Proceedings of the 2017 Winter Simulation Conference
W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, eds.

AN ITERATIVE REFINEMENT APPROACH TO FITTING CLEARING FUNCTIONS TO DATA FROM SIMULATION MODELS OF PRODUCTION SYSTEMS

Karthick Gopalswamy
Reha Uzsoy

North Carolina State University
Edward P. Fitts Department of Industrial and Systems Engineering
Campus Box 7906
Raleigh, NC 27695-7906, USA

ABSTRACT

We examine the problem of fitting clearing functions that estimate the expected output of a production resource as a function of its expected workload from empirical data. Unlike most regression problems, the independent variables are not directly controllable due to the presence of a planning model that controls releases to the production system, and the release decisions made by the planning model are themselves dependent on the estimated clearing function. We propose an iterative refinement procedure that uses simulation experiments to resample data from the production system as the parameters of the clearing function are iteratively updated. We compare the iterative procedure to previously used approaches with promising results.

1 INTRODUCTION

The *release planning* problem is that of determining the timing and quantity of releases of raw materials into the production system to ensure that output matches demand in an optimal or near-optimal manner. This requires modeling the cycle times, the delay between work being released into the production system and its emergence as finished product that can be used to meet demand. Both queueing theory (Buzacott and Shanthikumar 1993) and simulation models (Atherton and Atherton 1995) have demonstrated a nonlinear relation between mean cycle time and mean resource utilization. In general, the cycle time of a job through a production system is a random variable whose distribution depends on the level of resource utilization, among other factors. However, resource utilization is determined by the workload, the amount of work available to a resource in a planning period, which is, in turn, determined by the release decisions. This circularity, where the cycle time depends on release decisions that themselves require knowledge of the cycle times, has been a persistent issue in production planning for several decades.

There have been three different approaches to this issue in the literature to date. By far the most common is to represent cycle times as a workload-independent, exogenous parameter, or lead time. This approach is used in the widely used Material Requirements Planning (MRP) procedure (Vollmann et al. 2005) and the majority of linear (LP) and mixed integer (MIP) models in the literature (Missbauer and Uzsoy 2011). This approach yields computationally tractable models, but fails to capture the workload-dependent behavior of cycle times, particularly when resource utilization varies significantly over time. A second approach has been to decompose the problem into two subproblems, one that determines optimal releases for given cycle time estimates and another that estimates the cycle times that will be realized under given releases (Hung and Leachman 1996; Kim and Kim 2001; Byrne and Hossain 2005), and iterate between the two models until convergence is achieved. The first model is usually a LP, while simulation, queueing or regression models can be used for the latter (Hung and Hou 2001). However, the computational requirements of these procedures are high due to their use of a detailed simulation model,

and their convergence is not yet well understood (Irdem et al. 2010). The third approach, which motivates this paper, models production resources using nonlinear clearing functions (CFs) that represent the expected output of a production resource in a planning period as a function of the planned workload at the resource during that period (Missbauer and Uzsoy 2011). Hence the CF can be viewed as a metamodel of the queueing system describing the production resource. Most previous work has focused on developing CFs that can be incorporated into LP formulations of the release planning problem, generally using a single state variable, the total workload of all products at the resource in the planning period, as the independent variable. The use of a single state variable resulted in difficulties representing production systems with multiple products, which are largely, though not completely, resolved by the Allocated Clearing Function model of Asmundsson et al.(2009) in the absence of setups between products. A growing body of research has shown that when appropriately parameterized planning models using CFs can yield improved production plans over both fixed lead time and iterative multi-model approaches, especially when resource utilization varies over time (Asmundsson et al. 2009; Kacar et al. 2013, 2016).

The use of CFs has several advantages. If a suitable functional form is used, the resulting planning models can be solved using commercial software without requiring any time-consuming simulation runs, in contrast to simulation optimization or iterative multi-model approaches. The computationally intensive work of fitting the CF can be performed offline, outside the planning model. Finally, planning models using CFs yield dual prices for resources with utilization below 1, which LP models using exogenous, workload-independent lead times cannot do (Kefeli and Uzsoy 2016).

However, a satisfactory, general formulation of the problem of estimating CFs from data has not yet been developed. The most common approach in the literature is to fit CFs to all resources in the production system using data collected from direct observations of the production system or, much more commonly, a simulation model. However, there appears to be quite considerable room for improvement over current approaches: the simulation optimization approach of Kacar and Uzsoy (2015) resulted in improvements of up to 30% in the performance of the production system. Hence the improvement in production plans from better estimation of the CFs may be quite considerable. Various functional forms have been proposed, none of which has proved entirely satisfactory; there is also little agreement on what state variables (independent variables in regression terminology) should be included. This results in a highly unsatisfactory state of affairs, where it is hard to give practitioners a clearly stated, reliable procedure for fitting CFs that can be automated with minimal manual intervention.

After reviewing previous work in the next section, this paper explores the difficulties of the problem of fitting CFs to data from simulation models in Section 3, and suggests a heuristic solution analogous to the policy iteration algorithm used in stochastic optimization in Section 4. Section 5 presents computational results for some simple single-stage systems. We conclude the paper with a discussion of our principal findings and directions for future work.

