AN ABSTRACT OF THE THESIS OF

Akash Sharma for the degree of Master of Science in Industrial Engineering and Computer Science presented on December 12, 2003.

Title: Effectiveness of using Two and Three-parameter Distributions in Place of "Best-fit Distributions" in Discrete Event Simulation Models of Production Lines.

Abstract approved: [redacted]

David S. Kim

This study presents the results of using common two or three-parameter "default" distributions in place of "best fit distributions" in simulations of serial production lines with finite buffers and blocking. The default distributions used instead of the best-fit distribution are chosen such that they are non-negative, unbounded, and can match either the first two moments or the first three moments of the collected data. Furthermore, the selected default distributions must be commonly available (or easily constructed from) distributions in simulation software packages. The lognormal is used as the two-parameter distribution to match the first two moments of the data. The two-level hyper-exponential and three-parameter lognormal are used as three-parameter distributions to match the first three moments of the data. To test the use of these distributions in simulations, production lines have been separated into two major classes: automated and manual. In automated systems the workstations have fixed processing times and random time between failures, and random repair times. In manual systems, the workstations are reliable but have random processing times. Results for both classes of lines show that the differences in throughput from simulations using best-fit distributions and two parameter lognormal is small in some cases and can be reduced in others by matching the first three moments of the data. Also, different scenarios are identified which lead to higher differences in throughput when using a two-parameter default distribution.

Effectiveness of using Two and Three-parameter Distributions in Place of "Best-fit Distributions" in Discrete Event Simulation Models of Production Lines.

by

Akash Sharma

A THESIS

submitted to

Oregon State University

in partial fulfillment of

the requirements for the

degree of

Master of Science

Presented December 12, 2003

Commencement June, 2004

Master of Science thesis of Akash Sharma presented on December 12, 2003.

APPRO[redacted]

Co-Major Professor, representing Industrial Engineering

[redacted]

Co-Major Professor, representing Computer Science

[redacted]

Head of the Department of Industrial and Manufacturing Engineering

[redacted]

Director of the School of Electrical Engineering and Computer Science

[redacted]

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes the release of my thesis to any reader upon request.

[redacted]

Akash Sharma, Author

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF TABLES

LIST OF APPENDIX FIGURES

LIST OF APPENDIX FIGURES (Continued)

Effectiveness of using Two and Three-parameter Distributions in Place of "Best-fit Distributions" in Discrete Event Simulation Models of Production Lines.

1. Introduction

A drawback of conducting simulation analysis, from the perspective of companies, is that a complete simulation study requires skills in multiple areas (e.g., data collection, statistics, computer programming etc.), and usually takes a significant amount of time. If one or more of the steps can be simplified or eliminated, a simulation study can be conducted faster and may require less overall technical background to execute. Significant amounts of effort have been devoted to developing simulation software that essentially makes conducting steps such as model coding, input data preparation, and output analysis more efficient. However, most if not all of the steps (and what occurs in each step), remain the same. If, for certain identifiable situations, the work involved in executing some of the simulation steps is reduced, or if the step can be eliminated, then the overall study can be conducted in less time.

This research focuses on reducing the work involved with input data preparation in simulation studies of production lines (a more general description of these systems is *closed serial queuing systems with blocking*). This includes the class of systems that can be referred to as "tightly coupled" production lines (systems with a relatively small number of buffer spaces), where large amount of simulation effort is focused. Systems within this class may also be further segregated into *manual* and *automated* systems. In manual systems the processing times at each workstation are variable, whereas in automated systems the workstation processing times are fixed but each workstation is subjected to random failures and repair times. We contend that for simulations of this class of systems, it is sufficient in many cases to use "default distributions" that match either the first two or three moments of collected data, in place of a best fit distribution, thereby eliminating data preparation steps and analysis that normally follow data collection. By sufficient we mean that the difference in the simulated performance measures of the production line, when using default distributions is within 5% of the values obtained when simulating with best-fit distributions. Moreover the default distributions should be relatively simple. That is, they must be commonly available in simulation software packages, and have parameters that are easily estimated from sample data.

The input data preparation steps that we contend may be eliminated in production line simulations, after collecting raw data are: distribution selection, parameter estimation, and conducting goodness-of-fit tests (two goodness-of-fit tests often used are the Chi Square and Kolmogrov-Smirnov test). Moreover, this process may need to be performed for a number of processing time, failure time, and repair time distributions. Software exists that automates much of this; however, the individual data points must be saved and formatted properly, and the knowledge to use and interpret the software results is required. If our contention is justified, these input data preparation steps would be replaced by calculations of the sample mean, variance, and skewness. These are simple calculations once the data have been collected.

The approach used in this study is a careful evaluation and selection of the default distributions. This is followed by a designed computational experiment to test errors introduced by using the default distribution. A number of different manual and automated production lines are simulated assuming the *true* distribution of random components follow a variety of known distributions (e.g., exponential, uniform, discrete, etc.). A default distribution that matches the first two moments of this true distribution is then used as a replacement for the true distribution, and the throughput results are compared. The objectives are to assess the differences in throughput caused by using a two-parameter default distribution and identify when the use of two-parameter default distributions causes unacceptable errors. For those cases where using a two-parameter distribution causes unacceptable errors, three-parameter distributions are used and the errors are examined again using simulation.

In Section 2 we review the literature on the subject in question. Section 3 describes the configuration of the production systems examined in the study. The selection of the two-parameter and three-parameter default distributions is reviewed in Section 4. This is followed by a description of the experimental design in Section 5. Finally, results and concluding remarks are presented in Section 6 and 7, respectively.

2. Literature Review

There are several published papers that report on alternatives to the typical input data preparation steps of distribution selection, parameter estimation, and conducting goodness-of-fit tests. Shanker and Kelton (1991) list three different methods for input data modeling for any simulation study. These are: the use of best-fit distributions, the use of empirical distributions, and the use of a family of distributions like the Johnson and the generalized lambda distribution. The suggested use of empirical distributions in place of best-fit distributions will avoid the possible errors when selecting the wrong distribution, which may follow from the prevalent practice of approximating a "true" underlying distribution by a fitted distribution from a standard family. The major advantage of using empirical distributions is that they are easy to understand and implement. Further, Shanker and Kelton (1991) have claimed that when the results from the two methods were compared, the empirical approach did just as well as the best fitted standard distributions and sometimes better. The major disadvantage with this method is that an empirical distribution can only be used in situations where plenty of data is available. Another disadvantage of using empirical distributions is that they only take values within the range of the observations collected.

