Proceedings of the 2004 Winter Simulation Conference
R .G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, eds.

NONLINEAR REGRESSION FITS FOR SIMULATED CYCLE TIME VS. THROUGHPUT CURVES FOR SEMICONDUCTOR MANUFACTURING

Rachel T. Johnson
Feng Yang
Bruce E. Ankenman
Barry L. Nelson

Department of Industrial Engineering and Management Science
Northwestern University
Evanston, IL 60208-3119, U.S.A.

ABSTRACT

This paper illustrates an example of the use of a metamodeling approach to simulation through an example of two real world semiconductor manufacturing systems. The metamodel used was from Yang *et al.* (2004) and has similarities to Cheng and Kleijnen (1999). The approach aims at reducing the amount of simulation work necessary to generate high quality cycle time-throughput (CT-TH) curves. The paper specifically focuses on demonstrating that, in practice, CT-TH curves can deviate significantly from forms currently assumed in the literature (Cheng and Kleijnen 1999).

1 INTRODUCTION

One metric often used for the analysis of manufacturing factories is the cycle time of a lot or job. This cycle time is a random variable equivalent to the time it takes for a given lot or job to traverse a pre-determined path throughout the factory floor (Hopp and Spearman 1996). A manager can control this random variable by controlling the rate in which lots are introduced to the factory floor. This rate is known as the throughput (start rate) level and is often expressed as a decimal percentage from zero to one, with one being the maximum achievable throughput level; i.e. the system capacity. Each throughput level corresponds to a single estimate of mean cycle time and only provides a picture of the factory floor for a given throughput level. Therefore, it is preferable to characterize the system based on a series of consecutive throughput levels with corresponding cycle time value estimates. This type of characterization results in what is known as a cycle time-throughput (CT-TH) curve as shown in Figure 1.

The generation of CT-TH curves requires precise and accurate estimates of the average cycle time at a number of throughput levels. CT-TH curves can be generated in various ways. For simple systems, such as an M/M/1 queuing system, there are analytical equations that allow quick calculations of average cycle time given the throughput level. However, as the systems become more complex, mathematical calculations become unwieldy and too complex to solve in a reasonable amount of time. Therefore, simulation has become the tool of choice for the generation of CT-TH curves for most real-world manufacturing systems. One draw back to the use of simulation to create CT-TH curves is that the computational time for the generation of the curve can become arbitrarily large when trying to achieve very precise and accurate estimates of the average cycle time. The time it takes to run the simulation models is driven by the specific precision or level of accuracy an analyst is striving to achieve. This can often cause problems because the higher the level of precision the analyst wants, the longer the simulation must be run. Additionally the computation time it takes to run the simulation depends on the throughput levels the analyst chooses to select. CT-TH curves are known to follow a non-linearly increasing pattern with steep rises in cycle time variance as the throughput level reaches maximum capacity for systems without batch processing. Therefore, if an analyst is interested in the generation of mean cycle time estimates at high throughput levels with high pre-

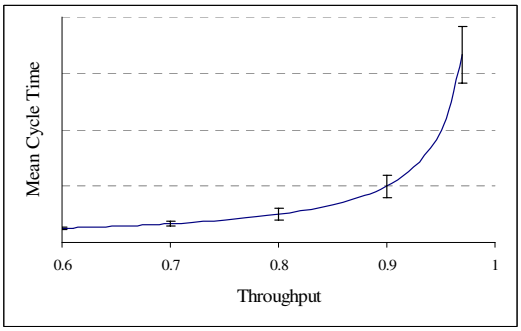

Figure 1: An Example of a CT-TH Curve without Batch Processing

cision and accuracy, large amounts of simulation time and effort can be expected.

One method of mitigating the time it takes to develop a simulation based CT-TH curve is the application of metamodels. Metamodeling can be described as using simulation to develop a tractable and (usually simple) equation-based model to represent the relationship between controllable system parameters and system performance. Yang *et al.* (2004) uses this concept of metamodeling to provide a methodology that requires nothing of the analyst beyond the simulation model, a range of throughput levels of interest, and a measure of the required precision for the estimated curve. The results derived from this method include a complete response profile like that provided by a queueing model, but with the fidelity of a simulation model.

