J Intell Manuf (2012) 23:2281–2294
DOI 10.1007/s10845-011-0572-y

Manufacturing intelligence to forecast and reduce semiconductor cycle time

Chen-Fu Chien · Chia-Yu Hsu · Chih-Wei Hsiao

Received: 1 August 2010 / Accepted: 20 June 2011 / Published online: 23 July 2011

Abstract Semiconductor manufacturing is one of the most complicated production processes with the challenges of dynamic job arrival, job re-circulation, shifting bottlenecks, and lengthy fabrication process. Owing to the lengthy wafer fabrication process, work in process (WIP) usually affects the cycle time and throughput in the semiconductor fabrication. As the applications of semiconductor have reached the era of consumer electronics, time to market has played an increasingly critical role in maintaining a competitive advantage for a semiconductor company. Many past studies have explored how to reduce the time of scheduling and dispatching in the production cycle. Focusing on real settings, this study aims to develop a manufacturing intelligence approach by integrating Gauss-Newton regression method and back-propagation neural network as basic model to forecast the cycle time of the production line, where WIP, capacity, utilization, average layers, and throughput are rendered as input factors for indentifying effective rules to control the levels of the corresponding factors as well as reduce the cycle time. Additionally, it develops an adaptive model for rapid response to change of production line status. To evaluate the validity of this approach, we conducted an empirical study on the demand change and production dynamics in a semiconductor foundry in Hsinchu Science Park. The approach proved to be successful in improving forecast accuracy and realigning the desired levels of throughput in production lines to reduce the cycle time.

Keywords Cycle time · Work in process (WIP) · Manufacturing intelligence · Gauss-Newton regression · Back-propagation neural network · Semiconductor

C.-F. Chien · C.-W. Hsiao
Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Hsinchu 30013, Taiwan
e-mail: cfchien@mx.nthu.edu.tw

C.-Y. Hsu (✉)
Department of Information Management, Yuan Ze University, Chungli 32003, Taiwan
e-mail: cyhsu@saturn.yzu.edu.tw

Introduction

Semiconductor manufacturing is one of the most complicated production processes because of the challenges of lengthy fabrication process, unrelated parallel machine, dynamic job arrival, non-preemptive, inseparable sequence-dependent setup time, multiple-resource requirements, general precedence constraint, job re-circulation, and shifting bottlenecks (Chien and Chen 2007). A semiconductor foundry provides a make-to-order production system with complex product mix, in which the bottleneck is often shifting with unbalanced machine loading caused by unplanned orders and thus results in WIP bubbles. Since the semiconductor industry is capital intensive, capacity is configured to highly load the critical and expensive equipment such as scanners for photolithography as the bottleneck. However, the near-bottleneck machines may become bottleneck due to bottleneck shifted and thus complicate the problem in addition to the conventional bottleneck. Most of the existing approaches are formulated under the assumption of a static production environment. In practice, special production actions are applied to respond to due dates or other customer requirements, especially in the foundry that often make the production line unbalanced.

With the increasing competition in global semiconductor market and declining average selling price of semiconductor products, an accurate cycle time forecast is critical to make

on-time delivery, reduce cycle time, which can reduce unnecessary buffer WIP in production lines and bullwhip effect in the whole semiconductor manufacturing supply chain. Indeed, wafer fabrication facility (fab) needs safety WIP level as the buffer to maintain the productivity of the expensive bottleneck equipment in light of dynamic production changes (Leachman et al. 2002). However, little research has been done to determine the corresponding WIP levels for various tool groups at different process steps to maintain overall production flow balanced.

Focusing on real settings, this study aims to develop a manufacturing intelligence approach via data mining and combine domain knowledge to forecast the cycle time with production line status such as WIP, movement (MOVE), and capacity. The derived rules can be used to control the production line status (e.g., WIP levels) to reduce the cycle time. Indeed, manufacturing intelligence approaches has been developed to extract useful information and derived patterns from production to support manufacturing decisions (Chien et al. 2010; Kuo et al. 2011). In particular, the proposed approach integrated Gauss-Newton regression (GNR) method and back-propagation neural network (BPNN) model into a two-phase manufacturing intelligence framework to provide accurately forecast, while sensitively detecting the change of production status. To evaluate the validity of this approach, we conducted an empirical study on a semiconductor foundry in Hsinchu Science Park on the basis of real data from the production line. The results have shown that this approach can induce effective models to reduce the forecast errors with the dynamic operation conditions.

The remainder of this paper is organized as follows. In the next section, this study reviews the production planning in semiconductor manufacturing. Then, proposed approach and integrated forecast model for production line performance is developed. Next, we conduct an empirical study on semiconductor manufacturing for validation. Finally, this study concludes with discussion of contribution and future research directions.

Production planning in semiconductor manufacturing

Semiconductor industry is very capital intensive in which a modern 300 mm wafer fab will need US$3 Billion capital investment that can produce about 30,000 wafers per month. In the past decades, semiconductor industry has closely followed the Moore's Law (Moore 1965) that the number of transistors on a wafer area will be doubled every 12–24 months. To maintain the shrinkage of Integrated Circuit (IC) features, a new generation of production technology will be developed and employed every 12–24 months. Therefore, multiple production technologies generally co-exist in a wafer fab with utilization of a pool of common tools for multiple technologies and critical tools dedicated for a specific technology. The wafer fabrication process flow in a semiconductor fab generally contains over 500 processing steps including oxidation, deposition, metallization, lithography, etching, ion implantation, photo-resist strip, cleaning, inspection, and metrology. The average production cycle time is approximately 30–60 days depending on the complexity of the technology. An average-sized factory containing average 400 pieces of equipment usually continuously operates 7 days a week, and 24 hours a day (Dabbas and Chen 2001). The production plan needs to be revised with dynamic re-entrant process flow and complex product-mix. In addition, the bottleneck also drifts with the unbalanced machine loading. Therefore, it is difficult to estimate production status in long production process of semiconductor foundry. Most existing approaches are based on rule of thumb and various heuristic indexes which are complicated and interrelated in semiconductor manufacturing management (Chien et al. 2004).

According to Little's Law, WIP and throughput are two key factors for cycle time. Throughput and cycle time are related to the line loading (WIP level) in the theoretical curves as illustrated in Fig. 1. This curve always lies in the positive region and is limited by some fixed capacity or resource constraints. Based on the Theory of Constraint (TOC), the relationship between fab line loading and relative throughput has three characteristics. Firstly, the line loading is proportional to the throughput and much lower than the capacity limitation. Secondly, throughput tends to grow gradually while line loading is close to the capacity limitation, where decision makers will start to reduce the line loading due to the decreasing turn ratio (i.e., the average throughput obtained by unit line loading). Thirdly, throughput cannot increase with the line loading while the capacity is over-loaded. At this stage, excess line loading only increases the production cycle time instead of throughput. The throughput level is critical for determining the optimal buffer allocation that minimizes the average WIP inventory (Papadopoulos and Vidalis 2001). Although a higher throughput and utilization can be achieved with sufficient buffer WIP for bottleneck tools, the cycle time may be increased because of queue time.