The idea behind using a family of distributions that belong to a flexible parametric family is similar to what is proposed in this study. These distributions can be parameterized such that they match multiple moments of the collected data. A method for using the Johnson distribution to model input processes in simulation experiments was suggested by De Brota, Roberts, Swain and Venkataraman (1989). They suggested that the Johnson system can be used for situations where little or no sample information is available. This aspect makes the approach attractive in the case of modeling conceptual systems.

Similar work has been done by Ramberg and Schmeiser (1974), involving Tukey's lambda distribution. They suggested a modification of Tukey's lambda distribution into a four-parameter distribution, the combination of which can result in variety of distribution shapes. They reported good results with this distribution. However, the values of the parameters are limited which limits the shapes of this distribution. A main advantage of using this distribution is that it is easier to generate sample observations since the distribution function has a straightforward inverse.

One drawback of both of the prior families of distributions is that the setting of distribution parameters is not as straightforward as with more common distributions. Additionally, when strictly non-negative random variates are required (not including common distributions that are included in these families, e.g., lognormal) these families are restricted to bounded distributions.

Given these alternatives to finding and using best-fit distributions in simulation, there is some evidence that the level of precision required in a default distribution depends on the type of performance measure of interest. Our objective is to use the simplest default distributions that give sufficient accuracy.

Studies have been done to compare measures of performance when different input distributions are used to represent the same random process. Gross and Juttijudata (1997) studied sensitivity of the average wait in queue to the particular type of input distribution. This was studied within the context of simulation modeling of queuing systems with gamma, lognormal, Weibull, and Pearson service time and inter-arrival time distributions. Sensitivity analysis of the average wait in queue, and the 95-percentile value of the queue wait distribution, indicated that matching only the first two moments of the data is not always sufficient for lower traffic intensities (0.2 and to some extent 0.5). Similar work was done by Gross and Masi (1998). In this case a two-node call center was considered and four types of service time and inter-arrival time distributions were used. The distributions tested were the gamma, lognormal, beta and Pearson type-5 distributions and mean queue wait was used as the performance measure. The results showed that the system performance measure (mean queue wait) is significantly different for different distributions even though the first two moments were matched. A similar study by Gross (1999), for a bank simulation model showed that the output performance measures (mean queue wait and the 95^{th} percentiles of the waiting time in queue) were sensitive to the particular distribution and just matching the first two moments is not always an acceptable way to model input data. Lau (1986) suggests that for a two stage directly coupled system (two-stage line with no buffer space between the stations), the mean and skewness of the service time distribution play an important role in minimizing the cycle time of the system where cycle time is defined as the time taken for a job to pass through a line from start to finish.

Harchol-Balter, Li, Osogami and Scheller-Wolf (2003), have shown good results for approximation models that match the first three moments of a general distribution. They devised an analytical method to approximate the expected wait time for long jobs for a distributed server system having both "long" and "short" jobs. An M/G/k system is modeled as a Markov Chain using 2-stage Coxian distributions (matching the first three moments) as the service time distribution. Whitt (1983), used Extremal distributions to approximate the upper and lower bounds of the mean queue length for GI/G/1 system in which both the inter-arrival times and service times are mixtures of the exponential distribution. The results obtained indicate that matching two moments can result in poor accuracy and additional shape constraints can reduce the range of the bounds. Johnson and Taffe (1991), study the effect of three performance measures: the probability of delay of an arbitrary customer, the mean queue length, and probability that an arriving customer is blocked for a GI/M/· type systems. They claim that two moment approximations are adequate when coefficients of variation (CV) values are less than one. However for higher CV values, three moments should be matched.

In contrast to waiting time and queue length measures, Muth (1973) asserts that the throughput for an n-station production line is a function of the first two moments of the service-time distribution. Other studies have been conducted to develop models for the performance of production systems with blocking and also provide an indication of the impact of distribution selection. Blumenfeld (1990) and Baker (1992), both developed analytical approximations for the throughput of production lines with blocking. Their work suggests that the throughput of a production line is primarily a function of only the mean workstation processing times, and the CV. This is referred to as the *two moment property* (Baker, 1992). In contrast to the prior work, Rao (1975) claims that the *two moment property* is not accurate for moderate values of CV when production rate is the performance measure.

The above studies present evidence that the accuracy of using distributions that match the first two moments of a "true" distribution depends on the performance measure of interest. In this research, we explore the effect of CV and skewness on throughput in production lines with finite buffers through a series of simulation experiments.

3. System Description

A production line system is shown in Figure 1. The system consists of four workstations (labeled a, b, c and d) and a fixed number of carriers (e.g., pallets, material handling devices etc.). The buffer between any two machines is finite and excludes the space at the machine. Thus in Figure 1, the total carrier capacity of the system is limited to 12. If each of the workstations has a deterministic processing time of one unit, then the system produces completed jobs at the rate of 1 part per time unit (if each workstation has a carrier and buffer transit times are negligible). A variation of this system may have random processing times (which we refer to as a manual system) in which case the throughput calculation of the system is not straightforward although some analytical solutions do exist for some small Markovian systems (Liu, Li and Buzacott, 1992). Another variation of this system is when the workstations have constant processing times but random failures and repairs (which we refer to as an automated system).

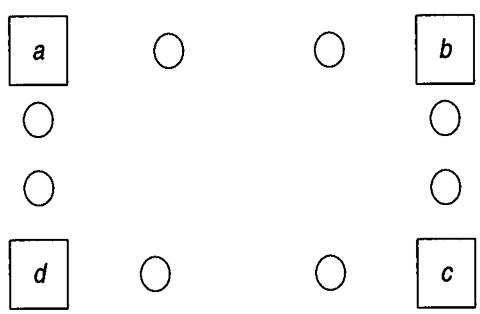

Figure 1: A closed production line with finite buffers.

In general let the system of interest have M workstations separated by finite buffers. Let the number of buffers preceding a workstation i (excluding the space at the workstation) be denoted by b_i. Further, let C denote the number of carriers in the system. Let the processing time distribution of workstation i in a manual system be denoted by P_i. Therefore, the throughput of the system through time t can be defined as :

$$T(M,B,P,C,t) = \frac{N(M,B,P,C,t)}{t} \tag{1}$$

Where $N(M,B,P,C,t)$ denotes the number of jobs completed by time t, $B = (b_1, b_2, \ldots, b_M)$ and $P = (P_1, P_2, \ldots, P_M)$.

