t+1} - \rho x_t)}{b}, \tag{11}$$

where $\rho^{(t)}$ is the estimate of ρ at Step t, $a^{(t)}$ is the estimate of a at Step t, and $\hat{y}_t$ is the estimate of y_t at Step t.

It should be noted that when the parameter estimations are already quite accurate, there could be no need to update the model parameters. In these cases, only a process adjustment is needed. At this time, because the parameters are no longer updated, the estimation $\hat{y}_t$ would not be obtained, either. Hence, the optimal recipe for the next run could not be generated by Equation (11). Instead, it would be calculated with the following formula:

$$u_t = \frac{T - \hat{\rho}E(y_t|F_t) - (1-\hat{\rho})\hat{a} + b\hat{\rho}u_{t-1} - c(x_{t+1} - \rho x_t)}{b}, \tag{12}$$

where $\hat{\rho}, \hat{a}$ are the final estimations, respectively, for ρ and a, before the updating ceases. Taking the treatment in Wang and Tsung (2007) in this equation, $E(y_t|F_t)$ can be written as follows:

$$E(y_t|F_t) = \frac{1}{2}(\hat{\gamma}_{Y_t-1} + \hat{\gamma}_{Y_t}), \tag{13}$$

where $\hat{\gamma}_j, j = 1, \ldots, c$ is the final estimate for the cutoff parameter γ_j. Replace $E(y_t|F_t)$ in Equation (12) with Equation (13), and we obtain the following:

$$u_t = \frac{T - \frac{\hat{\rho}}{2}(\hat{\gamma}_{Y_t-1} + \hat{\gamma}_{Y_t}) - (1-\hat{\rho})\hat{a} + b\hat{\rho}u_{t-1} - c(x_{t+1} - \rho x_t)}{b}. \tag{14}$$

Therefore, to adjust the process run-to-run, we should apply Equation (11) to generate the control action during the estimation of the model parameters, utilising Equation (14) instead when the whole estimation procedure ends.

4. Performance studies

In this section, we investigate the performance of the proposed method and compare it with existing methods for either estimation or process control. For all of the cases that are to be studied below, the true model is set to be the same as the model in Lin and Wang (2011). In other words, the target process follows Equation (1) with $a = 60$, $b = 2$, $c = 0.1$, and the disturbances follow Equation (2) with $\rho = 0.6$ and $\sigma = 3$. The process target T equals 400, and four cut-points,