Both throughput and cycle time are important factors for assessing manufacturing performance and helpful in giving directions for future improvement. Atherton and Dayhoff (1986) used signature analysis to characterize the curves of inventory, cycle time, and throughput as a function of wafer start rates. Chen et al. (1988) developed a queuing network model for predicting cycle time and throughput. Sattler (1996) used a queuing curve approximation to determine the productivity improvement of particular machine sets and provided a basis for cycle time adjustment with fab loadings. Miltenburg and Sparling (1996) developed three models, including a simple stochastic model, a Markov chain model, and a queuing model for cycle time reduction and

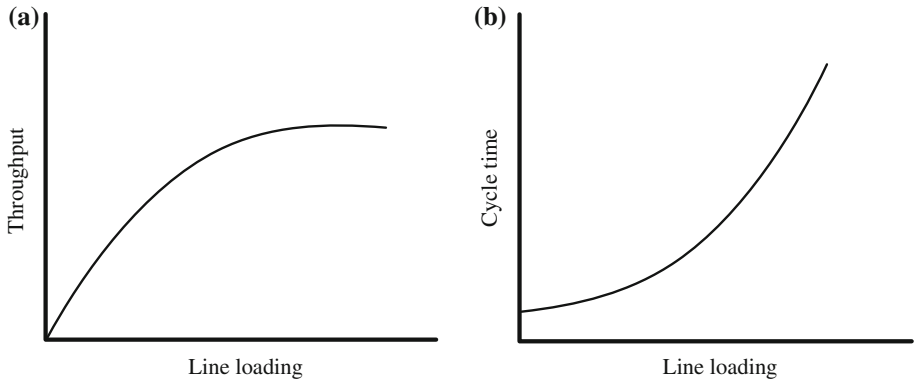


Fig. 1 Theoretical curves of (**a**) throughput (**b**) cycle time

management under different situations. Fowler et al. (1997) used a simulation-based analysis to evaluate the potential area for productivity and capacity improvement. In particular, the implementation helps a real fab reduce 25% cycle time and inventory. Fowler et al. (2001) developed sampling strategies to quickly generate the simulation-based cycle time-throughput curve. Yu and Huang (2002) used BPNN to form a relationship between product cycle time and the operation tool. Morrison and Martin (2007) incorporated four practical manufacturing realities into a closed-form approximation of a G/G/m-queue model to expand the practical applicability of existing approximations for the behavior of mean cycle time. However, most of the existing studies focused on clarifying the relationship among actual cycle time, throughput, and forecast cycle time. In practice, the feature of cycle time and throughput varies with operation conditions such as the change of production routings, preventative maintenance policies, throughput rates, line configurations (Fowler et al. 2001). Kuo et al. (2011) proposed a manufacturing intelligence approach based on neural networks to analyze production data and tool data for cycle time reduction. Little research has been done on incorporating dynamic operation conditions for cycle time forecast and deriving useful rules for WIP level management to reduce cycle time effectively.

Proposed approach

The terminologies and notations used in this study are listed as follows:

i: Index of the model types.

j: Index of the regression model types.

k: Index of the model periods.

m: Index of the data numbers in a specific model period.

b: Index of the basic model type.

a: Index of the adaptive model types.

e: Index of exponential regression model (for cycle time forecast).

l: Index of logistic regression model (for MOVE forecast).

p: Index of the model stage (period) number.

$GNR_{ij(k)}$: The Gauss-Newton regression function of model type i and regression model type j (in period k only for the adaptive models).

X: The variable of normalized WIP.

Y: The variable of normalized MOVE.

Z: The variable of cycle time.

$A_{ij(k)}$, $B_{ij(k)}$, $C_{ij(k)}$: Regression coefficient from model type i and regression model type j (in period k only for the adaptive models).

$D_{ij(k)}$: Dummy limit from model type i and regression model type j (in period k only for the adaptive models).

$c_{m(k)}$: The mth capacity data (in period k only for the adaptive models).

$x_{m(k)}$: The mth WIP data (in period k only for the adaptive models).

$y_{m(k)}$: The mth MOVE data (in period k only for the adaptive models).

$z_{m(k)}$: The mth cycle time data (in period k only for the adaptive models).

$x^*_{m(k)}$: The mth normalized WIP data (in period k only for the adaptive models).

$y^*_{m(k)}$: The mth normalized MOVE data (in period k only for the adaptive models).

$N_{i(k)}$: Data quantity used to construct model type i (in period k only for the adaptive models).

N_c: Data quantity used to calculate $\hat{C}_{bl}$.

RPT: Raw process time in this fab.

ε_{jmk}: The mth residual of regression model j in period k.

$\overline{\varepsilon}_{jk}$: The average bias of regression model j in period k.

α_{jk}: Adaptive coefficient of regression model j in period k.

β_{jk} : The best adaptive coefficients of regression model j in period k while the adaptive criteria are achieved.

W^*_L : A comparatively low level of normalized WIP.

W^*_H: A comparatively high level of normalized WIP.

η_j: Learning factor of regression model j.

q: Total trial numbers to test different adaptive coefficients.

Table 1 Summary of attributes

Attribute	Definition	Unit
Average layer	Average layers of one wafer need to be manufactured	Layers per wafer
Capacity	The possibly maximum wafers production quantity	Wafers output per month
Cycle time	Average time needed to finish one layer of a wafer	Days per layer
Fab utilization	The percentage of used capacity to the maximum useful capacity in a fab	Ratio
Movement (MOVE)	Total accomplished operations among all machines	Operations per day
Work in process (WIP)	Number of wafer that in processing	Wafers per day

This study proposes a two-phase approach consists of basic model and adaptive model to derive production line performance. In the first phase, the basic model is used to describe the static status of existing production environments. However, the latest situation of fab productivity should be continuously modified according to actual production line behavior which is influenced by the variance of fab WPH (wafers per hour), different dispatching criteria, operator learning effect, different product mix, and the new technology introduced. In order to reduce forecast deviation of line performance, an adaptive model is proposed for continuous bias detection and rapid model re-alignment. The adaptive models consider the physical properties of specific regression models and then construct an adaptive procedure with learning effect to adjust the regression models.

Data preparation

There are lots of performance indexes in the field of semiconductor manufacturing. Chien et al. (2004) reviewed related studies and made arrangement of these key indexes. The indexes such as capacity, cycle time, MOVE, and WIP, are selected to represent the production line status, as shown in Table 1. Theoretically, capacity can be described as the available production quantity for bottleneck machines. In a semiconductor fab, the capacity is measured by output quantity. Fab utilization can be accessed by the ratio of the total wafer output over the capacity during specific time period. MOVE is the major index for fab productivity measurement. In this study, the throughput is measured by MOVE instead of wafer output because wafer output can be affected by human and external factors, such as demand change, customer requirement, and urgent orders. WIP represents the wafers that are not fully finished.

The purpose of data preparation is to ensure the availability of production line data and data quality for analysis. The production data are automatically collected from various engineering databases by different engineers on a daily basis. To ensure that the relations of production attributes during the same time period are connected to each other, the missing values would be filled through discussion with domain experts.

Moreover, the WIP level under different capacity conditions may represent different meanings, and increases with capacity expansion. Thus, the original amount of WIP should be normalized by being divided with relative capacity to avoid incorrect information. Similarly, the MOVE is also influenced by different capacity. The WIP and MOVE can be normalized as follows:

$$x_m^* = \frac{x_m}{c_m} \quad m = 1 \ldots N_b \tag{1}$$

$$y_m^* = \frac{y_m}{c_m} \quad m = 1 \ldots N_b \tag{2}$$

In practice, the capacity sometimes should be adjusted based on different product mix, because different product types have different consumption of capacity. The difference could be estimated as a weighted relation by their manufacturing properties, fabricating difficulty, and proportion of their on-machine time.

Basic model

The basic model is to derive the contours of fab productive capability, which means the "existing" performance standard for productivity for the long-term history. The GNR method is used to forecast the global trend and BPNN is applied to improve the forecast accuracy based on the production line status.

Performance curve fitting

The performance curves among line loading, throughput and cycle time can be shown by nonlinear regression analysis. The iterative algorithm of GNR method is widely used to solve the nonlinear least square problems (Seber and Wild 1989). In particular, GNR model is one of the best iterative methods that perform well at converging speed and fitting accuracy for solving nonlinear least square problems (Kumar and Alsaleh 1996). The procedure of performance curve sketches some characteristics from the scatter plots as the theoretical curves first. In particular, cycle time increases

exponentially with WIP and MOVE increases proportionally with WIP. Therefore, the logistic and exponential functions are used to present the curves of "normalized WIP versus normalized MOVE" and "normalized WIP versus cycle time" respectively.

$$Y = \frac{C_{bl}}{1 + A_{bl} Exp(B_{bl} x)} \tag{3}$$

$$Z = A_{be} Exp(B_{be} x) + C_{be} \tag{4}$$

The parameter $\hat{C}_{bl}$ representing the higher limit of normalized MOVE in logistic function and the parameter $\hat{C}_{be}$ representing the lower limit of cycle time in exponential function are estimated as follows:

$$\hat{C}_{bl} = \frac{\sum_{m=1}^{N_c} y_{(m)}}{N_c} \tag{5}$$

$$\hat{C}_{be} = RPT \tag{6}$$

Meanwhile, the index (m) denoting normalized MOVE data is ranked by their normalized WIP (i.e., $m = 1$ refers to the data with maximally normalized WIP). N_c is the number of data that contains information about the boundary limitation of normalized MOVE. N_c changes from fab to fab based on the distribution of past WIP level. In addition, the cycle time tends toward its average raw process time (RPT) while the quantity of WIP decreases to an extremely low level. Once $\hat{C}_{bl}$ and $\hat{C}_{be}$ are determined, the other parameters $\left(\hat{A}_{bl}, \hat{B}_{bl}, \hat{A}_{be}, \hat{B}_{be}\right)$ could then be calculated iteratively by the GNR method. Furthermore, $\hat{C}_{bl}$ and $\hat{C}_{be}$ will also be the initial values in the adaptive models.

