International Journal of Production Research
Vol. 49, No. 13, 1 July 2011, 4103–4117

Online parameter estimation and run-to-run process adjustment using categorical observations

Jing Lin and Kaibo Wang*

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

(*Received 10 September 2009; final version received 26 April 2010*)

Categorical observations are frequently observed in run-to-run processes where obtaining accurate measurements of quality characteristics is difficult. In such circumstances, the use of categorical observations to estimate a process model and generate an adjustment recipe becomes inevitable. However, most conventional run-to-run controllers cannot be applied if no continuous observations are available; some parameter estimation methods that can handle categorical data only use historical dataset in an offline manner. In practice, it is common to see observations collected following a time sequence in a run-to-run process. Taking the lapping process in semiconductor manufacturing as an example, this paper develops an online approach for parameters estimation and run-to-run process adjustment using categorical observations. The proposed method optimises a penalised Maximum Likelihood (ML) function and updates parameters step by step when new categorical observations become available. A control strategy is also derived to generate receipts for process update between runs. The computational results of performance evaluation show that the proposed method is capable of estimating unknown parameters and control output quality online when initial bias exists.

Keywords: categorical observations; cut-points estimation; parameter estimation; statistical process adjustment; statistical process control

1. Introduction

Low-resolution categorical data are frequently seen in manufacturing processes. Some quality variables are inherently qualitative and cannot be measured on a numerical scale. These usually exist in manufacturing processes that rely heavily on visual quality inspection. One example is the production process of wafers in semiconductor manufacturing. Scratches on a wafer surface are difficult to quantify and can only be judged by operators visually. Spanos and Chen (1997) presented an example on a semiconductor plasma dry-develop process, in which sidewall roughness of etched wafers took only a few discrete values, labeled as ‘‘very rough’’, ‘‘rough’’, ‘‘smooth’’ and ‘‘very smooth’’. In some processes, practical constraints such as cost and instrument 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 can be measured by a scanning electron microscopy (SEM), which results in a bottleneck in large-volume manufacturing.

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

ISSN 0020–7543 print/ISSN 1366–588X online

DOI: 10.1080/00207543.2010.492806
http://www.informaworld.com

Instead, visual inspection, which produces quality-related readings of wafer trenches on a positive/normal/negative scale, can be used for quality control. Lu *et al.* (2009) also emphasised that analysis of new data type, including categorical data, is an important research topic in nanotechnology. Therefore, when only categorical data are available in a process, controlling process status and product quality based on this low-resolution information becomes necessary and important.

In this study, the lapping process in semiconductor manufacturing is used as an example (see Figure 1 for a schematic illustration of the process). A general wafer preparation process contains slicing, lapping, chemical vapour deposition (CVD), and polishing. In the lapping stage, the following tasks are done sequentially: (a) wafer loading, where a batch of wafers are mounted onto the lower lapping plate first; (b) machine setup, which is when controllable factors of the machine are adjusted based on operational instructions; (c) lapping, where the machine is started for a pre-determined amount of time; (d) wafer unloading, which is when finished wafers are unloaded; and (e) testing, which is when wafer thickness is measured and new settings for the next run are calculated.

As the first step in the mechanical treatment process on the wafer surface after slicing, lapping is a key step in forming quality characteristics for downstream processing. Thickness is an important geometric quality parameter and is largely dominated by lapping time and incoming thickness. Therefore, lapping time is usually adjusted between runs to control the amount of removal.

Ideally, if wafer thickness before and after lapping are both known online, it is possible to apply run-to-run controllers, such as EWMA controller and its extensions (Ingolfsson and Sachs 1993, Tseng *et al.* 2007), double EWMA controller (Chen and Guo 2001) and other controllers (He *et al.* 2009, Jin and Tsung 2009) to generate recipes and guide parameter updates between runs. However, to obtain accurate thickness values, measurement needs to be carried out in a special inspection room with the aid of an expensive testing machine, which is both time consuming and costly. In industrial practice, wafers are only moved to the inspection room for final testing before shipping. Therefore,

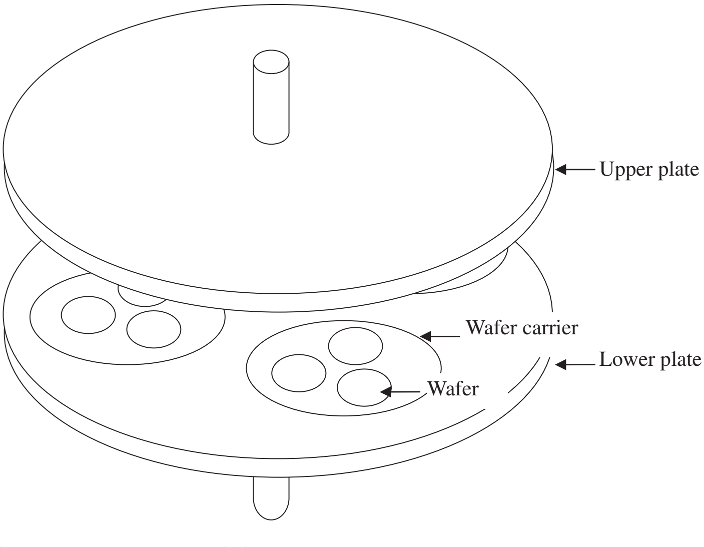

Figure 1. A schematic illustration of the lapping process.

no accurate data are immediately collected after each lapping run. Instead, a less expensive but inaccurate machine is equipped to help group lapped wafers into different categories, namely, very thin, thin, normal, thick to very thick.

