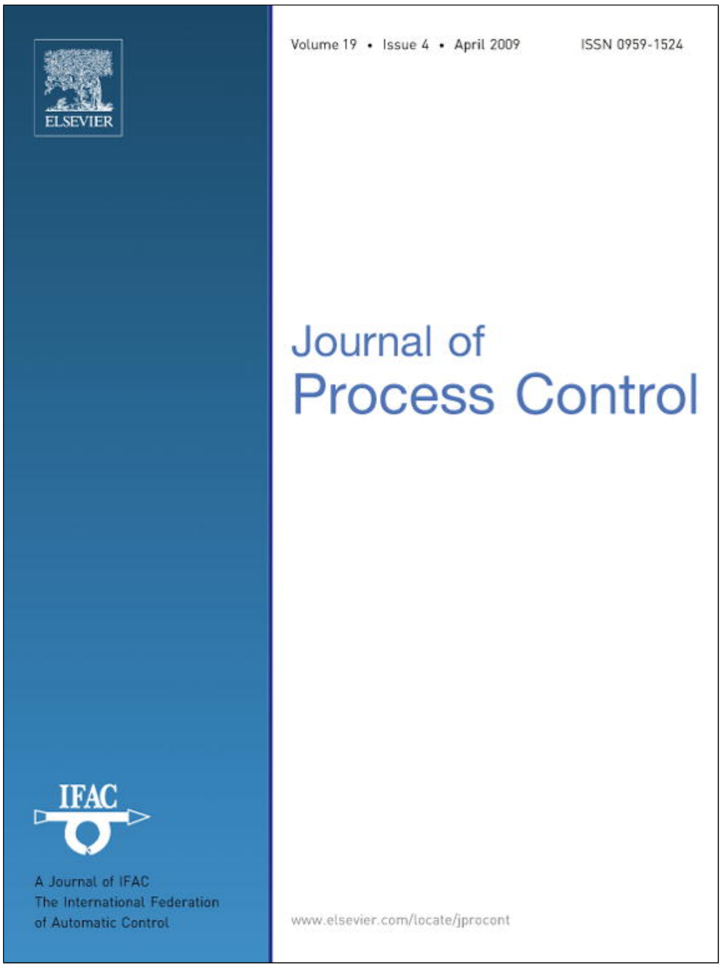
Volume 19 • Issue 4 • April 2009
ISSN 0959-1524
Journal of
Process Control
ELSEVIER
IFAC
A Journal of IFAC
The International Federation
of Automatic Control
www.elsevier.com/locate/jprocont

Journal of Process Control 19 (2009) 591–603

Contents lists available at ScienceDirect

Journal of Process Control

journal homepage: www.elsevier.com/locate/jprocont

Identification of tool and product effects in a mixed product and parallel tool environment

Ming-Da Ma [a], Chun-Cheng Chang [b], David Shan-Hill Wong [b,*], Shi-Shang Jang [b,*]

[a] Center for Control and Guidance Technology, Harbin Institute of Technology, Harbin 150001, China
[b] Department of Chemical Engineering, National Tsing-Hua University, Hsin-Chu, Taiwan 30043, Taiwan

A R T I C L E I N F O

Article history:
Received 20 February 2008
Received in revised form 27 May 2008
Accepted 26 July 2008

Keywords:
State estimation
ANOVA
Kalman filter
Run-to-run control

A B S T R A C T

In the semiconductor manufacturing industry, production resembles an automated assembly line in which many similar products with slightly different specifications are manufactured step-by-step, with each step being a complicated physiochemical batch process performed by a number of tools. This constitutes a high-mix production system for which effective run-to-run control (RtR) and fault detection control (FDC) can be carried out only if the states of different tools and different products can be estimated. However, since in each production run, a specific product is performed on a specific tool, absolute individual states of products and tools are not observable. In this work, a novel state estimation method based on analysis of variance (ANOVA) is developed to estimate the relative states of each product and tool to the grand average performance of this station in the fab. The method is formulated in the form of a recursive state estimation using the Kalman filter. The advantages of this method are demonstrated using simulations to show that the correct relative states can be estimated in production scenarios such as tool-shift, tool-drift, product ramp-up, tool/product-offline and preventive maintenance (PM). Furthermore, application of this state estimation method in RtR control scheme shows that substantial improvements in process capabilities can be gained, especially for products with small lot counts. The proposed algorithm is also evaluated by an industrial application.

1. Introduction

The semiconductor manufacturing industry is one of the fastest evolving industries in the world. As feature sizes shrink and wafer sizes increase, sophisticated control methods are needed to improve the product yield, throughput, and overall equipment effectiveness.

The run-to-run (RtR) controller is a model-based process control system that integrates concepts in statistical process control (SPC) and engineering process control (EPC). It is achieved by adjusting process inputs (recipes) at the beginning of each run based on the information obtained from previous runs. In the last decade, RtR control has been extensively deployed in the semiconductor industry. Research and development in this area have been summarized by many authors in books and review articles [2,3,6,11,15].

Most RtR control algorithms are based on the assumption that there is only a single product fabricated in the manufacturing line [1,4,17]. This is, however, far from reality. In the semiconductor manufacturing industry, production resembles an automated assembly line in which many similar products with different specifications are manufactured step-by-step, with each step being a complicated physiochemical batch process carried out by a number of tools. A specific combination of product and tool is known as a "thread" [7,18]. Single product RtR control algorithms can be applied to a thread. However, the number of threads can be very large, up to thousands in a foundry fab. It is cumbersome to maintain so many controllers. Moreover, the controller performance will be degraded for those infrequent threads since condition of the tool may be quite different from the last run of the same thread. Zheng et al. [22] showed that even if the actual root cause is the change in condition of the tool, a single tool-based EWMA (exponentially-weighted moving average) controller is unstable if the model uncertainties of different products are different. Pasadyn and Edgar [13] and Firth et al. [7] proposed to estimate contributions to biases states of each tool and product individually using previous runs and recombined them to determine recipe adjustments of future runs. However, in each production run, one product is manufactured on one tool. It can be shown that absolute individual states of products and tools are not observable. Pasadyn and Edgar [13] include the qualifying wafers and assumed that the performance of qualifying wafer is equivalent to tool performance. However, there is always the dilemma of reduced throughout capacity if there are frequent runs of non-product test wafers

* Corresponding authors. Tel.: +886 3 571 3697; fax: +886 3 571 5408.
E-mail addresses: dshwong@che.nthu.edu.tw (D.S.-H. Wong), ssjang@mx.nthu.edu.tw (S.-S. Jang).

0959-1524/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jprocont.2008.07.009

Nomenclature

a	absolute states vector
a_n^{tool}	absolute states for tools
a_m^{prod}	absolute states for products
b	process gain
$\mathbf{C}$	incidence matrix
d	number of products which are offline
I	identity matrix
k	time
N	number of tools
M	number of products
$\mathbf{O}$	observation matrix
p_m	ANOVA states of products
$\mathbf{P}$	covariance matrix of the ANOVA states
$\mathbf{Q}$	covariance matrix of the process noise
$\mathbf{R}$	covariance matrix of the measurement noise
$\mathbf{T}$	transition matrix
u	manipulating variable
v	prediction error
$\mathbf{W}$	weighting matrix
y	output quality variable
$\tilde{y}$	estimated thread
$\hat{y}$	observed thread
$\mathbf{Y}$	process bias vector
$\mathbf{Z}$	measurement matrix

Greek letters

$\boldsymbol{\alpha}$	ANOVA states vector
δ	Kronecker delta
ε	white noise
η	IMA (1,1) time series
μ	grand average of ANOVA states
σ	variance
τ	ANOVA states of tools
$\boldsymbol{\upsilon}$	measurement noise
ω	process noise
$\boldsymbol{\Phi}$	variance of prediction error

and reduced sensitivity of the identification algorithm if only a few qualifying wafers are used. Firth et al. [7] proposed a least square method known as just-in-time adaptive disturbance estimation (JADE) which includes additional constraints that the product and tool states remained unchanged from run-to-run and a proprietary weighting method. Wang et al. [21] gave a detailed discussion about the observability of non-threaded state estimation problem in high-mix production and showed unbiased observations of the overall state can be obtained.

