Proceedings of the 2011 Winter Simulation Conference
S. Jain, R.R. Creasey, J. Himmelspach, K.P. White, and M. Fu, eds.

OPTIMIZING SURGERY START TIMES FOR A SINGLE OPERATING ROOM VIA SIMULATION

Yang Sun

College of Business Administration
California State University, Sacramento
6000 J Street
Sacramento, CA 95819-6088, USA

Xueping Li

Department of Industrial & Information Engineering
The University of Tennessee
408 East Stadium Hall
Knoxville, TN 37996-0700, USA

ABSTRACT

Operating room scheduling is often done in steps. First, surgeries are assigned to an operating room's time blocks. Assigned surgeries are then sequenced. Idle time is often reserved at the end of the time block in order to buffer against possible overtime. This research focuses on the next step of determining the amount of time reserved for each of the pre-sequenced surgeries so that surgical teams know their exact start times. In this way the buffer time is redistributed to each of the surgeries in order to minimize total overtime and idling costs. The problem is modeled as a special periodic review inventory model and a simulation-based response surface method is used to optimize surgery start times for a single operating room with stochastic operation durations represented by an infinite set of stochastic scenarios. This purposed method does not require extensive computational effort and is easy for practitioners to implement.

1 INTRODUCTION

Operating rooms (ORs) are often considered among the most critical resources in a hospital, and OR scheduling is important for improving operations efficiency as well as service quality. OR scheduling is often determined in steps. First, surgeries are assigned to an operating room's time blocks. Assigned surgeries are then sequenced with the consideration of resource-related constraints and specifications (Jebali, Alouane and Ladet 2006). In practice, buffer time is often inserted at the end of the time block in order to buffer against possible overtime.

While overtime staffing is costly for hospitals, it is also desirable that a surgery can finish before the next surgery's scheduled start time in order not to keep the next surgical team waiting. (While in practice two consecutive surgeries can be performed by the same surgical team, it is still desirable that the first surgery can be completed on time so that the second patient can get into the OR as scheduled.) Such waiting times may lead to higher operations cost and surgical team fatigue (Denton, Viapiano and Vogl 2007). Lately some teaching hospitals have started to broadcast surgeries on the internet and interact with audiences on social networking sites (Cohen 2009), and on-time starts of surgeries are essential. A more important surgery may incur a heavier overtime penalty for the previous surgery that uses the same OR. Therefore it is necessary to redistribute the butter time into each of the sequenced surgeries. On the other hand, OR idling before the next surgery reduces OR utilization and incurs an opportunity cost. It is important to schedule the exact start times for surgeries in order to minimize both expected overtime and idling costs.

Assume that a surgery starts at time zero and needs to finish by time r. For a single surgery the problem is equivalent to a *newsvendor* problem of reserving a certain amount of time r in the OR in order to minimize total expected cost of overstocking and understocking.

Assume that the actual surgery duration time is a random variable t with pdf $f(t)$ and cdf $F(t)$. If $t > r$, a per time unit overtime penalty cost C^o incurs. On the other hand, if $t < r$, OR idling incurs an overstock (earliness penalty) cost at C^e per time unit. C^e can be considered a per time unit reservation cost. That is, if the surgery finished early and one less time unit were reserved, a cost C^e would have been saved. However, the cost $C^e t$ for using the OR in a duration t is a sunk cost and should not be considered in the decision problem. The objective function is

It is straightforward to show that the optimal solution has

Erdogan and Denton (2010) provide a comprehensive literature review on single OR surgery scheduling. Weiss (1990) studies a case in which there are $n = 2$ surgeries and $C_2^o = 0$ for the surgery that is scheduled to be second. That is, there is no overtime penalty for the second surgery. In general the more important surgery should be scheduled to be first in order to have a lower C_1^o that incurs from keeping the second surgical team waiting. Assume that the first surgery starts at time zero. The scheduling problem of determining the start time s_2 for the second surgery is equivalent to the above single surgery problem of determining r_1 for the first surgery and time reservation for the second surgery does not matter. However, if a positive overtime cost C_2^o exists and a convex ordering exists between surgery duration times, the smaller surgery should be scheduled first (Denton, Viapiano and Vogl 2007).

