International Journal of Production Research
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tprs20

Mixing macro and micro flowtime estimation model: Wafer fab example

T.-Y. Tseng , T.-F. Ho & R.-K. Li

Published online: 14 Nov 2010.

INT. J. PROD. RES., 1999, VOL. 37, NO. 11, 2447–2461

Mixing macro and micro flowtime estimation model: wafer fab example

T.-Y. TSENG†, T.-F. HO† and R.-K. LI†*

Accurately predicting the order flowtime for a wafer fab is a critical task. Previous investigations involving flowtime estimation of the regression approach have extensively applied the macro flowtime estimation concept. The 'macro' flowtime estimation concept constructs only one aggregate model to estimate the flowtime as a whole. In contrast, the 'micro' flowtime estimation approach constructs an individual regression model for each stage of operation to estimate its individual flowtime. The cumulative order estimated flowtime is then the sum of the individual estimated operation flowtime. Each approach has its own merits and limitations. Therefore, in this study we examine the feasibility of mixing the macro and micro flowtime estimation models. Comparing the proposed model with a variety of macro and micro models reveals that the mixed macro and micro flowtime estimation model can achieve a balance between flowtime estimation error and model complexity.

1. Introduction

Accurately predicting the order flowtime for a wafer fab is a critical task. Typical applications include assigning due-dates for customers, planning the order release, and evaluating the performance of scheduling policies for any unforeseen disturbances. Calculating the order flowtime allowance is not straightforward owing to the dynamic nature of wafer fab, in which new wafers are constantly arriving and order priorities are constantly changing. Although developing a system capable of always accurately predicting order flowtime is impossible, a relatively simple yet accurate method is highly desired.

The order flowtime is based on the processing time of operations and the order interoperation time, which consists of queue time, move time and waiting time. The queue delays are caused by resource contention due to factors such as machine status, variability in processing times, variability in arrival times, and variability in batch sizes. Investigations involving flowtime estimation have generally applied four different approaches. The first is the simulation approach (Weeks and Fryer 1977, Weeks 1979, Bertrand 1983, Baker 1984, Srivatsan and Kemf 1995, Lawrence 1995). By using current orders on shop and shop loads to simulate forward in time, the simulation is extremely general and can model nearly any conceivable production system; however, it is computationally expensive and time consuming (Connors *et al.* 1996).

The second approach is analytical research such as the queueing model. Previous works involving analytical models (Heard 1976, Seidmann and Smith 1981, Cheng

Revision received August 1998.

† Department of Industrial Engineering and Management, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu, Taiwan 30050.

‡ To whom correspondence should be addressed.

International Journal of Production Research ISSN 0020–7543 print/ISSN 1366–588X on line © 1999 Taylor & Francis Ltd
http://www.tandf.co.uk/JNLS/prs.htm
http://www.taylorandfrancis.com/JNLS/prs.htm

and Gupta 1989, Bookbinder and Noor 1985, Karmarkar 1987, Wein 1991, Lynes and Miltenburg 1994, Enns 1995, Connors *et al.* 1996) have elucidated problems that limit their practices. For example, they are valid only for a first-come, first-served (FCFS) dispatch rule. However, in practice, the FCFS is not generally recognized as the optimal one. Another problem is that the shop floor must be in a steady state, which is impossible in nearly every job shop (Baudin *et al.* 1992).

The third approach is the neural network model. By using historical data as the input variables, the neural network can capture the previous pattern of the system and predict on the basis of future events. Although a few cases have been implemented in manufacturing (Udo 1992, Philipoon and Fry 1992, Zhang and Huang 1995), the unknown process engine of a neural network increases the suspicion and unwillingness to accept in most practitioners.

The fourth approach is the regression model. By using historical data similar to how a neural network functions, the regression model differs from a neural network only in that the former appears to be more reasonable than the latter from the statistical perspective. Earlier works on flowtime estimation such as CON, RDM, TWK, NOP, JIQ (Eilon and Chowdhury 1976), JIS (Week 1979), OFS (Vig and Dooley 1991), and COFS (Vig and Dooley 1993) belong to this approach. However, these heuristic models can be viewed as flowtime prediction models only if regression analysis is used to determine the parameters.

Although numerous regression related heuristics have been developed and tested for estimating the flowtime or due-date (with some of these apparently quite effective), these rules have still not been fully developed owing to two features: the linear independence of explanatory variables and a macro view. The linear independent assumption prevents more explanatory variables to join the regression model because the mutually independent explanatory variables are rare in reality. Consequently, a trade-off arises between easy explanation and better fitting regardless of whether or not more dependent variables are involved. Therefore, the relative importance of the explanatory variables and their influence on the flowtime determination must be systematically investigated to minimize the prediction error. Stepwise regression is commonly used to identify the following: (1) the important variables in the flowtime estimation, (2) the interaction effects between the input variables, (3) the optimal form of explanatory variables and (4) the coefficients of estimation equation.

With the regression prediction approach, either the macro or micro flowtime estimation concept can be applied. The 'macro' flowtime estimation concept constructs only an aggregate regression model to estimate the flowtime as a whole. Its explanatory variables are total processing time, total number of jobs, or total number of operations. This concept prevails in conventional regression flowtime estimation approaches. However, in a wafer fab, over 100 operational stages are common, and each operational stage may have unique features. The coefficients of the regression flowtime estimation model for each operation may be not equivalent statistically; and then the forecasting error with the macro flowtime estimation approach may be too large to be acceptable. Thus, the validity of the macro concept is questionable. A 'micro' flowtime estimation approach, unlike a macro view, constructs an individual regression model for each operation stage to estimate its individual operational stage flowtime. Assume that only interaction effects among explanatory variables of the same stage are included, and that those of different stages are less significant and are thus neglected. Then the total order estimated

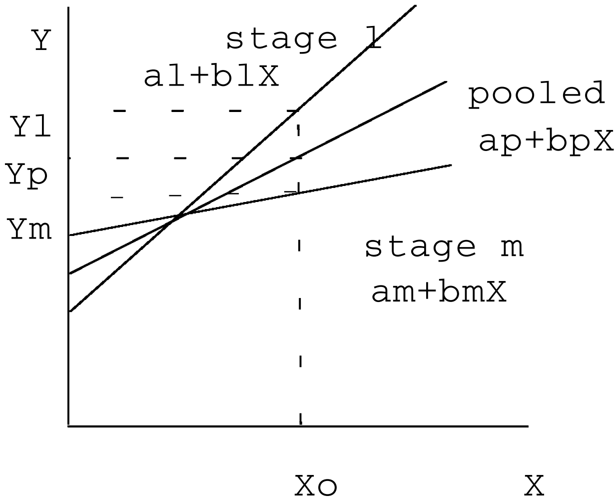

Figure 1. Pooled and distinct regression.

flowtime equals the sum of the individual estimated operation flowtime. An illustrative example presented here demonstrates that properly selecting the explanatory variables can reduce the interaction effects between operation stages.