In statistics, the problem of identification of different bias factors has been described as the analysis of variance (ANOVA) [12]. ANOVA has also been applied to semiconductor industries in many different areas such as control chart build-up [16] and feedback variable selections [14]. Vanli et al. [19] examined the problem of model selections in state identification of mixed run plant. The model structure involves an ANOVA model of main and interaction effects and auto-regressive dynamic terms that are specific to each thread. Again the emphasis is on the correct tracking of individual overall states for run-to-run control purposes. In this work, a novel state estimation method based on ANOVA is developed to estimate the difference of each product from the average of all products and difference of each tool to the average of all tools in the fab. It is shown that the relative states of each tool and product can be determined by introducing the ANOVA constraints after sufficient data were amassed. This method is formulated in a form of recursive state estimation using Kalman filter. The advantages of the proposed method are demonstrated using simulations to show that the correct ANOVA states can be estimated in production scenarios such as tool/product-shift, tool-drift, product ramp-up, tool/product-offline and preventive maintenance. Furthermore, application of this state estimation method in a deadbeat RtR control scheme shows that substantial improvements in process capability can be gained for specialized small lot counts products. Application to an industrial example will also be presented.

2. State estimation based on ANOVA

2.1. Linear plant with tool and product biases

Consider a multi-tool ($n = 1,\ldots,N$) multi-product ($m = 1,\ldots,M$) operation. Let us assume that the output of the kth run, which produces a product n_k on the tool m_k, is equal to

$$y_k = bu_k + a_{n_k}^{tool} + a_{m_k}^{prod} + \varepsilon_k. \tag{1}$$

Here, b is the process gain which relate the change in manipulating variable u_k to the change in output quality variable y_k; and $a_{n_k}^{tool}$ and $a_{m_k}^{prod}$ are two disturbance parameters associated with biases of the specific tool and product respectively. $\{\varepsilon_k\}$ is assumed to be a white noise process (i.e., $E[\varepsilon_k] = 0,\ \mathrm{Var}[\varepsilon_k] = \sigma^2,\ \mathrm{Cov}[\varepsilon_k, \varepsilon_{k+j}] = 0$ for $j \neq 0$). For simplicity, let us assume that b is a known constant. In the above model, it is assumed that there exist no interactions between tools $a_n^{tool}(n = 1, 2, \ldots, N)$ and parts $a_m^{prod}(m = 1, 2, \ldots, M)$. Given a set of historical production records, $(y_k, u_k, n_k, m_k,\ k = 1, \ldots, K)$ we can define an observed bias vector as

$$\hat{\mathbf{Y}} = \begin{bmatrix} y_1 - bu_1 \\ \vdots \\ y_k - bu_k \end{bmatrix} = \mathbf{Ca} = \begin{bmatrix} \delta_{1,n_1} & \ldots & \delta_{N,n_1}, \delta_{1,m_1} \ldots \delta_{M,m_1} \\ & \vdots & \\ \delta_{1,n_K} & \ldots & \delta_{N,n_K}, \delta_{1,m_K} \ldots \delta_{M,m_K} \end{bmatrix} \begin{bmatrix} a_1^{tool} \\ \vdots \\ a_N^{tool} \\ a_1^{prod} \\ \vdots \\ a_M^{prod} \end{bmatrix}, \tag{2}$$

where $\mathbf{C}$ is an incidence matrix in which $\delta_{n,n_k},\ \delta_{m,m_k}$ are Kronecker deltas, equal to 1 when n and m are equal to the tool n_k and product m_k of record k. Since each record contains one product being produced on one tool, we have

$$\sum_{j=1}^{n} \mathbf{C}_{ij} = 1 \qquad \sum_{j=N+1}^{N+M} \mathbf{C}_{ij} = 1. \tag{3}$$

The column rank of $\boldsymbol{C}$ is $N + M - 1$. Hence the $N + M$ absolute states of individual tool and parts cannot be estimated in an unbiased manner. If the Moore–Penrose pseudo-inverse of $\boldsymbol{C}$ is used:

$$\tilde{\mathbf{a}} = \mathbf{C}^{\sim 1}\hat{\mathbf{Y}}. \tag{4}$$

The value of $\tilde{\mathbf{a}}$ obtained corresponds to a least square fit estimate that would be biased. Firth et al. [7] suggested that given a set of past estimates of $\tilde{\mathbf{a}}_{t-1}$

$$\begin{bmatrix} \hat{\mathbf{Y}} \\ \tilde{\mathbf{a}}_{t-1} \end{bmatrix} = \mathbf{C}_1\tilde{\mathbf{a}}_t = \begin{bmatrix} \mathbf{C} \\ \mathbf{I} \end{bmatrix}\tilde{\mathbf{a}}_t. \tag{5}$$

A new estimate of $\tilde{\mathbf{a}}_t$ is obtained by a weighted least square solution:

$$\tilde{\mathbf{a}}_t = (\mathbf{C}_1^T\mathbf{W}\mathbf{C}_1)^{-1}\mathbf{C}_1^T\mathbf{W}\hat{\mathbf{Y}}, \tag{6}$$

where $\boldsymbol{W}$ is the weighting matrix.

2.2. Analysis of variance (ANOVA)

According to ANOVA, the effects of different factors are expressed as:

$$y_k - bu_k = \mu + \tau_{n_k} + p_{m_k} + \varepsilon_k, \tag{7}$$

where μ is the overall mean of all observed tool and product combinations, τ_n ($n = 1,\ldots,N$) represent the difference between the average results of all possible products on nth tool and the overall mean, and p_m ($m = 1,\ldots,M$) represent the difference between the average results on all possible tools of the mth product and the overall mean. Unlike a_n^{tool} and a_m^{prod}, which are absolute states of the particular tool and product, τ_m and p_m are relative contributions subject to the constraints [12]

$$\sum_{n=1}^{N} \tau_n = 0 \quad \sum_{m=1}^{M} p_m = 0. \tag{8}$$

Here, it is also assumed that there exist no interactions between tools τ_n ($n = 1, 2,\ldots,N$) and parts p_m ($m = 1, 2,\ldots,M$). In real practice, it should be noted that the output y is not only a function of present tool and product, but may also be influenced by other factors such as the specific tool used in the previous step. However, to include these factors is only a simple extension of this work. The matrix form of Eq. (7) is

$$\hat{\mathbf{Y}} = \mathbf{Z}\boldsymbol{\alpha} = \begin{bmatrix} 1 & \delta_{1,n_1} & \ldots & \delta_{N,n_1} & \delta_{1,m_1}\ldots\delta_{M,m_1} \\ & & \vdots & & \\ 1 & \delta_{1,n_K} & \ldots & \delta_{N,n_K} & \delta_{1,m_K}\ldots\delta_{M,m_K} \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \vdots \\ \tau_N \\ p_1 \\ \vdots \\ p_M \end{bmatrix}. \tag{9}$$

For the new incidence matrix $\boldsymbol{Z}$, there are $N + M + 1$ columns, but since

$$\sum_{j=2}^{N+1} \mathbf{Z}_{ij} = 1 \quad \sum_{j=N+2}^{N+M+1} \mathbf{Z}_{ij} = 1 \tag{10}$$

the actual column rank is $N + M - 1$, i.e. it is rank deficient by 2. However, given the two ANOVA constraints (8) we have

$$\hat{\mathbf{Y}}' = \begin{bmatrix} \hat{\mathbf{Y}} \\ 0 \\ 0 \end{bmatrix} = Z'\boldsymbol{\alpha} = \begin{bmatrix} & \mathbf{Z} & \\ 0 & \underbrace{1\ldots1}_{N} & \underbrace{0\ldots0}_{M} \\ 0 & \underbrace{0\ldots0}_{N} & \underbrace{1\ldots1}_{M} \end{bmatrix} \boldsymbol{\alpha} \tag{11}$$

which means that the equality constraints (8) should be satisfied when ANOVA states are updated each time. It should be pointed out that although the number of possible threads is NM, the possible number of different rows is greater than the number of independent columns. However, the actual number of threads is usually much less because some products with small lot counts may not be spread to all tools. There should be more than $N + M - 1$ threads to make $\mathbf{Z}'$ full rank so that an unbiased estimate of $\boldsymbol{\alpha}$ can be obtained. Furthermore, in some fabs, certain tools may be dedicated to produce certain products and vice versa. We must ensure that such groupings are treated as a separate system.