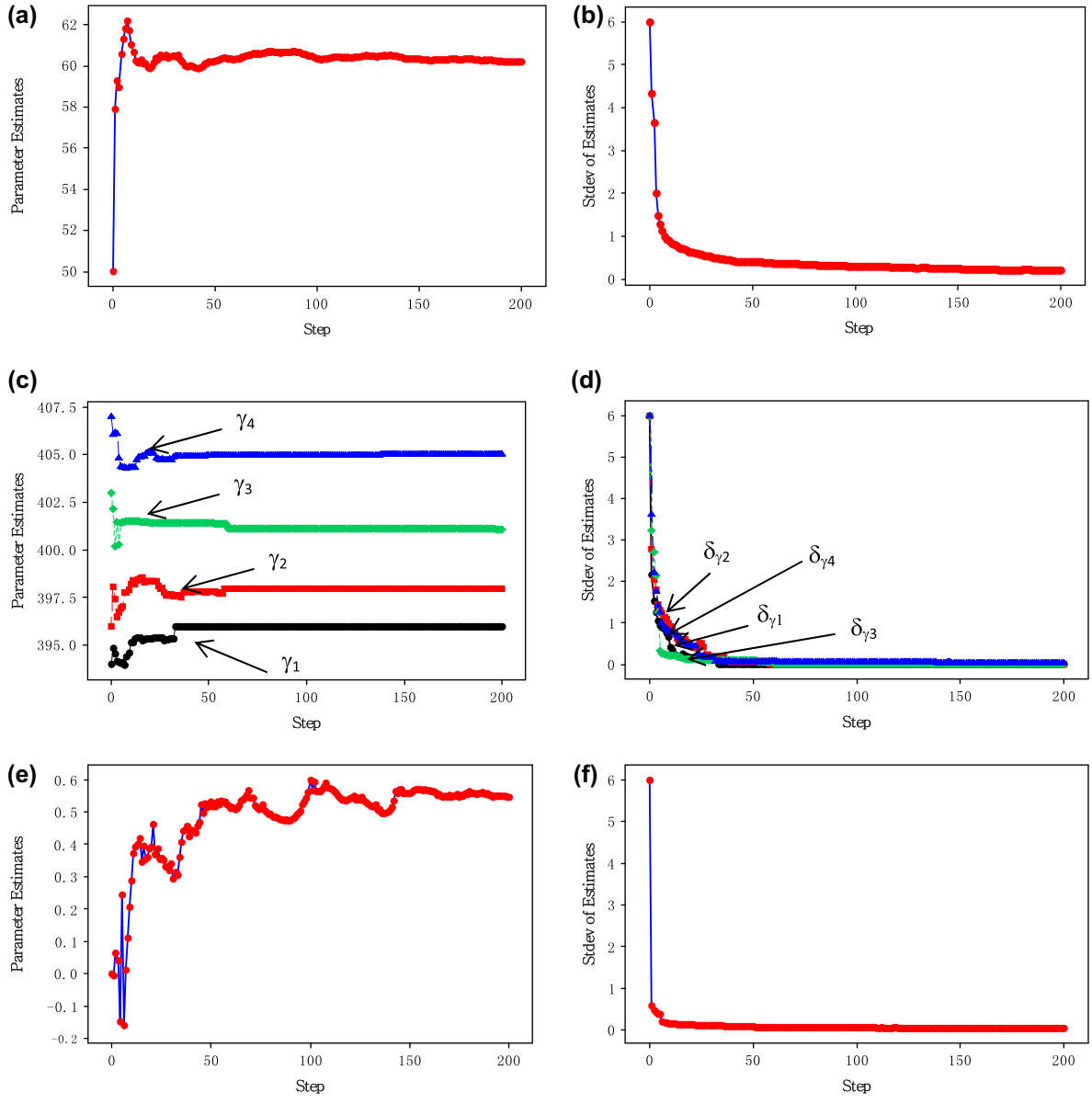

Figure 1. Trajectories of the estimated parameters with two-step delays. (a) The mean of a; (b) The standard deviation (SD) of a; (c) The mean vector of γ; (d) The SD of γ; (e) The mean vector of ρ; (f) The SD of ρ.

396, 398, 401 and 405, are used to classify the output y_t into five mutually exclusive categories, which are $\gamma = [396, 398, 401, 405]$.

4.1 *Parameter estimation and process adjustment performance*

In this calculation, we assume that the prior mean of a, γ and ρ are 50, $[394, 396, 403, 407]^T$ and 0, respectively, and their standard deviations are all 6, to investigate the issues that are caused by an initial bias. Additionally, the cut-points γ are restricted such that $392 < \gamma_1 < \gamma_2 < \gamma_3 < \gamma_4 < 409$. The Gibbs sampling is set to be repeated 10,000 times whenever a new observation is generated, and the last 5000 samples are used to calculate the marginal distribution of each unknown parameter. We simulate 200 categorical observations for each process.

Figures 1–3 show the trajectories of the means and standard deviations of the estimated a, γ and ρ with various delays (two steps, five steps and 10 steps). We can see from Figures 1–3 (a), (c) and (e) that the estimated parameters approach their true values gradually as categorical observations are collected run by run. An oscillation could exist in the early stage, because at the beginning, the samples are few and the information contained is comparatively rough.