The single OR scheduling problem with $n > 2$ is computationally challenging when combining both sequencing and start time problems. Denton, Viapiano and Vogl (2007) propose a two-stage strategy under which the sequence of surgeries is determined using heuristics. Start times for pre-sequenced surgeries are solved using stochastic programming where random surgery durations are represented by a finite set of stochastic scenarios. A set of artificial constraints are used in the stochastic programming model to balance overtime and idling.

While intuitive heuristics can be used to generate difference surgery sequences for comparison, in practice the sequencing decision should also be made by taking into account resource-related constraints and specifications. In this research, we assume that surgeries are already allocated to OR time blocks and pre-sequenced and focus on the stage of determining start times for pre-sequenced surgeries for a single OR. The assumption of having a finite set of stochastic scenarios is relaxed. In Section 2, the scheduling problem is modeled as a special periodic review inventory problem. Given the inclusion of multiple random variables and the complexity of the problem, in Section 3 we propose a simulation-based response surface method to help find the optimum. The purpose is to develop a straightforward, easy-to-implement method for practitioners to use. Numerical experiments are presented in Section 4 to show the effects of the cost structure and the variability of the random duration on the redistribution of the buffer time into each of the surgery jobs. Conclusions are drawn in Section 5.

2 PROBLEM FORMULATION

Consider a set of $J = \{1,2,\ldots n\}$ pre-sequenced surgery jobs in a time block T. The time duration of job j is a random variable t_j with pdf $f_j(t_j)$ and cdf $F_j(t_j)$. The decision variable is to reserve for job j. Assume that the first job starts at $a_1 = 0$, we have the following required finish time d_j for job j and scheduled start time s_{j+1} for job j+1.

The actual start time a_{j+1} for job j+1 can be recursively written as $a_{j+1} = \max\{a_j+t_j, s_{j+1}\}$, which depends on when job j actually finishes. A required finish time $d_n = T$ is also set for the last surgery. Assume that r_j is reserved at a cost of C^e per time unit for all jobs. However, depending on job j+1, each job has a different overtime cost rate . (depends on the hospital's overstaffing cost.) The objective function is

The problem is, in essence, a special case of the periodic review inventory problem where the lead time is zero and inventory items (time units) perish in a single period. Excess inventory cannot be carried to the next period (i.e., used by the next surgery); however, backorders are carried throughout the entire planning horizon. See Nahmias (1982) for a review on inventory problems with perishable products. The extreme case with inventory items that perish in one period is studied by Bulinskaya (1964). The problem in our research is a generalized case in which understock costs ( ) are heterogeneous for different periods. (In a more generalized case, the overstock cost C^e can also be heterogeneous.) Moreover, the problem has a finite planning horizon and the replenishment quantity of the last decision period depends on the decisions made for previous periods as well as the supplier's maximum capacity.

3 AN OPTIMIZATION VIA SIMULATION APPROACH

Given the inclusion of n random variables in the problem, the evaluation of the objective function is very challenging. In this section we propose a simulation-based approach to help optimize the cost function via designed simulation experiments executed under the response surface methodology (RSM). RSM fits well for stochastic optimization problems like ours where control variables (r_j) are numeric and the objective function is convex. See Montgomery and Myers (1995) for details on the RSM and Kleijnen(1987) for the RSM in a simulation context. RSM-based optimization via simulation is generally done in two phases (Fu 1994). We often start with a point that is remote from the optimum. First-order experimental designs are first used to estimate the steepest descent directions from fitted ordinary least square regression models to lead the control variables rapidly along a path of improvement towards the general optimum vicinity. This is repeated until the linear response surface has nearly zero slopes. Once the region of the optimum is found, a second-order polynomial model is then fitted in the second phase with a more detailed experimental design to analytically determine the optimum.