For the automated system, let F_i represent the distribution of the time between failures and let R_i be the distribution of the repair times on the i^{th} workstation. The throughput of the automated system is denoted by $T(M,B,P,C,F,R,t)$, is given by:

$$T(M,B,P,C,F,R,t) = \frac{N(M,B,P,C,F,R,t)}{t} \tag{2}$$

Where $F = (F_1, F_2, \ldots, F_M)$, $R = (R_1, R_2, \ldots, R_M)$, and P_i is now the fixed processing time at workstation *i*.

Certain aspects of the system that are worth mentioning here are:

1. The production line is subject to production blocking.
2. The buffers behave in First-in-first-out (FIFO) fashion.
3. The job flow is asynchronous. The job flow is based strictly on the availability of space downstream and hence the jobs can flow independently of one another.
4. The system is a closed system (a fixed number of carriers are in the system at all times). It should be noted that open systems are a special case of closed systems.
5. The performance measure of interest is system throughput.

In this study, we will use two and three parameter "default" distributions instead of the best-fit distributions for *P, F*, and *R*.

4. Default Distribution Selection

Since the objective of using default distributions is to simplify the process of developing production line simulation models, the default distribution should ideally have the following characteristics:

1. The two-parameter distribution should have the capability to match any CV.
2. The three-parameter distribution should also be able to match or come close to matching any skewness.
3. The default distributions should only generate positive random variates.
4. The default distributions should be available in most simulation software, or be easily constructed from distributions available in most simulation software.
5. Solving for the parameters of default distributions should be relatively simple (i.e., analytical equations as opposed to numerical procedures should exist).
6. The default distributions should be unbounded.

Characteristic 1 allows the default distribution to match any sample coefficient of variation obtained from actual data. Characteristic 2 permits matching three moments in cases where matching two moments may not be adequate. Characteristic 3 is important because in real life processes, negative process times are not possible. Characteristic 4 permits straightforward application of default distributions using commonly available simulation software. Characteristic 5 minimizes additional knowledge required to use default distributions. Finally, Characteristic 6 allows the default distribution to model the occurrence of less common events such as very long repairs, etc.

Various distributions were considered for use as default distributions. The generalized lambda distribution was considered, but was rejected as it does not meet the third criteria (it generates negative random variates for certain values of its parameters), and is not commonly found in simulation software. The unbounded Johnson distribution (JD) produces negative random variates. The suggested method to fit the bounded JD is the method of percentiles (Johnson, 1949), which would defeat the purpose of our study. The generalized beta distribution is a bounded distribution and may be a good choice for a default distribution when the true distribution is thought to be bounded.

The two-parameter distribution that meets the six criteria is the lognormal distribution. Two different three parameter distributions that meet most of the criteria are the two level hyper-exponential distribution, and the three-parameter lognormal distribution.

The lognormal distribution is a continuous distribution in which the logarithm of the random variable has a normal distribution. The probability density function of a lognormal random variable X is:

$$f(x)=\frac{1}{Sx\sqrt{2\pi}}e^{-(\ln x-M)^2/(2S^2)} \tag{3}$$

where S and M are the parameters of the distribution which can be used to define the mean, variance, and skewness of the distribution as shown below:

$$E[X]=e^{M+S^2/2} \tag{4}$$

$$Var[X]=e^{S^2+2M}\left(e^{S^2}-1\right) \tag{5}$$

$$Skewness[X]=\sqrt{e^{S^2}-1}\left(2+e^{S^2}\right) \tag{6}$$

The parameters of the lognormal distribution, expressed as a function of its mean and variance, are as follows:

$$S = \sqrt{\ln\left(\frac{\sigma^2}{\mu^2} + 1\right)} \tag{7}$$

where $\mu = E[X]$ and $\sigma^2 = Var[X]$. Having found S we can find M by using the formula:

$$M = \ln(\mu) - \frac{S^2}{2} \tag{8}$$

Thus by knowing the sample mean $(\hat{\mu})$ and sample variance $(\hat{\sigma}^2)$ of the data for a random system component we can compute the parameters of the lognormal default distribution that match the sample statistics.

To select the default three-parameter distribution, we analyze which distribution can most closely match the skewness values possible for non-negative random variables. Whitt (1982) summarizes necessary and sufficient conditions for moment consistency of non-negative random variables. Let m_i denote the i^{th} moment about the origin for a non-negative random variable. Then,

$$\begin{aligned} &m_1 \geq 0, \\ &m_2 - m_1^2 \geq 0, \\ &m_1 m_3 - m_2^2 \geq 0. \end{aligned} \tag{9}$$

The skewness of a random variable is given by the formula:

$$Skewness = \frac{m_3 - 3m_1 m_2 + 2m_1^3}{(m_2 - m_1^2)^{3/2}} \tag{10}$$

Equations 9 and 10 can be used to derive a bound for the skewness of a non-negative random variable as a function of its CV.

$$Skewness \geq CV - \frac{1}{CV} \tag{11}$$

For a three-parameter lognormal random variable X (which is a shifted lognormal distribution with τ denoting the added parameter):

$$E[\mathrm{X}] = \tau + e^{M+S^2/2} \tag{12}$$

$$Var[\mathrm{X}] = e^{S^2+2M}\left(e^{S^2} - 1\right) \tag{13}$$

$$Skewness[\mathrm{X}] = \sqrt{e^{S^2} - 1}\left(2 + e^{S^2}\right) \tag{14}$$

These formulas imply that $Skewness[X] \geq CV(CV^2 + 3)$ if $\tau \geq 0$, where $CV(CV^2 + 3)$ equals the skewness of a two-parameter lognormal random variable.

A two-level hyper-exponential random variable X with parameters λ_1, λ_2 and p has density function, $f(x) = p\lambda_1 e^{-\lambda_1 x} + (1-p)\lambda_2 e^{-\lambda_2 x}$, for $x \geq 0$, $0 \leq p \leq 1$. Here λ_1 and λ_2 represent the parameters of two different exponential distributions. For such a random variable, Whitt (1982) shows that $CV \geq 1$, and $m_3 \geq \frac{1.5m_2^2}{m_1}$. Using (10) it can be shown that $Skewness[X] \geq 1.5CV + \frac{0.5}{CV^3}$. A plot of minimum skewness versus CV for non-negative, three-parameter lognormal, and two-level hyper-exponential random variables is shown in Figure 2.