The upper and lower limits for the logistic and exponential regression model should be considered first to ensure that the adaptive model can form a reasonable shape. The parameters $\hat{C}_{bl}$ and $\hat{C}_{be}$ are used to control the upper limit and lower limit of logistic and exponential regression model respectively. Moreover, a relatively low WIP level for the logistic regression model and a relatively high WIP level for the exponential regression model are defined as W_L^* and W_H^*, respectively. According to the basic model, the dummy limits $\hat{D}_{bl}$ and $\hat{D}_{be}$ can be derived as follows:

$$\hat{D}_{bl} = \frac{\hat{C}_{bl}}{1 + \hat{A}_{bl} Exp\left(\hat{B}_{bl} W_L^*\right)} = GNR_{bl}\left(W_L^*\right) \tag{7}$$

$$\hat{D}_{be} = \hat{A}_{be} Exp\left(\hat{B}_{be} W_H^*\right) + \hat{C}_{be} = GNR_{be}\left(W_H^*\right) \tag{8}$$

Forecast error is used to evaluate the forecast performance. Mean Absolute Percentage Error (MAPE) is used to express the percentage error with the relative data scale and is presented as follows:

$$MAPE = \frac{1}{T} \sum_{t=1}^{T} \frac{|Actual_t - Forecast_t|}{Actual_t} \times 100\% \tag{9}$$

The smaller MAPE represents that the larger deviation can be explained from forecast model and higher accuracy for forecasting normalized MOVE and cycle time. If the GNR model can not provide good fitness for MOVE or cycle time, it needs to incorporate other factors in production line that may influence on MOVE or cycle time for enhancing forecast accuracy

Forecast accuracy enhancement

The GNR model only considers WIP level to present the global trend of MOVE and cycle time. In practice, the MOVE and cycle time are also influenced by the production line status. For example, the cycle time will be increased with manufactured layers. Moreover, fabs with larger capacity could provide more tolerance and better flexibility to avoid the throughput loss caused by improper job scheduling or machine failure. However, these existing relations may not be easily modeled via a mathematical formulation. BPNN model with powerful learning ability is applied to extract complicated patterns among various input factors and output variables. The residual of GNR model which defines as the difference between actual value and predicted value is used as the output variable. As listed in Table 1, the input factors related to production line status are used to construct BPNN model and enhance forecast accuracy.

BPNN model is widely used to extract the complex patterns from specific dataset by iterative learning approaches to minimize the squared error between forecast value and actual output. Basic topology of BPNN structure includes one input layer, one output layer and one or two hidden layers between input and output layers. In particular, the number of input nodes and output nodes can be determined directly based on the number of relative attributes. In general, BPNN model with single hidden layer can obtain better prediction performance. Each number of hidden nodes in hidden layer needs to be determined by experiment trials. In this study, conjugate gradient method is used for model learning and selection of number of hidden nodes with the minimum MAPE. In particular, 80% of the input data are randomly selected to train the BPNN model and the other 20% are used to test the model.

There are two advantages of the proposed basic model. The first is that decision makers can realize the relation among WIP, MOVE, and cycle time by using charts and tables to support their decisions about WIP level. The second is that they can incorporate the variation of other external factors and improve the forecast result via the BPNN models. There are also two disadvantages of the basic models. Firstly, it is difficult to determine the complex relation among models inputs (e.g., WIP and other external factors) and the outputs (e.g., MOVE and cycle time) in a regression model. Secondly, the BPNN are difficult to provide understandable rules or formulas for decision makers.

Adaptive model

The production line status may change dynamically with the external factors (such as short product life cycle, various market segments, and demand) and internal factors (such as WPH improvement, product-mix, yield, learning effect, and technology maturity). The basic model should adapt itself according to the actual production environment for increasing the forecast accuracy. Adaptive model is developed to forecast the recent production status accurately by a small number of newly observed data. This aggressive property is necessary for rapid response to productivity adjustment.

Judgment of model adaptation

The adaptive model determines whether the model should be adapted by both time and error criteria. Firstly, the time criterion determines the maximum time interval of the basic model with good forecast ability. In the semiconductor manufacturing industry, the rapid changes of production status, such as demand change, short product life cycle, market competition, and technology innovation, lead to a short model life cycle. Secondly, the error criteria are used to detect the changes in the production condition. In order to measure such situation, a simple control chart is applied to detect the assignable bias in the manufacturing process. The forecast errors between actual performance and corresponding predicted value from GNR models are recorded sequentially and monitored via the control charts. The upper control limit (UCL) and the lower control limit (LCL) should also be determined first (e.g., error rate within 5%) by the decision maker. The error criteria of model adaption are listed as follows:

(1) There are n_1 points that continuously locate in a row with out-of-limit.
(2) There are n_2 points that continuously locate in one side of the central line without out-of-limit.

Criterion (1) shows that the recent performance index starts with some obvious abnormal events and the guideline for performance target could be changed. Criterion (2) is used as a precaution against the expectable error when the system is biased but not yet out-of-limit.

Coefficient update

The parameters of the regression model have to be adjusted to incorporate the latest production status into models, while the adaptive criteria are achieved for stage p. Figures 2, 3 show how the adaptive coefficients (α_{lp} and α_{ep}) are applied to update parameters and construct adaptive models at stage $p+1$. As shown in Fig. 2, the new observations are performed

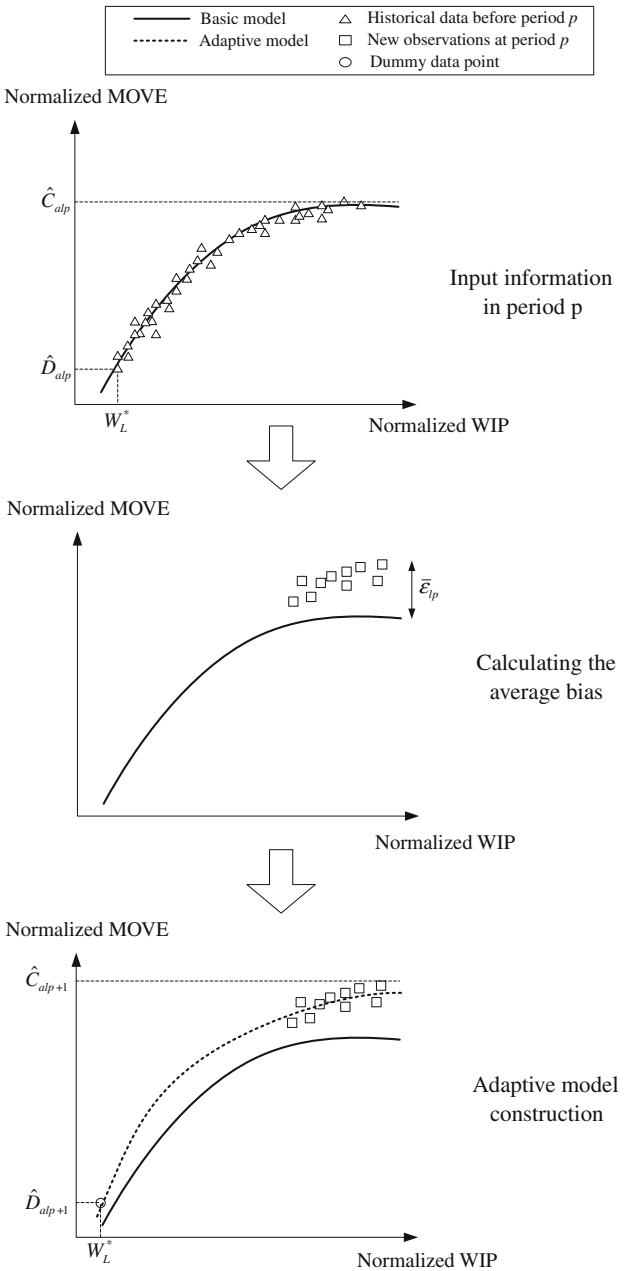


Fig. 2 Conception of adaptive logistic regression model

better at the previous stage p, the average bias of logistic regression model at stage p can be calculated to measure the average bias of normalized MOVE as Eq. (10). In particular, $\bar{\varepsilon}_{lp} < 0$ implies that the performance may have declined.