In order to demonstrate how the simulation-based response surface method works for our problem, a three-job case is discussed. Consider $n = 3$ surgery jobs that have been assigned to and sequenced in an OR time block with $T = 240$ (minutes). Denote the jobs in the given sequence as jobs 1, 2 and 3. Assume that random surgery durations t_j follow independent log-normal distributions with the following parameters. Strum et al. (2000) show that the log-normal distribution is superior to the normal distribution for modeling surgery duration times.

; ;

The parameters (μ, σ) are the mean and standard deviation of the associated normal distribution. The mean and standard deviation of the log-normal random variable are functions of μ and σ. With the above settings, we have log-normally distributed t_1, t_2, t_3 with means 60, 60, 90 and standard deviations 5, 10, 15, respectively. The decision variables are r_1 and r_2 and we automatically have $r_3 = T - r_1 - r_2$. Assume that job 1 starts at $a_1 = 0$. The Monte Carlo simulation is coded in the Matlab (http://www.mathworks.com/) computing environment and actual surgery durations t_j are randomly sampled from a log-normal random number generator. In the simulation we have $d_1 = s_2 = r_1$, $a_2 = \max\{a_1+t_1, s_2\}$, $d_2 = s_3 = r_1+ r_2$, $a_3 = \max\{a_2+t_2, s_3\}$, and $d_3 = T$. Assume $C^e = 1$, , and . The response variable is the total cost $C^T =$. The decision is to find a set of r_j that minimizes the expected total cost.

Without losing generality, we skip the description of the phase-I simulation and directly discuss the development of a phase-II experimental design for fitting the second-order polynomial response surface around the optimum. A rotatable central composite design (CCD) is used in this phase. (See Montgomery and Myers (1995) for details on the CCD.) Being rotatable means that the prediction power is the same at all points that are the same distance from the design center. From an optimization perspective, this is ne-

cessary since the location of the optimum is unknown and it is important to provide equal precision of estimation in all directions. Design points for simulation experiments are listed in Table 1.

Table 1: CCD of Simulation Experiments

Factor	**A: r_1**	**B: r_2**
Factorial Points	Level (-1) = 55 minutes	Level (-1) = 60 minutes
	Level (1) = 75 minutes	Level (-1) = 60 minutes
	Level (-1) = 55 minutes	Level (1) = 80 minutes
	Level (1) = 75 minutes	Level (1) = 80 minutes
Axial Points	Level (-1.4) = 51 minutes	Level (0) = 70 minutes
	Level (1.41) = 79 minutes	Level (0) = 70 minutes
	Level (0) = 65 minutes	Level (-1.41) = 56 minutes
	Level (0) = 65 minutes	Level (1.41) = 84 minutes
Center Point	Level (0) = 65 minutes	Level (0) = 70 minutes

For each of the factorial and axial points, 100 simulation replicates are executed. For the center point, 500 simulation replicates are executed. There are 1,300 simulation replicates in total. A statistical analysis on simulation results using Design-Expert 8.0 (http://www.statease.com/) leads to the following Analysis of Variance (ANOVA) output (Figure 1).