Chatterjee and Price (1991) analysed a situation in which a regression equation for a pooled data set can represent separated regression equations for subsets of the data collected. A serious bias may be incurred if one regression relationship is used to represent the pooled data set. Figure 1 depicts two flowtime estimation models: macro and micro models. Let Y represent the flowtime and let X be work in process (WIP). In the macro model, the operational stage distinction is ignored, the data are pooled and only one line exists. However, in the micro model two separate regression lines are available for the both operational stages, l and m, each with distinct regression coefficients. The graph indicates that if X_l and X_m have been set at certain levels of WIP for operational stages l and m respectively, by using the macro model, the total expected flowtime is $a_p + b_p(X_l + X_m)$ when total WIP on the shop is $X_l + X_m$. However, if the micro model is selected, the actual flowtime estimates are $Y_l(= a_l + b_l X_l)$ for stages l and $Y_m(= a_m + b_m X_m)$ for stage m when the individual stages WIP are X_l and X_m. Substituting $(a_p + b_p(X_i + X_m))$ for $(a_l + a_m + b_l X_l + b_m X_m)$ represents a bias for stages l and m. Whether the macro or micro model is used depends on if $(a_p + b_p(X_l + X_m))$ is within a certain level of confidence intervals of $(a_l + a_m + b_l X_l + b_m X_m)$.

Owing to the unavoidable bias incurred for the macro flowtime estimation from pooled data (Chatterjee and Price), predicting micro flowtime estimation is more accurate than estimating the macro flowtime. Although the micro concept reduces the forecasting error, the complexity of the estimation model is too high. This model is too difficult to implement in practice, particularly in a wafer fab, where over 100 operational stages are common. Therefore, in this study we examine the feasibility of mixing the macro and micro flowtime estimation models. The proposed model initially applies the micro approach to construct an individual regression equation for each operation stage. A macro approach is then used to group together those operational stage regression equations that have the same combination of explanatory variables but with different coefficients. A macro flowtime estimation equation is then fitted for those clustered operation stages. The criterion of whether those operation stages can be fitted into a macro regression equation is based on its prediction error. The mixed macro and micro flowtime estimation model is constructed and updated on the basis of the samples of the recently completed orders. Next, the total flowtime through all operational stages is estimated by merely summing the value of explanatory variables of each stage into the corresponding grouped macro equations instead of summing individual stage flowtimes estimated by micro equations. Thus, a balance can be achieved between flowtime estimation error and model complexity. In addition, the proposed model is also compared with various macro and micro flowtime estimation models in terms of factors such as mean absolute deviation, mean square error, mean absolute percent error, number of tardy jobs, percent of tardy jobs, average lateness, and standard deviation of lateness.

2. Fundamental concept and algorithm for the mixed macro and micro flowtime estimation model

Assume that a set of data consists of K explanatory variables, $X_1, X_2, \dots, X_k$, which may include interaction effects among input variables of the same operation stage and P operational stages, then the macro and micro flowtime estimations through the whole stages are Y and Y^*, respectively:

$$\left.\begin{aligned} \widehat{Y} &= \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K \\ \widehat{Y} &= \sum_{i=1}^{P} \widehat{Y}^* = \sum_{i=1}^{P} \left(\beta_{i0} + \beta_{i1} X_{i1} + \cdots + \beta_{iK} X_{iK}\right). \end{aligned}\right\} \quad (1)$$

The coefficients of the macro equation, $(\beta_0, \beta_1, \dots, \beta_K)$, are recalculated from pooled data of P operation stages. If one macro flowtime estimation can replace P micro flowtime estimation equations, then the micro approach total flowtime estimation Y^* can be replaced by the macro approach flowtime estimation Y which is simply calculated by summing the explanatory variables of P operation stages into the macro equation, where

$$X_1 = \sum_{i=1}^{P} X_{i1}, \dots, X_K = \sum_{i=1}^{P} X_{iK}.$$

Since the micro flowtime estimation model fits the shop well and the total flowtime estimation used in the micro flowtime estimation model is equal to the summation of the flow time estimation of each individual operation stage, no bias variance is incurred. Therefore, the sum of square error (SSE) for micro flowtime estimation is lower and the complexity is higher than that for macro flowtime estimation. On the other hand, the SSE for the macro flowtime estimation is in the upper bound and the complexity is in the lower (Chatterjee and Price).

Therefore, this work attempts to achieve a balance between estimation complexity and variability. Towards this end, we propose mixed macro and micro flowtime estimation. The proposed model initially applies micro flowtime estimation to construct a regression equation for each individual operation stage. Those individual operational stage regression equations that have the same combination of explanatory variables but different coefficients are then grouped together. A test is performed to observe whether those grouped operational stage regression equations can be represented as a macro regression equation. If not, we discard the stage regression equation that has the highest SSE and repeat the process again until a new fitted macro equation can be accepted. The algorithm involves two stages.

The first stage consists of regression equation fitting for each operation stage. This stage resembles any stepwise regression equation fitting. Stepwise regression attempts to achieve the following.

(a) Identify the most significant explanatory variables that influence flowtime, including interaction among input variables of the same stage.
(b) Perform the variable transformation. The forms of explanatory variables and equations for each stage could be entirely different, as attributed by shop status such as WIP distribution, and different workloads. The variable transformation adopted here attempts to increase the prediction accuracy.
(c) Estimate the coefficients of the estimated equation for each operation stage. In addition, any statistical package such as SAS or SPSS can easily perform this task.