To achieve a better thickness consistency, a controller is necessary to help generate optimal lapping time for each run. However, with only categorical observations available as the output of each run, a traditional EWMA controller is not applicable since it requires continuous data. Thus, a new controller able to generate control actions based on categorical observation is necessary for such a process.

To develop a controller that can work with categorical observations, two critical challenges need to be addressed. First, a process model should be built and unknown parameters estimated using categorical data; and second, optimal control actions must be generated using categorical data.

Different methods on model building using categorical observations can be found in the literature. Among these methods, cumulative logistic model is one of the most popular. Spanos and Chen (1997) used this model in the study of an etching process. In this model, parameters are usually attained using the maximum likelihood (ML) method. Agresti (1990) discussed the nonlinear optimisation problem related to maximising likelihood functions. McCullagh (1980) extended this model to a generalised linear model (GLM) by replacing the logit link function with other link functions, such as probit. Liu and Agresti (2005) discussed the choice between logit and probit functions. Some Bayesian-based methods can also be seen in the literature (Chipman and Hamada 1996, Girard and Parent 2001). However, all the aforementioned estimation strategies assume that all observations are already collected before model fitting. Parameters are estimated in an offline manner given all information available. However, in the lapping process and other run-to-run processes, products are produced batch by batch; data become available gradually in a continuous way. Therefore, an online parameter estimation method to incorporate categorical observations is needed.

To incorporate growing data streams in model building, Bauer *et al.* (1997) studied an update rule for parameter estimation in Bayesian networks and provided a unified framework for online learning. Parameters are updated by maximising a function containing a normalised log-likelihood in a model and a penalty. The authors obtained three rules under three different norms. Solla and Winther (1999) presented an online Bayesian learning approach. The exact posterior distribution in this approach is approximated using a simple parametric distribution and each new observation is utilised for posterior update. However, when categorical observations are encountered, the above algorithms could not be applied directly. In this work, a new method that can handle categorical observations and update parameters in an online manner is proposed.

Recently, research on process control and adjustment using qualitative information has drawn considerable attention. In an early work, Spanos and Chen (1997) demonstrated the importance and feasibility of implementing process monitoring and controlling with qualitative characteristics. Wang and Tsung (2007) introduced a two-phase R2R process control strategy in a semiconductor manufacturing process. They conducted experiments and collected data in Phase I for model parameters estimation. Shang *et al.* (2009) improved on this work by considering misclassification errors, in which misclassification possibilities were used to compensate for adjustment bias. Wang and Tsung (2010) studied the recursive parameter estimation issue when categorical observations are presented, and proposed a Bayesian Categorical Controller. However, the authors assumed all cut-points were known and only studied the estimation of a linear process model. Cut-points are threshold values used in generating categorical observations. To better utilise categorical observations in process control and quality improvement, a new algorithm that can estimate cut-points online and generate recipes for run-to-run control needs to be developed.

In this research, a strategy for online parameters estimation and process adjustment based on categorical observations is proposed. Compared with existing work, this paper tackles the following challenges: (a) online recursive estimation of cut-points using categorical observations; and (b) run-to-run control using categorical observations.

The rest 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 categorical data; Section 4 studies and analyses the performance of the proposed method. Finally, Section 5 concludes this paper and discusses topics related to categorical data that deserve further research.

2. Process modeling

In this section, the lapping process is still used in wafer production for illustration. In general semiconductor manufacturing and other manufacturing applications, linear models have been widely used to characterise processes with continuous inputs and outputs (e.g., Del Castillo and Hurwitz 1997, Wang and Tsung 2007, 2008, 2010, Shang *et al.* 2009).

To study the impact of controllable factors on lapping processes, Othman *et al.* (2006) conducted full-factorial designed experiments, the results of which could be used to build a linear model between removal rate and controller factors. In practice, we have conducted real experiments in a local wafer manufacturing company. Post-lapping thickness is chosen as a response variable. A fractional-factorial design that involves four controllable parameters in the lapping process (pressure, rotation speed, speed-up time, and lapping time) is chosen to guide our experiments. A linear model is then built from the experimental dataset. Three factors are optimised to improve robustness and one factor (lapping time) is left for online tuning to achieve better thickness control. Further historical data are collected and the effect of incoming thickness is incorporated into the model. The final model used for characterising wafer thickness after lapping is

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

where y_t is the output thickness obtained at time t; $u_{t-1}$is the lapping time set at time $t-1$, which is a controllable factor in lapping; and x_t is the incoming wafer thickness generated by the slicing stage. Therefore, x_t can be observed but not changed. ε_t represents process disturbance, and a, b, and c are the coefficients. Without loss of generality, $\varepsilon_t \sim N(0, \sigma^2)$ is assumed.

Essentially, if the process is already on target and is only affected by white noise, process adjustment is unnecessary. However, real processes usually have bias due to the following reasons: (a) true process parameters are never known. If parameters were already estimated with bias, the correct settings would not be obtained initially; (b) true parameters themselves can change due to sudden process shifts. Therefore, it is always necessary to implement process controllers to remove initial deviations and ensure quality (Del Castillo 2006, He *et al.* 2009). Initial bias is a more serious problem in short-run processes. In fact, the lapping process has to stop each day when the shift switches.

Production break causes the machine to cool down, which leads to significant bias compared with normal operational status after warm up. Physical and mechanical properties of silicon wafers also change several times a week when raw single-crystal silicon material and wafer size change.