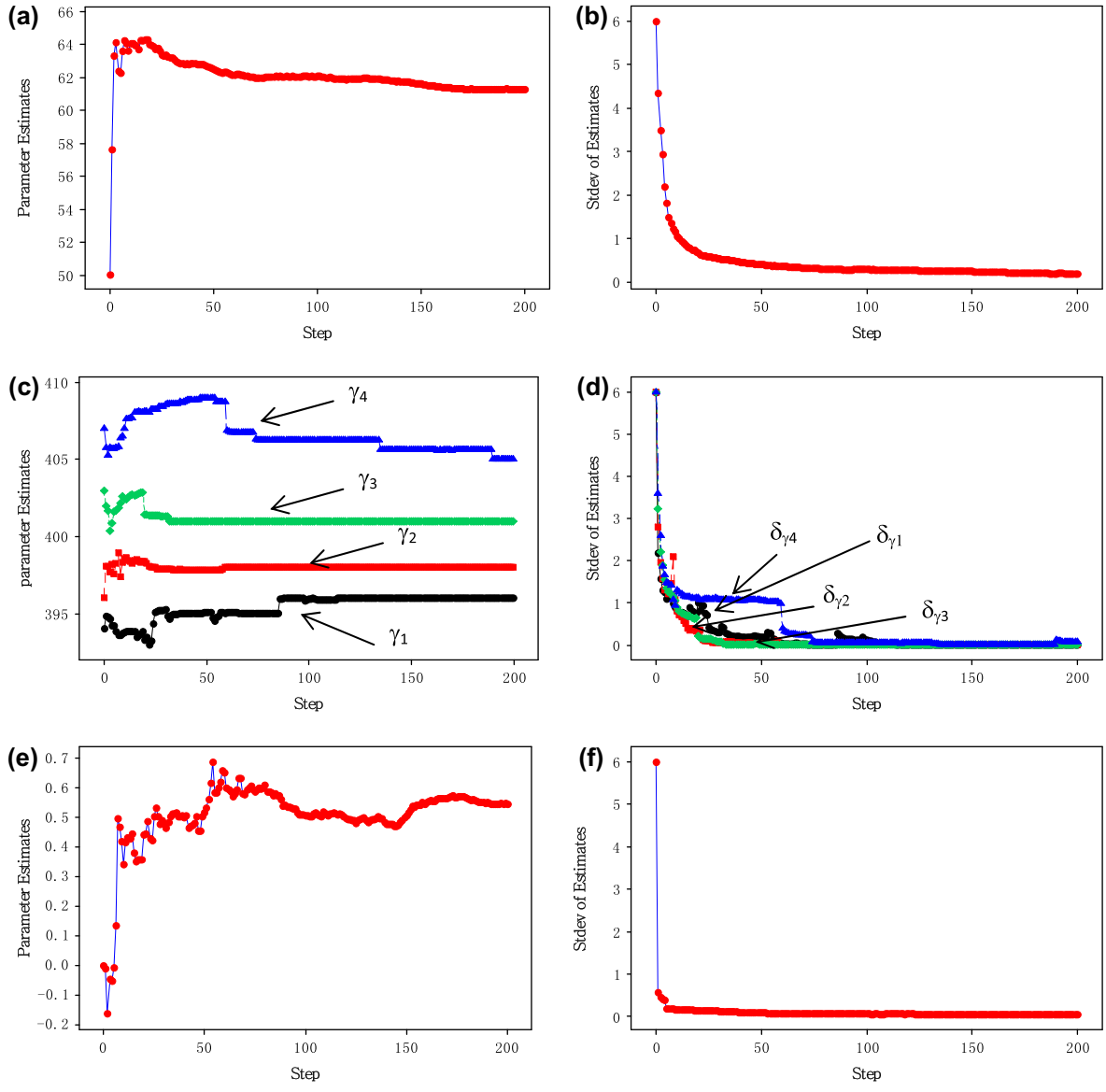

Figure 2. Trajectories of estimated parameters with five-step delays. (a) The mean of a; (b) The SD of a; (c) The mean vector of γ; (d) The SD of γ; (e) The mean vector of ρ; (f) The SD of ρ.