$$\bar{\varepsilon}_{lp} = \frac{\sum_{m=1}^{N_{ap}} \left[y_{mp}^* - GNR_{alp}\left(x_{mp}^*\right)\right]}{N_{ap}} \tag{10}$$

The upper limit $\left(\hat{C}_{alp}\right)$ and lower limit $\left(\hat{D}_{alp}\right)$ at stage p should also be adjusted with the change. In general, the complexity of management and capacity constraints of bottleneck

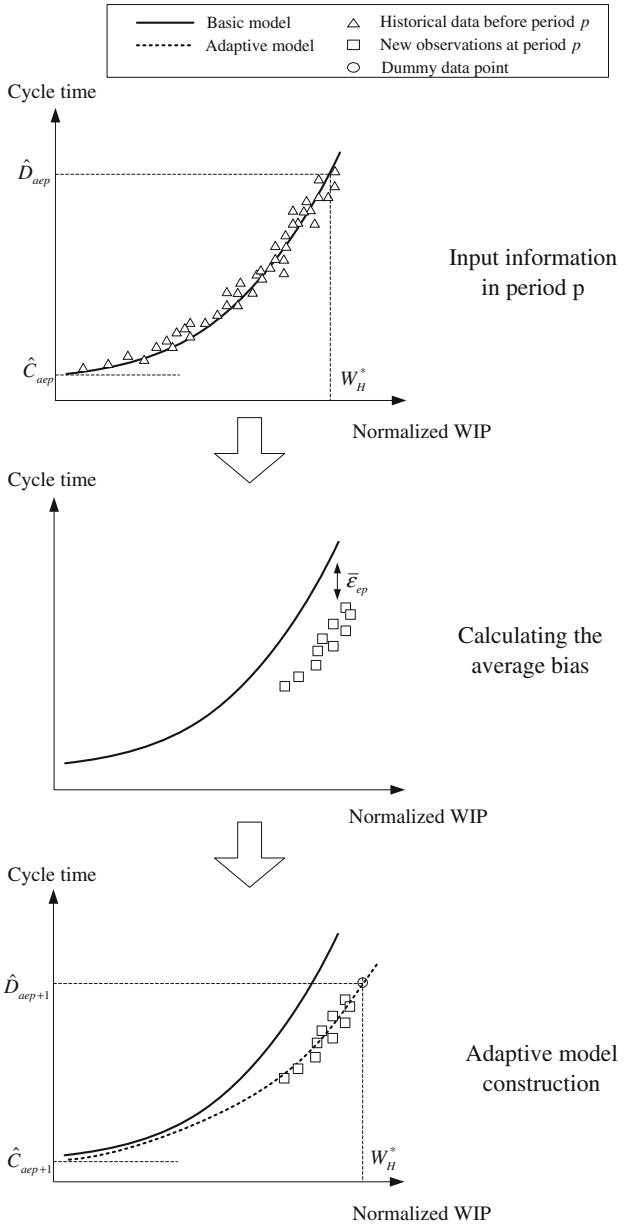


Fig. 3 Conception of adaptive exponential regression model

machines will depress the growth of throughput at a higher WIP level. The growth of throughput at a lower WIP level should also be obtained. Thus, $\hat{C}_{alp}$ and $\hat{D}_{alp}$ can be updated to $\hat{C}_{alp+1}$ and $\hat{D}_{alp+1}$ with $\overline{\varepsilon}_{lp}$ and α_{lp} as follows.

$$\hat{C}_{alp+1} = \hat{C}_{alp} + \alpha_{lp}\overline{\varepsilon}_{lp} \tag{11}$$

$$\hat{D}_{alp+1} = \hat{D}_{alp} + \overline{\varepsilon}_{lp} \tag{12}$$

α_{lp} is the adaptive coefficient of logistic regression model at stage p with a value between 0 and 1. After $\hat{C}_{alp}$ is replaced by $\hat{C}_{alp+1}$ and a dummy data point $\left(W_L^*, \hat{D}_{alp+1}\right)$ is inserted into the training dataset, other estimators in the adapted logistic regression model ($\hat{A}_{alp+1}$ and $\hat{B}_{alp+1}$) can be derived iteratively by the GNR method according to the updated estimators and new observations.

Similarly, as shown in Fig. 3, the conception of the adaptive exponential regression model for cycle time at time period p can also be updated by Eqs. (13)–(15):

$$\overline{\varepsilon}_{ep} = \frac{\sum_{m=1}^{N_{ap}} \left[z_{mp} - GNR_{aep}\left(x_{mp}^*\right)\right]}{N_{ap}} \tag{13}$$

$$\hat{C}_{aep+1} = \hat{C}_{aep} + \overline{\varepsilon}_{ep} \tag{14}$$

$$\hat{D}_{aep+1} = \hat{D}_{aep} + \alpha_{ep}\overline{\varepsilon}_{ep} \tag{15}$$

α_{ep} is the adaptive coefficient of exponential regression model at stage p with a value between 0 and 1. After $\hat{C}_{aep}$ is replaced by $\hat{C}_{aep+1}$ and a dummy data point $\left(W_H^*, \hat{D}_{aep+1}\right)$ is inserted into the training dataset, other estimators in the adapted exponential regression model ($\hat{A}_{aep+1}$ and $\hat{B}_{aep+1}$) can be derived iteratively.

The adaptive coefficients are designed to revise the regression models at a high WIP level by the average bias between actual value and predicted value for new observations. The learning factor η that represents the learning ability of adaptive models is a constant ratio between 0 and 1. The larger learning factor leads parameters of adaptive model to a rapid change. The small learning factor makes the adjustment of parameters slowly that could result in a local optimal solution. The learning factor η is usually set between 0.5 and 0.8. At the initial stage ($p = 0$), the value of adaptive coefficients (α_{l0} and α_{e0}) is set as an initial value. Once the adaptive criteria are achieved, the average bias ($\overline{\varepsilon}_{lp}$ and $\overline{\varepsilon}_{ep}$) can be calculated by Eqs. (10) and (13). The parameters of regression models at stage $p+1$ can be updated immediately.

Once the adaptive criteria are achieved again, a heuristic method is used to derive the optimal adaptive coefficients at stage p, i.e., β_{lp} and β_{ep}, to minimize the forecasting error at stage $p+1$. First, we determine the comparison trials q $(q \geq 1)$. Next, the different values of α_{lp} and α_{ep} are set as $0, 1/q, 2/q, \ldots, (q-1)/q$, and then we reconstruct the GNR_{alp} and GNR_{aep} according to the data collected at stage p. Then, the new observations collected at stage $p+1$ are applied to each regression model to calculate the sum of square errors (SSE, namely, square forecasting errors). The adaptive coefficients β_{lp} and β_{ep} with the minimum SSE are selected for updating new adaptive coefficients α_{lp+1} and α_{ep+1} as follows:

$$\alpha_{lp+1} = \eta_l \times \beta_{lp} + (1-\eta_l) \times \alpha_{lp} \tag{16}$$

$$\alpha_{ep+1} = \eta_e \times \beta_{ep} + (1-\eta_e) \times \alpha_{ep} \tag{17}$$

where η_l is the discount factor for the logistic regression models, and η_e is the discount factor for the exponential regression models. The new adaptive coefficients (α_{lp+1} and α_{ep+1}) are the weighted combinations of the latest adaptive coefficients (α_{lp} and α_{ep}) and the best adaptive coefficients

(β_{lp} and β_{ep}) while the adaptive criteria are achieved at stage p. Meanwhile, the new adaptive coefficients are used to construct the new regression models at stage $p+1$. The estimators $\hat{C}_{alp+1}$, $\hat{D}_{alp+1}$, $\hat{C}_{aep+1}$ and $\hat{D}_{aep+1}$ can be derived from previous Eqs. (11), (12), (14), and (15), respectively by translating the time period from stage p to stage $p+1$. The estimators ($\hat{A}_{alp+1}$, $\hat{B}_{alp+1}$, $\hat{A}_{aep+1}$ and $\hat{B}_{aep+1}$) can also be derived from the iterative GNR method, and then the regression model can be reconstructed to forecast the relative throughput and cycle time.