Usually, y_t is assumed to be measured on a continuous scale. Under such circumstances, conventional process controllers can be applied to control the process. However, in the aforementioned lapping process, continuous measurements cannot be obtained; thus y_t becomes unknown. To handle this challenge, y_t is treated as a latent variable. 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)

$$Y_t = j \Leftrightarrow \gamma_{j-1} < y_t \leq \gamma_j, j = 1, \ldots, c, \tag{2}$$

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 worst and best values of y_t), we assume that $\gamma_0 = -\infty$ and $\gamma_c = \infty$. If y_t was known, the value of Y_t is generated by comparing y_t against γ to decide which category it should be assigned to. If y_t was unknown but γ is known, the above equation will be useful for estimating y_t given Y_t. The detailed method is introduced in the following section.

In Equation (1), we assume b and c are known, and a is unknown. This happens because a easily fluctuates with the temperature of lapping pans and machine setup when a new order comes, while b and c are dominated by physical mechanism and are therefore relatively stable. Wafer thickness before lapping, x_t, is to be measured precisely after slicing, hence it is available to the lapping stage. The cut-points γ in Equation (2) are the hidden rules used by operators. However, these rules are neither known to the operators nor can they be measured directly. Therefore, γ needs to be estimated.

3. Online estimation and adjustment using categorical observations

At the end of each production run, the following tasks are done sequentially: (a) measure samples to obtain categorical data; (b) update estimates of unknown parameters; and (c) recommend parameter settings for the next run. Task (a) is done manually by operators. In the following, the treatment for tasks (b) and (c) is introduced.

3.1 *Online parameters estimation*

In this section, a recursive method to estimate the unknown parameters a and γ online is presented. In Section 1, we mentioned that Bauer *et al.* (1997) investigated a rule for parameter estimation using continuous data. This method is straightforward and easy to implement. In this section, the ML method is first introduced followed by an Adjusted ML method derived for parameter estimation using categorical data.

3.1.1 *The ML method*

Kivinen and Warmuth (1997) first proposed a supervised online learning framework, which was successfully used in a large number of supervised and unsupervised problems (Singer and Warmuth 1999). Bauer *et al.* (1997) extended this framework to parameter online update in Bayesian networks when new data continuously arrive, and proposed to maximise the following function:

$$F(\hat{\theta}) = \eta L_D(\hat{\theta}) - d(\hat{\theta}, \bar{\theta}), \tag{3}$$

where θ is used to represent any unknown parameters to be estimated, $\bar{\theta}$ is an initial estimate of θ, $L_D(\hat{\theta})$ is the normalised log-likelihood of data cases D, $d(\hat{\theta}, \bar{\theta})$ is a function of the distance between old and new estimates, and η is a parameter that controls learning rate.

Bauer *et al.* (1997) mainly discussed the case of when a group of data, D, is available for model update. In this research, it is simplified into a scenario in which only one observation is available for model update each time. This is more suitable for the lapping process and other run-to-run processes studied. The above objective function is then modified as follows:

$$F(\hat{\theta}) = \eta \cdot \log P_{\hat{\theta}}(Y = j) - \frac{1}{2}\left(\hat{\theta} - \bar{\theta}\right)^2, \tag{4}$$

where $\log P_{\hat{\theta}}(Y = j)$ is the log-likelihood function of seeing an observation fall into category j under parameters setting $\hat{\theta}$, which is used to grasp information in the new observation D, $D = \{Y = j\}$. The learning rate η could now be interpreted as how far away the new estimate $\hat{\theta}$ can be expected to deviate from the original one $\bar{\theta}$. The penalty, $(\hat{\theta} - \bar{\theta})^2/2$, is the distance between $\bar{\theta}$ and $\hat{\theta}$ under L-2 norm, which could be interpreted as a distance measure between the original and updated model. The penalty is used to prevent $\hat{\theta}$ from staying too far away from $\bar{\theta}$, thus ensuring that the new estimate preserves part of the original information. Therefore, a large η places more weight on new observations and forces parameters to update quickly, while a small η places more emphasis on the difference between updated and old estimates and ensures that the difference will not be exaggerated too much.

Following the treatment in Bauer *et al.* (1997), the objective function is expanded using the one order Taylor expansion to simplify $F(\hat{\theta})$

$$F(\hat{\theta}) \approx \eta \cdot \log P_{\bar{\theta}}(Y = j) + \eta \cdot \frac{\mathrm{d}\left(\log P_{\bar{\theta}}(Y = j)\right)}{\mathrm{d}\theta}\left(\hat{\theta} - \bar{\theta}\right) - \frac{1}{2}\left(\hat{\theta} - \bar{\theta}\right)^2. \tag{5}$$

Then, the derivative of $F(\hat{\theta})$ is obtained as

$$F'(\hat{\theta}) = \eta \cdot \frac{\mathrm{d}\left(\log P_{\bar{\theta}}(Y = j)\right)}{\mathrm{d}\theta} - \left(\hat{\theta} - \bar{\theta}\right). \tag{6}$$

Since $F'(\hat{\theta}) = 0$ is the prerequisite for maximising $F(\hat{\theta})$, we therefore have

$$\hat{\theta} = \bar{\theta} + \eta \cdot \frac{\mathrm{d}\left(\log P_{\bar{\theta}}(Y = j)\right)}{\mathrm{d}\theta}. \tag{7}$$

Following Equation (7), θ can be updated using information contained in new observations.

The above derivation provides a recursive method of parameter estimation. As no assumptions are applied on θ, the above procedure can be extended to general situations for parameter estimation. In the following, the scenario is constrained in categorical data modeling and specific equations are developed to estimate parameters in Equation (1).

3.1.2 *The adjusted ML method based on categorical observations*

In this section, we use the update rule to our application to estimate parameters in the process model using categorical measurements.