2.3. ANOVA state space model

If we assume that the ANOVA states are stationary over several periods of time, then ANOVA model of the multi-tool and multi-product plant can be expressed in the following state space form:

$$\boldsymbol{\alpha}_{t+1} = \mathbf{T}\boldsymbol{\alpha}_t + \boldsymbol{\omega}_t, \tag{12}$$

$$\hat{\mathbf{Y}}'_t = \mathbf{Z}'\boldsymbol{\alpha}_t + \boldsymbol{v}_t, \tag{13}$$

where ω_t and $\boldsymbol{v}_t$ are independent, zero-mean, Gaussian noise processes of covariance matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$, respectively. $\boldsymbol{T}$ is the transition matrix

$$\mathbf{T} = \begin{bmatrix} 1 & 0_{1\times N} & 0_{1\times M} \\ 0_{N\times 1} & 1_{N\times N} & 0_{N\times M} \\ 0_{M\times 1} & 0_{M\times N} & 1_{M\times M} \end{bmatrix}. \tag{14}$$

The observability matrix for the above ANOVA state space model is

$$\mathbf{O} = (\mathbf{Z}', \mathbf{Z}'\mathbf{T} \cdots \mathbf{Z}'\mathbf{T}^{N+M})^{\mathrm{T}}. \tag{15}$$

The system is observable if the observability matrix is full rank. In this case, the transition matrix $\mathbf{T}$ is an identity matrix, the observability matrix $\boldsymbol{O}$ is equal to the output matrix $\mathbf{Z}'$ which is of full rank $N + M + 1$.

2.4. Recursive estimation

Estimation can then be carried out in a recursive manner from interval to interval. At the start of any time interval t, given an estimated ANOVA state vector $\tilde{\boldsymbol{\alpha}}_{t-1}$ and a estimated covariance matrix of the ANOVA states $\tilde{\mathbf{P}}_{t-1}$, then the predicted values of the ANOVA state vector $\tilde{\boldsymbol{\alpha}}_{t|t-1}$ and predicted the covariance matrix for this period $\tilde{\mathbf{P}}_{t|t-1}$ are given by

$$\tilde{\boldsymbol{\alpha}}_{t|t-1} = \mathbf{T}\tilde{\boldsymbol{\alpha}}_{t-1}, \tag{16}$$

$$\tilde{\mathbf{P}}_{t|t-1} = \mathbf{T}\tilde{\mathbf{P}}_{t-1}\mathbf{T}^{\mathrm{T}} + \mathbf{Q}. \tag{17}$$

After the operating records $(y_{t,k}, u_{t,k}, n_{t,k}, m_{t,k})$, $k = 1 \ldots K_t$ of this period are collected, the minimum mean square estimator of the ANOVA states and the covariance matrix can be updated by the following equations (e.g., [5]):

$$\hat{\boldsymbol{\alpha}}_t = \tilde{\boldsymbol{\alpha}}_{t|t-1} + \tilde{\mathbf{P}}_{t|t-1}\mathbf{Z}_t'^{\mathrm{T}}\boldsymbol{\Phi}_t^{-1}(\hat{\mathbf{Y}}_t' - \mathbf{Z}_t' - \tilde{\boldsymbol{\alpha}}_{t|t-1}), \tag{18}$$

$$\tilde{\mathbf{P}}_t = \tilde{\mathbf{P}}_{t|t-1} - \tilde{\mathbf{P}}_{t|t-1}\mathbf{Z}_t'^{\mathrm{T}}\boldsymbol{\Phi}_t^{-1}\mathbf{Z}_t'\tilde{\mathbf{P}}_{t|t-1}, \tag{19}$$

$$\boldsymbol{\Phi}_t = \mathbf{R} + \mathbf{Z}_t'\mathbf{P}_{t|t-1}\mathbf{Z}_t'^{T}. \tag{20}$$

In these circumstances a set of prediction error or 'innovations'

$$\boldsymbol{v}_t = \hat{\mathbf{Y}}_t' - \mathbf{Z}_t'\tilde{\boldsymbol{\alpha}}_{t|t-1} \tag{21}$$

are produced by the Kalman filter. These are independently and normally distributed with mean zeros and variance $\sigma^2\boldsymbol{\Phi}_t$ if the process follows (12) and (13).

It is not possible to guarantee that $\mathbf{Z}_t'$ contains all the threads and of full rank $N + M + 1$ during the data collection interval. For example, in the extreme case, the model can be updated whenever the result of a single run is reported. However, it is important to ensure that $\mathbf{Z}'$ over an extended history is of full rank $N + M + 1$. This assumption may be invalid if some products are terminated or a certain tool is offline for an extended period of time. The assurance of the above full rank assumption can be achieved by monitoring the condition number of the matrix $\tilde{\mathbf{P}}_t$.

2.5. Run-to-run control

The control objective is to maintain the process output as close to target as possible. Given a set of predicted states $\tilde{\boldsymbol{\alpha}}_{t|t-1}$, the dead-beat control action in the period t, before new data arrive are given by:

$$u_{t,k} = \frac{T_k - \tilde{\mu}_t - \tilde{\tau}_{t,n_k} - \tilde{P}_{t,m_k}}{b}. \tag{22}$$

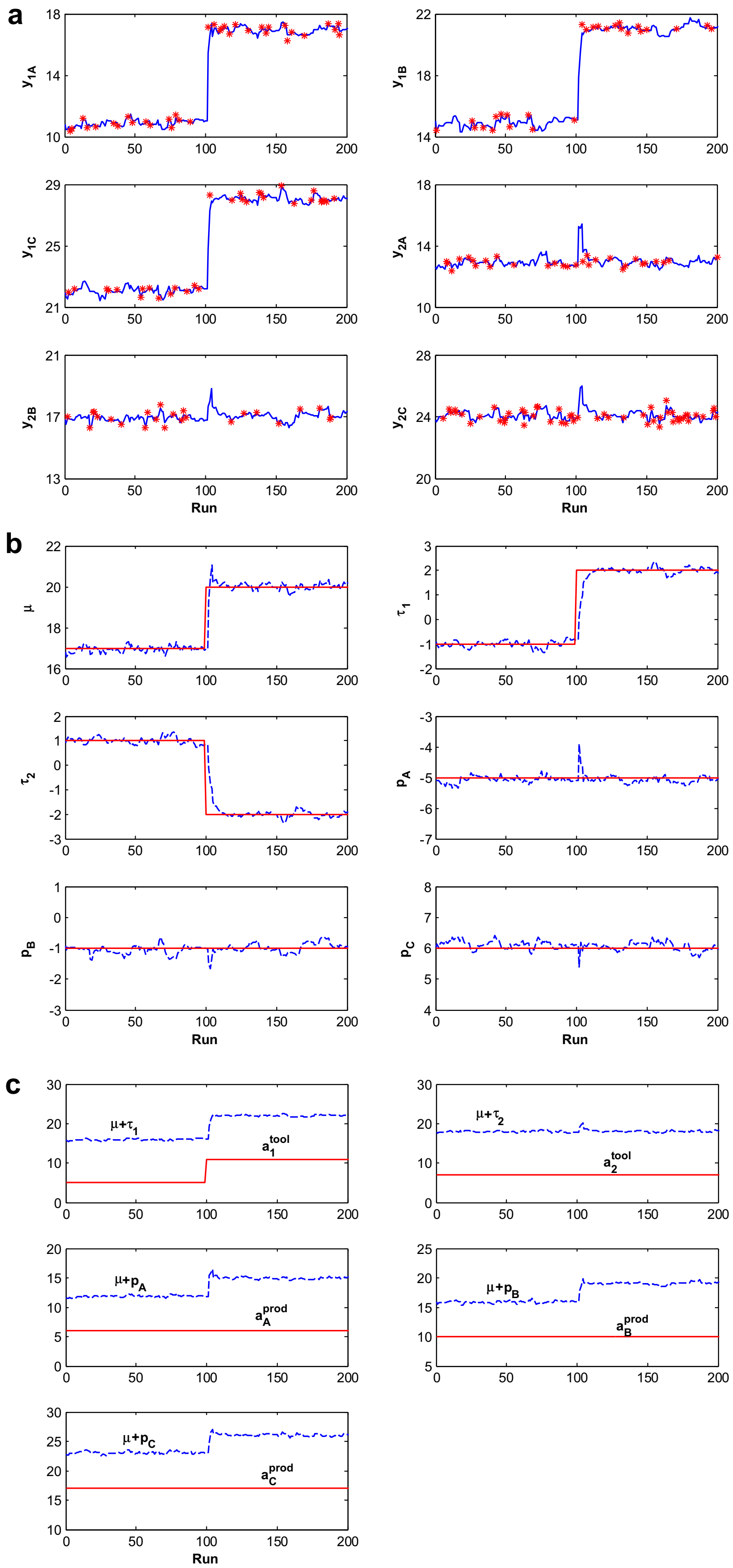

Fig. 1. Changes in actual and estimated (a) states of thread, (b) ANOVA parameters and (c) absolute and relative factor biases in a tool shift. (— actual values, --- estimated values, ∗ observed values).