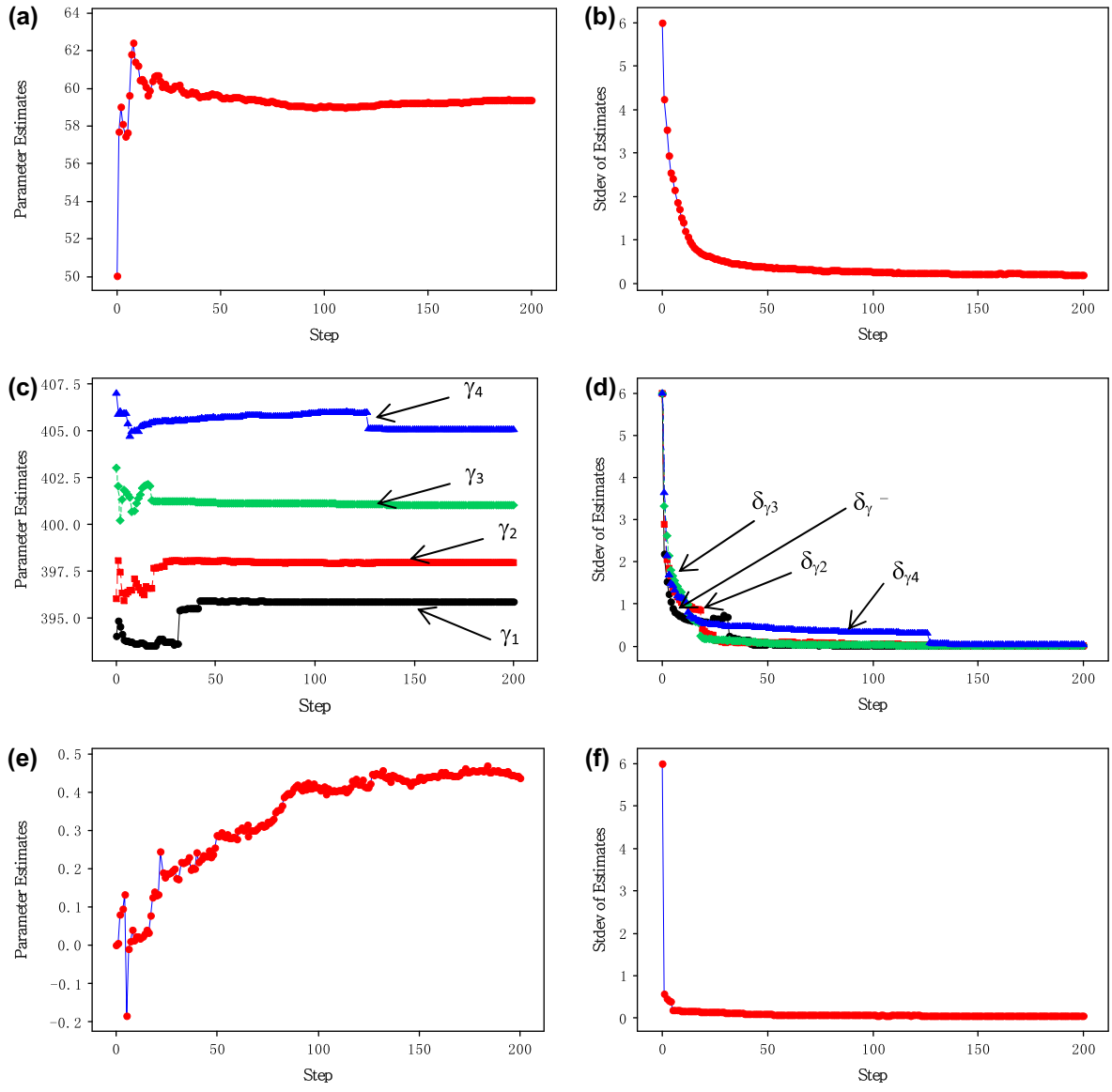

Figure 3. Trajectories of estimated parameters with 10-step delays. (a) The mean of a; (b) The SD of a; (c) The mean vector of γ; (d) The SD of γ; (e) The mean vector of ρ; (f) The SD of ρ.

Nevertheless, after approximately 40 steps, the estimates already become notably close to their true values. From Figures 1–3 (b), (d) and (f), it is clearly seen that the standard deviations of the estimated parameters decrease gradually until reaching certain small values, which show that the online algorithm could provide more and more accurate estimates with a small variance using a continuous mixed-resolution data stream.

The sequences of controlled and uncontrolled outputs y_t with different delay times are shown in Figure 4. Figure 4 (a), (b) and (c) suggest that, for all of the delay cases, the uncontrolled outputs have deviated from the target of 400 to a large extent, whereas the controlled output is maintained almost on target. Therefore, we can conclude that the proposed Bayesian method is quite effective in process output control using mixed-resolution observations.