Response: Total Cost

Transform: Natural Log

ANOVA for Response Surface Quadratic Model

Source	Sum of Squares	df	Mean Square	F Ratio	p-value Prob > F	
Model	31.92458	5	6.384915	36.42696	< 0.0001	**significant**
A-r1	0.569885	1	0.569885	3.251286	0.0716	
B-r2	4.551929	1	4.551929	25.96948	< 0.0001	
AB	3.548465	1	3.548465	20.24456	< 0.0001	
A^2	19.36257	1	19.36257	110.4665	< 0.0001	
B^2	6.400279	1	6.400279	36.51461	< 0.0001	
Residual	226.8123	1294	0.17528			
Lack of Fit	0.510044	3	0.170015	0.969893	0.4061	not significant
Pure Error	226.3022	1291	0.175292			
Total	258.7368	1299				

Figure 1: ANOVA for simulation outputs

A natural log transformation on the response variable is used in order to obtain normally distributed residuals. Figure 2 graphically shows the fitted quadratic response surface model . This model follows the hierarchical principle. If a higher order term is significant, all lower order terms that compose it are also included in the model to provide internal consistency. With an F-Ratio of 36.4, the chance that a model F-Ratio this large could occur due to noise is less than 0.01%. The Lack of Fit F-Ratio of 0.97 implies that the Lack of Fit is not significant relative to the pure error. In other words, the model built is significant and adequately fits the data. An adequate precision ratio can also be calculated to measures the signal to noise ratio - a ratio greater than 4 is desirable. The above study has an adequate precision ratio of 16.14 that indicates an ade-

quate signal. The model built can be used to navigate the design space. The optimum is found at $r_1 = 65$ and $r_2 = 74$ with an expected total cost of 36. ($r_3 = 240 - 65 - 74 = 101$.)