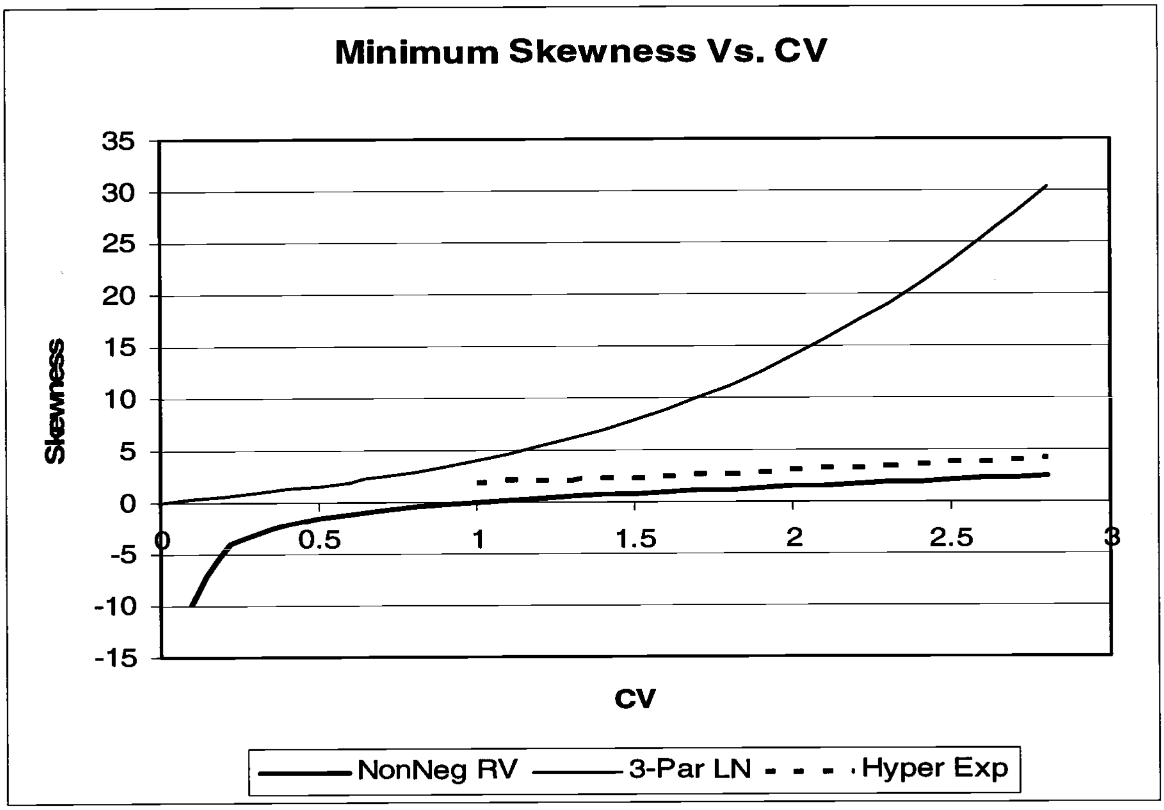

Figure 2: Minimum skewness versus CV for three-parameter distributions

As can be seen from Figure 2, it is not possible to use a two-level hyper-exponential distribution when the input data has a CV of less than 1. Also, the three-parameter lognormal is a heavy tailed distribution. Hence to maximize the ability to match the skewness of sample data, the three-parameter default distributions used are three-parameter lognormal for $CV < 1$, and the two-level hyper-exponential when $CV \geq 1$.

The parameters of the three-parameter lognormal (M, S, τ) can be computed from the sample mean $(\hat{\mu})$, sample variance $(\hat{\sigma}^2)$, and sample skewness $(\hat{K})$ using the following formulas (Crow and Shimizu, 1988):

$$S = \left(\ln\left[\frac{2^{1/3}}{(2+\hat{K}^2+\sqrt{4\hat{K}^2+\hat{K}^4})^{1/3}} + \frac{(2+\hat{K}^2+\sqrt{4\hat{K}^2+\hat{K}^4})^{1/3}}{2^{1/3}} - 1 \right] \right)^{1/2} \qquad (15)$$

$$M = \ln\left[\frac{\hat{\sigma}}{\sqrt{e^{S^2}(e^{S^2}-1)}}\right] \tag{16}$$

$$\tau = \hat{\mu} - e^{M+S^2/2} \tag{17}$$

In those cases where the skewness cannot be matched (indicated by $\tau < 0$), set $\tau = 0$ That is, use the two-parameter lognormal to match the first two moments of the data. This will result in the closest match of the sample skewness

The parameters of the two-level hyper-exponential can be computed from the sample moments $(\hat{m}_1, \hat{m}_2, \hat{m}_3)$ using the formulas from Whitt (1982).

$$\lambda_i^{-1} = \frac{(x+1.5y^2+3\hat{m}_1^2 y) \pm \sqrt{(x+1.5y^2+3\hat{m}_1^2 y)^2 + 18\hat{m}_1^2 y^3}}{6\hat{m}_1 y} \tag{18}$$

$$p = (\hat{m}_1 - \lambda_2^{-1})/(\lambda_1^{-1} - \lambda_2^{-1}) \tag{19}$$

$$x = \hat{m}_1\hat{m}_3 - 1.5\hat{m}_2^2 \tag{20}$$

$$y = \hat{m}_2 - 2\hat{m}_1^2 \tag{21}$$

If the skewness of the data is too small to match exactly (see Figure 2), which is indicated by a negative value for a parameter, then replace m_3 with a value slightly larger than $1.5m_2^2/m_1$ (Whitt 1982).

5. Experimental Design

In order to explore the impact of using default distributions, different production line parameters that affect throughput were identified as experimental factors, and experiments were conducted over different levels of these factors. The factors and levels are as listed in Table 1. The factor "carrier concentration" can be defined as the fraction of the total capacity of the systems that has been occupied by the carriers. Thus eight different systems were simulated for each distribution. Each system was replicated 30 times each until 30,000 jobs were obtained per replication.

Table 1: Factors and levels used to examine the use of default distributions.

Factor	Factor Description	Levels
1	Number of machines	5 and 20 workstation
2	Buffer space available between two machines	0 and 4 buffer spaces between any two workstations
3	Carrier concentration	0.2S and 0.8S *
4	The type of distribution assumed to be the true distribution of *P, F,* and *R*	16 different distributions for *P*, 15 different distributions for *F* and *R*, and the default distribution

* S represents the total carrier capacity of a production line

The first set of simulation experiments used the two-parameter lognormal as the single default distribution in place of best-fit distributions. Based on the results, a second set of experiments were conducted that matched the first three moments of the best-fit distribution using either the hyper-exponential distribution or the three parameter lognormal distribution. The methodology used to carry out the simulation experiments in this study is summarized in Figure 3.