Furthermore, the forecast accuracy of adaptive models can also be improved by BPNN model as the basic model. All the historical data are used to construct the BPNN model for the adaptive model. For example, BPNN model is applied to stage $p+2$, and all the historical data including the basic model and adaptive model at stage p and stage $p+1$ are integrated into a single dataset to construct the BPNN model. According to the proposed forecast model, detailed tabulation about WIP level with corresponding MOVE and cycle time in basic and adaptive models can be analyzed as a useful tool for WIP level decision and cycle time forecast. Decision maker can determine the WIP level decisions based on the latest productivity forecast considering various kinds of parameter setting.

Empirical study

To evaluate the validity of the proposed approach, an empirical study was conducted in a semiconductor wafer fab in Hsinchu Science Park. This fab was faced with the challenge of a lengthy production process, such as Make to Order (MTO) production type, complicated job re-circulation flow, various product mix, and serious bottleneck shifts. However, the appropriate WIP level is crucial for cycle time reduction in light of the dynamic nature of a wafer fab. The analysis results provide useful information for wafer release plan and reduce the production cycle time.

Data preparation

Without losing the generality, all the data showed in this study were transformed for further analysis. The production line data of one technology among 28 months were daily collected to validate the proposed approach. The production line data, including WIP, MOVE, cycle time, capacity, average layer, and fab utilization, were integrated into an analysis dataset. In order to eliminate the effect of capacity expansion, all the collected WIP and MOVE data have to been normalized by dividing the relative capacity at the same period. According to the time sequence, the first 80% of these data were used to construct the basic model for a long-term basis. The other 20% data were applied to construct the adaptive models to fine-tune the dynamic production line.

Basic model

To capture the trend of production line status and avoid the disturbance of daily data noise, the daily data were sampled in every 10 days to ensure that the minimum data quantity contains various patterns in the beginning, middle, and end of a month. According to the sampling strategy, historical data in the first 22 months were transformed to 66 data to construct the basic model for evaluating the long-term behavior. Low fab utilization which indicates the abnormal production environment will be excluded from the adaptive model. Therefore, 6 production data whose fab utilization is lower than 60% were excluded and the other 60 production data were used to derive the regression models of MOVE and cycle time as shown in Fig. 4. The value of MAPE was used to evaluate the forecast performance. The MAPE of MOVE and cycle time were 3.14 and 4.59%, respectively. Furthermore, the remaining 90% daily data were input into the constructed regression models to evaluate the effectiveness of sampling strategy. As listed in Table 2, by using 90% daily data, the MAPE of MOVE and cycle time are 3.43 and 4.96%, respectively, which were close to the 10 sampled daily data. The

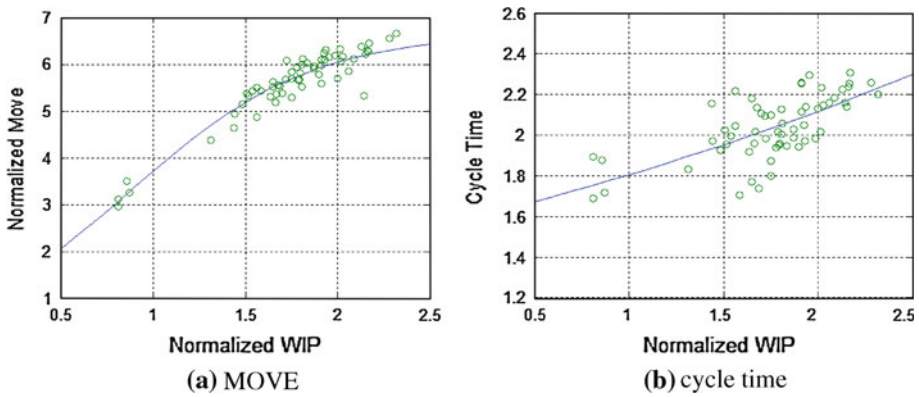


Fig. 4 Training result of GNR model

Table 2 Forecast result with only GNR method in basic model

	Number of data	MOVE MAPE (%)	Cycle time MAPE (%)
Sampled data	60	3.14	4.59
Un-sampled data	533	3.43	4.96

results showed that the main trend can be extracted by using the sampling interval in the basic model as follows.

$$\hat{GNR}_{bl}(x) = \frac{\hat{C}_{bl}}{1 + \hat{A}_{bl} Exp\left(\hat{B}_{bl} x\right)} = \frac{6.655}{1 + 6.029 \times Exp(2.039x)}$$

$$\hat{GNR}_{be}(x) = \hat{A}_{bl} Exp\left(\hat{B}_{be} x\right) + \hat{C}_{be} = 1.073 \times Exp(0.204x) + 0.5$$

BPNN model was used to improve the forecast accuracy of MOVE and cycle time by the related factors of production line status. To model the remainder of MOVE in the logistic regression model, the input factors are defined as the normalized WIP, capacity, average layers and expected cycle time. To model the remainder of cycle time in the exponential regression model, the input factors are defined as the normalized WIP, capacity, average layers and expected MOVE. A three-layer network topology was applied, where the sigmoid activation function was used in the hidden layers. The numbers of input nodes is four and the number of output nodes is one. The number of hidden nodes needs to be experimented by conjugate gradient algorithm. In particular, we randomly select 80% training data to construct the model and then employ the derived BPNN model to forecast the 20% remaining data. Each BPNN model with different number of hidden nodes is trained and tested iteratively with ten times. Table 3 lists the forecast result for MOVE and cycle time for different BPNN models. In particular, BPNN model under each number of hidden nodes can obtain lower MAPE than only GNR model. To compare the results in MOVE and cycle time, respectively, the structure of four hidden nodes with smaller MAPE is selected for BPNN model construction and enhancement of forecast accuracy.

Sensitivity analysis is conducted to support WIP level decision based on the forecast model. Although the increase of WIP is useful for increasing throughput and machine utilization, the cycle time is also extended with the increasing WIP. In addition, machine capacity is wasted and loss of order due to fewer WIP. The tabulations that summarize the WIP level with the corresponding MOVE and cycle time are useful for WIP level decision and cycle time forecast. Decision makers can determine the cycle time based on the due date and demand trend. Moreover, MOVE requirement can also be derived through the order list and current WIP level. Therefore, the feasible range of WIP level for the specific production plan is limited. Then the WIP level can be easily adjusted based on the manufacturing strategy. For example, assuming the average layers in the next period are 32.5 layers per lot and the monthly capacity is estimated as 37,300 wafers, the expected MOVE and cycle time under different WIP levels can be estimated through the result of combining the outputs of GNR models and BPNN models. Table 4 lists the generated information about the combined forecast result of relative expected MOVE and cycle time under different planning WIP levels by sensitivity analysis. As the WIP level is close to the production capacity, the rate of increased MOVE becomes slowly and gets the extended cycle time. The decision makers can estimate the required MOVE based on the aggregated information such as capacity and average layers, the corresponding minimal WIP level can be determined through the proposed forecast model. The cycle time under determined WIP level can facilitate the decision-making process of wafer starts for on-time delivery and reduce the unnecessary wastes of fab production cycle time.

Adaptive models for MOVE and cycle time forecast

Total 182 daily data in the last 6 months were used to validate the proposed adaptive model. After removing the low-utilization data through employing the basic model, 139 historical daily production data were applied to illustrate the effectiveness of the adaptive models. Each model was adjusted according to its specific adaptive criteria and procedure. The threshold of adaptive criteria was set by discussing with the domain experts in semiconductor manufacturing. In practice, the production cycle time should never be longer than 2 months. Therefore, the maximum life cycle of each constructed model was set as 60 days. The UCL and LCL for error criteria were determined as 5% under or below the forecast target. The parameters of n_1 and n_2 were determined as 10 based on the response effectiveness and minimum data requirement for new model construction. In addition, cycle time usually has higher variance than MOVE because of the man-made operations. In order to detect the change of real production status rapidly, the parameters of n_1 and n_2 for cycle time will be reduced to 8 days.