2 PREVIOUS RELATED WORK

Clearing functions for capacitated production resources were first suggested by Graves (1986), Karmarkar (1989) and Srinivasan et al. (1988). These CFs use a single state variable, either the work in process inventory (WIP) at the resource at the start of period t, denoted by W_{t-1}, or the workload Λ_t, in units of time, available to the resource over the planning period. The workload Λ_t denotes the total amount of work, measured in time units, that becomes available to the resource in period t, and is given by $\Lambda_t = W_{t-1} + R_t$, where R_t denotes the amount of material released to the resource in period t, again in units of time. The CF of Graves (1986) differs somewhat from the others in that it assumes the production resource will be able to convert a constant fraction of its available workload into output in any period, implying that the rate of production can be varied. Motivated by steady-state queueing models, Karmarkar (1989) suggests the functional form

$$f(\Lambda_t) = \frac{K_1 \Lambda_t}{K_2 + \Lambda_t} \ , \ \Lambda_t \geq 0 \tag{1}$$

where K_1 and K_2 are parameters to be estimated from data. Motivated by results from traffic modeling (Carey and Bowers 2012), Srinivasan et al. (1988) suggest the form

$$f(W_t) = K_3(1 - e^{-K_4 W_t}) \ , \ W_t \geq 0 \tag{2}$$

where K_3 and K_4 denote parameters to be estimated. Missbauer (2002) proposes the form

$$f(\Lambda_t) = \frac{1}{2}\left[C + k + \Lambda_t - \sqrt{C^2 + 2Ck - 2C\Lambda_t + 2k\Lambda_t + \Lambda_t^2} \right] \tag{3}$$

where C denotes the length of the planning period and $k = 0.5(\sigma^2 / t + t)$, where t and σ denote the mean and variance of the processing time, respectively. All three functional forms are concave non-decreasing in Λ_t for positive values of the estimated parameters.

These functional forms are all based on steady state queueing models. However, the CF seeks to represent the behavior of the production resource during a finite planning period, over which the system may not attain steady state. Missbauer (2009, 2011) shows that the shape of the CF for a given value of Λ_t changes as a function of R_t and W_{t-1}, demonstrating that the shape of the CF can vary over time based on system state. The incorporation of transient CFs results in nonlinear integer optimization models that are difficult to solve, and remains the subject of ongoing research.

Given the difficulty of obtaining tractable CFs from queueing analysis, the prevalent approach in the literature has been to postulate a functional form, usually one of (1) - (3) above, and estimate its parameters from data, usually obtained from a simulation model of the production system, using linear regression (LR). The CF thus obtained is assumed to represent the expected behavior of the production system at any point in time. However, this approach has encountered unexpected difficulties. Asmundsson et al. (2009) found that using LR with a single state variable representing the aggregate workload systematically underestimated the CFs; a heuristic percentile fit yielded much improved results. Kacar and Uzsoy (2015) used simulation optimization to estimate the CF that optimized the performance of the production system, instead of the fit to the data; they obtained substantial improvements in performance over a piecewise linearized LS fit. Kacar et al. (2012; 2013; 2015; 2016) visually partition the range of data into two segments containing approximately equal numbers of data points, and then use linear regression to fit linear functions of the workload in each partition. They then add a third segment with slope of zero representing the estimated maximum output of the resource in a period. This approach produces a CF that is a piecewise linear function of a single state variable, and thus ideal for use in the Allocated Clearing Function formulation (Asmundsson et al. 2009). However, the improved results from simulation optimization (Kacar and Uzsoy 2015) show that this approach will not always lead to the best possible fit. It is important to note that Kacar et al. do not consider the impact of the planning model on the CF; the CF is fit to data collected by setting the releases in each period equal to that period's demand.

Missbauer's (2009, 2011) results on transient CFs suggest that a single state variable such as the workload Λ_t may not be sufficient to describe the behavior of the production resource accurately. Kacar and Uzsoy (2014) compared a number of LR models for fitting workload-based CFs, again without considering the impact of the planning model on the simulation data collected. They experiment with a range of independent variables in their regression models including releases R_t, the entering WIP W_{t-1} and the same quantities for earlier time periods. They find that at high utilization including state variables from earlier periods improves performance, but no single model yields consistently higher performance across all experimental conditions. Haeussler and Missbauer (2014) present extensive regression analyses using both simulation and empirical data obtained from a manufacturer of optical storage media. They find that incorporating additional independent variables leads to better fits, noting substantial differences in results between empirical and simulation data. For bottleneck machines they find that incorporating simple quadratic terms of a single workload variable yields marked improvements in fit, as measured by

adjusted R^2. Multidimensional CFs have been proposed for systems with setups and lot-sizing decisions (Albey et al. 2014; Kang et al. 2014; Albey et al. 2017), but lead to non-convex optimization models.

The picture that emerges from this body of work is far from clear, except that simple linear regression with a single state variable leaves considerable room for improvement. The choice of an appropriate functional form remains, largely, open; the state of the art in traffic modeling, where very similar functions relate the flow of traffic through a road segment to the number of vehicles on the segment in each period suggests that this is a difficult problem in its own right (Carey and Bowers 2012). We give some evidence below that functional forms derived from steady state queueing analysis may not be appropriate, while those based on transient analysis yield intractable planning models. In the following section we discuss a number of issues that we believe render the problem of fitting CFs difficult even when the functional form is specified: the selection of an appropriate fitting technique, and the fact that the independent variables in the regression are not directly controllable in simulation experiments since the planning model determines the releases, introducing a circularity into the problem.