The methodology developed in Yang *et al.* (2004) is similar to previous work by Cheng and Kleijnen (1999). The procedure proposed in Yang *et al.* (2004) describes the input-output relationship that is inherit in simulation of manufacturing facilities, by a metamodel known as the expected cycle-time model. This expected cycle time model was developed from the work in Cheng and Kleijnen, but deviates from it by changing a known exponential term in the denominator of the equation to an unknown (called p in this research). This is done for three reasons. First is to get the exponent right, thus simplifying the model. Second, the addition of p as an exponent in the denominator saves computational effort. Finally, the addition of p in the equation reserves the properties of the true CT-TH curve. This paper addresses the addition of this unknown term and provides empirical evidence that its value deviates from the assumed value of one by applying the methodology to several models of real world semiconductor manufacturing facilities.

2 SUMMARY OF YANG *ET AL.* (2004)

As described previously, CT-TH curves are known to follow a monotonically increasing path as the throughput levels reach a maximum capacity. Maximum capacity can be described as the maximum throughput level of the system, which takes on a value of one. The range of throughput level interests used to describe a system and thus create the base for the CT-TH curve can be described as $[x_L, x_U]$ and range from 0, which is the minimum capacity and at this point equal to the pure processing time of a product, to 1, which is the maximum capacity of the system where the cycle time value goes to infinity.

The object of the research in Yang *et al.* (2004) is to estimate the CT-TH curve via sequential experimentation. There has been extensive literature on fitting CT-TH curves to simulation responses, two papers include; Fowler *et al.* (2001) and Park *et al.* (2002). In the paper by Yang *et al.* (2004) an input-output relationship for throughput levels and average cycle time values is represented by the following metamodel, which is referred to as the expected cycle-time (ECT) model:

$$Y_j(x) = \mu_t(x,c,p) + \varepsilon_j(x) \qquad j = 1, 2, \ldots, n(x) \qquad (1)$$

where,

$$\mu_t(x,c,p) = \frac{\sum_{l=0}^{t} c_l x^l}{(1-x)^p} \qquad (2)$$

This input-output relationship is derived from the fact that the output response, Y(x), average cycle time, is directly dependent on the input value, x, the throughput level. $Y_j(x)$ is the output from the j^{th} replication at throughput level x and n(x) is the number of replications placed at the input level x. The term, $\varepsilon_j(x)$, is an error term with an expected value of zero and a variance of $\sigma^2(x)$. In (1) and (2) t is the degree of the polynomial factor in x and p, t, and the vector $c = (c_1, c_2, \ldots, c_t)$ are unknown parameters. The equations given in (1) and (2) were derived from Cheng and Kleijnen (1999); however, the response of the metamodel differs in the fact that Cheng and Kleijnen investigated a response variable of expected waiting time as opposed to expected cycle time.

The expectation function given in (2) is composed of two parts, the polynomial function in the numerator and the exponential denominator which accounts for the unbounded behavior of the CT-TH curve as the system reach-es capacity. Notice the limit of the numerator as the system reaches a maximum of one is infinity. The form of this model was motivated by queueing results for some elementary stochastic models and heavy-traffic analysis found in Whitt (1989). The linear regression model referred to as the expected cycle time model given in (1), is the same model used in Cheng and Kleijnen (1999), but with the p value unknown. In fact, Cheng and Kleijnen (1999) use a default value for p as one and rely on the polynomial numerator of (2) to adjust the fitted curve for the misspecification in the denominator. They argue that the error in p can be eliminated or corrected by adding terms in the polynomial numerator. Yang *et al.*(2004) argues that this meth-od is undesirable, not only because of the addition of unknown terms, but also because the incorporation of a higher-order polynomial may not preserve the monotonicity of the CT-TH curve. The goal of this paper is to provide evidence that the equations used in Cheng and Kleijnen (1999) can be incorrect in a real manufacturing setting and often the value of p can vary greatly from the assumed value of one.

3 METHODOLOGY

The research presented in this document follows the YAN procedure found in Yang *et al.* (2004). In a very brief summary the steps are as follows:

- Choose a set of throughput levels

- Simulate each design point to within a specified precision level, while accounting for initialization bias through truncation methods
- Use a statistical program to sequentially fit the non-linear regression equation proposed in (1).

The set of design points (throughput levels) used in the experiments presented in this paper were chosen to span a range between 0.5 and 1.0. These points were chosen to span the portion of the curve that captured the sharp increase in estimated cycle time as the system reaches capacity. In order to ensure that the simulation was run to within the specified precision level two short trial runs were conducted. The first trial run was run for a long enough portion of time to capture the warm-up period of the system. The warm-up length was determined by studying the WIP graph (WIP plotted against time) and watching when the increase stopped and the WIP began to oscillate around a horizontal line. Once the warm-up period was determined, the next trial run was conducted with a run length ten times the warm-up period. This was done to insure that the bias from the warm up period did not affect the results. From the second trial run, calculations were made on the mean cycle time half-widths to predict the number of replications needed to obtain a specified precision level. In the case of this experiment, a precision level of 1% was chosen.