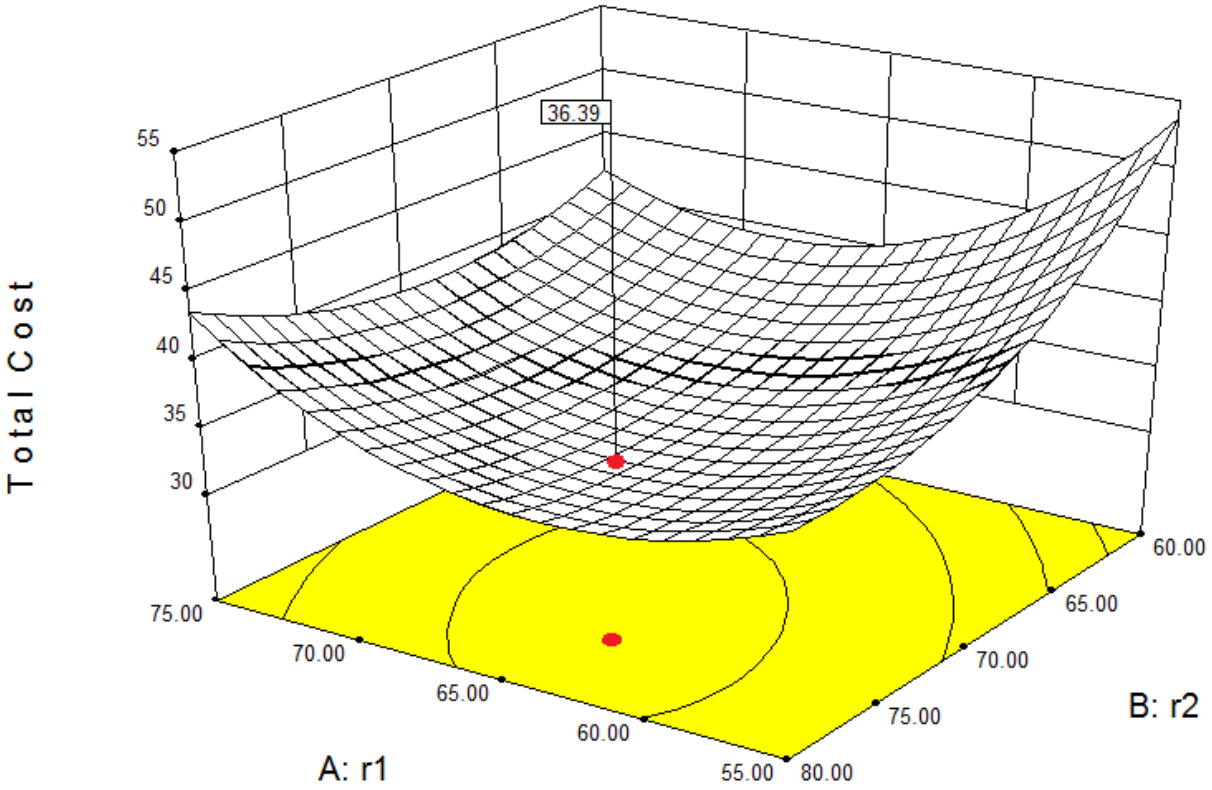

Figure 2: Fitted response surface

From an implementation perspective, RSM-based optimization via simulation is generally a very quick method since data are only collected at design points. (With manual transposing of data from Matlab to Design-Expert on a regular PC laptop, it takes the authors less than 2 minutes to conclude the procedure. Due to the nature of Monte Carlo simulation and RSM-based methods, this will not be a lot longer for $n > 3$ cases or longer time blocks.) In order to verify our method, the same simulation model is also built in the Arena environment and OptQuest (http://www.arenasimulation.com/Products_OptQuest.aspx) is used to search for the optimum. OptQuest combines several metaheuristics into a single search algorithm; however the exact algorithm is unknown. For the above case, it takes OptQuest approximately 2 minutes to get to the optimum vicinity with a replication setting of 50 and a random start point, but over 1.5 hours to converge at the optimum $\{r_1,r_2\}= \{65,74\}$ on a high-end server.

4 NUMERICAL STUDIES

In order to gain insights into the surgery start time problem, we focus on the $n = 3$ case. The purposed method can be easily implemented in $n > 3$ cases. Assume that $n = 3$ surgery jobs are allocated and presequenced for a $T = 240$ minutes (half day) time block of a single OR. In the first set of experiments, we assume that are i.i.d. random variables following . That is, all jobs have a random duration time with the same mean of 75 and standard deviation of 15. If , then conceptually a total buffer time of minutes is reserved at the end of the block. Let be the optimal reservation found for job j and be the percentage out of

the total buffer time that is allocated to job j. We start with . The effect of is tested and presented in Figure 3.

A counter intuitive result can be seen that, when the overtime costs are equal for all jobs, the buffer time is not equally distributed into each of the jobs. This is in accordance with the perishable nature of the reserved time that overstocking of time from the first job, if any, cannot be carried and used by the second job and is a sunk loss. On the other hand, if the first job finishes late, there is still a chance for succeeding jobs to catch up with the schedule. Wang (1993) shows that, in single server scheduling, for identical jobs with exponentially distributed durations, the middle job should always get the largest share of time allowance in order to minimize delays. Our results show that this is generally true for log-normally distributed durations unless the idling cost is largely higher than the overtime cost.

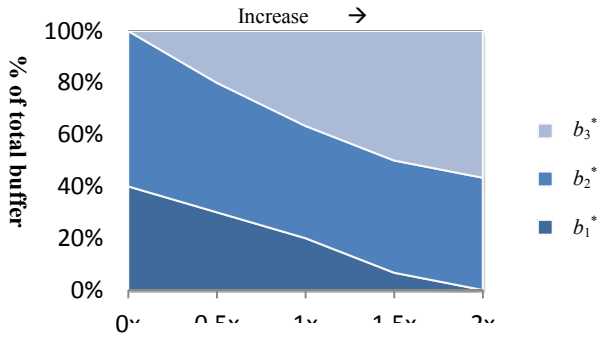

Figure 3: The effect of C^e on the allocation of the buffer time

In the second set of experiments, we assume that ; i.e., general OR utilization has been addressed by previous decisions and buffering for overtime is more important than buffering for idling when determining start times. Figure 4 shows the effect from increasing one of the overtime costs (while remaining the other two unchanged at) on the redistribution of the buffer time. As one of the gets larger, i.e., as one of the jobs faces a more important successor, that job quickly gains a bigger share of the buffer time.