3 ISSUES IN FITTING CLEARING FUNCTIONS

In this section we first propose a simple formulation of the problem of fitting a CF to data from a production system, specifying the data to which the fitting procedure will be applied, how the CF is used in the production planning model, and then addressing some difficulties that arise. For simplicity of exposition we consider a single production resource producing a single product whose behavior can be described as a queueing system. We seek to fit the functional form (1) of Karmarkar (1989); the issues we raise remain valid regardless of the specific functional form we seek to fit. We also assume that we have access to a simulation model of the system that represents its behavior to the desired degree of accuracy.

3.1 Planning Model

Releases into the system are computed by a planning model that seeks to determine the amount of material R_t released to the resource at the start of period t over a planning horizon of T periods such that the sum of WIP holding, finished inventory holding and backorder costs are minimized over all periods. The decision variables are I_t, the amount of finished inventory on hand at the end of period t; B_t, the number of backlogged units at the end of period t; W_t, the WIP at the end of period t; and X_t, the output of the resource in period t. Demand for the product in period t is denoted by D_t, treated as a deterministic parameter in the planning model, which can be written as the following convex optimization problem:

$$\min \sum_{t=1}^{t} \left[w_t W_t + h_t I_t + b_t B_t \right] \tag{4}$$

subject to

$$I_t - B_t = I_{t-1} - B_{t-1} + X_t - D_t \ , \ t = 1,...,T \tag{5}$$

$$W_t = W_{t-1} + R_t - X_t \ , \ t = 1,...,T \tag{6}$$

$$X_t \le \frac{K_1 (R_t + W_{t-1})}{K_2 + (R_t + W_{t-1})} \ , \ t = 1,...,T \tag{7}$$

$$R_t, I_t, X_t, W_t \ge 0 \tag{8}$$

where w_t, h_t and b_t denote unit WIP holding, inventory holding and backordering costs, respectively.

The use of the CF within this model raises some interesting questions. When the model is run, the only parameters known with certainty are the initial WIP and finished goods inventory levels W_0 and I_0 and the cost parameters. The optimal values $\hat{R}_t, \hat{X}_t, \hat{I}_t$ and $\hat{W}_t$ of the decision variables obtained from the model (4) - (8) thus represent planned, or predicted, values of the quantities they represent, which are random variables whose distribution depends, potentially, on the entire history of the production system up to the time they are observed as well as the distribution of the processing times, the dispatching or

scheduling procedures used on the shop floor, and other work practices that are abstracted away in the model. Thus the workloads $\Lambda_t = R_t + W_{t-1}$ that form the argument of the CF, and hence the independent variables in the regression model used to estimate the parameters of the CF, are not directly controllable in an experimental design.

It should also be noted that the model (4) - (8) assumes the same CF for all periods in the planning horizon. The distribution of the demand D_t represents the state of the world in which the production resource must operate. Hence the CF we fit seeks to represent the behavior of the production system aimed at meeting this demand, and depends on the distributions of processing times and failures, the manner in which work is released into the system over time, shop floor scheduling and staffing policies, and so on. The use of the planning model to determine release schedules results in a dependency of the CF on the demand distribution, since the planning model will produce different release patterns, and hence different patterns of workload over time, under different demand distributions. Thus the CF we seek represents the expectation of output in a planning period over all these random variables. If we denote the random variables representing workload over time by Λ, those representing demand by D and all internal random variables such as processing times and failures as P, the CF we seek can be written as

$$X = E_{\Lambda,D,P}\left[\bar{X} \mid \Lambda, D, P\right] \tag{9}$$

It is immediately apparent that this very aggregate formulation of the CF may result in substantial inaccuracies when estimating output for a particular system state in a particular time period, especially when the arguments of the CF used in the planning model represent predictions of random variables that will be realized some number of periods in the future. The work on multivariate CFs (Haeussler and Missbauer 2014; Kacar and Uzsoy 2014) attempts to address this issue by including additional state variables in the functional form. However, whatever functional form is used, the presence of the planning model prevents direct control of the independent variables in the regression.

3.2 Sampling Issues

The process by which we obtain the data required to fit the CF is in principle unremarkable. We perform a number of simulation replications $g = 1,...,G$ that simulate the operation of the production resource over a time horizon of T discrete periods. This yields TG observations (X_{gt}, Λ_{gt}) denoting the observed output X_{gt} and workload Λ_{gt} of the resource in period t in replication g, to which the CF can be fitted. The question is how to obtain these observations in an appropriate manner.

In order to obtain data from the simulation model to which we can fit the CF, we must sample from the distributions of Λ, D and P. Although sampling from the distributions of D and P is straightforward, the presence of the planning model prevents us from sampling from the distribution of Λ directly. This creates a circularity in that the planning model uses the estimated CF to determine releases, but the releases determine the workloads Λ_t used as the independent variables from which the CF is estimated.

In much previous research (Kacar et al. 2012; Kacar et al. 2013; Kacar and Uzsoy 2014, 2015; Kacar et al. 2016), CFs were fit using data obtained by simulating the system without any planning model, simply setting releases equal to demand in each period to obtain the observations (X_{gt}, Λ_{gt}). However, ignoring the presence of the planning model in this way is likely to distort the sample of observations obtained. In periods of high demand, the planning model will not release all of a period's demand in the period it is needed, since this will cause high resource utilization and long cycle times; it will release some of this material earlier, building finished goods inventory from which demand can be met later.