The second stage consists of individual regression grouping and new macro regression equation fitting. Based on the above fundamental concept, the grouping and fitting processes can be performed in six steps.

Step 1. Temporally groups together those operational stages that have the same combination of explanatory variables but different coefficients. For each group, perform steps 2 to 6.

Step 2. For each group two sets S and $\bar{S}$ and a function $|S|$ are given, where S = all operation stages that intend to be grouped into the same group, $\bar{S}$ = all operation stages that are deleted from S $|S|$ = number of elements in S. The group of regression equations expressed as follows is called model 1.

$$\textit{Model}\ 1:\left.\begin{array}{l} \widehat{Y_1^*} = \beta_{10} + \beta_{11}X_{11} + \cdots + \beta_{1k}X_{1k} + \varepsilon_1 (\text{operation stage 1}) \\ \widehat{Y_2^*} = \beta_{20} + \beta_{21}X_{21} + \cdots + \beta_{2k}X_{2k} + \varepsilon_2 (\text{operation stage 2}) \\ \quad\vdots \\ \widehat{Y_P^*} = \beta_{P0} + \beta_{P1}X_{P1} + \cdots + \beta_{Pk}X_{Pk} + \varepsilon_P\ (\text{operation stage P}), \end{array}\right\} \tag{2}$$

where the explanatory variables, $\left(X_{i1}, X_{i2}, \ldots, X_{iK}, i = 1, 2, ..., P\right)$, of P operation stages may include interaction among input variables of the same operation stage.

Step 3. For each group, construct a new macro regression equation called model 2. In model 2, operation stage distinction is ignored, the data are pooled, and there is one re-estimated regression line.

$$\textit{Model}\ 2:\ \widehat{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K + \varepsilon\ (\text{pooled}). \tag{3}$$

Step 4. Test the hypothesis to see whether those operational stages of each group can be represented as a macro regression equation. Assume that a macro group has p operation stages $\left(S_1, S_2, \ldots, S_P\right)$ and that each operation stage has n observations with K identical explanatory variables but different coefficients. The model 2 is the macro regression equation of the P operation stages. The operational stage distinction is ignored and nP observations are included. Model 1 is an individual regression equation for each p operation stage with n observations individually. Formally, we want to test the null hypothesis

$$H_0: \quad \beta_{10} = \beta_{20} = \cdots = \beta_{p0},$$
$$\beta_{11} = \beta_{21} = \cdots = \beta_{p1}, \ldots, \beta_{1k} = \beta_{2k} = \cdots = \beta_{pk}$$

against the alternative that there are substantial differences among them. The test can be performed using P indicator variables $(T_1, T_2, \ldots, T_P)$, where,

$$\left.\begin{aligned} T_1 &= \begin{cases} 1, & \text{if observation was from stage 1} \\ 0, & \text{otherwise} \end{cases} \\ T_p &= \begin{cases} 1, & \text{if observation was from stage } P \\ 0, & \text{otherwise.} \end{cases} \end{aligned}\right\} \tag{4}$$

A new model, model 3, which represents any stage regression equation in model 1 is then expressed as:

$$\left.\begin{aligned} \textit{Model 3:} \quad \widehat{Y}^* = {} & \beta_0 + \beta_1 X_{i1} + \cdots + \beta_K X_{iK} \\ & + \gamma_1 T_1 + \gamma_2 T_2 + \cdots \gamma_P T_P \\ & + \delta_{11} T_1 X_{i1} + \delta_{12} T_1 X_{i2} + \cdots + \delta_{1k} T_1 X_{ik} \\ & + \delta_{21} T_2 X_{i1} + \delta_{22} T_2 X_{i2} + \cdots + \delta_{2k} T_2 X_{ik} \\ & \qquad \vdots \\ & + \delta_{P1} T_P X_{i1} + \delta_{P2} T_P X_{i2} + \cdots + \delta_{Pk} T_P X_{ik}, \end{aligned}\right\} \tag{5}$$

where $\beta_{i0} - \beta_0 = \gamma_i, \beta_{ij} - \beta_j = \delta_{ij}, i = 1, 2, \ldots, P, \ j = 1, 2, \ldots, K$ and by changing indicator variables the model 3 becomes

$$\left.\begin{aligned} \widehat{Y}_1^* &= (\beta_0 + \gamma_1) + (\beta_1 + \delta_{11})X_{11} + \cdots + (\beta_K + \delta_{1K})X_{1K} + \varepsilon_1 \ (\text{operation } 1) \\ \widehat{Y}_2^* &= (\beta_0 + \gamma_2) + (\beta_1 + \delta_{21})X_{21} + \cdots + (\beta_K + \delta_{2K})X_{2K} + \varepsilon_2 \ (\text{operation } 2) \\ & \qquad \vdots \\ \widehat{Y}_P^* &= (\beta_0 + \gamma_P) + (\beta_1 + \delta_{P1})X_{P1} + \cdots + (\beta_K + \delta_{PK})X_{PK} + \varepsilon_P \ (\text{operation } P). \end{aligned}\right\} \tag{6}$$

Assume that Var $(\varepsilon_1) =$ Var $(\varepsilon_2) = \cdots =$ Var (ε_p) and the null hypothesis test of H_0 becomes

$$\gamma_1 = \gamma_2 = \cdots = \gamma_1 = \gamma_2 = \cdots = \gamma_P = \delta_{11} = \delta_{21} = \cdots = \delta_{PK} = 0$$

whether each operation stage regression equation of model 3 can be combined into one macro equation of model 2. Model 3 is referred to here as the full model (FM) and model 2 is referred to here as the reduced model (RM). Let $\hat{Y}^*$ and $\hat{Y}$ be the estimates for actual flowtime Y by the full model and the reduced model respectively. In the full model there are P estimation equations and each estimate has n observations. While in the reduced model, there is only an estimation equation with nP observations. The SSE for the full model and reduced model are as follows:

$$\left.\begin{aligned} SSE(FM) &= \sum^{P}\sum^{n}\left(Y-\hat{Y}^*\right)^2 \\ SSE(RM) &= \sum^{nP}\left(Y-\hat{Y}\right)^2. \end{aligned}\right\} \tag{7}$$