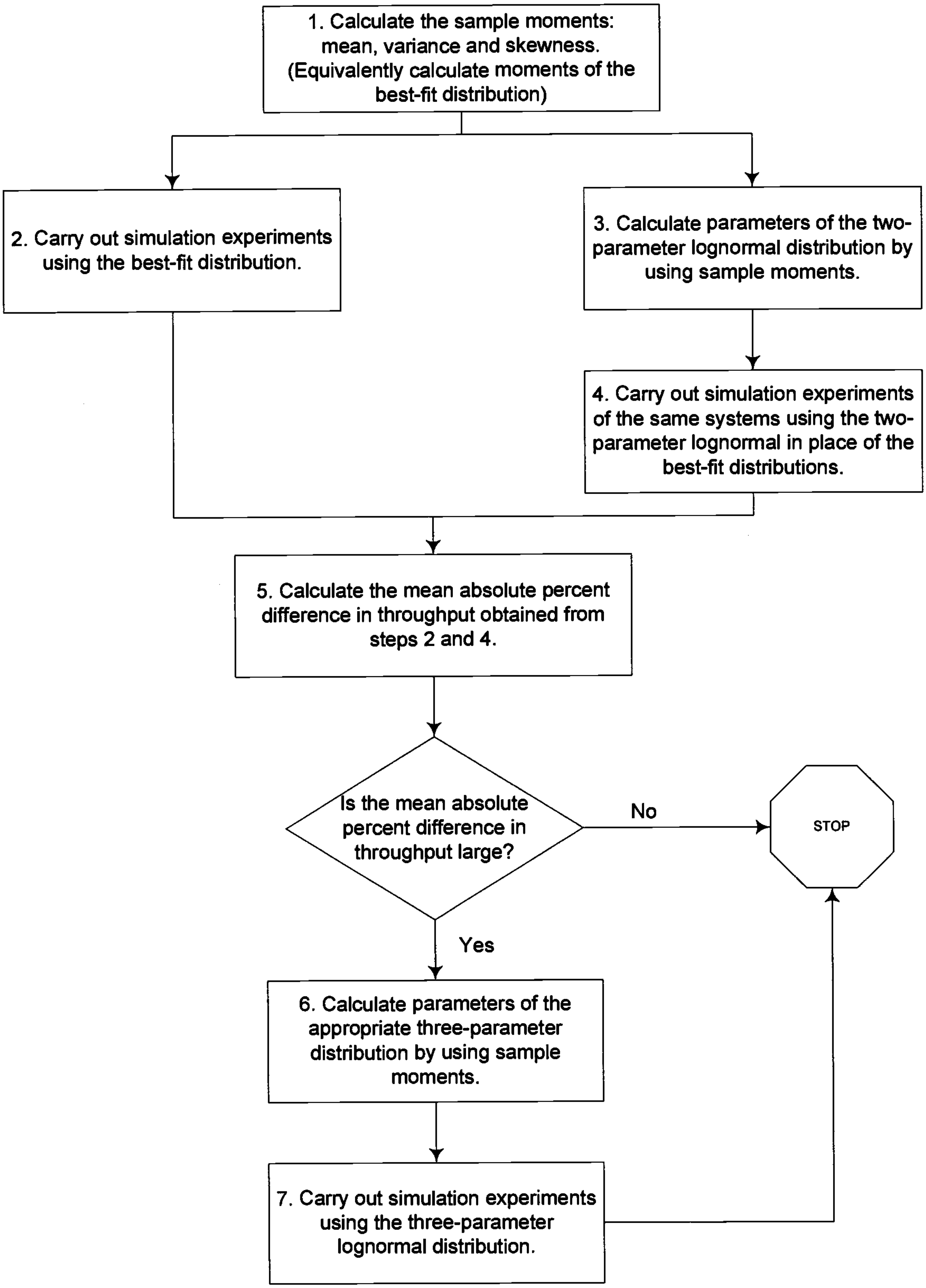

Figure 3: Methodology for conducting the simulation experiments.

The distributions that were assumed to be the true distributions of P, F, and R had a variety of shapes ranging from flat for a uniform distribution to highly skewed for a Weibull distribution. This was done to test the robustness of the two-parameter default distribution used, and to isolate characteristics of the true distributions that may lead to higher errors. The response was taken to be the mean absolute percent difference between the throughput of systems simulated with best-fit distributions, and the same systems simulated by matching the first two or three moments of the best-fit distributions. The different distributions that were used as true or best-fit distributions for P are given in Table 2, and those used for F and R are given in Table 3.

Table2: "True" distributions used for process times (P) in simulations of manual production lines.

Distribution Number	Name of Distribution	Parameters of the Process Time Distribution	CV of the Process Time Distribution
1	Triangular	TRIA(8,10,12)	0.08
2	Uniform	UNIF(8,12)	0.12
3	Bimodal	4*BETA(0.5,0.5)+8	0.14
4	Normal	NORM(10,2)	0.20
5	Beta	20*BETA(0.05,0.95)+9	0.31
6	Exponential	EXPO(10)	1.00
7	Weibull	WEIB(0.75,8.3988)	1.36
8	Gamma	GAMMA(0.5,20)	1.41
9	Weibull	WEIB(0.7,7.9)	1.46
10	Transformed Beta	480*BETA(0.05,40)+20	0.13
11	Transformed Beta	480*BETA(0.1,30)+20	0.23
12	Transformed Beta	480*BETA(0.025,3)+20	0.90
13	Transformed Beta	980*BETA(0.025,5)+20	1.13
14	Weibull	WEIB(0.5,5)	2.24
15	Gamma	GAMMA(0.1,100)	3.16
16	Gamma	GAMMA(0.05,200)	4.47

Table 3: "True" distributions used to for failure (F) and repair times (R) in simulations of automated production lines.