Following the above framework, $\theta = \{a, \gamma\}$, $D = \{Y_t = j \,|\, u_{t-1}, x_t\}$ is set, where Y_t is a categorical variable at step t, j represents the category the sample falls into $1 \le j \le c$, and, u_{t-1} and x_t are controllable and uncontrollable input variables, respectively. Estimated parameters at step t are denoted as $a^{(t)}$ and $\gamma^{(t)}$. Initial parameter values at step t are the updated estimates obtained at step t-1 (i.e. $\bar{\theta} = \{a^{(t-1)}, \gamma^{(t-1)}\}$). Recursively, the updated estimate at step t, $\hat{\theta} = \{a^{(t)}, \gamma^{(t)}\}$, becomes the initial values of step $t+1$.

From this we learn that the probability a sample falls into category j is

$$P(Y_t = j) = P(\gamma_{j-1} < y_t \le \gamma_j) = P(y_t \le \gamma_j) - P(y_t \le \gamma_{j-1}). \tag{8}$$

Therefore, the corresponding log-likelihood given u_{t-1} and x_t (the condition in the equation is omitted for clarity) is

$$\log P_{\bar{\theta}}(Y_t = j) = \log\left(\Phi\left(\frac{\gamma_j^{(t-1)} - a^{(t-1)} - bu_{t-1} - cx_t}{\sigma}\right) - \Phi\left(\frac{\gamma_{j-1}^{(t-1)} - a^{(t-1)} - bu_{t-1} - cx_t}{\sigma}\right)\right). \tag{9}$$

Its derivatives with respect to all unknown parameters are given as follows

$$\frac{\mathrm{d}\left(\log P_{\bar{\theta}}(Y_t = j)\right)}{\mathrm{d}a} = -\frac{1}{\sqrt{2\pi}} \times \frac{\exp\left(-\left(\gamma_j^{(t-1)} - a^{(t-1)} - bu_{t-1} - cx_t\right)^2 / 2\sigma^2\right) - \exp\left(-\left(\gamma_{j-1}^{(t-1)} - a^{(t-1)} - bu_{t-1} - cx_t\right)^2 / 2\sigma^2\right)}{\sigma\left(\Phi\left(\frac{\gamma_j^{(t-1)} - a^{(t-1)} - bu_{t-1} - cx_t}{\sigma}\right) - \Phi\left(\frac{\gamma_{j-1}^{(t-1)} - a^{(t-1)} - bu_{t-1} - cx_t}{\sigma}\right)\right)} \tag{10}$$

$$\frac{\mathrm{d}\left(\log P_{\bar{\theta}}(Y_t = j)\right)}{\mathrm{d}\gamma_k} = \begin{cases} \frac{1}{\sqrt{2\pi}} \cdot \frac{\exp\left(-\left(\gamma_j^{(t-1)} - a^{(t-1)} - bu_{t-1} - cx_t\right)^2 / 2\sigma^2\right)}{\sigma\left(\Phi\left(\frac{\gamma_j^{(t-1)} - a^{(t-1)} - bu_{t-1} - cx_t}{\sigma}\right) - \Phi\left(\frac{\gamma_{j-1}^{(t-1)} - a^{(t-1)} - bu_{t-1} - cx_t}{\sigma}\right)\right)}, & k = j \\ -\frac{1}{\sqrt{2\pi}} \cdot \frac{\exp\left(-\left(\gamma_{j-1}^{(t-1)} - a^{(t-1)} - bu_{t-1} - cx_t\right)^2 / 2\sigma^2\right)}{\sigma\left(\Phi\left(\frac{\gamma_j^{(t-1)} - a^{(t-1)} - bu_{t-1} - cx_t}{\sigma}\right) - \Phi\left(\frac{\gamma_{j-1}^{(t-1)} - a^{(t-1)} - bu_{t-1} - cx_t}{\sigma}\right)\right)}, & k = j - 1 \\ 0, & o.w. \end{cases} \tag{11}$$

Substituting Equations (11) and (10) for the corresponding terms in Equation (7), and ignoring the constant item $1/\sqrt{2\pi}$ as it could be absorbed into the learning rate η, the online estimation algorithm of categorical data modeling is obtained as below.

$$a^{(t)} = a^{(t-1)} - \eta \cdot \frac{\exp\left(-\left(\gamma_j^{(t-1)} - a^{(t-1)} - bu_{t-1} - cx_t\right)^2 / 2\sigma^2\right) - \exp\left(-\left(\gamma_{j-1}^{(t-1)} - a^{(t-1)} - bu_{t-1} - cx_t\right)^2 / 2\sigma^2\right)}{\sigma\left(\Phi\left(\frac{\gamma_j^{(t-1)} - a^{(t-1)} - bu_{t-1} - cx_t}{\sigma}\right) - \Phi\left(\frac{\gamma_{j-1}^{(t-1)} - a^{(t-1)} - bu_{t-1} - cx_t}{\sigma}\right)\right)} \tag{12}$$

$$\begin{cases} \gamma_{j-1}^{(t)} = \gamma_{j-1}^{(t-1)} - \eta \cdot \dfrac{\exp\left(-\left(\gamma_j^{(t-1)} - a^{(t-1)} - bu_{t-1} - cx_t\right)^2 / 2\sigma^2\right)}{\sigma\left(\Phi\left(\frac{\gamma_j^{(t-1)} - a^{(t-1)} - bu_{t-1} - cx_t}{\sigma}\right) - \Phi\left(\frac{\gamma_{j-1}^{(t-1)} - a^{(t-1)} - bu_{t-1} - cx_t}{\sigma}\right)\right)}, \\ \gamma_j^{(t)} = \gamma_j^{(t-1)} + \eta \cdot \dfrac{\exp\left(-\left(\gamma_j^{(t-1)} - a^{(t-1)} - bu_{t-1} - cx_t\right)^2 / 2\sigma^2\right)}{\sigma\left(\Phi\left(\frac{\gamma_j^{(t-1)} - a^{(t-1)} - bu_{t-1} - cx_t}{\sigma}\right) - \Phi\left(\frac{\gamma_{j-1}^{(t-1)} - a^{(t-1)} - bu_{t-1} - cx_t}{\sigma}\right)\right)}, \\ \gamma_k^{(t)} = \gamma_k^{(t-1)}, k \neq j - 1 \text{ and } j \end{cases} \tag{13}$$

Using Equations (12) and (13), a and γ can be updated whenever a new categorical observation becomes available.