3. Results and discussion

3.1. Simulation example

In this section, a series of simulation tests are designed to investigate the effectiveness of the proposed algorithm in various operation scenarios. A simple two-tool (1, 2) and three-product (A, B, C) example is used in the following simulation studies. Each tool and product is assigned a unique bias. Therefore, there are six ANOVA states in the model. The initial states of the ANOVA states are assumed to be on target before the start of the simulations in different operation scenarios. The tool and product adopted for each run is randomly selected based on a given probability of occurrence. The probability distributions of two tools are 0.6 and 0.4, respectively. The probability distributions of product A, B, and C for the examples of Sections 3.2–3.5 are 0.4, 0.3 and 0.3, respectively.

In these simulations, the state noise and measurement noise are white noise with standard deviation 0.1. The adjustable parameters of the observer $\boldsymbol{Q}$ and $\boldsymbol{R}$ can be obtained by offline analysis in real applications. They could be different at different production stages. However, they remained the same in the simulation examples below. It is also assumed that there is no metrology delay and the states are updated whenever the measurement arrives.

For each operation scenarios, three different comparisons were made. First estimated ANOVA states $\tilde{\boldsymbol{\alpha}}$ are compared with the actual ANOVA states $\boldsymbol{\alpha}$. The estimated thread states $\tilde{y}_{mn,k}$ and observed performance of all the threads $\hat{y}_k$ are also compared. The actual biases of each tool a_n^{tool} and products a_m^{prod} are compared with $\mu + \tau_n$, the average performance of all products on the nth tool, and $\mu + p_m$, the average performance of the mth products on all tools. Note that it is easy to identify the changes of tools or products using the ANOVA estimator proposed in this paper. If the ANOVA states of the tools or products remain stable, the corresponding conditions of the tools or products are confirmed to be unchanged. In case of a change of the ANOVA state are observed, the change of $\tilde{\mu} + \tilde{\tau}_n$ or $\tilde{\mu} + \tilde{p}_m$ confirms the change of the tool n or the product m. Furthermore, the changes of $\tilde{\mu} + \tilde{\tau}_n$ or $\tilde{\mu} + \tilde{p}_m$ are consistent with the actual states a_n^{tool} and a_m^{prod}.

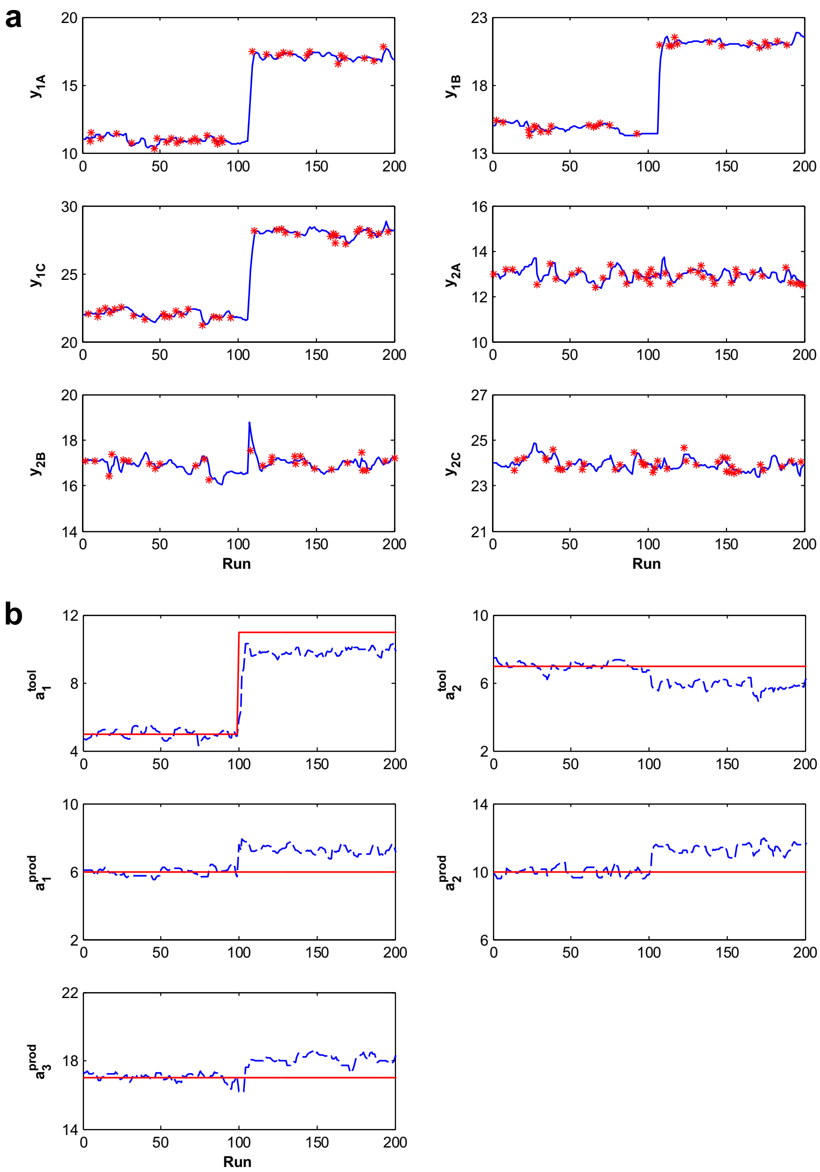

Fig. 2. Changes in actual and estimated (a) states of thread and (b) absolute factor biases using JADE in a tool shift. (— actual values, - - - estimated values, ∗ observed values).

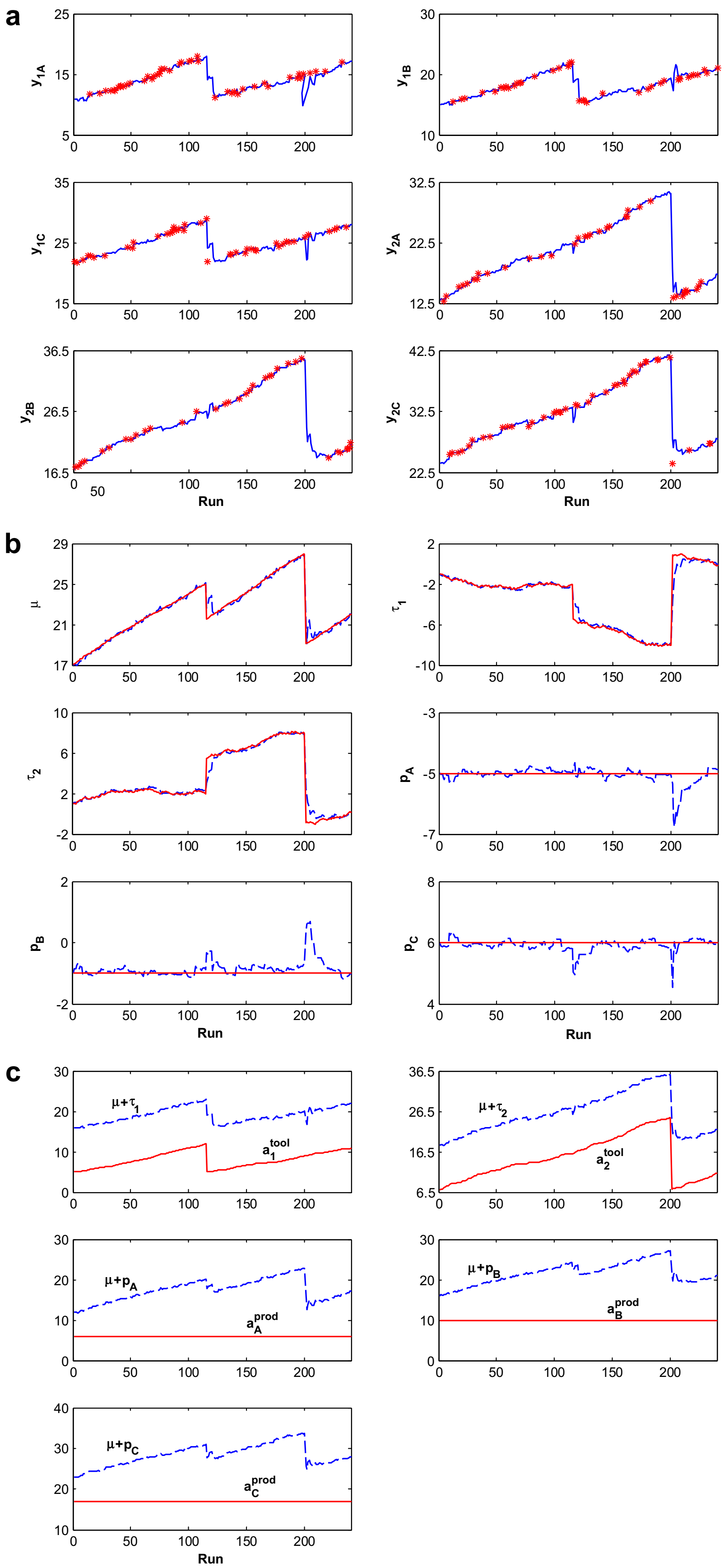

Fig. 3. Changes in actual and estimated (a) states of thread, (b) ANOVA parameters and (c) absolute and relative factor biases with tool drifts and preventive maintenances (— actual values, --- estimated values, ∗ observed values).