4.2 *Performance comparison*

Next, we will compare the performance of our proposed approach with that of other methods for handling online parameter estimation or process adjustment.

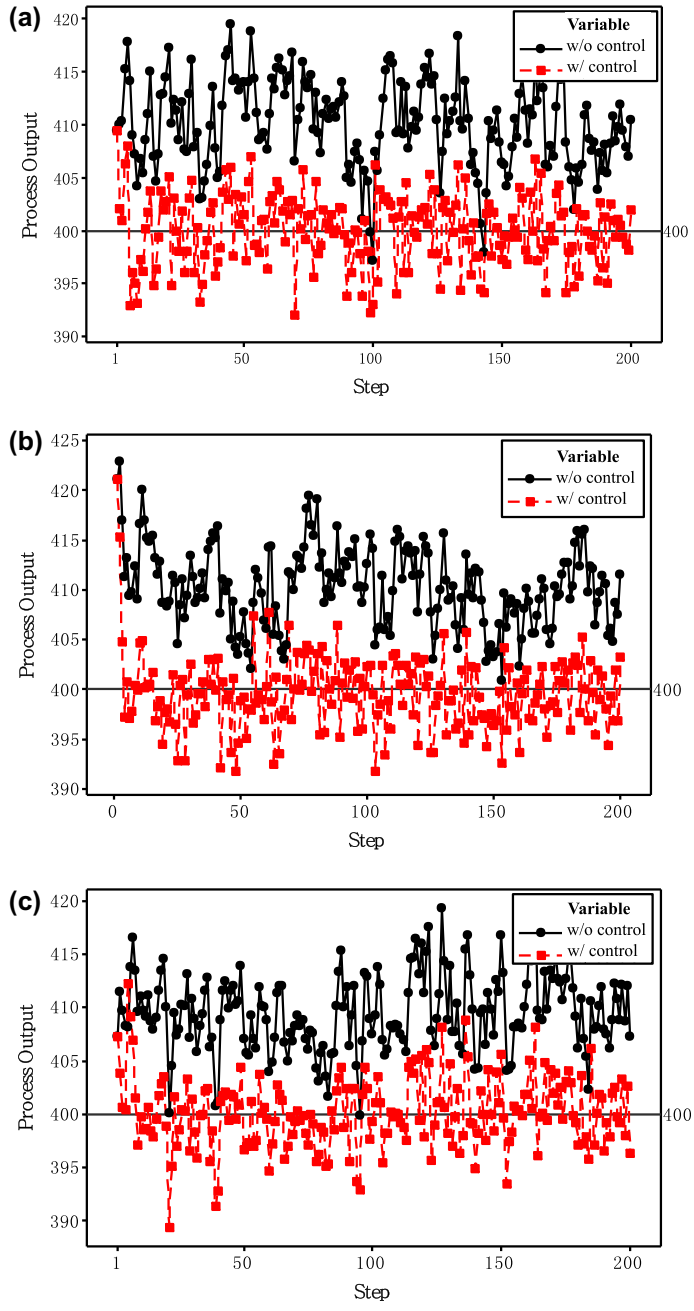

Figure 4. Trajectories of process outputs with a controller versus without a controller. (a) Two-step delay; (b) Five-step delay; (c) Ten-step delay.

With respect to parameter estimation performance, our online estimation method will be compared with the Bayesian estimation method, which applies the Bayes' theorem directly to calculate the posterior distributions of the parameters that are based on online quantitative data and on the same method based on delayed quantitative data.

For the continuous data in the studied mixed-resolution and the delayed continuous scenarios, both are assumed to have 10-step delay. The simulation is repeated 100 times, and the average MSE of all of the parameter estimates is calculated at each step to show their estimation accuracy. Because the delayed precise observations would not be available before the tenth step, and the estimates are quite accurate after the one-hundredth step, only the MSE for between the tenth and the one-hundredth steps will be calculated in this study. We will study different initial value cases, and for each case, the prior distributions of the unknown parameters in all of the three approaches will be set to be equal to the

same initial values. The scenario of having online precision data is studied solely for the purpose of performance investigation. The results are shown in Figure 5.