As the above method is derived from the ML method, it is named the adjusted ML method. The new online algorithm is expected to utilise sampled categorical data continuously and provide an effective estimate of unknown parameters.

3.2 *The process adjustment algorithm*

The adjusted ML method outlined above routinely incorporates new categorical samples to update parameter estimates. Given estimated parameter values, strategies can be developed to control process output by generating a recipe for each run. Such a control strategy should be developed to minimise output deviation and variation.

Denote the target of the process (1) as T, and the cumulative information up to the present step t as $F_t, F_t = \{Y_t, \ldots, Y_1, u_{t-1}, \ldots, u_0, x_t, \ldots, x_1\}$. The following quadratic loss function conditioning is defined on all historical information as the objective of process adjustment at step t

$$L = E\left[(y_{t+1} - T)^2 | F_t\right]. \tag{14}$$

The control goal then is to find a u_t to minimise the above objective function.

As Equation (1) shows, $y_{t+1} = a + bu_t + cx_{t+1} + \varepsilon_{t+1}$ and $E(\varepsilon_{t+1}^2) = \sigma^2$, it follows that

$$L = (a - cx_{t+1} - T)^2 + b^2 u_t^2 + 2(a - cx_{t+1} - T)bu_t + \sigma^2. \tag{15}$$

Taking the partial derivative of the above equation with respect to u_t as zero and replacing unknown parameters with their estimates lead to the following optimal control action

$$u_t = \frac{T - a^{(t)} - cx_{t+1}}{b}, \tag{16}$$

where $a^{(t)}$ is the estimate of a at step t.

It is interesting to note that the above equation resembles the traditional EWMA controller in generating control actions. The EWMA controller is a popular method in run-to-run process control. It is designed to use continuous observations for parameter estimation and process adjustment, while the above scheme works on categorical observations.

4. Performance studies

In this section, the performance of the newly proposed method is investigated and the conditions under which it can work more efficiently are identified.

Since Wang and Tsung (2007) and Shang *et al.* (2009) have both worked on a problem similar to what we have tackled, the current study's method is compared with the methods they proposed. However, a direct comparison is difficult. Wang and Tsung (2007) and Shang *et al.* (2009) did not consider online parameter estimation issues. The authors assumed that historical Phase I experimental data are available for off-line parameter estimation. Such estimates are then used to build a model for online process control. However, online parameter estimation is one major part of this work due to a real demand from our application. In addition, in the Phase I estimation in Wang and Tsung (2007), the authors assume both categorical and corresponding continuous data are available. In this work, accurate process outputs not available in all phases are allowed.

Therefore, the performance of the proposed method is examined and compared with other works. As the whole work is divided into two steps, online parameter estimation and online process adjustment, we study the performance of the Adjusted ML method from these two aspects. Since the first step focuses on unknown parameter estimation, estimated values are compared with true values in simulation studies. In the second step, the Adjusted ML method is compared with the Categorical Controller proposed by Wang and Tsung (2007) and its control performance is investigated.

Since the online estimation of unknown parameters in the process model is already complicated, measurement errors (misclassification) are not considered in this work and are left as a topic for future research. Therefore, it is not compared with Shang *et al.* (2009).

All simulation studies are designed based on the lapping process in Equation (1). Due to confidentiality issues, model parameters have been transformed from the original values obtained from designed experiments to the following settings: $a = 60$, $b = 2$, $c = 0.1$ and standard deviation $\sigma = 3$. Such transformation has no impact on the current performance study. The process target is $T = 400$. Four cut-points, 396, 398, 401, and 405, are treated as true values for classifying output y_t into five mutually exclusive categories, that is, $\gamma = [396, 398, 401, 405]$. The incoming thickness before lapping, x_t, is assumed to follow a uniform distribution in interval [450, 550]. In practice, extremely thin and thick wafers are treated as defects and recycled after slicing; such wafers cannot enter the lapping process.