3.2. Tool shift

There are many events which can result in an apparent immediate shift in the operating conditions. For instance, such a disturbance might occur when a tool undergoes a maintenance event (e.g. [8] reported sudden change in trench depth after a wet clean operation). This event would be seen by the process as a step disturbance in the output variable. The disturbance is often not measurable. Therefore, the controller should learn from the process output and compensate for the effect of the disturbance.

The comparisons of actual and estimated ANOVA states, the actual and estimated states of each thread, and actual biases and relative bias of the each tool and products are shown in Fig. 1. In this case, there is an abrupt change for the bias value of tool 1 at the 100th run. In Fig. 1a, we observed that all the threads on tool 1 experience an abrupt change while the states of all the threads associated with tool 2 remain unchanged. In Fig. 1b, we found that the ANOVA states $\tilde{\mu}$ and $\tilde{\tau}_1$ experience positive shifts while the ANOVA tool state $\tilde{\tau}_2$ experience a shift in the other direction. In Fig. 1c, it is found that the average performance of all products on tool 1 $\tilde{\mu}+\tilde{\tau}_1$ experienced a shift; the average performance $\tilde{\mu}+\tilde{\tau}_2$ of all products on tool 2 remained unchanged while the average performance $\tilde{\mu}+\tilde{p}_{A,B,C}$ of all products on tool 2 experienced a shift too.

Fig. 2 illustrated the results of JADE estimates. In the simulation, an identity weighting matrix is used for comparison with ANOVA estimates. It is interesting to note that when a shift is induced to tool 1, shifts are also observed to other factors as shown in Fig. 2b. The biases obtained are not true estimates of these factors. However, the estimated states of all the threads are correct as shown in Fig. 2a since the method is a least square fit of all the threads. Hence, use of the recombined thread states for controlling existing threads is not a problem. Application of these individual factor states to estimate new thread may lead to errors.

3.3. Tool drift and PM

If a manufacturing process is known to drift due to equipment aging, then a deterministic drift exists in the system. Aging can be found in wafer etching process and chemical mechanical

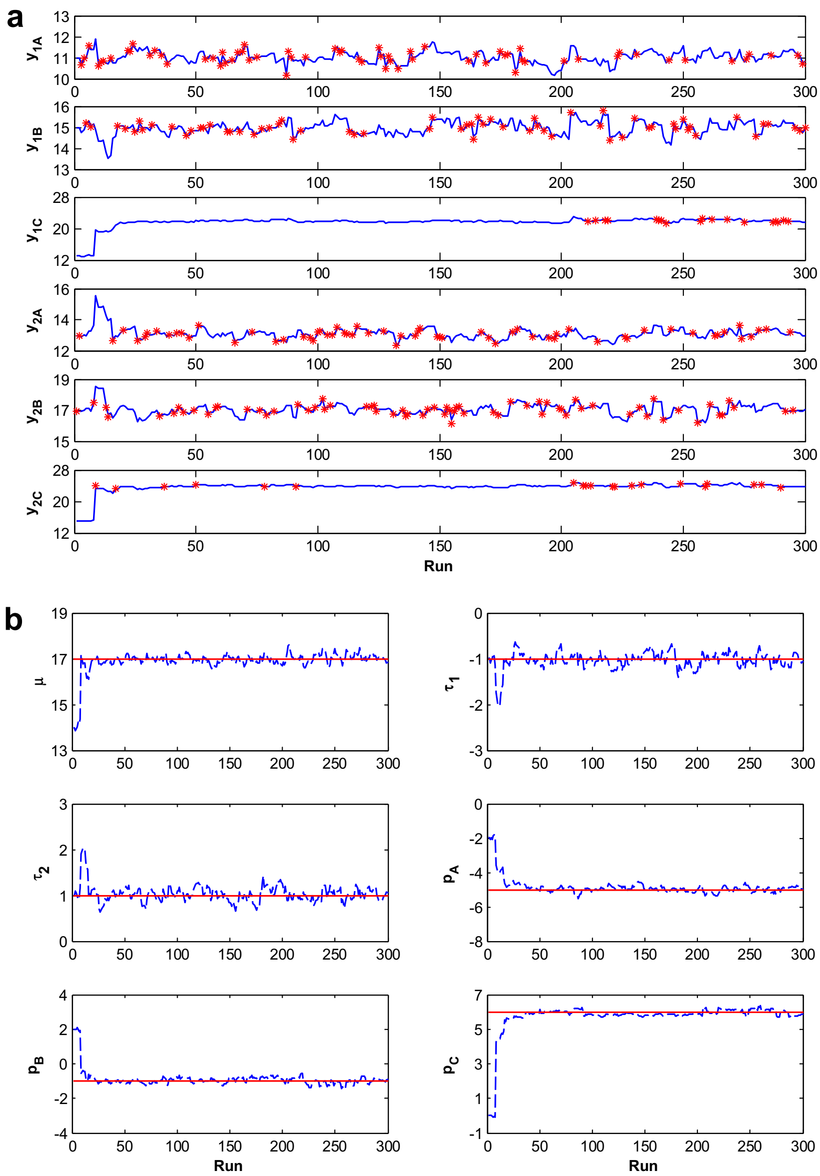

Fig. 4. Changes in actual and estimated (a) states of thread, (b) ANOVA parameters with product ramp-up (— actual values, - - - estimated values, * observed values).

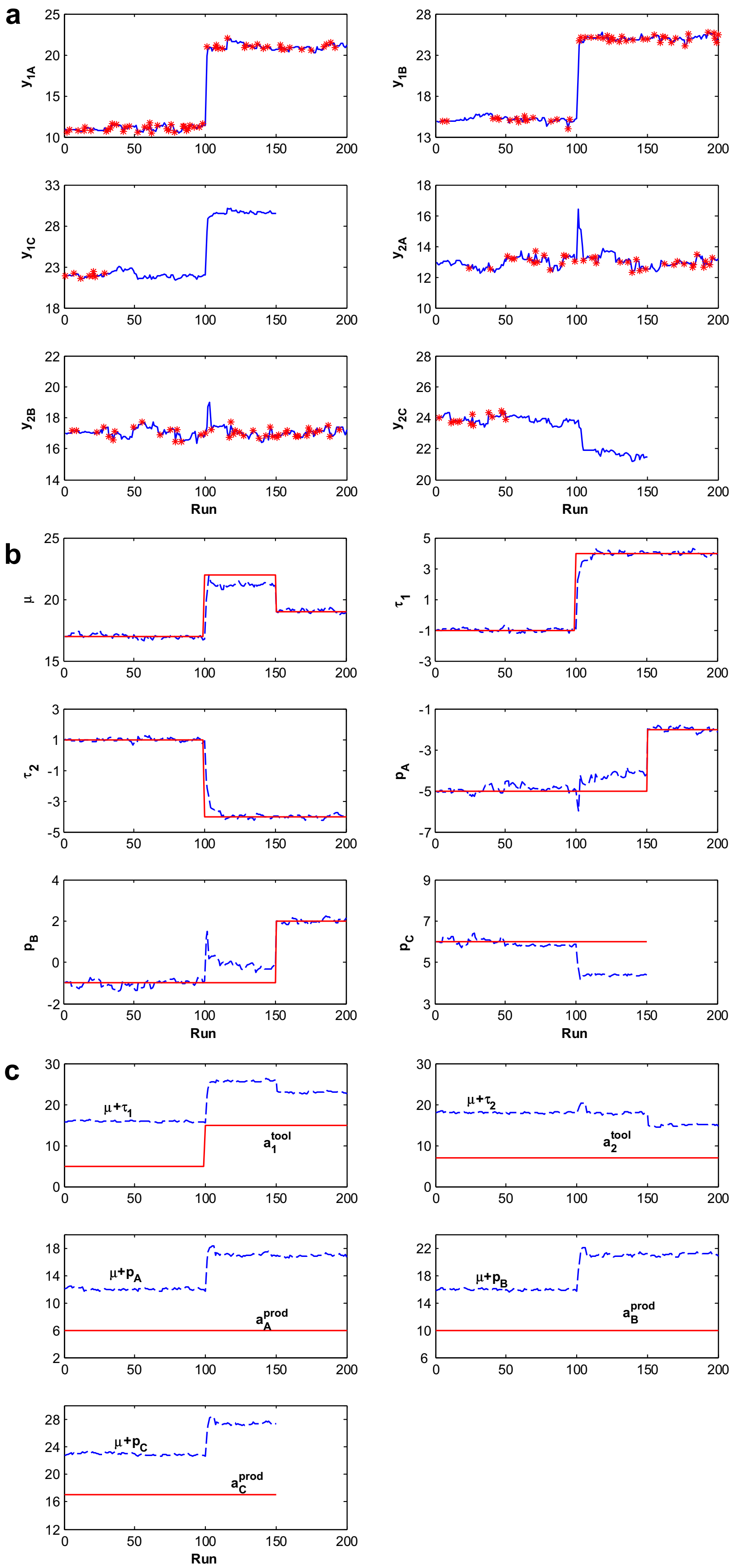

Fig. 5. Changes in actual and estimated states (a) states of thread, (b) ANOVA parameters and (c) absolute and relative factor biases with product offline (— actual values, - - - estimated values, ∗ observed values).