A similar problem arises in our context due to the presence of the planning model. For a given sample q from the distributions of D and P, the observed output and workload (X_{qt}, Λ_{qt}) are determined by the releases R_{qt} which, in turn, are determined by the parameters K_1 and K_2 of the estimated CF. Errors in the estimates of K_1 and K_2 can lead to release patterns that would not occur with the correct CF, resulting in observations (X_{qt}, Λ_{qt}) that are unlikely to occur with the correct CF. We seek to address this circularity using the iterative refinement procedure discussed in the next section.

3.3 Impact of the Functional Form

Another set of issues arises from the choice of a functional form derived from a steady state queuing model. Our adoption of the CF (1) in constraint (7) implies a regression model of the form

$$\bar{X}_t = f(\hat{R}_t, \hat{W}_{t-1}) + \varepsilon_t = \frac{K_1(\hat{R}_t + \hat{W}_{t-1})}{K_2 + (\hat{R}_t + \hat{W}_{t-1})} + \varepsilon_t \tag{10}$$

where ε_t is, ideally, normally distributed with a mean of zero and standard deviation σ. The time independence of σ assumes the absence of heteroscedasticity, which is often not the case. The queueing analysis leading to the functional form makes a different statement, which is that

$$E\left[\bar{X}_t\right] = f(E[\bar{R}_t], E[\bar{W}_{t-1}]) = \frac{K_1(E[\bar{R}_t] + E[\bar{W}_{t-1}])}{K_2 + (E[\bar{R}_t] + E[\bar{W}_{t-1}])} \tag{11}$$

Taking the expectation of (10) and assuming that the optimal values of the decision variables represent unbiased estimators of the realized quantities - a highly questionable assumption - we obtain

$$\hat{X}_t = E[\bar{X}_t] = E\left[f(E[\bar{R}_t], E[\bar{W}_{t-1}])\right] \leq f(E[\bar{R}_t], E[\bar{W}_{t-1}]) \tag{12}$$

by Jensen's inequality and the observation that (1) is concave in both variables. This suggests the distinct possibility that, even under the highly idealized condition that the optimal values of the decision variables yield unbiased estimates of the random variables they represent and the variance of the residuals is time stationary, the regression model implied by the use of (10) may underestimate the expected output in a planning period. This argument remains valid for any functional form derived from steady-state queueing analysis that relates the expected output to the expected value of some workload-related random variable. This again calls into question the desirability of using functional forms derived from steady-state queueing models, suggesting the need for an alternative approach.

4 AN ITERATIVE REFINEMENT PROCEDURE

The previous section has raised three specific issues regarding the use of least squares regression to estimate clearing functions from simulation data: the inability to observe the independent variables directly due to the use of the planning model to compute releases; the dependence between the data required for fitting the CF and the CF itself; and the possibility of systematic underestimation when using functional forms whose arguments are expectations of an underlying random variable. The work in this paper addresses the first two issues by proposing an iterative approach that explicitly considers the planning model in the data collection process, beginning with an initial estimate of the CF which is then iteratively refined until convergence in the CF parameters is achieved. Since the iterative approach retains the functional form (1), the possibility of biased estimates resulting from the particular functional form is not addressed in this work.

The circularity described in Section 3.2 suggests viewing the sampling problem as that of learning the correct values of the CF parameters K_1 and K_2. The Iterative Refinement procedure develops an initial estimate of the CF parameters following Kacar et al. (2012), which we shall refer as the Unplanned Estimation, since it does not use a planning model in developing the release schedules.

Unplanned Estimation:

Step 1: Identify a set of target mean utilization levels ρ_j, $j = 1,...,J$ that span the range of utilizations the system is expected to experience under its routine operating conditions.

Step 2: For each value of ρ_j, generate M independent realizations of demand that yield a mean utilization level of ρ_j. This yields a total of JM independent demand realizations

Step 3: For each demand realization, set releases in each period equal to demand in that period, i.e., $R_t = D_t$, and simulate the execution of this release plan for G independent replications, yielding a total of MGT observations (X_{gt}, Λ_{gt}).

Step 4: Use the MGT observations obtained in Step 3 to obtain initial estimates of the CF parameters estimates $K^U = (K_1^U, K_2^U)$ using least-squares regression.

The Iterative Refinement procedure starts with the parameter estimates K^U obtained by the Unplanned CF Estimation procedure and refines it iteratively as follows:

Iterative Refinement Procedure:

Step 1: Identify a set of target mean utilization levels ρ_j, $j = 1,...,J$ that span the range of utilizations the system is expected to experience under its routine operating conditions.

Step 2: For each value of ρ_j, generate M independent realizations of demand that yield a mean utilization level of ρ_j, for a total of JM independent demand realizations.

Step 3: Set $i = 0$, $K^i = K^U$.