In the full model there are $P(K+1)$ parameters $(\gamma_1, \gamma_2, \ldots, \gamma_P, \delta_{11}, \delta_{12}, \ldots, \delta_{PK})$ and in the reduced model there are $K+1$ parameters $(\beta_0, \beta_1, \ldots \beta_k)$. To assess the feasibility of the reduced model, we compare SSE(RM) − SSE(FM) with SSE(FM). The ratio is as follows:

$$F = \frac{[\mathrm{SSE(RM)} - \mathrm{SSE(FM)}]/(K+1)(P-1)}{\mathrm{SSE(FM)}/P(n-K-1)}. \tag{8}$$

If the observed F value is larger than the tabulated value of $F(1-\alpha, (K+1)(P-1), P(n-K-1))$ at α percent level, then the reduced model is unsatisfactory, and the null hypothesis is therefore rejected (implying that those operation stages cannot be grouped together); in this case, go to step 5. Otherwise, the null hypothesis is accepted, implying that those operation stages can be grouped together; then go to step 6. Recall that it was assumed that variances were identical in the p subgroups. A plot of residuals versus the indicator variables verifies this assumption. If the p sets do not significantly differ, the equal variance assumption is accepted. Otherwise, this implies that the remaining P stages do not correlate with the grouping, go back to step 1 and fit another group.

Step 5. Delete the operation stage m with maximum $SSE(FM_i)$ defined as

$$\mathrm{Max}\left(\mathrm{SSE}(\mathrm{FM}_i)\right) = \mathrm{Max}\left(\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i^*\right)^2\right) \tag{9}$$

from set S and put it into set $\bar{S}$. A situation in which the number of elements in S is less than 2 implies no grouping; then go to step 6. Otherwise, redefine the full model and the reduced model without stage m; then go back to step 2.

Step 6. Reconsider the possibility of grouping to one regression equation from all stages in $\bar{S}$. Although the operational stages in $\bar{S}$ are those deleted from S, they may still be grouped together. Therefore, put all stages in $\bar{S}$ back to S and repeat step 2 again. If the number of elements in $\bar{S}$ is less than 2, leave that operation alone.

After all groups formed in step 1 complete their fitting, the mixed macro and micro flowtime estimation model is constructed. The total flowtime through all operational stages is estimated by summing the value of explanatory variables for each stage into corresponding grouped macro equations (model 2) instead of summing individual stage flowtime estimated by micro equations (model 1).

3. Illustrative examples

To demonstrate the feasibility of our developed procedure, two examples are presented. Details of the test are provided for example one and the second example provides the results only.

3.1. *Example 1:*

The data used to perform the developed concept are accumulated from Mosel Vitelic Inc., one of Taiwan's largest semiconductor-manufacturing companies. The data are automatically downloaded from Mosel's real-time shop floor control software (WORKSTREAM) to our model. The data collection spans from March to December 1996 on a weekly basis. Data from March to August are used to determine the regression parameters, while data from September to December are used for performance evaluation. Since this research is intended to demonstrate the feasibility of the mixed model, only the final 21 operation stages of the 4MB DRAM fabrication process are demonstrated.

Here, let Y be the cycle time, let X_1 be the WIP in lot (1 lot = 50 wafers), let X_2 be the process time, and let X_3 be the move volume in lots. The mixed model initially performs regression equation fitting for each operational stage. Table 1 lists those 21 regression equations after performing stepwise regression. The forms of explanatory

No.	Stage name	Regression equation
1	BPSG2 CVD	$Y = 23.4 - 0.07(X_1) + 152.6\log(X_2) + 1.08(X_3)^{1/2} + 0.4 \times \log(X_2) \times (X_3)^{1/2}$
2	BPSG2 FLOW	$Y = 8.23 - 0.9(X_1)^{1/2} + 3.3\log(X_2) + 1.98(X_3)$
3	3C PHOTO	$Y = 120.23 - 27.7\log(X_1) + 0.74(X_2)^{1/2} + 0.69(X_3)$
4	3C ETCHING	$Y = 43.7 - 108.3\log(X_1) + 115.1\log(X_2) + 1.36(X_3)$
5	RTA	$Y = 12.2 - 10.3\log(X_1) + 20.3\log(X_2) + 0.5(X_3)$
6	TI SPUTTER	$Y = 21.1 - 1.5(X_1)^{1/2} + 143.6\log(X_2) + 0.44(X_3)$
7	PLANKET W CVD	$Y = 3.4 - 0.02(X_1) + 103.3\log(X_2) + 1.1(X_3)^{1/2} + 0.5 \times \log(X_2) \times (X_3)^{1/2}$
8	W ETCH BACK	$Y = 32.2 - 20.1\log(X_1) + 0.5(X_2)^{1/2} + 0.7X_3$
9	BAKING	$Y = 23.7 - 93.1\log(X_1) + 89.3\log(X_2) + 1.05(X_3)$
10	WEB SCRUBBING	$Y = 2.1 - 9.7\log(X_1) + 27.1\log(X_2) + 23.1\log(X_3)$
11	AlSiCu SPUTTER	$Y = 13.43 - 1.7(X_1)^{1/2} + 154.1\log(X_2) + 0.65(X_3)$
12	METAL PHOTO	$Y = 9.3 - 1.0(X_1) + 3.7\log(X_2) + 2.1(X_3)$
13	METAL ETCHING	$Y = 34.7 - 18.3\log(X_1) + 19.2\log(X_2) + 1.54(X_3)$
14	SINTERING	$Y = 14 - 1.4(X_1)^{1/2} + 3.0\log(X_2) + 1.86(X_3)$
15	PAD PSG CVD	$Y = 12.6 - 10.2\log(X_1) + 94\log(X_2) + 10.7(X_3)$
16	PE-NITRIDE CVD	$Y = 3.1 - 9.3\log(X_1) + 24.3\log(X_2) + 30.7\log(X_3)$
17	PV PHOTO	$Y = 99.1 - 0.8(X_1)^{1/2} + 123.3\log(X_2) + 0.3(X_3)$
18	PV ETCH	$Y = 13.7 - 116.2\log(X_1) + 125.4\log(X_2) + 2.76(X_3)$
19	PI PHOTO	$Y = 2.6 - 8.30\log(X_1) + 26.5\log(X_2) + 26.9\log(X_3)$
20	UV03 ASHING	$Y = 6.6 - 117\log(X_1) + 104(\log(X_2) + 9.3(X_3)$
21	PI CURE	$Y = 17.6 - 1.6(X_1)^{1/2} + 164.3\log(X_2) + 0.54(X_3)$

Table 1. Regression equations for 21 operational stages.

variables and equations for each stage are not exactly the same, which may possibly be attributed to factors such as WIP distribution, and different workload ratios. As mentioned earlier, variable transformations are adopted herein to increase prediction accuracy.