It is clearly seen from Figure 5 that under both of the two settings of initial values, our proposed method outperforms the Bayesian method using timely numerical data and outperforms the same method using delayed numerical data with smaller average MSE values. Meanwhile, the method using timely data performs better than the method using delayed data, which can be easily explained because delayed information naturally leads to comparatively less efficiency.

The EWMA controller is widely applied in industrial engineering, which utilises continuous data to generate recipes run by run. To address a measurement delay, Jin and Tsung (2009) developed a specific Smith–EWMA control algorithm. To investigate the performance of the method that was proposed in this paper, the EWMA controller and the Smith–EWMA controller are also set up to control the same process. These two controllers are both operated with the parameter λ, which could influence the control performance. Therefore, we study here the cases that have three different λ values, which are 0.2, 0.4 and 0.6. Two hundred lapping runs are simulated in one simulation, and the simulation is repeated 100 times. The initial values of the parameter estimates are all assumed to be $a^{(0)} = 50, \gamma^{(0)} = [394, 396, 403, 407], \rho^{(0)} = 0$. The MSE of the output is calculated at each step to show the control accuracy, which is shown in Figure 6.

Figure 6 shows the MSE at 200 steps. We can see that in the first 40 steps, the EWMA controller and the Smith–EWMA controller perform better than the proposed method, with smaller MSE values and variances being observed. However, after approximately step 40, our R2R adjustment method begins to reach the EWMA controller with MSE values of the same level and outperforms the Smith–EWMA controller with a much smaller MSE. It is also found that the standard deviation of the MSE for our method is maintained at approximately three after step 40, which equals σ of the white noise item ε_t in the disturbance d_t. Therefore, we can conclude that our control method could effectively compensate for the deviations and fluctuations that were induced by the initial bias and the autocorrelation between the outputs.

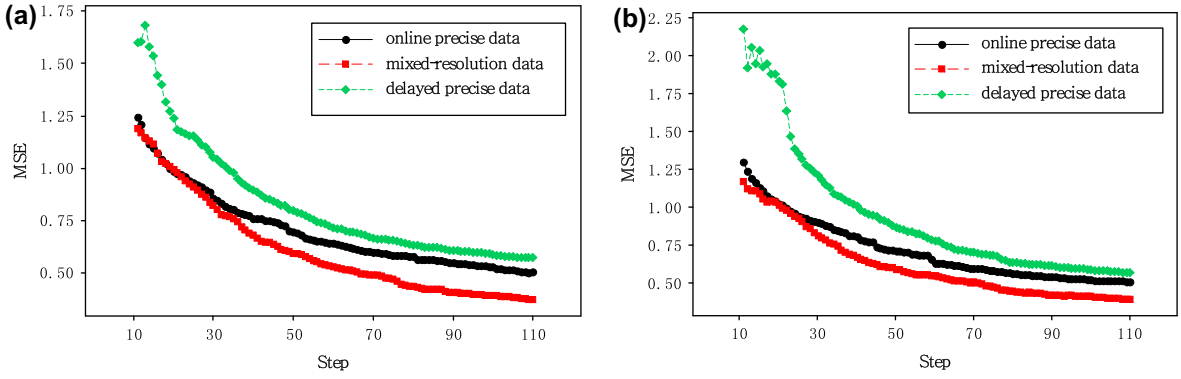

Figure 5. Parameter estimation performance comparison with different initial values. (a) $a^{(0)} = 50, \gamma^{(0)} = [394, 396, 403, 407], \rho^{(0)} = 0$; (b) $a^{(0)} = 45, \gamma^{(0)} = [393, 395, 404, 408], \rho^{(0)} = -0.2$.