Distribution Number	Name of Distribution	Fixed Processing Times	Parameters of Distribution for Failures	CV of Failure Distribution	Parameters of Repair Times Distribution	CV of Repair Time Distribution
1	Bimodal	10	10*BETA(0.5,0 5)+45	0.07	4*BETA(0.5,0.5)+8	0.14
2	Triangular	10	TRIA(40,50,60)	0.08	TRIA(8,10,12)	0.08
3	Normal	10	NORM(50,5)	0.10	NORM(10,2)	0.20
4	Uniform	10	UNIF(40,60)	0.12	UNIF(8,12)	0.12
5	Exponential	10	EXPO(50)	1.00	EXPO(10)	1.00
6	Weibull	10	WEIB(0.75,41.9942)	1.35	WEIB(0.75,8.3988)	1.36
7	Gamma	10	GAMMA(0.5,100)	1.41	GAMMA(0.5,20)	1.41
8	Weibull	10	WEIB(0.7,39.5)	1.46	WEIB(0.7,7.9)	1.46
9	Transformed Beta	20	2400*BETA(0.05,40)+100	0.13	480*BETA(0.05,40)+20	0.13
10	Transformed Beta	20	2400*BETA(0.1,30)+100	0.23	480*BETA(0.1,30)+20	0.23
11	Transformed Beta	20	2400*BETA(0.025,3)+100	0.90	480*BETA(0.025,3)+20	0.90
12	Transformed Beta	20	4900*BETA(0.025,5)+100	1.13	980*BETA(0.025,5)+20	1.13
13	Weibull	10	WEIB(0.5,25)	2.24	WEIB(0.5,5)	2.24
14	Gamma	10	GAMMA(0.1,500)	3.16	GAMMA(0.1,100)	3.16
15	Gamma	10	GAMMA(0.05,1000)	4.47	GAMMA(0.05,200)	4.47

6. Results and Discussion

A total of 16 best-fit distributions for manual systems and 15 best-fit distributions for automated systems (Tables 2 and 3) were simulated. These same systems were simulated using both two-parameter and three-parameter default distributions (wherever required). The results tabulated were the mean absolute percent differences in throughput between the best-fit simulation and the default distribution simulation. The simulation results of systems with different configurations are summarized in Tables 4 to 7. These tables are divided into four quadrants each corresponding to a high or low coefficient of variation (CV) of the best-fit distribution, and the difference in skewness between the best-fit and corresponding default distribution. The low CV level corresponds to a CV less than or equal to 1.33, and a high level are CVs greater than 1.33. A low level of difference in skewness corresponds to values less than or equal to 5 and a high level for differences in skewness greater than 5. The quadrants will be referenced according to the scheme mentioned in Table 4:

Table 4: Scheme for referencing quadrants for displaying results.

CV	Difference in Skewness	Quadrant
Low	Low	1
High	Low	2
Low	High	3
High	High	4

Table 5 shows the mean simulation results obtained by simulating manual systems using a two parameter lognormal distribution as the default distribution. As can be seen from the results, quadrant 1 has the lowest absolute mean error and quadrant 4 has the highest. Quadrant 2 has moderate values of mean absolute error. Quadrant 3 has mixed results, and hence displays no significant trend. Matching the first two moments lead to low error in most of the cases. Out of 16 best-fit distributions used in the manual systems simulations, the mean absolute difference in throughput was below 5% in 11 of the cases.

Table 5: Mean absolute percent difference in throughput obtained by using a two-parameter lognormal in place of best-fit distributions for manual systems.

		Difference in skewness							
		Low				High			
CV		Distribution Number	CV	Diff. in Skewness	Mean absolute % difference in throughput	Distribution Number	CV	Diff. in Skewness	Mean absolute % difference in throughput
	Low	1	0.08	0.25	0.02	10	0.13	-8.22	0.38
		2	0.12	0.35	0.37	11	0.23	-5.30	0.76
		3	0.14	0.07	0.09	12	0.90	-5.22	7.56
		4	0.20	0.61	0.17	13	1.13	-5.01	10.55
		5	0.31	2.94	3.33				
		6	1.00	2.00	1.68				
		Type	CV	Diff. in Skewness	Mean absolute % difference in throughput	Type	CV	Diff. in Skewness	Mean absolute % difference in throughput
	High	7	1.36	3.41	2.24	14	2.24	11.27	5.67
		8	1.41	4.24	2.86	15	3.16	34.79	10.63
		9	1.46	4.02	2.55	16	4.47	93.91	17.66

In those cases where high differences in throughput were observed, a second set of experiments were conducted using the same manual systems but this time using the default distributions matching the first three moments of the best-fit distributions. Hence the second set of experiments were carried out for quadrants 2, 3, and 4 . The results are shown in Table 6.

Table 6 suggests that matching the first three moments of the data works most of the time except for some cases in quadrant 3 where the value of the mean absolute percent error is high.

Table 6: Mean absolute percent difference in throughput obtained by matching the first three moments of the best-fit distribution for manual systems.

		Difference in skewness							
		Low				High			
CV		Distribution Number	CV	Diff. in Skewness	Mean absolute % difference in throughput	Distribution Number	CV	Diff. in Skewness	Mean absolute % difference in throughput
	Low					10	0.13	0.00	-
						11	0.23	0.00	-
						12	0.90	0.00	6.22
						13**	1.13	0.00	10.70
		Distribution Number	CV	Diff. in Skewness	Mean absolute % difference in throughput	Distribution Number	CV	Diff. in Skewness	Mean absolute % difference in throughput
	High	7	1.36	0.00	0.43	14	2.24	0.00	2.22
		8	1.41	0.00	0.77	15	3.16	0.00	2.49
		9	1.46	0.00	0.53	16	4.47	0.00	2.56

** Three-parameter lognormal instead of hyper-exponential distribution has been used for this case even though the CV is greater than 1.

Table 7 shows the mean absolute percent difference in simulation results obtained by simulating automated systems by matching the first two moments of F_i and R_i. Table 8 shows the results of the same systems by matching the first three moments of F_i and R_i. The results indicate similar patterns as obtained with manual systems. As can be seen from the results of Table 7, out of the 15 best-fit distributions tested, the difference in throughput was below 5% in 10 of the cases. Matching the first three moments helps in decreasing the mean absolute percent difference in throughput most of the time (except for some cases in quadrant 3).

Thus from the above results we conclude that matching the first two moments works for quadrants 1 and 2. Also if the CV is high, and the skewness difference versus a lognormal distribution is high (CV>1.33 and difference in skewness >5) then it is better to use a three parameter default distribution.

Table 7: Mean absolute percent difference in throughput obtained by simulating automated systems with best-fit and lognormal distribution

		Difference in skewness									
		Low					High				
CV	Low	Distribution Number	CV of TBF	Diff. in Skew. of Time between Failure distribution	Diff. in Skew. of Repair Time distribution	Mean absolute % difference in throughput	Distribution Number	CV of TBF	Diff. in Skew. of Time between Failure distribution	Diff. in Skew. of Repair Time distribution	Mean absolute % difference in throughput
		1	0.07	0.21	0.07	0.26	9	0.13	-8.22	-8.22	2.35
		2	0.08	0.25	0.25	0.08	10	0.23	-5.30	-5.30	1.35
		3	0.10	0.30	0.61	0.06	11	0.90	-5.22	-5.22	6.31
		4	0.12	0.35	0.35	0.10	12	1.13	-5.01	-5.01	8.46
		5	1.00	2.00	2.00	1.78					
	High	Distribution Number	CV of TBF	Diff. in Skew. of Time between Failure distribution	Dif. in Skew. of Repair Time distribution	Mean absolute % difference in throughput	Distribution Number	CV of TBF	Diff. in Skew. of Time between Failure distribution	Diff. in Skew. of Repair Time distribution	Mean absolute % difference in throughput
		6	1.35	3.41	3.41	3.32	13	2.24	11.27	11.27	6.79
		7	1.41	4.24	4.24	3.58	14	3.16	34.79	34.79	11.49
		8	1.46	4.02	4.02	3.64	15	4.47	93.91	93.91	17.99

Table 8: Mean absolute percent difference in throughput obtained by matching the first three moments of the failure time and repair time distributions for automated systems.