polishing [1,4,8]. A drift persisted for a long period would normally be followed by a maintenance event and corresponding process reset, resulting a saw-tooth pattern in an uncontrolled quality characteristic.

The simulation results are shown in Fig. 3. The two tools experienced deterministic drifts of slopes 0.1 and 0.2, respectively. Tool 1 was reset at the 115th run and tool 2 was reset at the 200th run. As shown in Fig. 3a and b, the states of each thread and the ANOVA states of each tool and product can be estimated correctly throughout the simulation. Furthermore, the ANOVA states of the products remained unchanged. Fig. 3c illustrated that the average performances $\tilde{\mu}+\tilde{\tau}_{1,2}$ of tool 1 and tool 2 show patterns that are consistent with the actual states. The average performances $\tilde{\mu}+\tilde{p}_{A,B,C}$ of the products A–C show saw-tooth patterns that are consistent with the changes in average performance of the whole plant. Since there are no changes in the relative performance of different product, the ANOVA product states $\tilde{p}_{A,B,C}$ remained unchanged as shown in Fig. 3b.

3.4. Product ramp-up

In a foundry, commissioning of new products is a regular activity. Before a new product is produced in large quantities, it is usually test tried in small quantities on a limited number of tools. If they are deemed satisfactory, large orders will be placed and the product will spread to other tools. Optimum recipes on different tools are expected in a short period. This phenomenon is called ramp-up in the semiconductor industry.

Simulation results for this case are shown in Fig. 4. In the first 100 runs, product C, namely the new product, was produced on tool 2 in small quantity. In the second 100 runs, the product C did not appear and it was produced in large quantity in the third 100 runs. It is assumed that no information about the ANOVA state of the new product is available before it is produced. The initial ANOVA state of product C is set to be zero $\tilde{p}_c = 0$. Once the new product is produced, a value of $\tilde{p}_C$ is estimated. It should be noted that values of $\tilde{p}_A$ and $\tilde{p}_B$ and $\tilde{\mu}$ also change as product C is introduced to the system (Fig. 4a). The correct thread performances of C on both tool 1 and tool 2 can be estimated as soon as the production of C ramped up (Fig. 4b).

3.5. Product offline

In a semiconductor foundry, as many new products may be commissioned, many old products may be terminated from production. When a product is stopped being produced and the corresponding state remains in the system, the system will be unobservable. At this time, the ANOVA states will be biased when a disturbance happens.

Simulations of such an operation scenario are presented in Fig. 5. The production of C is terminated at the 50th run. Although the estimates of thread performances are correct (Fig. 5a), estimates of the ANOVA parameters p_A, p_B, and p_C become biased if large changes occur, e.g. at 100th run, a tool shift is found in tool 1 (Fig. 5b).

It should be noted that there may be hundreds of products in a fab and the operator probably does not know whether certain product has been offline for a long time so that the ANOVA states estimates may become biased. This can be done by monitoring the changes in condition number of the matrix $\tilde{\mathbf{P}}_t$ over time. As shown in Fig. 6, the condition number of the matrix $\tilde{\mathbf{P}}_t$ increases rapidly when the product C is offline.

It is therefore desirable to remove products that will be permanently offline from the list. To do so, a set of new ANOVA parameters need to be estimated. This can be done by equating threads estimated by the new ANOVA parameters to the product effects estimated by the existing ANOVA parameters. The procedure is shown for the example in which the product C is removed from a 2-tool-3-product system to give the ANOVA parameters of a 2-tool-2-product system:

$$\begin{bmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \mu' \\ \tau'_1 \\ \tau'_2 \\ p'_A \\ p'_B \end{bmatrix} = \begin{bmatrix} \tilde{\mu}+\tilde{\tau}_1+\tilde{p}_A \\ \tilde{\mu}+\tilde{\tau}_1+\tilde{p}_B \\ \tilde{\mu}+\tilde{\tau}_2+\tilde{p}_A \\ \tilde{\mu}+\tilde{\tau}_2+\tilde{p}_B \\ 0 \\ 0 \end{bmatrix}, \tag{23}$$

where $[\mu' \ \tau'_1 \ \tau'_2 \ p'_A \ p'_B]^T$ is the new ANOVA state vector with product C removed and $[\tilde{\mu} \ \tilde{\tau}_1 \ \tilde{\tau}_2 \ \tilde{p}_A \ \tilde{p}_B \ \tilde{p}_C]^T$ is the current state vector. Fig. 5b show that unbiased estimations of the new ANOVA states can be

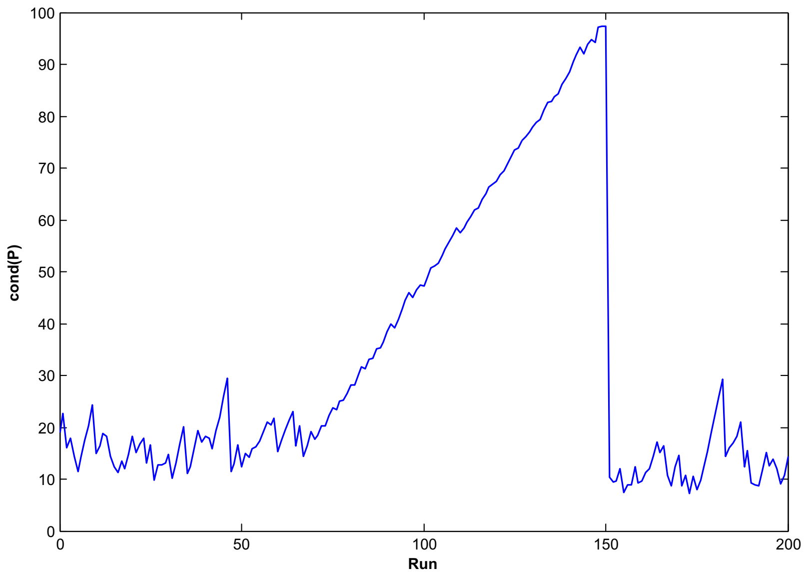

Fig. 6. Changes in condition number of matrix $[\tilde{P}_t]$ when a product is offline.

Table 1
Comparison of RtR control performances for different products

Method	MSE_A	MSE_B	MSE_C	MSE
ANOVA	0.2400	0.2833	0.4124	0.2642
JADE	0.2557	0.3182	0.5423	0.2964
Threaded EWMA	0.2746	0.4103	2.2348	0.4152

found if product C is removed from the list at the 150th run. As soon as the product is removed, the condition number of the matrix $\tilde{\mathbf{P}}_t$ becomes normal again.

3.6. Controller performance

In a highly-mixed foundry, some products are fabricated infrequently in small quantities. However, it should be noted that products which are produced with large quantity are usually of marginal profits and the products which are produced infrequently are often high-value added and contribute a substantial portion of the profit. Therefore, it is highly desirable for a control algorithm to have comparable performance for products with different run counts.