Step 4: For each demand realization, solve the planning model (4) - (8), augmented with the following constraints which we have found improve the fit:

$$X_t \le C_t \ , t = 1,...,T \tag{13}$$

$$X_t \le R_t + W_{t-1} \ , t = 1,...,T \tag{14}$$

Constraint (13) ensures that expected output (in units of time) does not exceed the expected capacity of the system, given by the expected time the resource is available during the planning period, while (14) ensures that the output in any period cannot exceed the planned workload. Simulate the execution of each release plan for G independent replications, obtaining a total of MGT observations (X_{gt}, Λ_{gt}).

Step 4: Use the MGT observations obtained in Step 3 to obtain revised estimates $K^{i+1} = (K_1^{i+1}, K_2^{i+1})$ using least-squares regression.

Step 5: If $i > maxIter$ or $0 \le \dfrac{\left\| K^{i+1} - K^i \right\|}{\left\| K^i \right\|} \le \varepsilon$, stop and return K_{i+1}. Otherwise set $i = i+1$, go to Step 3.

The intuition behind the Iterative Refinement procedure is to progressively refine the sample of observations used to fit the CF. The Unplanned Estimation procedure does not consider the effects of the planning model on releases, and hence may well create release schedules, and hence workload trajectories, that are unlikely to be encountered when the planning model is used. Use of the planning model will result in the elimination of workload trajectories that the planning model will not generate, resulting in a sample that better represents he behavior of the system under the planning model.

5 COMPUTATIONAL EXPERIMENTS

We examined the performance of the Iterative Refinement procedure on four simple single-state single-item production systems whose characteristics are summarized in Table 1. All times are given in minutes. External demand was assumed to follow a binomial distribution whose parameters were adjusted to yield the desired mean utilization levels and coefficient of variation in the table. We considered a planning horizon of $T = 26$ periods, each of length 1440 minutes. Separate, independent data sets were used to fit the CFs and to evaluate their performance. Demand follows a binomial distribution with $n = 100$ and p chosen to yield the desired mean utilization level. Hence the variability of demand is not constant across all utilization levels.

Table 1: Experimental Design.

Factor	Values	Levels
Service Time Distribution	Exponential(20), Erlang(4)	2
Failure Distribution	None; Time to Failure: Gamma(14400,1) Time to Repair: Gamma(2400,1.5)	2
Mean Utilization Level	0.3, 0.45, 0.7, 0.9	4
Independent Demand Replications		5
Simulation Replications per Release Plan		5

We compare the performance of three different release planning procedures. The first two use the planning model (4) - (8) with the CFs obtained from the Unplanned Estimation and Iterative Refinement procedures. The final procedure, included as a benchmark, simply sets releases equal to demand for each period. As in Section 5, we generate 5 independent demand realizations for each level of mean utilization; a release schedule is generated for each demand replication, and its execution then simulated for 5 independent replications to estimate the realized costs of each planning procedure. The planning model assumes a unit backordering cost of $b = 50$, a unit WIP holding cost of $w = 6$, and a finished goods inventory holding cost of $h = 5$. The values of these cost parameters will affect the releases produced by the planning model, and hence the fitting of the CF in the Iterative Refinement approach.

6 RESULTS OF EXPERIMENTS

The results of the experiments are summarized in Figures 1 through 8 below. Figures 1 through 4 compare the planning models with the three different CFs: the Best Case CF of Hopp and Spearman (2008), which represents the expected throughput of a deterministic system; the Unplanned CF obtained using the Unplanned Estimation procedure, and the Iterative CF obtained from the Iterative Refinement procedure. A consistent pattern emerges in these figures: Unplanned Estimation yields CFs that overestimate output at low workloads, and significantly underestimate it at higher workloads. In contrast, Iterative Refinement appears to exhibit high accuracy at lower workloads but to overestimate quite drastically at higher workloads. However, this initial impression is misleading, because the planning model will not allow high workloads that fall in the region where the CF is relatively flat; the marginal increase in output cannot offset the additional WIP costs. Hence the planning model will eliminate the regions where the CF from Iterative Refinement appears to have very poor accuracy, causing the system to operate in the region where its accuracy is highest.

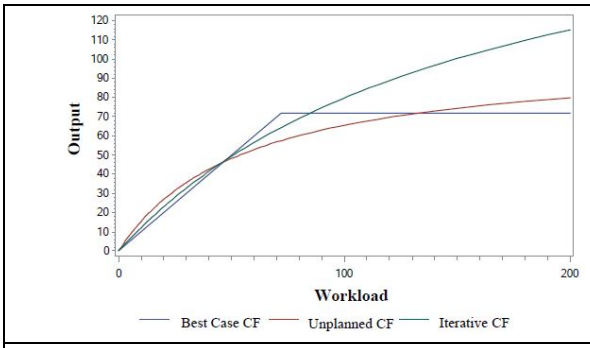

Figure 1: Estimated CF for Exp(20), No Failures.

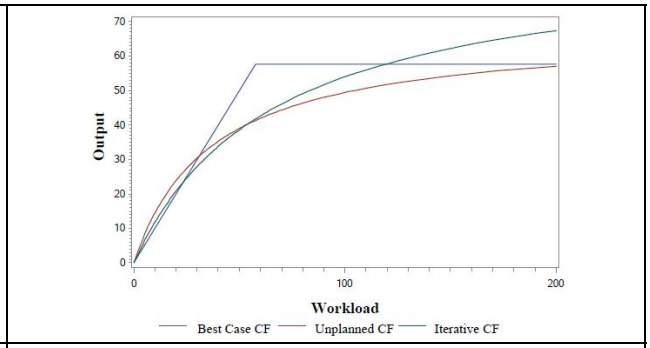

Figure 2: Estimated CF for Exp(20), Failures.