By performing step 1 of stage 2, those operational stages with the same combination of explanatory variables but different coefficients are then placed into five groups, as follows:

Group 1: operation stages 1, 7
Group 2: operation stages 2, 6, 11, 12, 14, 17, 21
Group 3: operation stages 3, 8
Group 4: operation stages 4, 5, 9, 13, 15, 18, 20
Group 5: operation stages 10, 16, 19.

Next, the feasibility of combining those operation stages grouped into one macro model is examined. Group 4 is considered as an example. For the full model (model 3), 7 $(P = 7)$ regression equations exist, each with 4 $(K = 3)$ parameters and 25 $(n = 25)$ observations, yielding a total of 28 degrees of freedom. However, for the reduced mode (model 2), there are 4 parameters after merging 175 observations of operation stages 4, 5, 9, 13, 15, 18, and 20. The reduced model (or macro model) for group 4 is then expressed as follows:

$$Y = 32.2 - 106.1Z_1 + 102.4Z_2 + 1.45Z_3$$

$$Z_1 = \log(X_1), \quad Z2 = \log(X_2), \quad Z3 = X_3.$$

Now, assume that

$$\mathrm{Var}(\varepsilon 1) = \mathrm{Var}(\varepsilon 2) = \mathrm{Var}(\varepsilon 3) = \cdots = \mathrm{Var}(\varepsilon 7)$$

and by adding $P(=7)$ indicator variables $(T_1, T_2, \ldots, T_7)$, the model 3 becomes

$$\begin{aligned}\text{Model 3:}\ \widehat{Y^*} = {} & \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 \\ & + \gamma_1 T_1 + \gamma_2 T_2 + \cdots + \gamma_7 T_7 \\ & + \delta_{11} T_1 Z_1 + \delta_{12} T_1 Z_2 + \delta_{13} T_1 Z_3 \\ & + \delta_{21} T_2 Z_1 + \delta_{22} T_2 Z_2 + \delta_{23} T_2 Z_3 \\ & \quad \vdots \\ & + \delta_{71} T_7 Z_1 + \delta_{72} T_7 Z_2 + \delta_{73} T_7 Z_3.\end{aligned}$$

Table 2 lists each of the parameter values.

The SSE for the full model is 699 230 and for the reduced model it is 1 097 690, and the F ratio is then computed as:

$$\begin{aligned}F &= \frac{[\mathrm{SSE(RM)} - \mathrm{SSE(FM)}]/(K+1)(P-1)}{\mathrm{SSE(FM)}/P(n-K-1)} \\ &= \frac{(1\,097\,690 - 699\,230)/[(3+1)(7-1)]}{699\,230/7(25-3-1)} = 3.49.\end{aligned}$$

Since

$\gamma_1 = 11.5$, $\gamma_2 = -20$, $\gamma_3 = -8.5$, $\gamma_4 = 2.5$, $\gamma_5 = -19.6$, $\gamma_6 = -18.5$, $\gamma_7 = -25.6$
$\beta_0 = 32.2$, $\beta_1 = -106.1$, $\beta_2 = 102.4$, $\beta_3 = 1.45$
$\delta_{11} = -2.2$, $\delta_{12} = 12.7$, $\delta_{13} = -0.09$
$\delta_{21} = 95.8$, $\delta_{22} = -82.1$, $\delta_{23} = -0.95$
$\delta_{31} = 13$, $\delta_{32} = -13.1$, $\delta_{33} = -0.4$
$\delta_{41} = 87.8$, $\delta_{42} = -83.2$, $\delta_{43} = 0.09$
$\delta_{51} = 4.1$, $\delta_{52} = -8.4$, $\delta_{53} = 9.25$
$\delta_{61} = -10.1$, $\delta_{62} = 23$, $\delta_{63} = 1.31$
$\delta_{71} = -10.9$, $\delta_{72} = 1.6$, $\delta_{73} = 7.85$

Table 2. The parameter value of group 4.

$$F_\alpha((K+1)(P-1), P(n-K-1)) = F_{0.05}(24, 147) = 1.6$$

it is larger than $F_\alpha(24, 147)$. Therefore, the null hypothesis is rejected, implying that those operational stages in group 4 cannot be combined into one macro model. It is then initiated to compute the $\mathrm{SSE}(\mathrm{FM}_i)$ for each operational stage in step 5. Operational stage 5 is then deleted from the group because its $\mathrm{SSE}(\mathrm{FM}_i)$ is the maximum. Perform step 2 again and the operation stage 13 is further deleted from the group. Finally, the operational stages 4, 9, 15, 18, 20 can be grouped into the equation

$$Y = 7.2 - 117.1Z_1 + 105.4Z_2 + 1.03Z_3$$

for F ratio is equal to 1.69 and less than $F_{0.05}(16,105) = 1.80$.

Step 6 then considers the possibility of grouping operational stages 5, 13 in $\overline{S}$ and finds that they can be grouped to the equation

$$Y = 23.4 - 14.2Z_1 + 19.8Z_2 + 1.02Z_3$$

for F ratio equal to 1.43 and less than $F_{0.05}(4, 42) = 2.61$. Finally, after passing residuals analysis that verifies equal variances among seven stages, all seven operational stages in group 4 can be statistically grouped into the two following regression equations:

$$Y = 7.2 - 117.1Z_1 + 105.4Z_2 + 1.03Z_3$$

$$Y = 23.4 - 14.2Z_1 + 19.8Z_2 + 1.02Z_3.$$

The total flowtime through all operational stages in group 4 is the sum of two flowtime estimations. The first one is estimated by summing the value of explanatory variables for stages 4, 9, 15, 18, and 20 into the macro equation

$$Y = 7.2 - 117.1Z_1 + 105.4Z_2 + 1.03Z_3$$

and the second one is those for stages 5 and 13 into macro equation

$$Y = 23.4 - 14.2Z_1 + 19.8Z_2 + 1.02Z_3.$$

The same procedure is performed for the groups 1, 2, 3, and 5. The results are listed as follows:

Group 1: operation stages 1 and 7 can be grouped into the equation

$$Y = 13.6 - 0.05(X_1) + 130.1\log(X_2) + 1.09(X_3)^{1/2} + 0.45 \times \log(X_2) \times (X_3)^{1/2}$$

Group 2: operation stages 2, 12 and 14 can be grouped into the equation

$$Y = 10.7 - 1.1(X_1)^{1/2} + 3.4\log(X_2) + 1.97(X_3)$$

Operation stages 6, 11 and 21 can be grouped into the equation

$$Y = 17.3 - 1.59(X_1)^{1/2} + 152.1\log(X_2) + 0.51(X_3)$$

Operation stage 17 becomes

$$Y = 99.1 - 0.8(X_1)^{1/2} + 123.3\log(X_2) + 0.3(X_3)$$

Group 3: operation stages 3 and 8 can be grouped into the equation

$$Y = 129.2 - 24.1\log(X_1) + 0.6(X_2)^{1/2} + 0.7X_3$$

Group 5: operation stages 10, 16 and 19 can be grouped into the equation

$$Y = 99.1 - 0.8(X_1)^{1/2} + 123.3\ \log(X_2) + 0.3(X_3).$$

In summary, the regression equations for all 21 operational stages can be reduced to 8.

The total estimated flowtime of mixed macro and micro model is then the sum of estimated flowtime of 21 operational stages by summing the value of explanatory variables into 8 macro models instead of 21 micro models. Next, the performance evaluation for our mixed macro and micro model is compared with various macro and micro models (with and without variable transformations). The reasons why both macro and micro models with and without variable transformation are included in the performance evaluation is to verify whether or not variable transformation increases the prediction accuracy at the cost of the difficulty to comprehend. The macro models include seven conventional flowtime estimation models, as proposed by Cheng and Gupta (1989), and two new macro models with and without variable transformations. All definitions of explanatory variables for seven conventional macro models are the same as those of previous studies except WIP. Apparently, WIP in lots is used to replace the number of jobs in the system. Both new macro models use the same input variables with mixed model except the one without transformation, which uses the original input variables. Table 3 presents the equations and coefficients of the seven models and two new macro flowtime estimation models in which the coefficients are acquired by summing the value of explanatory variables of 21 stages into the macro equation. Table 1 and 4 present the parameters of micro flowtime estimation with and without variable transformation, respectively. The evaluation criteria are based on the factors related to due date performance, e.g. mean absolute deviation, mean square error, mean absolute percent error, number of tardy jobs, percent of tardy jobs, average lateness and standard deviation of lateness. Detailed performance measures are taken from Vig and Dooly (1993). Table 5 lists the performance equations and compares the results. The coefficients of regression equations are determined from data from March to August. Meanwhile, seven performance results are generated on the basis of data from September to December where the total observations are $n\ (= 8)$. Four important findings are observed.

(a) The selection of explanatory variables is critical to the performance of macro models. The relative priorities of the explanatory variables are as follows: flowtime distribution (OFS, COFS, JIS), WIP (JIQ, COFS), processing time (TWK, SLK), and the number of operations (NOP). Selecting the proper combination of explanatory variables can increase the prediction accuracy.

	Model	Equation	Coefficients
1	TWK	$\mathrm{CT} = KPO_1$	$K = 1.82$
2	NOP	$\mathrm{CT} = Kn_u$	$K = 159$
3	SLK	$\mathrm{CT} = P_i + K$	$K = 1385$
4	JIQ	$\mathrm{CT} = K_lP_i + K_2Q_i$	$K_1 = 345,\ K_2 = 28$
5	JIS	$\mathrm{CT} = P_i + D + a(J_i)\sigma_D$	$D = 1390\ \sigma_D = 34.9$
6	OFS	$\mathrm{CT} = K_1F_i + K_2N_i + K_3P_i$	$K_1 = 33,\ K_2 = 43,\ K_3 = 33$
7	COFS	$\mathrm{CT} = K_1F_1 + K_2n_i + K_3P_i + K_4Q_i$	$K_1 = 33,\ K_2 = 2.45,\ K_3 = 22,\ K_4 = 2.3$
8	Macro without trans.	$\mathrm{CT} = K_0 + K_1Q_i + K_2P_i + K_3O_i$	$K_0 = 31, K_1 = 6.12,\ K_2 = 12.4,\ K_3 = 10.2$
9	Macro with trans.	$\mathrm{CT} = K_0K_1 \log Q_i + K_2 \log P_i + K_3 \log O_i + K_4(Q_iP_i)^{1/2}$	$K_0 = 2.7,\ K_1 = 50.4,\ K_2 = 24.4,$ $K_3 = 11.2,\ K_4 = 2.41$

CT = cycle time estimation, P_i = total processing time of product i, n_i = number of operations of product i, Q_i = WIP in lots of product i, D = Mean waiting in the system, σ_D = standard deviation of waiting time in the system, J_i = WIP in the system when job I arrives, J = mean WIP in the system, σ_j = standard deviation of WIP in the system, $a(J_i) = -1$ if $J_i < J - \sigma_j$, $= 0$ if $J - \sigma_j < J_i < J + \sigma_j$, $= 1$ if $J_i > J + \sigma_j$, $F_i = T \cdot n_i$, T = average flowtime per operation for newly completed batches, O_i = move out quantity of product i.

Table 3. Macro models and the model equations.