In this section, the performance for "infrequent" products for the three control algorithm, threaded EWMA algorithm with ($\lambda = 0.2$), JADE and the ANOVA method proposed in this paper are investigated. A two-tool-three-product simulation example is used. Runs are evenly distributed between the two tools, and the probability distributions of products A, B, C are 60%, 35% and 5%, respectively. The fluctuations in conditions of the two tools are assumed to be two constant biases plus integrated moving average (IMA) stochastic time series:

$$a_{i,n_k}^{\text{tool}} = c_{n_k} + \eta_{i,n_k},$$

$$\eta_{i,n_k} = \eta_{i-1,n_k} + \varepsilon_{i,n_k} - \theta_{n_k}\varepsilon_{i-1,n_k}, \tag{24}$$

where i is an index of the number of runs on the tool n_k. c_{n_k} is constant. η_{i,n_k} is an IMA(1,1) disturbance which represents the dynamic behavior of the tool. $\varepsilon_n \epsilon N(0,\sigma^2)$ is Gaussian distributed random

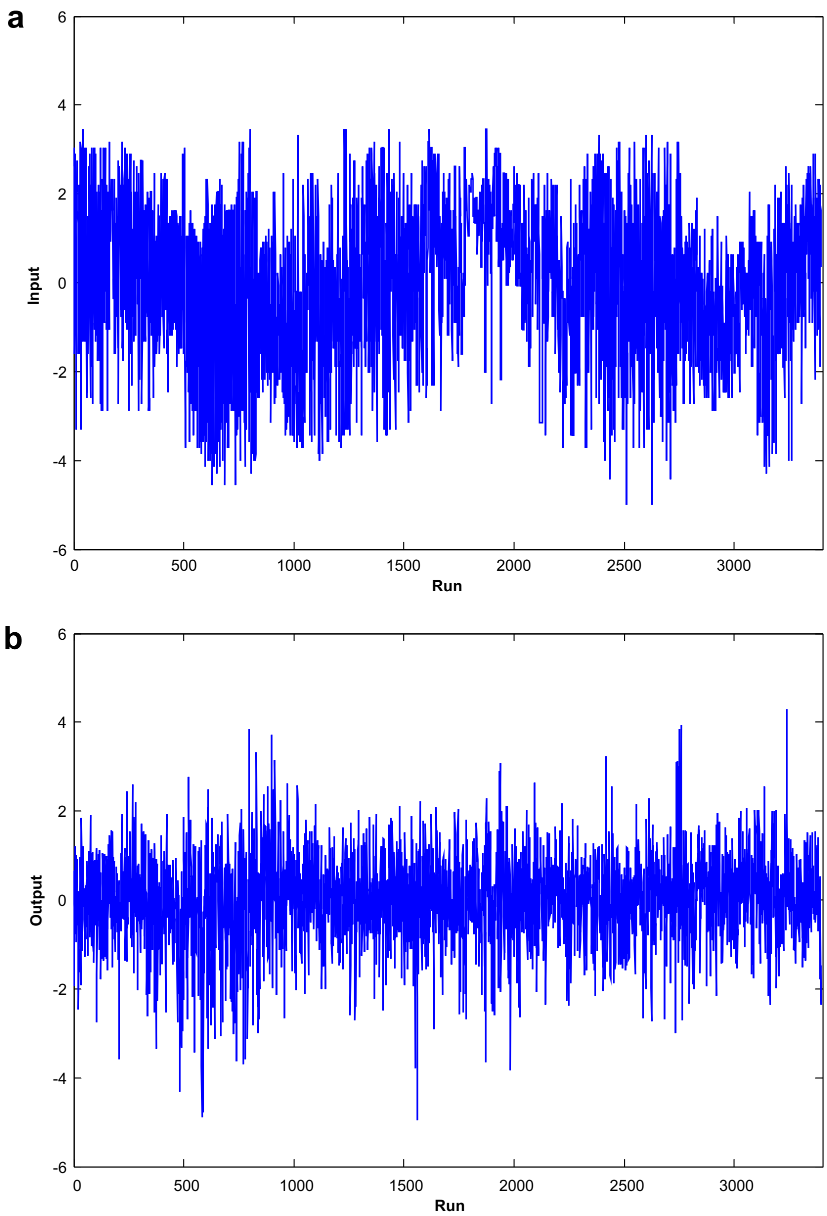

Fig. 7. Changes output and manipulated variables of a mixed run plant (a) output and (b) input.

noise with zero mean and variance $\sigma^2 = 0.16$. θ_{n_k} are assumed to be 0.4 for both tools. If $\theta = 1$ the disturbance is white noise, if θ = zero, the disturbance is a random walk. The mean squared errors (MSE) for different products are used as the performance index:

$$\text{MSE}_m = \frac{\sum_k \delta_{m,m_k} (T_k - y_k)^2}{\sum_k \delta_{m,m_k}}. \tag{25}$$

The simulation results are shown in Table 1. For all three control algorithms, it can be found that Product A, which is the most frequent produced product, had the best performance, while Product C, which is the produced in small quantities, had the worst performance. It is not surprising that the performance of threaded EWMA for product C is poor because it is produced only occasionally and the condition of the tool may be quite different from the last run it is produced. The performance of threaded EWMA for products A and B are acceptable due to their higher running frequencies. The performance of JADE for all three products is acceptable. The ANOVA result is the best among the three methods for all three products with different run counts. This makes the proposed method an appropriate candidate for high-mix production.

4. Industrial example

Test results of application of the method outlined in the previous sections of this paper to an industrial shallow trench isolation (STI) process is presented here. The STI is used to prevent electrical current leakage between adjacent semiconductor device components. One of the key steps in STI is the use of a reactor ion etch process to etch the trench. The resulting trench depth variation due to seasoning of the etch chamber; and use of etching time as the manipulated variable in run-to-run control to control the trench depth have been reported by several authors (e.g. [8,9,10,20]. However, the effects of different chambers on different products are not the same. Hence, an ANOVA based controller was used to estimate the etch time required for different products on different chambers.

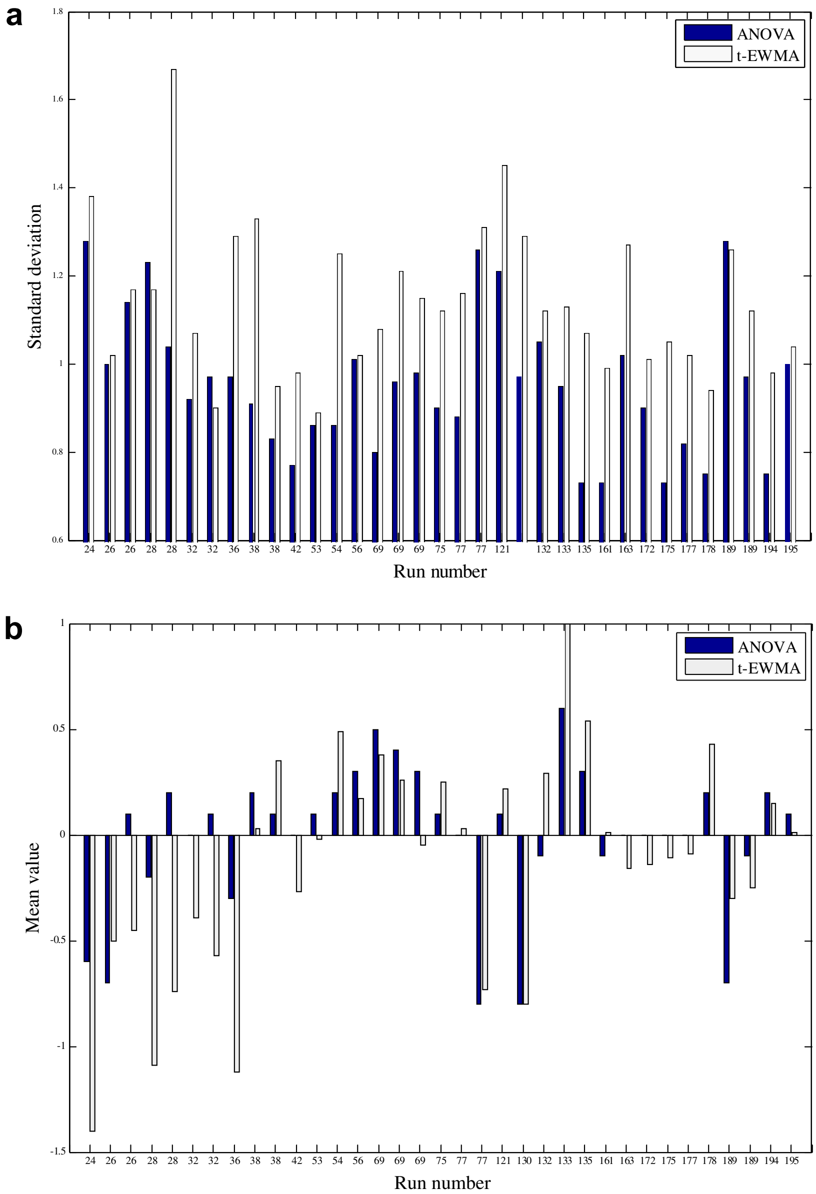

Fig. 8. Comparison of plant performance of different products with stimulated threaded EWMA control (a) standard deviations vs. run counts and (b) mean offsets vs. run counts.