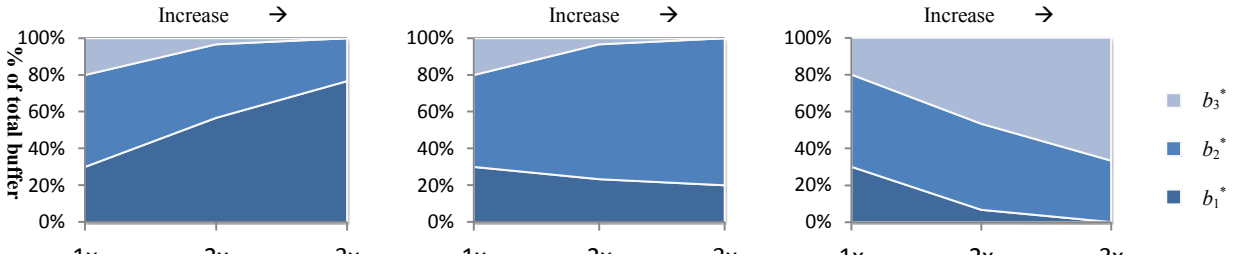

Figure 4: The effect of on the allocation of the buffer time

The third set of experiments assumes that . However, the variability of is heterogeneous for different jobs. Different levels of the standard deviation are tested for one of the while the standard deviations for the other two duration variables are kept at 15.

Interestingly, Figure 5 shows that increasing the duration variability of the last job does not affect the allocation of the buffer time. This is in accordance with the fact that r_3 is not a decision variable. The middle job (job 2) always gets a larger share of the buffer for risk pooling unless its duration variability is very low. When the variability of t_2 gets larger, a bigger share of the buffer is allocated to job 2 and is mostly taken from r_1. Increasing the variability of t_1, on the other hand, makes both r_1 and r_2 larger.

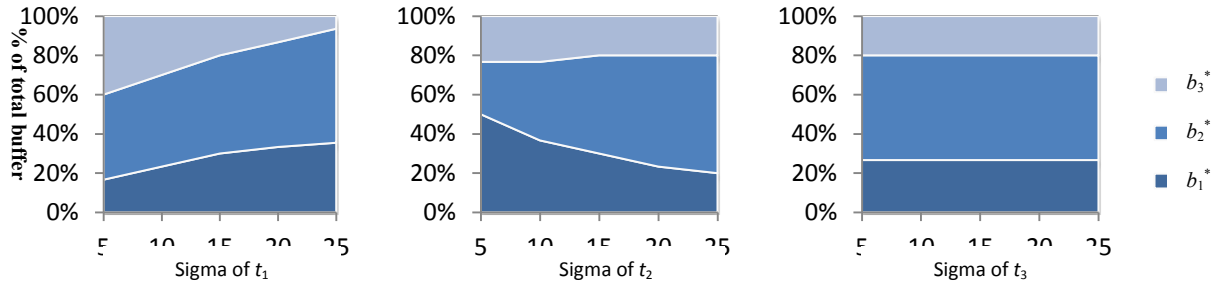

Figure 5: The effect of duration variability on the allocation of the buffer time

5 CONCLUSION REMARKS

In this research, we consider the problem of determining start times for pre-sequenced surgery jobs for a single OR, where stochastic surgery durations are represented by an infinite set of stochastic scenarios. The problem is modeled as a special periodic review inventory problem where stock items (reserved time units) perish in a single period and the supplier's capacity needs to be fully utilized in n periods. A simulation-based response surface method is used to find the optimum of distributing the available buffer time into different jobs so that the time reservation, or the start time, of each job is determined. Numerical studies show that the levels of the per time unit idling cost, per time unit overtime cost, and variability of random duration have effects on such allotments of the buffer time.