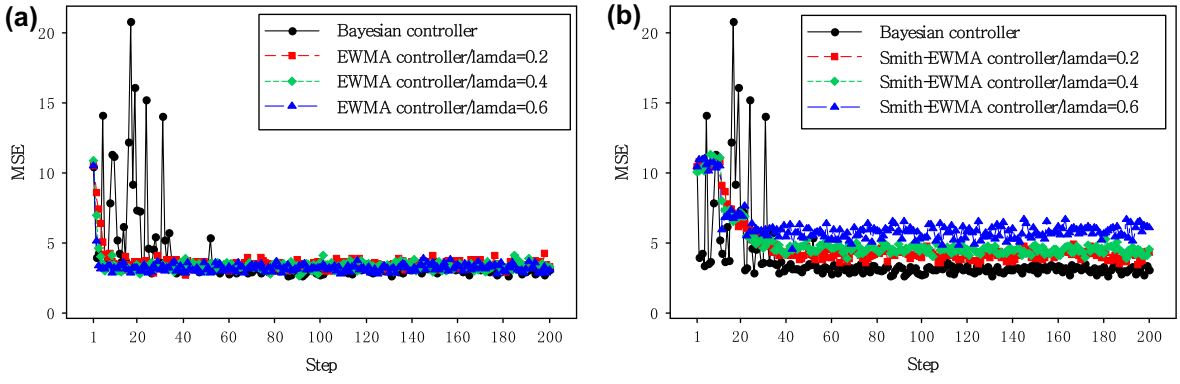

Figure 6. Parameter estimation performance comparison with different initial values.

The above study demonstrates the efficiency of the proposed method in calibrating an initial bias in the unknown parameters and in providing accurate estimates online based on mixed-resolution observation time series. This study also proves that the proposed algorithm is effective for controlling the process using mixed-resolution data.

5. Conclusions

It is a common practice in some manufacturing processes that timely categorical observations and delayed continuous information are available for R2R process adjustment. This paper investigated the online estimation and control of a run-to-run process with AR(1) disturbances when mixed-resolution observations are available. A Bayesian method, which utilises Gibbs sampling to estimate the posterior distributions, was proposed to update the model parameters when mixed-resolution data were collected sequentially from the process. A control algorithm that functioned based on the mixed-resolution observations was also proposed to adjust the process output to be on target.

Simulation studies showed that when an initial bias existed, the proposed method could move parameter estimates toward their respective true values quickly. The proposed scheme, together with the control algorithm, was proved to be effective in controlling processes that had an initial bias.

This paper used a simple model with AR(1) noises to characterise a lapping process. When the disturbance series became more complicated, for example, following a general autoregressive integrated moving average (ARIMA) time series, the statistical process adjustment algorithm had to be modified accordingly. In addition, in this study, we assume that the cut-points are fixed. For future work, we may investigate the estimation of the cut-points as a random variable, i.e. when misclassification exists and human factors are involved. Finally, this paper discussed this approach based on only simulation results. Therefore its performance in controlling real processes deserves more in-depth study and could be the focus of future research.

Acknowledgement

We greatly thank the editor and the anonymous referees for their very helpful comments which have significantly assisted us in improving this manuscript. This work was supported by National Natural Science Foundation of China (NSFC) under grants No. 70802034 and No. 71072012.

References

Agresti, A. 1990. *Categorical Data Analysis*. New York: Wiley.

Chamness, K., G. Cherry, R. Good, and S. J. Qin. 2001. “A Comparison of Run to Run Control Algorithms for the CMP with Metrology Delays”. AEC/APC XIII Symposium, 1-4.

Chen, A., and R. S. Guo. 2001. “Age-based Double EWMA Controller and its Application to CMP Processes.” *IEEE Transactions on Semiconductor Manufacturing* 14 (1): 11–19.

Chipman, H., and M. Hamada. 1996. “Bayesian Analysis of Ordered Categorical Data from Industrial Experiments.” *Technometrics* 38 (1): 1–10.

Del Castillo, E., and A. M. Hurwitz. 1997. “Run-to-run Process Control: Literature Review and Extensions.” *Journal of Quality Technology* 29 (2): 184–196.

Fan, S. K. S. 2005. “Multiple-input Single-output (MISO) Ridge-optimizing Quality Controller for Semiconductor Manufacturing Processes.” *International Journal of Production Research* 43 (22): 4745–4770.

Fan, S. K. S., B. C. Jiang, C. H. Jen, and C. C. Wang. 2002. “SISO Run-to-run Feedback Controller Using Triple EWMA Smoothing for Semiconductor Manufacturing Processes.” *International Journal of Production Research* 40 (13): 3093–3120.

Girard, P., and E. Parent. 2001. “Bayesian Analysis of Autocorrelated Ordered Categorical Data for Industrial Quality Monitoring.” *Technometrics* 43 (2): 180–191.