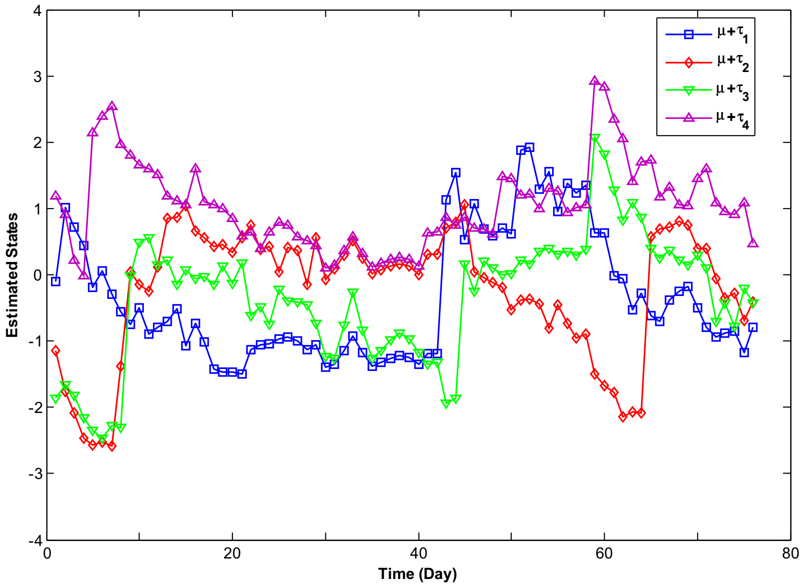

Fig. 9. Estimated relative states of 4 different tools by the proposed ANOVA method.

The example includes 3393 runs during a time period of 76 days. There were four tools (etch chambers), 35 products and 134 threads (combination of etch-chamber and products). The data was standardized to have overall mean zero and overall variance 1. The controlled output (trench depth) and manipulated input (etch time) are plotted in Fig. 7. It is found that the run-to-run correlations of controlled output is very small (Fig. 7a) because such run-to-run variations are eliminated by more systematic variations in the manipulated input (Fig. 7b). A simulation is carried out to evaluate the performance of the plant under threaded EWMA control. It was found that the optimal output is achieved at $\lambda = 0.2$. The overall standard deviation of the plant is 1.19 which is about 19% higher than the ANOVA control.

Fig. 8 shows the standard deviations and means of each of the 35 products obtained using EWMA simulation and plant results using ANOVA. Each product is labeled by its run counts during this period. It is found that the standard deviations (Fig. 8a) of and the mean offset (Fig. 8b) of different products are generally higher for threaded EWMA control, especially so for products with low run counts. This shows that control actions based on estimated ANOVA states is an effective framework for RtR implementation in a high-mixed production.

It is also of interest to study the behaviors of tools. The variation of corresponding four ANOVA tool states are shown in Fig. 9. There was a maintenance event around the 40th day which caused a shift for the four tool states. The state values were closed to each other after maintenance, indicating that the chambers conditions are relatively the close. However, Tool 2 began to deteriorate and differs substantially from other tools. A maintenance procedure was executed at the 63rd day. The performance of the tools become close to each other again. This shows that the state estimation method is very valuable in monitoring the health of individual tools and as well relative differences among similar tools.

5. Conclusion

In this paper a novel state estimation method based on statistics method ANOVA is developed to estimate the relative states of each product and the relative states of each tool to the grand average of this station in the fab. The method is formulated in a form of recursive state estimation using Kalman filter. Simulation results show that the correct ANOVA states can be estimated in production scenarios such as tool-shift, tool-drift, product ramp-up and offline, or other non-stationary stochastic disturbances. Furthermore, application of this state estimation method in a RtR control of trench depth by etch time shows that substantial improvement of quality of products with small run counts. This makes the proposed method highly suitable for mixed product control system.

Acknowledgement

The authors thank the financial support provided by National Science Council through the grant NSC95-2221-E-007-027.

References

[1] S.W. Bulter, J.A. Stefani, Supervisory run-to-run control of polysilicon gate etch using in situ ellipsometry, IEEE Trans. Semicond. Manuf. 7 (1994) 193–201.
[2] E.D. Castillo, A.M. Hurwitz, Run-to-run process control: literature review and extensions, J. Quality Technol. 29 (1997) 184–196.
[3] E.D. Castillo, Statistical Process Adjustment for Quality Control, John-Wiley and Sons, New York, 2002.
[4] A. Chen, R.S. Guo, Age-based double EWMA controller and its application to CMP processes, IEEE Trans. Semicond. Manuf. 14 (2001) 11–19.
[5] D.B. Duncan, S.D. Horn, Linear dynamic regression estimation from the viewpoint of regression analysis, J. Am. Statist. Assoc. 67 (1972) 815–824.
[6] T.F. Edgar, S.W. Butler, W.J. Campbell, C. Pfeiffer, C. Bode, S.B. Hwang, K.S. Balakrishnan, J. Hahn, Automatic control in microelectronics manufacturing: practices, challenges, and possibilities, Automatica 36 (11) (2000) 1567–1603.
[7] S.K. Firth, W.J. Campbell, A. Toprac, T.F. Edgar, Just-in-time adaptive disturbance estimation for run-to-run control of semiconductor processes, IEEE Trans. Semicond. Manuf. 19 (2006) 298–315.
[8] S. Gaddam, M.W. Braun, Etch chamber condition-based process control model for shallow trench isolation trench depth control, in: Proceedings, 2005 IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop – Advancing Semiconductor Manufacturing Excellence, 2005, pp. 17–20.
[9] P. Jowett, V. Morozov, Shallow trench isolation run-to-run control project at Infineon technologies Richmond, in: IEEE/SEMI Advanced Semiconductor Manufacturing Conference, Advancing the Science and Technology of Semiconductor Manufacturing, 2002, pp. 107–112.
[10] K. Miwa, T. Inokuchi, T. Takahashi, A. Akira Oikawa, K. Imaoka, Reduction of depth variation in an Si etching process by applying an optimized run-to-run control system, IEEE Trans. Semicond. Manuf. 18 (2005) 309–318.
[11] J. Moyne, Run-to-run Control in Semiconductor Manufacturing, CRC Press, Florida, 2001.
[12] D.C. Montgomery, Design and Analysis of Experiments, fourth ed., John Wiley & Sons, New York, 1997.
[13] A.J. Pasadyn, T.F. Edgar, Observability and state estimation for multiple product control in semiconductor manufacturing, IEEE Trans. Semicond. Manuf. 18 (2005) 592–604.

[14] O.D. Patterson, X.B. Dong, P.P. Khargonekar, V.N. Nair, Methodology for feedback variable selection for control of semiconductor manufacturing processes – Part 1: Analytical, and simulation results, IEEE Trans. Semicond. Manuf. 16 (2003) 575–587.

[15] S.J. Qin, G. Cherry, R. Good, J. Wang, C.A. Harrison, Semiconductor manufacturing process control and monitoring: a fab-wide framework, J. Process. Control 16 (2006) 179–191.

[16] G.C. Runger, J.W. Fowler, Run-to-run control charts with contrasts, Quality and Reliability Engineering International 14 (1998) 261–272.

[17] E. Sachs, A. Hu, A. Ingolfsson, Run by run process control: combining SPC and feedback control, IEEE Trans. Semicond. Manuf. 8 (1995) 26–43.

[18] A. Toprac, Y. Wang, Solving the high-mix control problem, in: AEC/APC Asia Conference, Hsinchu, Taiwan, 2004.

[19] O.A. Vanli, N.S. Patel, M. Janakiram, E.D. Castillo, Model context selection for run-to-run control, IEEE Trans. Semicond. Manuf. 20 (2007) 506–516.

[20] J. Wang, Q.P. He, S. Joe, C.A. Bode, M.A. Purdy, Recursive least squares estimation for run-to-run control with metrology delay and its application to STI etch process, IEEE Trans. Semicond. Manuf. 18 (2005) 309–318.

[21] J. Wang, Q.P. He, T.F. Edgar, A general framework for state estimation in high-mix semiconductor manufacturing, Proc. Am. Control Conf. (2007) 3636–3641.

[22] Y. Zheng, Q.H. Lin, D.S.H. Wong, S.S. Jang, K. Hui, Stability and performance analysis of mixed product run-to-run control, J. Process. Control 16 (2006) 431–443.