No.	Stage name	Regression coefficients			
1	BPSG2 CVD	$\beta_0 = 3.2,$	$\beta_1 = -1.2,$	$\beta_2 = 5.6,$	$\beta_3 = -11.3$
2	BPSG2 FLOW	$\beta_0 = 1.4,$	$\beta_1 = -3.2,$	$\beta_2 = 4.4,$	$\beta_3 = -3.41$
3	3C PHOTO	$\beta_0 = 0.5,$	$\beta_1 = -12.2,$	$\beta_2 = 21.1,$	$\beta_3 = -0.19$
4	3C ETCHING	$\beta_0 = 2.1,$	$\beta_1 = -132.2,$	$\beta_2 = 0.9,$	$\beta_3 = -42.3$
5	RTA	$\beta_0 = 0.4,$	$\beta_1 = -142.2,$	$\beta_2 = 3.6,$	$\beta_3 = -7.5$
6	TI SPUTTER	$\beta_0 = 90.2,$	$\beta_1 = -13.9,$	$\beta_2 = 5.0,$	$\beta_3 = -23.2$
7	PLANKET W CVD	$\beta_0 = 12.2,$	$\beta_1 = -321.7,$	$\beta_2 = 6.2,$	$\beta_3 = -9.63$
8	W ETCH BACK	$\beta_0 = 4.2,$	$\beta_1 = -42.2,$	$\beta_2 = 0.77,$	$\beta_3 = -31.2$
9	BAKING	$\beta_0 = 23.1,$	$\beta_1 = -101.2,$	$\beta_2 = 7.6,$	$\beta_3 = 21.3$
10	WEB SCRUBBING	$\beta_0 = 42.2,$	$\beta_1 = -11.2,$	$\beta_2 = 3.5,$	$\beta_3 = 7.3$
11	AlSiCu SPUTTER	$\beta_0 = 203.2,$	$\beta_1 = -1.2,$	$\beta_2 = 44.1,$	$\beta_3 = 52.1$
12	METAL PHOTO	$\beta_0 = 12.2,$	$\beta_1 = -42.2,$	$\beta_2 = 9.1,$	$\beta_3 = 1.23$
13	METAL ETCHING	$\beta_0 = 93.0,$	$\beta_1 = -31.1,$	$\beta_2 = 32.1$	$\beta_3 = 62.7$
14	SINTERING	$\beta_0 = 42.2,$	$\beta_1 = -25.2,$	$\beta_2 = 1.14$	$\beta_3 = 122.0$
15	PAD PSG CVD	$\beta_0 = 52.6,$	$\beta_1 = -47.2,$	$\beta_2 = 15.2$	$\beta_3 = 8.9$
16	PE-NITRIDE CVD	$\beta_0 = 24.1,$	$\beta_1 = -34,$	$\beta_2 = 3.4,$	$\beta_2 = 7.4$
17	PV PHOTO	$\beta_0 = 52.7,$	$\beta_1 = -541.7,$	$\beta_2 = 4.31,$	$\beta_3 = 19.3$
18	PV ETCH	$\beta_0 = 23.2,$	$\beta_1 = -182.3,$	$\beta_2 = 2.77,$	$\beta_3 = 6.6$
19	PI PHOTO	$\beta_0 = 96.2,$	$\beta_1 = -121.6,$	$\beta_2 = 5.1,$	$\beta_3 = 7.23$
20	UV03 ASHING	$\beta_0 = 104.4,$	$\beta_1 = -51.1,$	$\beta_2 = 0.76,$	$\beta_3 = 11.7$
21	PI CURE	$\beta_0 = 9.2,$	$\beta_1 = -47.1,$	$\beta_2 = 3.0,$	$\beta_3 = 33.3$

Table 4. The parameter value of micro regression without variable transformations. Model 1: $Y_1^* = \beta_{10} + \beta_{11}X_{11} + \cdots + \beta_{13}X_{13} + \varepsilon_1$ (operation 1). $Y_2^* = \beta_{20} + \beta_{21}X_{21} + \cdots + \beta_{23} + \varepsilon_2$ (operation stage 2) ... $Y_{21}^* = \beta_{21,0} + \beta_{21,1}X_{21,1} + \cdots + \beta_{21,3} + \varepsilon_{21}$ (operation stage 21).

(b) The micro models outperform macro models even if interaction effects among stages are neglected in this case.
(c) The models with variable transformations outperform those without variable transformations, regardless of whether they are in macro or micro models.

	MAD	MSE	MAP	NT	PT	Mean	SE
TWK	962	1 163 900	23.76	5	20	− 467.88	5394
NOP	2051	6 024 619	36.14	21	84	1821.96	12 273
SLK	1137	1 721 310	25.51	18	72	445.36	6560
JIQ	881	1 161 106	22.17	3	12	− 628.88	5388
JIS	598	512 055	14.87	7	28	− 300.84	3578
OFS	405	304 129	10.15	9	36	− 151.92	2757
COFS	474	382 917	11.13	9	36	− 298.2	3094
Macro without trans.	271	100 217	17.13	5	20	−210.16	2931
Macro with trans.	150	37 464	6.63	4	16	−60.99	894
Micro without trans	200	69 846	12.05	4	16	190.23	1929
Micro with trans	88	12 757	1.94	3	12	41.91	4.19
Mixed	91	13 435	1.99	3	12	−45.16	580

MAD = mean absolute deviation $=\sum^n |Y-\widehat{Y}|/n$; MSE = mean square error $=\sum^n (Y-\widehat{Y})^2/n$; MAP = mean absolute percentage error $=\sum^n |Y-Y|/\sum^n Y$; NT = no. of tardy jobs $=\{i|i=1,2,\ldots,n \text{ and } Y>Y\}$; PT = percentage of tardy jobs $=(1/n)\{i|i=1,2,\ldots,n \text{ and } Y>Y\}$; Mean = average lateness $=(1/n)\sum^n (Y-Y)$; SD = standard deviation of lateness $=(1/n)\sum^n [(Y-Y)-(1/n)\sum^n (Y-Y)]^2$.

Table 5. Performance evaluation for Example 1.

(d) The performance of the mixed flowtime estimation model ranks between the micro and macro models. In this example, the model complexity reduced by 62% (from 21 equations to 8). Although the prediction error only increases 4% comparing with micro models, two models do not significantly differ. This finding confirms that our mixed model not only reduces the estimation complexity, but also does not degrade the prediction variability.

3.2. *Example 2*

The test procedures for Example 2 are the same as in Example 1 except that the test data are collected from the first 21 stages (instead of the final 21 stages of Example 1) of the 4MB DRAM fabrication process. Table 6 presents the comparative performance results.