Firstly, the regression parameters in stage 0 were followed by the basic model. The comparatively low level of normalized WIP $\left(W_L^*\right)$ is set to 0.5. The initial adaptive coefficient (α_{l0}) and learning factor (η_l) for MOVE forecast model are set to be 0.5 and 0.8 respectively. The test trial q is set to 10. As shown in Fig. 5a, the new MOVE data violated the error criterion (1) from day 15 to day 24. The average bias

Table 3 Comparisons of MAPE in MOVE and cycle time for different BPNN models

Number of hidden node	MOVE MAPE (%)		Cycle Time MAPE (%)	
	Training	Testing	Training	Testing
2	2.88	3.50	4.49	4.13
3	2.86	3.48	4.48	4.15
4	2.93	3.41	4.45	4.09
5	2.85	3.49	4.46	4.13
6	2.85	3.44	4.44	4.15
7	2.93	3.53	4.43	4.11
8	2.83	3.50	4.47	4.12

Table 4 Sensitivity analysis for the relative MOVE and cycle time under different WIP level in basic model

Normalized WIP	WIP level	△ WIP	Move forecast	△ MOVE	Cycle time forecast	△ Cycle time
0.5	18,650		86,134.74		1.6214	
0.6	22,380	3,730	106,570.20	20,435.46	1.6489	0.0275
0.7	26,110	3,730	127,015.73	20,445.53	1.6793	0.0305
0.8	29,840	3,730	146,628.33	19,612.59	1.7125	0.0331
0.9	33,570	3,730	164,713.51	18,085.19	1.7417	0.0292
1.0	37,300	3,730	180,812.50	16,098.99	1.7709	0.0292
1.1	41,030	3,730	194,719.53	13,907.03	1.8015	0.0306
1.2	44,760	3,730	206,444.07	11,724.54	1.8338	0.0323
1.3	48,490	3,730	216,142.12	9,698.05	1.8682	0.0343
1.4	52,220	3,730	224,048.47	7,906.35	1.9047	0.0366
1.5	55,950	3,730	230,423.90	6,375.43	1.9436	0.0389
1.6	59,680	3,730	235,522.95	5,099.05	1.9826	0.0389
1.7	63,410	3,730	239,580.52	4,057.58	2.0155	0.0330
1.8	67,140	3,730	242,813.07	3,232.55	2.0593	0.0438
1.9	70,870	3,730	245,432.53	2,619.45	2.1043	0.0450
2.0	74,600	3,730	247,663.88	2,231.35	2.1469	0.0426
2.1	78,330	3,730	249,761.82	2,097.95	2.1829	0.0359
2.2	82,060	3,730	251,993.91	2,232.09	2.2278	0.0449
2.3	85,790	3,730	253,437.34	1,443.43	2.2669	0.0392
2.4	89,520	3,730	254,440.43	1,003.09	2.3055	0.0386
2.5	93,250	3,730	255,051.56	611.13	2.3510	0.0455

of the newly collected data from day 1 to day 24 were calculated by Eq. (10); the regression coefficients $\hat{C}_{al1}$ and $\hat{D}_{al1}$ can be adjusted through Eqs. (11) and (12) as shown in Fig. 5a. Next, the new regression model for stage 1 was constructed based on the newly adjusted regression coefficients ($\hat{C}_{al1}$ and $\hat{D}_{al1}$) and collected data from day 1 to day 24. Although the MOVE from day 25 to day 84 did not violate the error criteria, the adaptive model should be employed again based on the time criterion. Before the adaptive procedure, a comparison test was used to determine the best adaptive coefficient at stage 0. As shown in Fig. 5b, the result showed the lowest SSE at stage 0 while β_{l0} is set to 0. The new adaptive coefficient α_{l1} is updated by Eq. (16). Then, the new regression coefficients ($\hat{C}_{al2} = 6.732$ and $\hat{D}_{al2} = 2.360$) were used to adjust the new regression model of stage 2. Finally, the MOVE data collected from day 85 to day 139 did not meet the time criteria and error criteria, as shown in Fig. 5c.

Similarly, the adaptive models for cycle time were constructed according to the newly collected dataset. In the initial stage 0, the comparatively high level of normalized WIP $\left(W_H^*\right)$ is set to be 2.5, and the initial adaptive coefficient (α_{e0}) and learning factor (η_e) are set to be 0.5 and 0.8, respectively. The adaptive processes of cycle time forecast were illustrated in Fig. 6 a–f with the five stages, respectively. Firstly, the new cycle time data violated the error criterion (1) from day 11 to day 18 as shown in Fig. 6a. The average bias of collected data from day 1 to day 18 by Eq. (13), and then the regres-

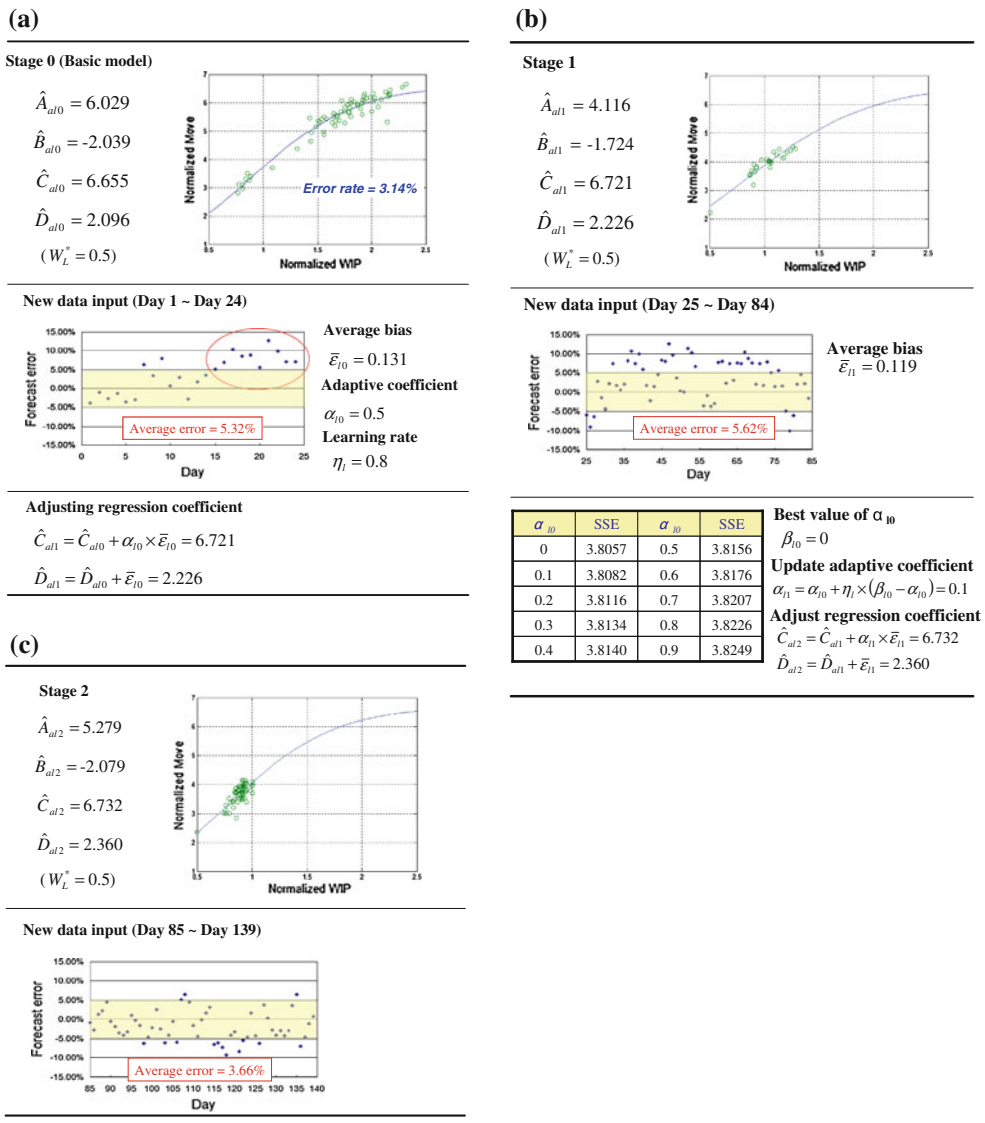


Fig. 5 **a** Adaptive procedure for MOVE forecast at stage 0. **b** Adaptive procedure for MOVE forecast at stage 1. **c** Adaptive procedure for MOVE forecast at stage 2