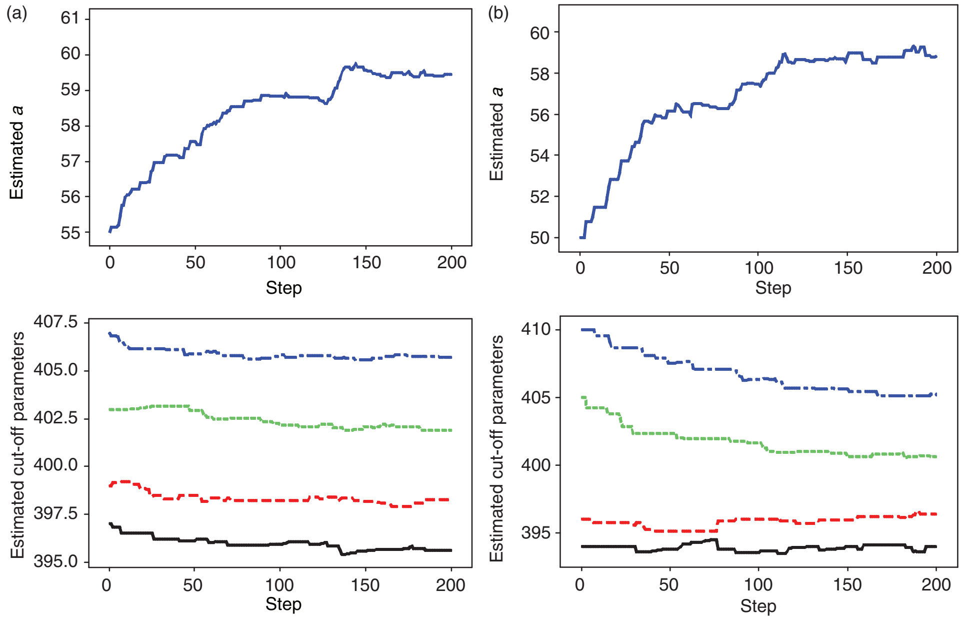

Figure 2. Trajectories of estimated parameters without process control. (a) $a^{(0)} = 55$, $\gamma^{(0)} = [397, 399, 403, 407]$, $\eta = 0.13$ (b) $a^{(0)} = 50$, $\gamma^{(0)} = [397, 396, 403, 407]$, $\eta = 0.18$.

4.1 *Online estimation performance*

As emphasised in Section 1, initial bias is a major problem in the lapping process. In addition, since the distance between initial and true values usually affect the convergence performance of an algorithm, different initial values are investigated in the following studies. In this section, the control algorithm (16) is not applied so that the impact of the feedback-loop on parameter estimation is removed. The control action, u_t, is sampled from a uniform distribution within [130, 150]. Different learning rates η are also tried to test its influence on estimation performance. In the next section, the estimation algorithm is combined with the feedback-controller and its performance is tested when observations are collected from an online-controlled process.

Figure 2 shows the trajectories of estimated parameters with different initial values and learning rates. Each process is run for 200 steps. From Figure 2 (a) and (b), we can see that when initial values deviate from their true values, estimated values approach their true value gradually as new categorical samples are collected step by step. This shows that the Adjusted ML method is capable of estimating the unknown parameters in Equation (1) using categorical observations.

The convergence rate of the method is influenced by parameter η. In Figure 3, the mean square error (MSE) of estimated parameters is used at step 200 of 100 trajectories as a performance index to measure estimation accuracy. A lower value means the parameters are estimated more accurately. The solid line shows the MSE average of five unknown parameters in α and γ.

From Figure 3, it is interesting to see that for all the cases studied, there exists an optimal η that achieves the minimum MSE. For example, in case (a), the minimum MSE is achieved when $\eta = 0.13$; in case (b), the minimum MSE is achieved when $\eta = 0.18$.

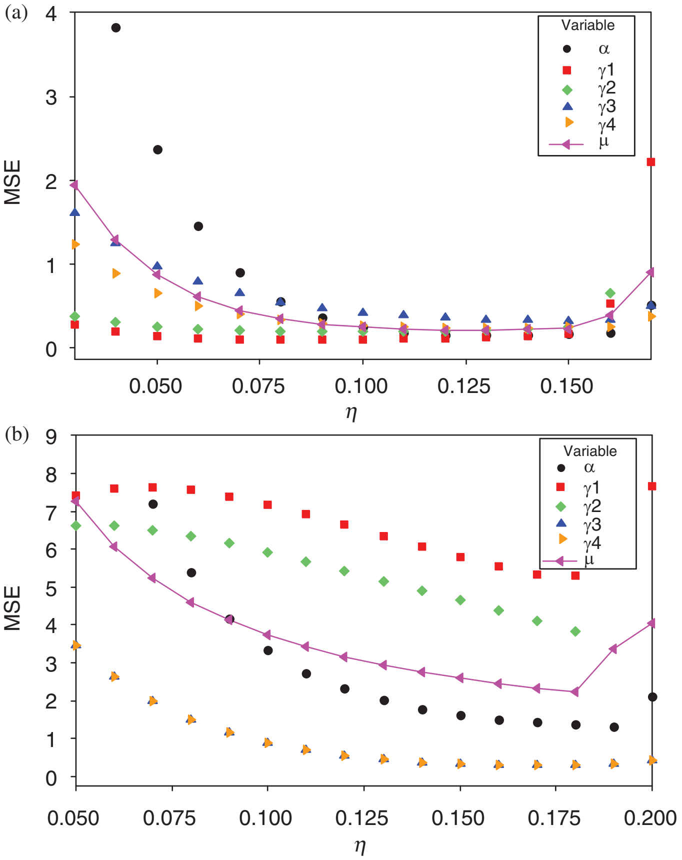

Figure 3. Characteristics curves about different learning rate with the corresponding performance. (a) $a^{(0)} = 55$, $\gamma^{(0)} = [397, 399, 403, 407]$; (b) $a^{(0)} = 50$, $\gamma^{(0)} = [394, 396, 405, 410]$.

The optimal η may also be different for the different sizes of initial bias. As stated earlier, a small η tends to limit the difference between updated and initial values. As categorical observations are less reliable, the magnitude of η should be relatively small. Extensive simulations show that $\eta = [0.1, 0.2]$ gives reasonably good performance for moderate initial bias. Furthermore, it is noticed that the MSE curve is flat around its minimum, which means the performance of the adjusted ML estimation method is stable around the optimal η.

The above study shows that when there is no feedback-loop, the Adjusted ML method updates the unknown parameters gradually and leads them to their true values. In the following section, the Adjusted ML method with feedback-control is applied and its performance is validated in process control scenarios.