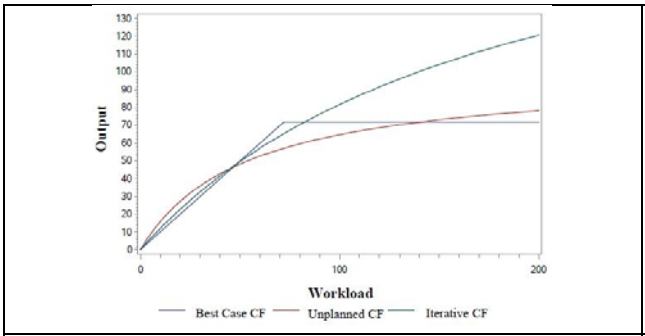

Figure 3: Estimated CF for Erlang (4), No Failures.

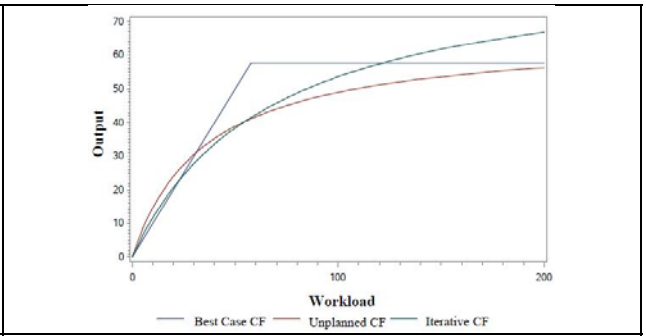

Figure 4: Estimated CF for Erlang (4), Failures.

Figures 5 through 8 show the expected costs obtained using the Unplanned and Iterative CFs, with the cost of a naive release policy that sets releases equal to demand in each period included as a baseline. In the absence of failures, there is no significant different in costs at the lower utilization levels of 0.32 and 0.45. At u = 0.7, the Unplanned CF yields worse performance than the naive release policy; Iterative Refinement significantly outperforms Unplanned Estimation, but is not quite as good as the naive policy. At high utilization, however, Iterative Refinement outperforms both its competitors by a wide margin. This behavior is due to the fact that the planning model plans releases that keep the system operating in the area where the Iterative Refinement procedure provides the best fitting CF.

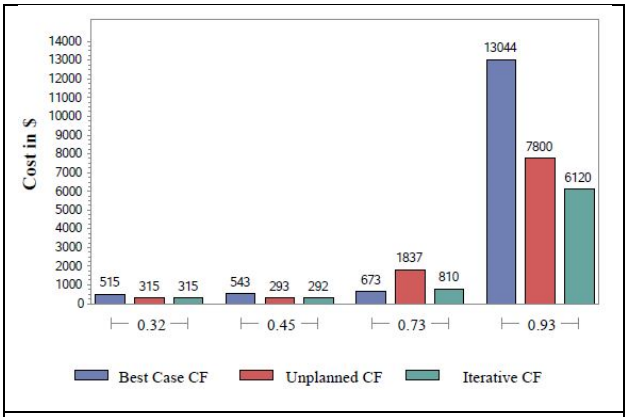

Figure 5: Cost Comparison for Exp(20), No Failures.

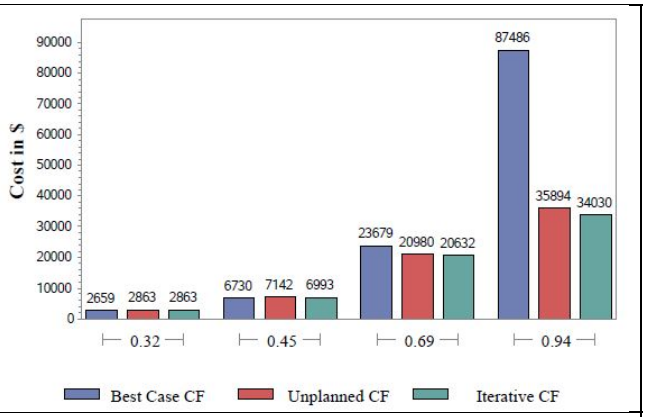

Figure 6: Cost Comparison for Exp(20), Failures.

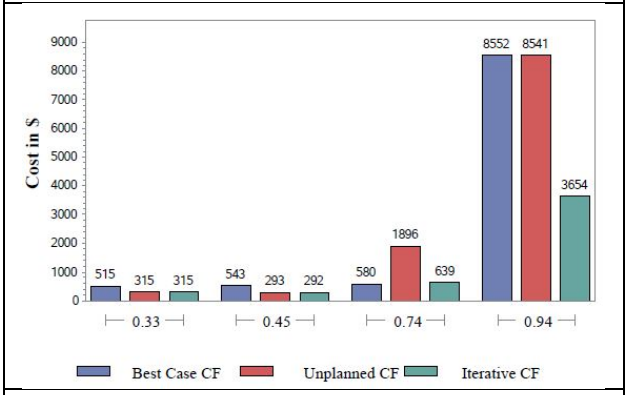

Figure 7: Cost Comparison for Erlang(4), No Failures.