sion coefficients $\hat{C}_{el1}$ and $\hat{D}_{el1}$ can be adjusted through Eqs. (14) and (15) as shown in Fig. 6a. The new regression model for stage 1 was constructed based on the data from day 1 to day 18. Secondly, the cycle time from day 19 to day 26 is lower than the LCL as shown in Fig. 6b, which implies that the performance of fab cycle time is getting improved. Then, the best adaptive coefficient α_{e0} with lowest SSE is set as 0.9 based on the result of comparison test. The new adaptive coefficient α_{e1} is updated as 0.82 by Eq. (18) and the new regression coefficients ($\hat{C}_{ae2} = 0.051$ and $\hat{D}_{ae2} = 1.951$) can be adjusted for constructing the adaptive model in stage 2. Thirdly, the adaptive model need to be adapted as shown in Fig. 6c due to the 8 data continuously below the LCL from day 30 to day 37. The new adapting coefficient α_{e2} is set as 0.884 based on the best adaptive coefficient α_{e1} which has the lowest SSE in stage 1. Then, the new adaptive model in stage 3 is constructed by the new coefficients ($\hat{C}_{ae3} = -0.138$ and $\hat{D}_{ae3} = 1.784$) and new collected data from day 27 to day 37. Fourthly, the average error rate from day 38 to day 97 is about 4.25% without violating the error criteria as shown in Fig. 6d. According to the determined maximum model life cycle, the basic model should be adapted. By the result of comparison test for α_{e2}, the new adapting coefficient α_{e3} is set as 0.177. Then, the new adaptive model in stage 4 is constructed by the new regression coefficients ($\hat{C}_{ae4} = -0.169$ and $\hat{D}_{ae4} = 1.779$) and data collected from day 38 to day 97. Fifthly, the cycle time from day 102 to day 109 is lower than the LCL. The process of model adoption is shown in Fig. 6e. Finally, the new adaptive model in stage 5 is constructed based on the new adjusted

regression coefficients ($\hat{C}_{ae5} = -0.314$ and $\hat{D}_{ae5} = 1.669$) and new cycle time data from day 98 to day 109. In stage 5, all the newly collected data did not violate the error criteria or time criteria, with an average error rate about 3.18% from day 110 to day 139.

Result evaluation and discussion

To improve the forecast accuracy of the adaptive model in each stage, BPNN model was also integrated with GNR model as the basic model. All the historical data in basic

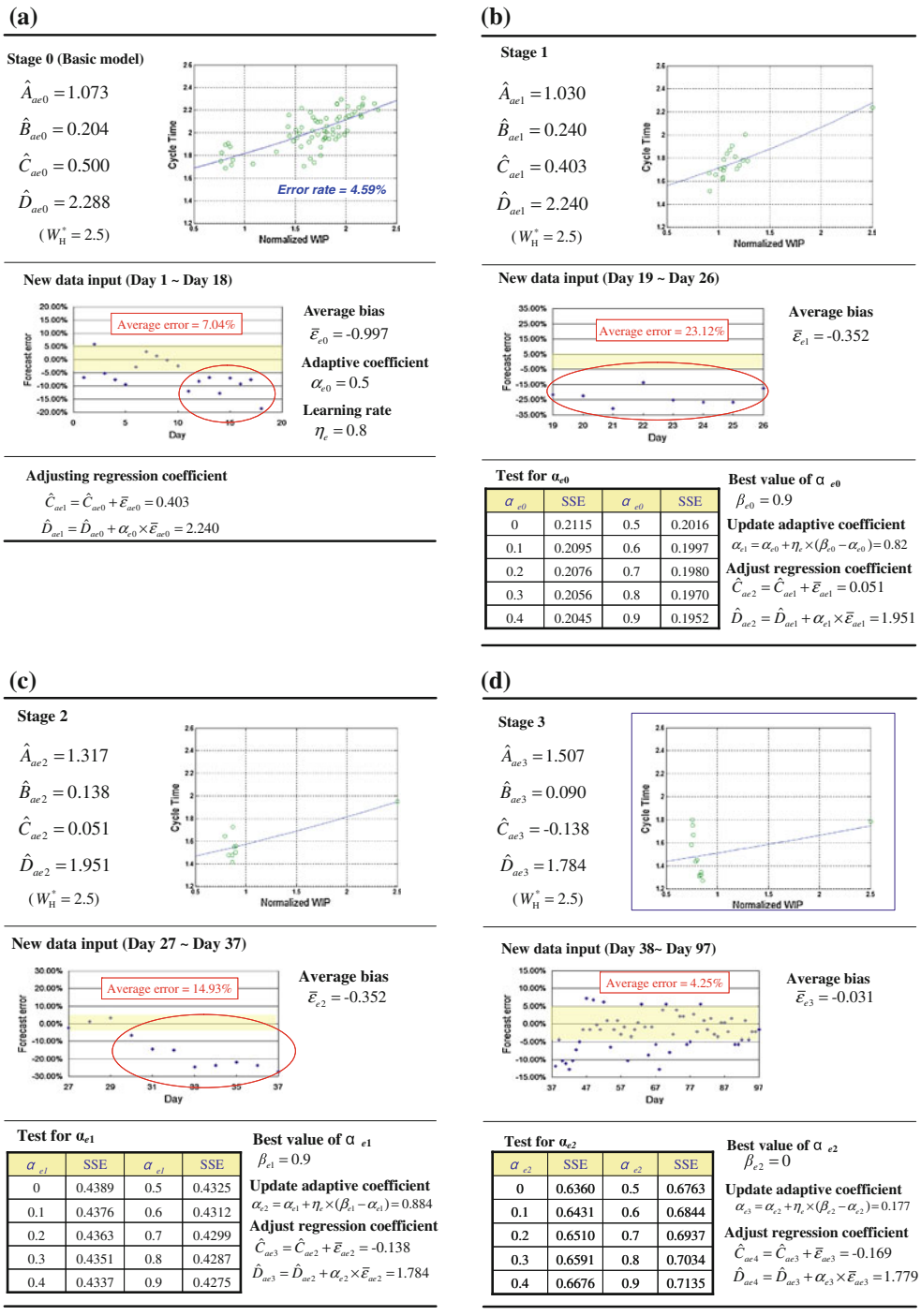
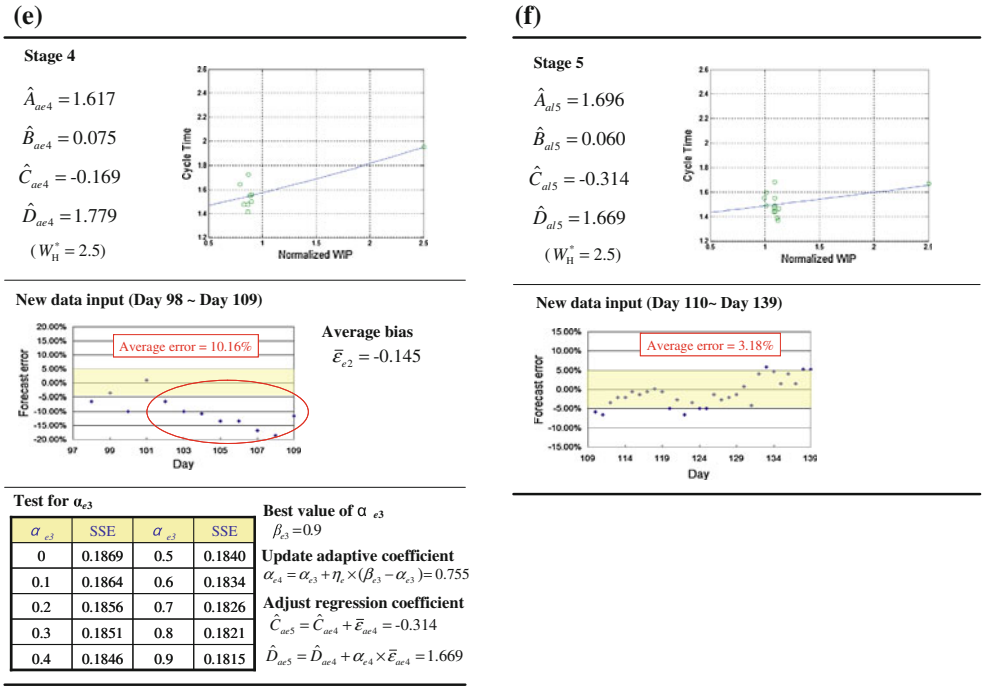