4.2 *Process adjustment performance*

In the following study, unknown parameters in the lapping process are assumed to take the following initial values, $a^{(0)} = 50, \gamma^{(0)} = [394, 396, 405, 410]$. The learning rate in

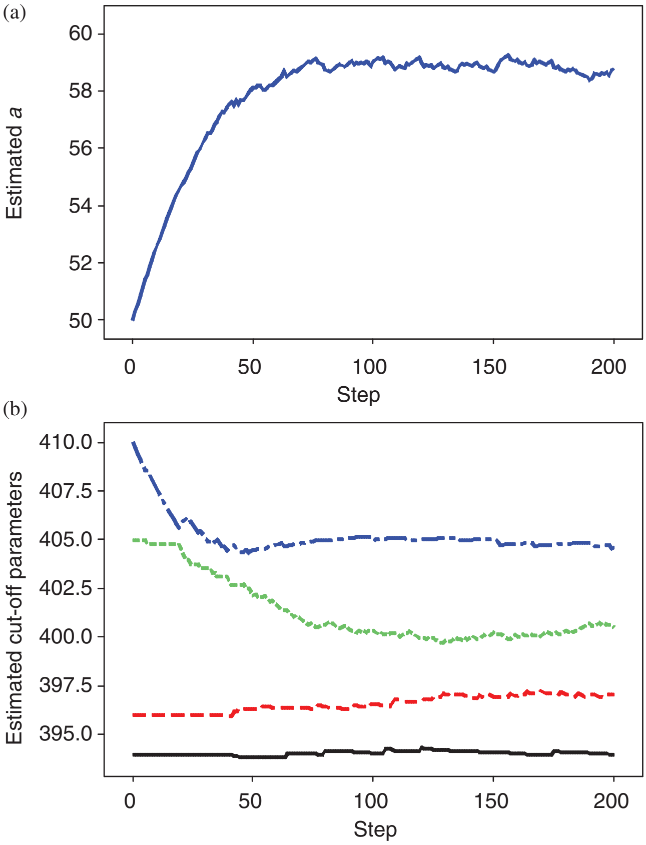

Figure 4. Trajectories of estimated parameters. (a) Estimates of a; (b) Estimates of γ.

Equations (12) and (13) is set to $\eta = 0.1$. The control algorithm in Equation (16) is used to generate recipe u_t at step t to guide the setup for run $t+1$. Two-hundred lapping runs are simulated in total. The evolvement of unknown parameters is shown in Figure 4.

Figure 4 (a) shows the path of estimated a; Figure 4 (b) shows the trajectories of four cut-points. It is clearly seen that the initial values are deviated from their respective true levels. As the process evolves, more observations are collected, and unknown parameters are continuously updated. All parameters are seen approaching their respective true values gradually. After 200 runs, the deviations between estimated and true values are already rather small.

Wang and Tsung (2007) proposed a Categorical Controller for process adjustment when only categorical observations are available for online process adjustment. To study the performance of the proposed Adjusted ML method, the Categorical Controller is also set up to control the same process. Since the Categorical Controller requires accurate observations for online adjustment, true observations are used to allow it to generate control actions. The simulation is repeated 100 times and the MSE is calculated at each step to show their control accuracy. The results are shown in Figure 5.

It is clearly seen from Figure 5 that in the first 60 steps, the Categorical Controller performs better than the proposed method. However, after around step 60, the Adjusted

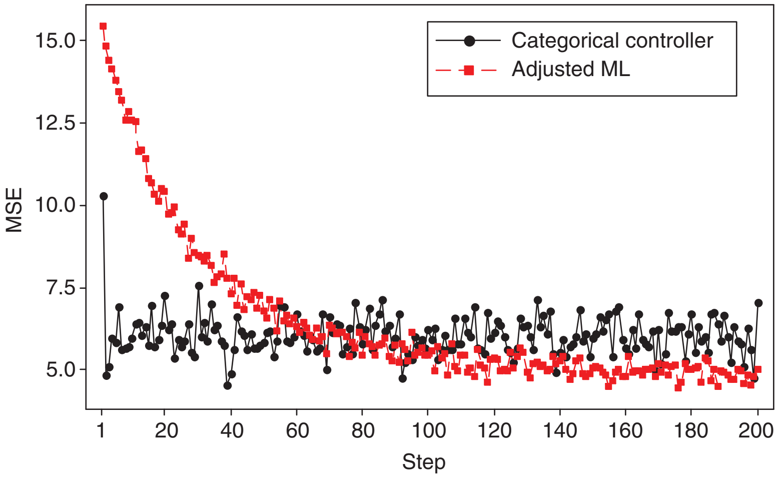

Figure 5. Control performance comparison.

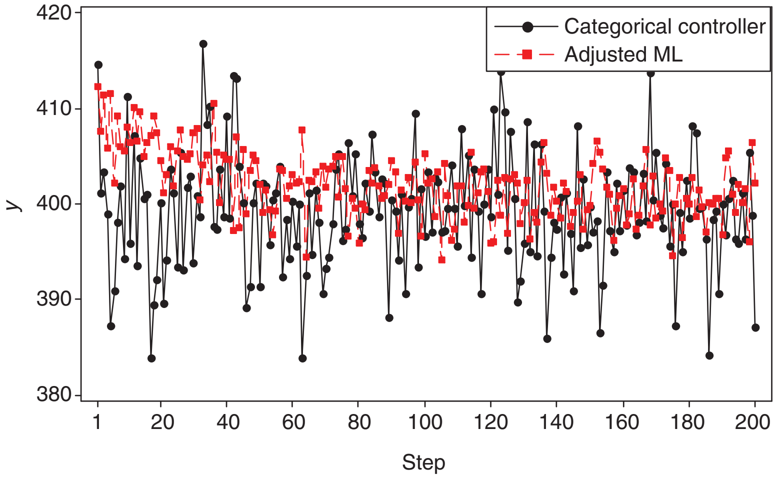

Figure 6. Trajectories of process output.