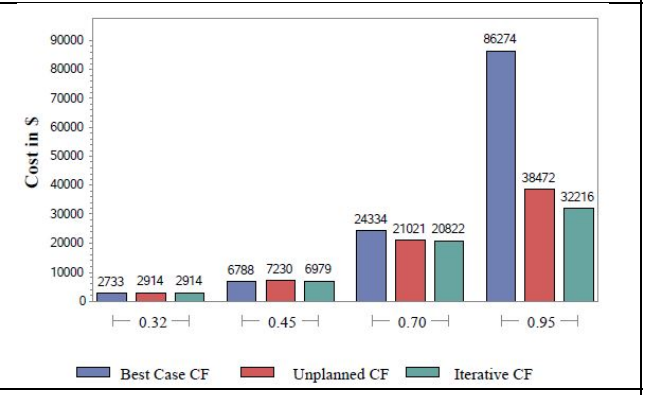

Figure 8: Cost Comparison for Erlang(4), Failures.

7 CONCLUSIONS AND FUTURE DIRECTIONS

The results presented above are limited in scope and exploratory in nature, and hence must be treated with some caution. However, they suggest that significant improvements over the Unplanned Estimation procedure used to fit CFs in the literature to date are possible, and give some insight into why this is the case. The use of the planning model allows the Iterative Refinement procedure to obtain better fits in the region in which the production system will operate, effectively weighting observations in that region more heavily than in the Unplanned Estimation procedure. The Iterative Refinement procedure yields little benefit at low workloads, but results in considerably better cost performance at high utilization. We conjecture that the benefits of the Iterative Refinement procedure may actually be greater at utilization levels between 0.7 and 0.93. At low utilization, the system is generally capable of processing all work released. At very high utilization the system is running at the maximum utilization compatible with meeting demand, so there is little scope for improved decision making in a single-product system where product mix is not a concern. Intermediate utilization levels correspond to the regions in Figures 1 - 4 where the curvature of the CF is highest, and hence an accurate fit is most important. Further experiments on both single and multiproduct systems are needed to further explore this issue.

Theoretical results for transient queues (Ingolfsson et al. 2007; Schwarz et al. 2016) suggest that different functional forms may be appropriate for different workload levels. The use of specialized fitting techniques for piecewise linear functions (Magnani and Boyd 2009; Toriello and Vielma 2012) offers a promising future direction in this regard; different segments can be fitted for different regions to improve the overall fit, and the resulting piecewise linear functions can be implemented directly in the ACF model of Asmundsson et al. (2009), which yields a straightforward, albeit large, LP formulation.

Finally, the approach adopted in this paper uses classical least squares regression to fit the selected functional form to data. However, the success of the ad hoc percentile fitting approach adopted in Asmundsson et al. (2009) suggests that a data-driven, risk-averse approach using an asymmetric loss function may be appropriate. The examination of such techniques and the formulation of appropriate loss functions might address some of the issues of biased estimates due to specific functional forms raised in Section 3.3, suggesting another interesting direction for future work.

ACKNOWLEDGMENTS

The authors would like to thanks Dr. Baris Kacar of SAS Corporation for his assistance with the simulation experiments.

REFERENCES

Albey, E., U. Bilge, and R. Uzsoy. 2014. "An Exploratory Study of Disaggregated Clearing Functions for Multiple Product Single Machine Production Environments." *International Journal of Production Research* 52 (18):5301-5322.

Albey, E., U. Bilge, and R. Uzsoy. 2017. "Multi-Dimensional Clearing Functions for Aggregate Capacity Modelling in Multi-Stage Production Systems." *International Journal of Production Research* 55 (14):4164 - 4179.

Asmundsson, J. M., R. L. Rardin, C. H. Turkseven, and R. Uzsoy. 2009. "Production Planning Models with Resources Subject to Congestion." *Naval Research Logistics* 56 (2):142-157.

Atherton, L. F., and R. W. Atherton. 1995. *Wafer Fabrication: Factory Performance and Analysis*. Norwell, MA: Kluwer Academic Publishers

Buzacott, J. A., and J. G. Shanthikumar. 1993. *Stochastic Models of Manufacturing Systems*. Englewood Cliffs, NJ: Prentice-Hall

Byrne, M. D., and M. M. Hossain. 2005. "Production Planning: An Improved Hybrid Approach." *International Journal of Production Economics* 93-94:225-229.

Carey, M., and M. Bowers. 2012. "A Review of Properties of Flow–Density Functions." *Transport Reviews* 32 (1):49-73.

Graves, S. C. 1986. "A Tactical Planning Model for a Job Shop." *Operations Research* 34:522-533.

Haeussler, S., and H. Missbauer. 2014. "Empirical Validation of Meta-Models of Work Centres in Order Release Planning." *International Journal of Production Economics* 149:102-116.

Hopp, W. J., and M. L. Spearman. 2008. *Factory Physics : Foundations of Manufacturing Management*. 3rd ed. Boston: Irwin/McGraw-Hill

Hung, Y. F., and M. C. Hou. 2001. "A Production Planning Approach Based on Iterations of Linear Programming Optimization and Flow Time Prediction." *Journal of the Chinese Institute of Industrial Engineers* 18 (3):55-67.