4. Conclusion

The fact that interaction effects occur among operational stages accounts for why most investigations on flowtime estimation concentrate on the macro approach instead of the micro approach. From a practical perspective, macro approach models usually do not fit in long operational stages manufacturing environments, especially in wafer fabrication. Although the micro model reduces the forecast error, the complexity of the estimation model is too high. This study presents a novel mixed macro and micro flowtime estimation model, capable of overcoming the complexity (micro approach) and variability (macro approach) of flowtime estimation, particularly for environments such as wafer fab in which hundreds of operational stages are

	MAD	MSE	MAP	NT	PT	Mean	SE
TWK	993	1 376 290	25.32	5	20	− 12 287	6031
NOP	2672	7 936 214	40.22	21	84	47 825	15 369
SLK	1329	1 934 326	27.39	18	72	11 686	7790
JIQ	827	1 276 213	24.12	6	24	− 14 867	5733
JIS	804	497 637	22.76	7	28	− 7902	3759
OFS	721	344 219	20.81	7	28	− 4839	3215
COFS	698	41 621	20.72	7	28	− 8423	3367
Macro without trans.	294	133 921	19.71	6	24	−28 738	3251
Macro with trans.	176	19 391	7.34	4	16	−2800	932
Micro without trans	193	71 392	15.32	6	24	−40 334	2137
Micro with trans	132	16 793	2.31	3	12	−1499	530
Mixed	159	18 472	2.96	4	16	−2313	695

Table 6. Performance evaluation for Example 2.

common. According to the result, the proposed model outperforms other conventional flowtime estimation models and does not significantly differ from the micro approach. This finding confirms that the mixed model proposed here not only reduces the estimation complexity, but also does not degrade the prediction variability.

The proposed model is developed while assuming that only interaction effects among explanatory variables of the same stage are considered and those of different stages are neglected, Although the assumption is verified in this case, a future investigation should develop a more general model. In addition, variable transformation through stepwise regression obviously increases prediction accuracy at the cost of difficulty in comprehension. However, it is unnecessary for the mixed model.

In this study, cycle times are the only input variables. However, more data can be obtained and new factors that can influence production performance can be included. By doing so, more promising and meaningful models can be developed. Based on the mixed model construction procedure presented in this study, Mosel Vitelic Inc. is planning to construct and implement the performance prediction system.

Acknowledgments

The authors would like to thank the production control department of Mosel Vitelic Inc. for technically supporting this research.

References

Baker, K. R., 1984, Sequencing rules and due-date assignments in a job shop. *Management Science*, **30**, 1093–1104.

Baudin, M., Mehrotra, V., Tullis, B., Yeaman, D. and Hughes, R. A., 1992, From spreadsheets to simulations: a comparison of analysis methods for IC manufacturing performance. *IEEE/SEMI International Semiconductor Manufacturing Science Symposium*, pp. 94–99.

Bertrand, J. W. M., 1983, The effect of workload dependent due-date on job shop performance. *Management Science*, **29**, 799–816.

Bookbinder, J. H. and Noor, A. I., 1985, Setting job-shop due-dates with service-level constraints. *Journal of Operation Research Society*, **36**, 1017–1026.

Chatterjee, S. and Price, B., 1991, *Regression Analysis by Example*, 2nd edn (New York: Wiley-Interscience), pp. 107–110, 227–251.

Cheng, T. C. E. and Gupta, M. C., 1989, Survey of scheduling research involving due date determination decisions. *European Journal of Operational Research*, **38**, 156–166.

Connors, P., Feigin, G. E. and Yao, D. D., 1996, A queueing network model for semiconductor manufacturing. *IEEE Transactions on Semiconductor Manufacturing*, **9**, 412–427.

Conway, R. W., Maxwell, W. L. and Miller, L. W., 1967, *Theory of Scheduling* (MA: Addison-Wesley).

Eilon, S. and Chowdhury, I. G., 1976, Due-date in job shop Scheduling. *International Journal of Production Research*, **14**, 223–238.

Enns, S. T., 1995, A dynamic forecasting model for job shop flow time prediction and tardiness control. *International Journal of Production Research*, **33**, 1295–1312.

Heard, E. L., 1976, Due-dates and instantaneous load in the one-machine shop. *Management Science*, **23**, 444–449.

Karmarkar, U. S., 1987, Lot sizes, lead time and in-process inventories. *Management Science*, **33**, 409–418.

Lawrence, S. R., 1995, Estimating flowtimes and setting due-dates in complex production systems. *IIE Transactions*, **27**, 657–668.

Lynes, K. and Miltenburg, J., 1994, The application of an open queueing network to the analysis of cycle time, variability, throughput, inventory and cost in the batch production system of a microelectronics manufacturer. *International Journal of Production Economics*, **37**, 189–203.

Philipoon, P. R. and Fry, T. D., 1992, Capacity-based order review/release strategies to improve manufacturing performance. *International Journal of Production Research*, **30**, 2559–2572.

Ragatz, G. L. and Mabert, V. A., 1984, A simulation analysis of due-date assignment rules. *Journal of Operations Management*, **5**, 27–39.

Seidmann, A. and Smith, M. L., 1981, Due date assignment for production system. *Management Science*, **27**, 571–581.

Srivatsan, N. and Kempf, K., 1995, Effective modeling of factory throughput times. *IEEE/CPMT International Electronics Manufacturing Technology Symposium*, pp. 377–383.

Udo, G. J., 1992, Neural networks applications in manufacturing processes. *Computers and Industrial Engineering*, **23**, 97–100.

Vig, M. M. 1989, Dynamic estimation of job flow time. Masters thesis, University of Minnesota, Twin Cities, USA.

Vig, M. M. and Dooly, K. J., 1991, Dynamic rules for due-date assignment. *International Journal of Production Research*, **29**, 1361–1377. 1993, Mixing static and dynamic flow time estimates for due-date assignment. *Journal of Operations Management*, **11**, 67–79.

Weeks, J. K. and Fryer, J. S., 1977, A methodology for assigning minimum cost due-dates. *Management Science*, **23**, 872–881.

Weeks, J. K., 1979, A simulation study of predictable due-dates. *Management Science*, **25**, 363–373.

Wein, L. M., 1991, Due-date setting and priority sequencing in a multiclass M/G/1 queue. *Management Science*, **37**, 834–850.

Zhang, H. C. and Huang, S. H., 1995, Applications of neural networks in manufacturing: a state-of-the-art survey. *International Journal of Production Research*, **33**, 705–728.