ML method begins to outperform the Categorical Controller and shows smaller MSE values. The improved performance can be explained using the online estimation method that the Adjusted ML used. The Categorical Controller uses accurate and continuous observations. Therefore, it can quickly overcome the initial bias. However, the Adjusted ML method uses more steps to compensate for the initial bias. Once the initial bias is compensated, the Adjusted ML becomes better since it provides more precise control actions than the Categorical Controller.

The sequences of process output, y_t, are also shown in Figure 6. It is seen that when controlled by the Adjusted ML method, the initially biased parameters lead to a large deviation between the process output and the target thickness 400 μm. As the unknown parameters are estimated more accurately, deviations are also reduced.

After around 100 steps, the process output finally wanders around the target value. The Categorical Controller can bring the process output to the target faster than the Adjusted ML. However, it shows a larger variation than the Adjusted ML. Therefore, in Figure 5, the Categorical Controller shows a larger MSE when the Adjusted ML evolves.

The above study demonstrates the efficiency of the proposed method in calibrating initial bias in unknown parameters. It also proves that the proposed algorithm is effective in controlling the process using categorical observations.

5. Conclusions

This paper investigated the online estimation and control of a run-to-run process when only categorical observations are available. A modified method based on maximised likelihood, named the Adjusted ML method, was proposed to update model parameters when categorical data were collected sequentially from the process. A control algorithm that functioned based on categorical observations was also proposed to adjust the process output on target.

Simulation studies showed that when initial bias existed, the adjusted ML method can move parameter estimates toward their respective true values gradually. The proposed scheme, together with the control algorithm, was proved effective in controlling processes with initial bias.

This paper used a simple model with white noises to characterise a lapping process. When the disturbance series became more complicated, for example, following an IMA(1, 1) or a general ARIMA time series, the likelihood of observations can be calculated using one-step-ahead prediction. The update equations will be changed accordingly. Its performance in controlling real processes deserves more in-depth study and could be the focus of future research.

Acknowledgements

The authors are grateful to the editor and the referees for carefully reading this manuscript and giving valuable comments, which have helped improve this work greatly. This research was supported by the National Natural Science Foundation of China (NSFC) under grant No. 70802034.

References

Agresti, A., 1990. *Categorical data analysis*. New York: Wiley.

Bauer, E., Koller, D., and Singer, Y., 1997. Update rules for parameter estimation in Bayesian networks. *Proceedings of the Thirteenth Annual Conference on Uncertainty in Artificial Intelligence*, Providence, RI, 3–13.

Chen, A. and Guo, R.S., 2001. Age-based double EWMA controller and its application to CMP processes. *IEEE Transactions on Semiconductor Manufacturing*, 14 (1), 11–19.

Chipman, H. and Hamada, M., 1996. Bayesian analysis of ordered categorical data from industrial experiments. *Technometrics*, 38 (1), 1–10.

Del Castillo, E., 2006. Statistical process adjustment: a brief retrospective, current status, and some opportunities for further work. *Statistica Neerlandica*, 60 (3), 309–326.

Del Castillo, E. and Hurwitz, A.M., 1997. Run-to-run process control: Literature review and extensions. *Journal of Quality Technology*, 29 (2), 184–196.

Girard, P. and Parent, E., 2001. Bayesian analysis of autocorrelated ordered categorical data for industrial quality monitoring. *Technometrics*, 43 (2), 180–191.

He, F., Wang, K., and Jiang, W., 2009. A general harmonic rule controller for run-to-run process control. *IEEE Transactions on Semiconductor Manufacturing*, 22 (2), 232–244.

Ingolfsson, A. and Sachs, E., 1993. Stability and sensitivity of an EWMA controller. *Journal of Quality Technology*, 25 (4), 271–287.

Jin, M. and Tsung, F., 2009. Smith-EWMA run-to-run control schemes for a process with measurement delay. *IIE Transactions*, 41 (4), 346–358.

Kivinen, J. and Warmuth, M., 1997. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132 (1), 1–64.

Liu, I. and Agresti, A., 2005. The analysis of ordered categorical data: an overview and a survey of recent developments. *Test*, 14 (1), 1–73.

Lu, J.C., Jeng, S.L., and Wang, K., 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., *et al.*, 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., Wang, K., and Tsung, F., 2009. An improved run-to-run process control scheme for categorical observations with misclassification errors. *Quality and Reliability Engineering International*, 25 (4), 397–407.

Singer, Y. and Warmuth, M., 1999. Batch and on-line parameter estimation of Gaussian mixtures based on the joint entropy. *In*: *Proceedings of the 1998 conference on advances in neural information processing systems*, 578–584.

Solla, S.A. and Winther, O., 1999. Optimal online learning: a Bayesian approach. *Computer Physics Communications*, 121–122, 94–97.

Spanos, C.J. and Chen, R.L., 1997. Using qualitative observations for process tuning and control. *IEEE Transactions on Semiconductor Manufacturing*, 10 (2), 307–316.

Tseng, S.T., Tsung, F., and Liu, P.Y., 2007. Variable EWMA run-to-run controller for a drifted process. *IIE Transactions*, 39 (3), 291–301.

Wang, K. and Tsung, F., 2007. Run-to-run process adjustment using categorical observations. *Journal of Quality Technology*, 39 (4), 312–325.

Wang, K. and Tsung, F., 2008. An adaptive T2 chart for monitoring dynamic systems. *Journal of Quality Technology*, 40 (1), 109–123.

Wang, K. and Tsung, F., 2010. Recursive parameter estimation for categorical process control. *International Journal of Production Research*, 48 (5), 1381–1394.