Once the simulations were complete, the values for the estimated mean cycle time for each throughput level were extracted and used to fit the non-linear regression model. The metamodel proposed in (1) was fitted in a step by step process. The order of the polynomial in the numerator of (2) was increased in each step, until the T-value from the results was no longer significant. At this point, the metamodel with highest number of polynomial terms, with the T-value still significant, was chosen as the fitted model.

4 RESULTS AND DISCUSSION

The experiments in this report were conducted on two data sets provided by the Modeling and Analysis for Semiconductor Manufacturing Lab at Arizona State University (`www.eas.asu.edu/~masmlab/`). The data sets include real world wafer fab processing and product information. The first data set (MASM Lab Set #1) included two products (MASM Lab Set #5) included twenty-one products. All of the products had unique factory production routes in which they followed; resulting in differing expected cycle time values. The simulations were conducted in Factory Explorer and the statistical analysis was preformed using the program S-Plus.

The implementation of the YAN procedure, described briefly in the methodology section of this report, included choosing the design points from a range of values spanning [0.5, 1.0]. In order to test the efficiency of the metamodel, more simulation effort was allocated to the simulation of each design point than required by the YAN procedure to ensure that the values for the expected cycle times were "nearly true" estimates. Essentially the simulation was run until the standard error of the expected cycle time estimate was nearly zero.

4.1 Results from MASM Lab Set #1

The results from the first data set contained two products. Table 1 shows a comparison between the simulated values for expected cycle time and the values obtain from the YAN procedure.

Table 1: Comparison of the Estimated Expected Cycle Time Values to the True Values

		Prod 1			Prod 2	
Check Points	μ'	μ*	Error	μ'	μ*	Error
0.52	471.9	467.9	0.80%	617	606.1	1.80%
0.58	479.7	471.7	1.70%	618	608	1.60%
0.64	481.1	478.2	0.60%	626.7	615.7	1.80%
0.7	493.4	488.8	0.90%	638	630.7	1.10%
0.76	511.4	505.8	1.10%	661.4	656	0.80%
0.82	540.8	534.1	1.20%	698.7	697.1	0.20%
0.88	595.9	586.1	1.60%	767.3	767.2	0.01%
0.94	703	708.6	0.80%	896.2	912.9	1.90%
	ABS Average:		1.09%	ABS Average:		1.15%

In Table 1, the first column for each product, μ', represents the "true" values for cycle time (estimates from the simulation with nearly zero confidence intervals), the second column, μ*, gives the estimates from the YAN procedure, and the third column shows the relative error between the two. The last row in Table 1 shows the absolute value of the average percent error for each product. The relative error is at most no greater than 2% and the fitted expected cycle time (ECT) model values are given in (3) and (4) for products 1 and 2, respectively.

$$Y(x) = 480.43 - 225.12x/(1 - x)^{0.34} \quad (3)$$

$$Y(x) = 736.80 - 624.29x + 337.72x^2/(1 - x)^{0.25} \quad (4)$$

As seen from equations (3) and (4), the p values differ greatly from a value of one. The value of p for product 1 is 0.34, while the value of p for product 2 is 0.25.

4.2 Results from MASM Lab Set #5

The results from the second model are summarized in Table 2. The second column in the table contains the value of p obtained from fitting the ECT model (equation (1)) to the

simulated product values over the same range of throughput values used on data set 1. As seen in Table 2, the values for p range from 0.04806 to 1.17277. These values clearly differ from a value of one and they show a wide range of values supporting the need for the addition of p in the numerator of the ECT equation. The third column in Table 2 shows the average absolute percent error. This value was obtained by calculating the absolute of the percent error between the "true values" obtained from the simulation and the values obtained through the YAN procedure and then averaging the values across the range of throughput values selected. This column corresponds to the last row in Table 1. In the case of this set (MASM Lab Set #5), the greatest average % error is 1.26%. This low number signifies good fits from the YAN procedure.

Table 2: Values for p and Average Percent Error from Data Set 5 Containing 21 Products.