		Difference in skewness									
		Low					High				
CV	Low	Distribution Number	CV of TBF	Diff. in Skew. of Time between Failure distribution	Diff. in Skew. of Repair Time distribution	Mean absolute % difference in throughput	Distribution Number	CV of TBF	Diff. in Skew. of Time between Failure distribution	Diff. in Skew. of Repair Time distribution	Mean absolute % difference in throughput
							11	0.90	0.00	0.00	4.40
							12***	1.13	0.00	0.00	6.09
	High	Distribution Number	CV of TBF	Diff. in Skew. of Time between Failure distribution	Diff. in Skew. of Repair Time distribution	Mean absolute % difference in throughput	Distribution Number	CV of TBF	Diff. in Skew. of Time between Failure distribution	Diff. in Skew. of Repair Time distribution	Mean absolute % difference in throughput
		6	1.35	0.00	0.00	0.63	13	2.24	0.00	0.00	3.82
		7	1.41	0.00	0.00	0.48	14	3.16	0.00	0.00	1.63
		8	1.46	0.00	0.00	1.50	15	4.47	0.00	0.00	2.24

*** Three-parameter lognormal instead of hyper-exponential distribution has been used for this case even though the CV is greater than 1.

6.1 Effect of Carrier Concentration

We plotted scatter plots for the values of the mean absolute percent error at different levels of machines and carrier concentration. The graphs are shown in Figures 4 to 7. Figure 4 shows the graph for manual systems using the two-parameter lognormal as the default distribution. As can be seen from Figure 4, the error is higher in case of systems with higher carrier concentration (i.e. at 80% of the system capacity).

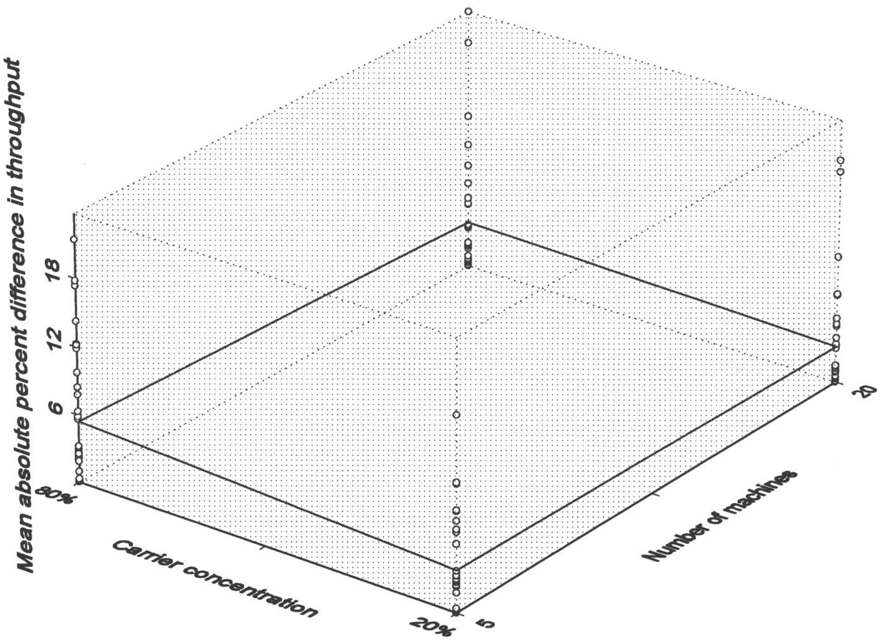

Figure 4: 3-D scatter plot for mean absolute percent difference in throughput over different levels of machines and carrier concentration for manual systems using lognormal as the default distribution.

A similar scatter plot is shown in Figure 5 for automated systems using best-fit and lognormal distributions. The pattern of the graph suggests that the error is higher in case of systems with more carrier concentration.

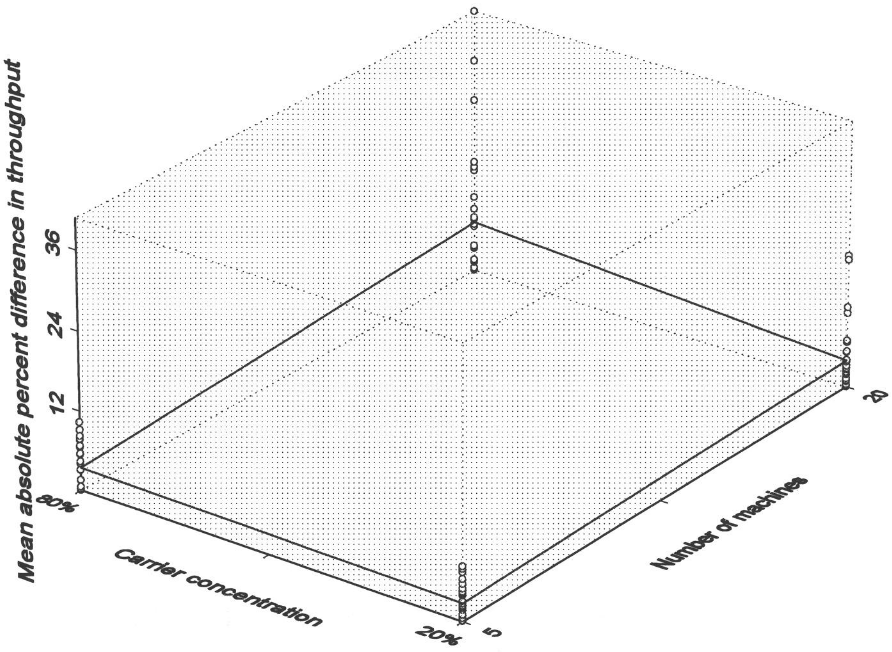

Figure 5: 3-D scatter plot for mean absolute percent difference in throughput over different levels of machines and carrier concentration for automated systems using lognormal as the default distribution.