Future research is suggested to address the multiple OR problem that combines surgery-to-OR allocation, sequencing and start time decisions, where not only the overtime cost varies depending on the sequence of surgeries but also the idling cost is unequal for different ORs and different times of the day. Methods should be developed to overcome the possible loss of global optimality. Sequence-dependent setup times, emergency surgeries, possible surgery cancellations and non-log-normally distributed duration times can be considered. Since the purposed method does not require extensive computational effort and is very easy for practitioners to implement, similar scheduling methods can also be applied to other services (e.g., clinical appointments).

REFERENCES

Bulinskaya, E. V. 1964. "Some Results Concerning Optimum Inventory Policies." *Theory of Probability and its Applications* 9:389-403.

Cohen, E. 2009. "Surgeons Send 'Tweets' from Operating Room." *CNN Technology*. February 17.

Denton, B. T., J. Viapiano, and A. Vogl. 2007. "Stochastic Optimization of Surgery Sequencing and Start Time Scheduling Decisions." *Health Care Management Science* 10:13-24.

Erdogan, S. A., and B. T. Denton. 2010. "Surgery Planning and Scheduling: A Literature Review." *Wiley Encyclopedia of Operations Research and Management Science*. John Wiley & Sons.

Fu, M. C. 1994. "Optimization via Simulation: A Review." *Annals of Operations Research* 53:199-248.

Jebali, A., A. B. H. Alouane, and P. Ladet. 2006. "Operating Rooms Scheduling." *International Journal of Production Economics* 99:52-62.

Kleijnen, J. P. C. 1987. *Statistical Tools for Simulation Practitioners*. New York: Marcel Dekker.

Montgomery, D. C., and R. H. Myers. 1995. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. 1st ed. John Wiley & Sons.

Nahmias, S. 1982. "Perishable Inventory Theory: A Review." *Operations Research* 30:680-708.

Strum, D. P., A. R. Sampson, J. H. May, and L. E. Vargas. 2000. "Surgeon and Type of Anesthesia Predict Variability in Surgical Procedure Times." *Anesthesiology* 92:1454–1466.

Wang, P. P. 1993. "Static and Dynamic Scheduling of Customer Arrivals to a Single-Server System." *Naval Research Logistics* 40:345-360.

Weiss, E. N. 1990. "Models for Determining the Estimated Start Times and Case Orderings." *IIE Transactions* 22:143–150.

AUTHOR BIOGRAPHIES

YANG SUN is an Assistant Professor in the College of Business Administration at California State University, Sacramento. In addition to his Ph.D. in Industrial Engineering from Arizona State University, he has a Six-Sigma Black Belt. He also received an engineering degree from Tsinghua University. He worked as a precursor of Application Service Provider in the IT industry and as a lean and six-sigma consultant for several organizations. He teaches Decision Sciences and Operations Management courses at undergraduate, MBA, and executive levels. Yang Sun's research interests focus on the application of quantitative methods in supply chain, production and service systems. He has worked on various funded or self-supported research projects and is a recipient of the University's Outstanding Scholarly and Creative Activities Award. He is on the editorial board of Sacramento Business Review and International Journal of Operations Research and Information Systems, and is a member of INFORMS, M&SOM, POMS, IIE, DSI, Omega Rho and Alpha Pi Mu. His email address is suny@csus.edu.

XUEPING LI is an Assistant Professor of Industrial and Information Engineering and the Director of the Intelligent Information Engineering Systems Laboratory (IIESL) at the University of Tennessee, Knoxville. He holds a Ph.D. in Industrial Engineering from Arizona State University. His research areas include complex system modeling, simulation and optimization, information assurance, scheduling, web mining, supply chain management, lean manufacturing, and sensor networks. Dr. Li is a recipient of the I&IE Outstanding Researcher Award and named as a QUEST Scholar of the University of Tennessee. He is a member of IIE, IEEE and INFORMS. His e-mail is Xueping.Li@utk.edu.