Product	Value of p	Average ABS % Error
Product 1	p = 1.08517	0.20%
Product 2	p = 1.0117	0.07%
Product 3	p = 1.10568	0.30%
Product 4	p = 1.01164	0.07%
Product 5	p = 1.03686	0.08%
Product 6	p = 1.07985	0.15%
Product 7	p = 1.04368	0.04%
Product 8	p = 1.0296	0.05%
Product 9	p = 0.04806	0.27%
Product 10	p= 0.0496296	0.35%
Product 11	p = 1.01991	0.05%
Product 12	p = 1.14058	0.93%
Product 13	p = 1.04359	0.04%
Product 14	p = 1.034	0.04%
Product 15	p = 1.13779	0.95%
Product 16	p = 0.162968	1.13%
Product 17	p = 0.17277	1.26%
Product 18	p = 1.05862	0.02%
Product 19	p = 1.04393	0.13%
Product 20	p = 1.04222	0.05%
Product 21	p = 1.16398	0.61%

5 CONCLUSIONS

The results in this paper demonstrate an example of a real world semiconductor manufacturing setting where the fitted metamodel results in a p value that differs greatly from one. In the case of this demonstration, setting the p value to one would have often overestimated its true value and resulted in a curve that did not adequately represent the system. It is clear that the p term directly relates to the variability in the system. While CT-TH curves are known to follow a sharply increasing trend as throughput levels increase, not all CT-TH curves follow the exact same increases. Less variability in the system will cause a slightly less drastic increase in the estimated cycle times as the system throughput level is increased, resulting in p values that can be drastically less than one.

REFERENCES

Cheng, R. C. H., and J. P. C. Kleijnen. 1999. Improved design of queueing simulation experiments with highly heteroscedastic responses. *Operations Research* 47: 762-777.

Yang, F., B.E. Ankenman, and B.L. Nelson, S. 2004. Efficient Generation of Cycle Time-Throughput Curves through Simulation and Metamodeling. Working Paper, Industrial Engineering, Northwestern University.

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

Hopp, W. J., and M. L. Spearmean. 1996. *Factory Physics: Foundations of Manufacturing Management.* Chicago: Irwin.

Park, S., J.W. Fowler, G. T. Mackulak, J.B. Keats, and W. M. Carlyle. 2002. D-Optimal Sequential Experiments for Generating Simulation-Based Cycle Time – Throughput Curve. *Operations Research* 50 (6):981-990.

Whitt, W. 1989. Asymptotic formulas for Markov process with applications to simulation. *Operations Research* 40 (2):279-291.

AUTHOR BIOGRAPHIES

RACHEL T. JOHNSON is a Masters student in the Industrial Engineering department at Arizona State University. Her research interest is in discrete event simulation methodologies. She currently serves as secretary in the local area chapter of INFORMS. She received her B.S. in Industrial Engineering from Northwestern University. She recently was awarded the SRC/Intel Fellowship for the duration of her Masters program.

FENG YANG is a Ph.D candidate in the Department of Industrial Engineering at the Northwestern University. Her dissertation has focused on the efficient generation of Cycle time-Throughput curves for manufacturing systems via simulation and metamodeling. Her research interests include experimental design and quality control, simulation of manufacturing systems, and simulation input analysis. Her e-mail and web addresses are <ffyang@northwestern.edu> and <http://pubweb.northwestern.edu/~fya287/>.

BRUCE ANKENMAN is an Associate Professor in the Department of Industrial Engineering and Management Sciences at the McCormick School of Engineering at Northwestern University. His current research interests include response surface methodology, design of experiments, robust design, experiments involving variance

components and dispersion effect, and experimental design for simulation experiments. He is past chair of the Quality Statistics and Reliability Section of INFORMS, is on the editorial board for *IIE Transactions: Quality and Reliability Engineering* and is an Associate Editor for *Naval Research Logistics*. His e-mail and web addresses are <ankenman@northwestern.edu> and <www.iems.northwestern.edu/~bea>.

BARRY L. NELSON is the Krebs Professor of Industrial Engineering and Management Sciences at Northwestern University, and is Director of the Masters of Engineering Management Program. His research centers on the design and analysis of computer simulation experiments on models of stochastic systems. He has published numerous papers and two books. Dr. Nelson has served the profession as the Simulation Area Editor of *Operations Research* and President of the INFORMS (then TIMS) College on Simulation. He has held many positions for the Winter Simulation Conference, including Program Chair in 1997 and current membership on the Board of Directors. His e-mail and web addresses are <nelsonb@northwestern.edu> and<www.iems.northwestern.edu/~nelsonb/> .