Hung, Y. F., and R. C. Leachman. 1996. "A Production Planning Methodology for Semiconductor Manufacturing Based on Iterative Simulation and Linear Programming Calculations." *IEEE Transactions on Semiconductor Manufacturing* 9 (2):257-269.

Ingolfsson, A., E. Akhmetshina, S. Budge, Y. Li, and X. Wu. 2007. "A Survey and Experimental Comparison of Service-Level-Approximation Methods for Nonstationary M(T)/M/S(T) Queueing Systems with Exhaustive Discipline." *INFORMS Journal on Computing* 19 (2):201-214.

Irdem, D. F., N. B. Kacar, and R. Uzsoy. 2010. "An Exploratory Analysis of Two Iterative Linear Programming-Simulation Approaches for Production Planning." *IEEE Transactions on Semiconductor Manufacturing* 23:442-455.

Kacar, N. B., D. F. Irdem, and R. Uzsoy. 2012. "An Experimental Comparison of Production Planning Using Clearing Functions and Iterative Linear Programming-Simulation Algorithms." *IEEE Transactions on Semiconductor Manufacturing* 25 (1):104-117.

Kacar, N. B., L. Moench, and R. Uzsoy. 2013. "Planning Wafer Starts Using Nonlinear Clearing Functions: A Large-Scale Experiment." *IEEE Transactions on Semiconductor Manufacturing* 26 (4):602-612.

Kacar, N. B., L. Moench, and R. Uzsoy. 2016. "Modelling Cycle Times in Production Planning Models for Wafer Fabrication." *IEEE Transactions on Semiconductor Manufacturing* 29 (2):153-167.

Kacar, N. B., and R. Uzsoy. 2014. "A Comparison of Multiple Linear Regression Approaches for Fitting Clearing Functions to Empirical Data." *International Journal of Production Research* 52 (11):3164-3184.

Kacar, N. B., and R. Uzsoy. 2015. "Estimating Clearing Functions for Production Resources Using Simulation Optimization." *IEEE Transactions on Automation Science and Engineering* 12 (2):539-552.

Kang, Y. H., E. Albey, S. Hwang, and R. Uzsoy. 2014. "The Impact of Lot Sizing in Multiple Product Environments with Congestion." *Journal of Manufacturing Systems* 33:436-444.

Karmarkar, U. S. 1989. "Capacity Loading and Release Planning with Work-in-Progress (Wip) and Lead-Times." *Journal of Manufacturing and Operations Management* 2 (1):105-123.

Kefeli, A., and R. Uzsoy. 2016. "Identifying Potential Bottlenecks in Production Systems Using Dual Prices from a Mathematical Programming Model " *International Journal of Production Research* 54 (7):2000-2018.

Kim, B., and S. Kim. 2001. "Extended Model for a Hybrid Production Planning Approach." *International Journal of Production Economics* 73:165-173.

Magnani, A., and S. P. Boyd. 2009. "Convex Piecewise-Linear Fitting." *Optimization and Engineering* 10:1-17.

Missbauer, H. 2002. "Aggregate Order Release Planning for Time-Varying Demand." *International Journal of Production Research* 40:688-718.

Missbauer, H. 2009. "Models of the Transient Behaviour of Production Units to Optimize the Aggregate Material Flow." *International Journal of Production Economics* 118 (2):387-397.

Missbauer, H. 2011. "Order Release Planning with Clearing Functions: A Queueing-Theoretical Analysis of the Clearing Function Concept." *International Journal of Production Economics* 131 (1):399-406.

Missbauer, H., and R. Uzsoy. 2011. "Optimization Models of Production Planning Problems." In *Planning Production and Inventories in the Extended Enterprise: A State of the Art Handbook*, 437-508. Boston: Springer.

Schwarz, J. A., G. Selinka, and R. Stolletz. 2016. "Performance Analysis of Time-Dependent Queueing Systems: Survey and Classification." *OMEGA* 63:170 - 189.

Srinivasan, A., M. Carey, and T. E. Morton. 1988. Resource Pricing and Aggregate Scheduling in Manufacturing Systems. In *Graduate School of Industrial Administration, Carnegie-Mellon University*. Pittsburgh, PA

Toriello, A., and J. P. Vielma. 2012. "Fitting Piecewise Linear Continuous Functions." *European Journal of Operational Research* 219 (86 - 95).

Vollmann, T. E., W. L. Berry, D. C. Whybark, and F. R. Jacobs. 2005. *Manufacturing Planning and Control for Supply Chain Management.* . New York: McGraw-Hill

AUTHOR BIOGRAPHIES

KARTHICK GOPALSWAMY is a doctoral student in the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. He holds a Masters in Industrial and Systems Engineering from North Carolina State University. His research focuses on data oriented decision analysis in production systems using stochastic simulation. His e-mail address is kgopals@ncsu.edu.

REHA UZSOY is Clifton A. Anderson Distinguished Professor in the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. He holds BS degrees in Industrial Engineering and Mathematics and an MS in Industrial Engineering from Bogazici University, Istanbul, Turkey. He received his Ph.D. in Industrial and Systems Engineering in 1990 from the University of Florida. His teaching and research interests are in production planning, scheduling, and supply chain management. He was named a Fellow of the Institute of Industrial Engineers in 2005, Outstanding Young Industrial Engineer in Education in 1997, and has received awards for both undergraduate and graduate teaching. His email address is ruzsoy@ncsu.edu.