Fig. 6 **a** Adaptive procedure for cycle time forecast at stage 0. **b** Adaptive procedure for cycle time forecast at stage 1. **c** Adaptive procedure for cycle time forecast at stage 2. **d** Adaptive procedure for cycle time forecast at stage 3. **e** Adaptive procedure for cycle time forecast at stage 4. **f** Adaptive procedure for cycle time forecast at stage 5



α_{e3}	SSE	α_{e3}	SSE
0	0.1869	0.5	0.1840
0.1	0.1864	0.6	0.1834
0.2	0.1856	0.7	0.1826
0.3	0.1851	0.8	0.1821
0.4	0.1846	0.9	0.1815

Fig. 6 continued

Table 5 Comparisons in adaptive models

	MOVE			Cycle time					
	Stage 0	Stage 1	Stage 2	Stage 0	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5
Number of days in stage	24	60	55	18	8	11	60	12	30
GNR MAPE (%)	5.32	5.62	3.66	7.04	23.12	14.93	4.25	10.16	3.18
GNR+BPNN MAPE (%)	5.30	3.88	2.92	8.95	9.90	9.32	3.74	8.36	2.56

model and previous stage p of adaptive model were used to construct the BPNN model for the adaptive model in stage p. In particular, 80% of the data are randomly selected for BPNN model training and 20% are used for evaluation.

As shown in Table 5, the integrated GNR and BPNN model have better forecast accuracy via MAPE than only GNR model, except the cycle time forecast at stage 0. The result may be caused from the gap between the long-term sampled data and the short-term daily data. The forecast performance of the integrated model is significantly improved at the following stages by increasing more daily data. The overall MAPE for MOVE and cycle time using GNR model are 4.79% and 6.13%. The overall MAPE from the integrated GNR and BPNN model are 3.75% and 4.80%. The integrated model has an improvement of 21.7% of MAPE in forecasting MOVE and cycle time. With the ability of detecting productivity change immediately, the GNR model can be re-aligned again within 12 days. Although the forecast error of GNR model will be larger given a rapid change in production line status, the forecast accuracy was significantly improved by BPNN model and was less than 10% of few daily data. The forecast result can be presented as tabulations based on the assumption of the calculated average layers and monthly capacity. The corresponding WIP release plans can be made according to the due date and the corresponding target of MOVE and cycle time.

Conclusion

To determine and control appropriate WIP levels is important to maintain the productivity of various tool groups in fabs, which enables the flexibility to meet the demand, while reducing the cycle time and inventory days in the supply chain. The proposed two-phase approach for cycle time forecast can provide valuable information for production decision making to control the corresponding WIP levels. In addition, production managers can quickly identify the latest status of production capability of the fabs, and determine the wafer release plan and scheduling to control WIP. The production

line can also be effectively balanced by aligning with the targets to reduce the cycle time and avoid WIP bubbles causing unnecessary inventory in the supply chain.

In the first phase, the basic model is constructed for deriving a foundation to forecast cycle time and throughput based on the long-term historical production line data. However, the forecast models need be adapted based on the dynamic production status. In the second phase, in order to align with the production change, the adaptive models are designed for detecting the rapid change by measuring the difference of forecast errors. In addition, they can also re-align the forecast models with a small number of newly collected data and still maintain acceptable forecast accuracy. The results showed that the cycle time can be forecasted accurately by the proposed approach even though the status of production line changed dramatically. Considering the complexity of this problem and the practicability in real business, the forecast procedure was designed into certain rule formats or flows to be embedded into the decision support system (DSS). Additionally, the forecast results are presented with charts and tables to support effective decision making. Considering the current fab information such as capacity and number of average layer, the proposed framework can be a useful tool for making WIP level decisions, performance benchmarking of fabs.

The parameter settings of adaptive criteria directly influence the frequency of the adaptive model. Few data will be selected into the adaptive model based on the strict adaptive criteria, in which the models may become sensitive to productivity change and increase the adjustment frequency. On the contrary, loose adaptive criteria decrease the frequency and may lower the forecast quality once the production status is changed. The other parameters such as the learning factor of adaptive coefficients have the similar issues. Further research can be done to discuss how to obtain the optimal parameters considering the quality of forecast result and adapting frequency. Furthermore, other attributes can also be considered as the input of BPNN models, such as seasonality, product mix, performance and capacity of specific high loading machine groups.

Acknowledgments This research is supported by National Science Council (NSC97-2221-E-007-111-MY3) and Taiwan Semiconductor Manufacturing Company (95A0132J8).

References

Atherton, R. W., & Dayhoff, J. E. (1986). Signature analysis: Simulation of inventory, cycle time, and throughput trade-offs in wafer fabrication. *IEEE Transactions on Components, Hybrids, and Manufacturing Technology, 9*(4), 498–507.

Chen, H., Harrison, J. M., Mandelbaum, A., van Ackere, A., & Wein, L. M. (1988). Empirical evaluation of a queueing network model for semiconductor wafer fabrication. *Operations Research, 36*(2), 202–215.

Chien, C.-F., & Chen, C. (2007). Using genetic algorithms (GA) and a coloured timed Petri net (CTPN) for modelling the optimization-based schedule generator of a generic production scheduling system. *International Journal of Production Research, 45*(8), 1763–1789.

Chien, C.-F., Chen, Y., & Peng, J. (2010). Manufacturing intelligence for semiconductor demand forecast based on technology diffusion and product life cycle. *International Journal of Production Economics, 128*(2), 496–509.

Chien, C.-F., Hsiao, A., & Wang, I. (2004). Constructing semiconductor manufacturing performance index and applying data mining for manufacturing data analysis. *Journal of Chinese Institute of Industrial Engineering, 21*(4), 313–327.

Dabbas, R. M., & Chen, H. N. (2001). Mining semiconductor manufacturing data for productivity improvement—an integrated relational database approach. *Computers in Industry, 45*(1), 29–44.

Fowler, J. W., Brown, S., Gold, H., & Schoemig, A. (1997). Measure improvement in cycle-time-constrained capacity. In *Proceedings of the 6th IEEE international symposium on semiconductor manufacturing* (pp. 21–24).

Fowler, J. W., Park, S., Mackulak, G. T., & Shunk, D. L. (2001). Efficient cycle time-throughput curve generation using a fixed sample size procedure. *International Journal of Production Research, 39*(12), 2595–2613.

Kumar, K., & Alsaleh, M. A. (1996). A comparative study for the estimation of parameters in nonlinear models. *Applied Mathematics and Computation, 77*(2–3), 179–183.

Kuo, C., Chien, C., & Chen, C. (2011). Manufacturing intelligence to exploit the value of production and tool data to reduce cycle time. *IEEE Transactions on Automation Science and Engineering, 8*(1), 103–111.

Leachman, R. C., Kang, J., & Lin, V. (2002). SLIM: Short cycle time and low inventory in manufacturing at samsung electronics. *Interfaces, 32*(1), 61–77.

Miltenburg, J., & Sparling, D. (1996). Managing and reducing total cycle time: Models and analysis. *International Journal of Production Economics, 46–47*, 89–108.

Moore, G. E. (1965). Cramming more components onto integrated circuits. *Electronics, 38*(8), 114–117.

Morrison, J. R., & Martin, D. P. (2007). Practical extensions to cycle time approximations for the G/G/m-Queue with applications. *IEEE Automation Science and Engineering, 4*(4), 523–532.

Papadopoulos, H. T., & Vidalis, M. I. (2001). Minimizing WIP inventory in reliable production lines. *International Journal of Production Economics, 70*(2), 185–197.

Sattler, L. (1996). Using queuing curve approximation in a fab to determine productivity improvement. In *Proceedings of 1996 advanced semiconductor manufacturing conference and workshop* (pp. 140–145).

Seber, G. A. F., & Wild, C. J. (1989). *Nonlinear regression*. New York: Wiley.

Yu, C., & Huang, H. (2002). On-line learning delivery decision support system for highly product mixed semiconductor foundry. *IEEE Transactions on Semiconductor Manufacturing, 15*(2), 274–278.