Good, R., and S. J. Qin. 2002 “Stability Analysis of Double EWMA Run-to-run Control with Metrology Delay”. In: *Proceedings of the American Control Conference, 2002,* 3: 2156–2161.

Good, R. P., and S. J. Qin. 2006. “On the Stability of MIMO EWMA Run-to-run Controllers with Metrology Delay.” *IEEE Transactions on Semiconductor Manufacturing* 19 (1): 78–86.

He, F., K. Wang, and W. Jiang. 2009. “A General Harmonic Rule Controller for Run-to-run Process Control.” *IEEE Transactions on Semiconductor Manufacturing* 22 (2): 232–244.

Ingolfsson, A., and E. Sachs. 1993. “Stability and Sensitivity of an EWMA Controller.” *Journal of Quality Technology* 25 (4): 271–287.

Jen, C. H., B. C. Jiang, and C. C. Wang. 2011. “Integration of Run-to-run Control Schemes and On-line Experiment to Deal with the Changes in Semiconducting Dynamic Processes.” *International Journal of Production Research* 49 (19): 5657–5678.

Jin, M., and F. Tsung. 2009. "Smith-EWMA Run-to-run Control Schemes for a Process with Measurement Delay." *IIE Transactions* 41 (4): 346–358.

Lawrence, E., D. Bingham, C. Liu, and V. N. Nair. 2008. "Bayesian Inference for Multivariate Ordinal Data Using Parameter Expansion." *Technometrics* 50 (2): 182–191.

Li, B., K. Wang, and A. Yeh. 2013. "Monitoring Covariance Matrix via Penalized Likelihood Estimation". *IIE Transactions* 45 (2): 132–146.

Lin, J., and K. Wang. 2011. "Online Parameter Estimation and Run-to-run Process Adjustment Using Categorical Observations." *International Journal of Production Research* 49 (13): 4103–4117.

Lin, J., and K. Wang. 2012. "A Bayesian Framework for Online Parameters Estimation and Process Adjustment Using Categorical Observations." *IIE Transactions* 44: 291–300.

Liu, I., and A. Agresti. 2005. "The Analysis of Ordered Categorical Data: An Overview and a Survey of Recent Developments." *Test* 14 (1): 1–73.

Lu, J. C., S. L. Jeng, and K. Wang. 2009. "A Review of Statistical Methods for Quality Improvement and Control in Nanotechnology." *Journal of Quality Technology* 41 (2): 148–164.

McCullagh, P. 1980. "Regression Models for Ordinal Data." *Journal of the Royal Statistical Society, Series B* 42 (2): 109–142.

Othman, M. K., A. Dolah, N. A. Omar, and M. R. Yahya. 2006. "Design of Experiment (DOE) for Thickness Reduction of GaAs Wafer using Lapping Process". *2006 IEEE International Conference on Semiconductor Electronics, Proceedings*, 583–585.

Shang, Y., K. Wang, and F. Tsung. 2009. "An Improved Run-to-run Process Control Scheme for Categorical Observations with Misclassification Errors." *Quality and Reliability Engineering International* 25: 397–407.

Spanos, C. J., and R. L. Chen. 1997. "Using Qualitative Observations for Process Tuning and Control." *IEEE Transactions on Semiconductor Manufacturing* 10 (2): 307–316.

Tseng, S. T., F. Tsung, and P. Y. Liu. 2007. "Variable EWMA Run-to-run Controller for a Drifted Process." *IIE Transactions* 39: 291–301.

Vanli, O. A., N. S. Patel, M. Janakiram, and E. Castillo. 2007. "Model Context Selection for Run-to-run Control." *IEEE Transactions on Semiconductor Manufacturing* 20 (4): 506–516.

Wang, K., and F. Tsung. 2007. "Run-to-run Process Adjustment Using Categorical Observations." *Journal of Quality Technology* 39 (4): 312–325.

Wang, K., and F. Tsung. 2008. "An Adaptive T^2 Chart for Monitoring Dynamic Systems." *Journal of Quality Technology* 40: 109–123.

Wang, K., and F. Tsung. 2010. "Recursive Parameter Estimation for Categorical Process Control." *International Journal of Production Research* 48 (5): 1381–1394.