Figures 6 and 7 show the 3-D scatter plot for mean absolute percent error for manual and automated systems respectively when the first three moments of the F_i and R_i have been matched. Similar scatter plots for each quadrant in Tables 5 to 8 are shown in Appendix 1. The pattern suggests that the higher the concentration of carriers is in the systems, the higher resulting mean absolute error is.

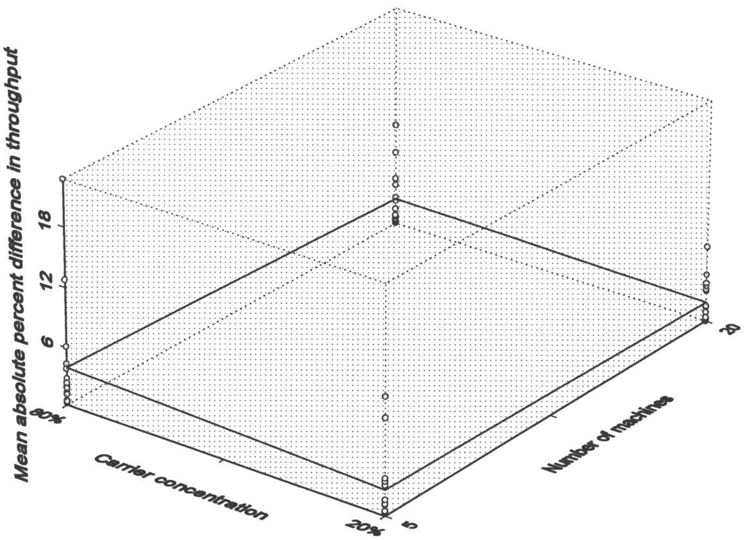

Figure 6: 3-D scatter plot for mean absolute percent difference in throughput over different levels of machines and carrier concentration for manual systems by matching the first three moments of the true distribution.

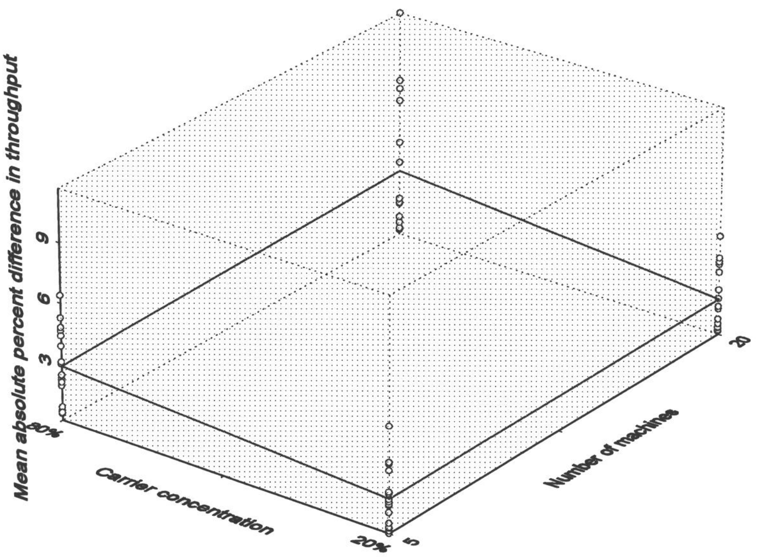

Figure 7: 3-D scatter plot for mean absolute percent difference in throughput over different levels of machines and carrier concentration for automated systems by matching the first three moments of the true distribution.

We performed ANOVA by keeping the mean absolute percent difference in throughput as the response variable and using number of machines, number of buffer spaces between machines, type of distribution and carrier concentration as factors. The ANOVA was performed for both manual and automated systems when only the first two moments of the best-fit distributions were matched using a two-parameter lognormal distribution. The results from the ANOVA table show that all the factors have a significant effect on the response variable. The interaction plot between carrier concentration and number of distributions for manual systems is shown in Figure 8.

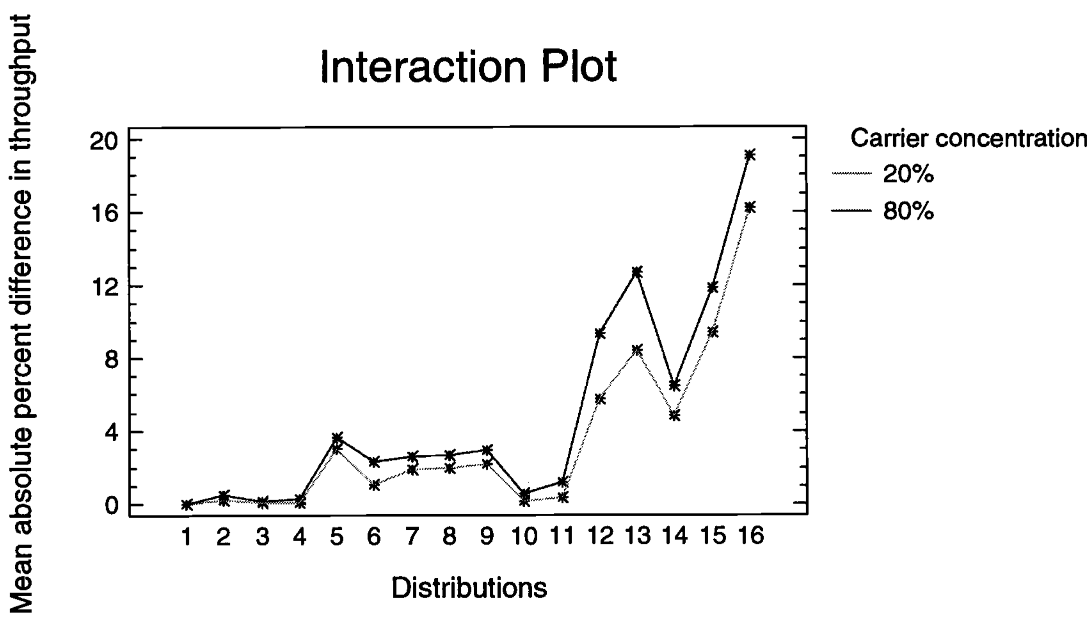

Figure 8: Interaction plot between carrier concentration and different types of distributions for manual systems.

The interaction plot also suggests that the mean absolute percent error is higher in case of systems having high carrier concentration at 80%. A similar plot for automated systems, shown in Figure 9, also exhibits the same trend for the most part. The 95% confidence intervals in Figure 10 show that the mean absolute percent error at different carrier concentrations differs significantly and is higher for higher carrier concentrations of automated systems.

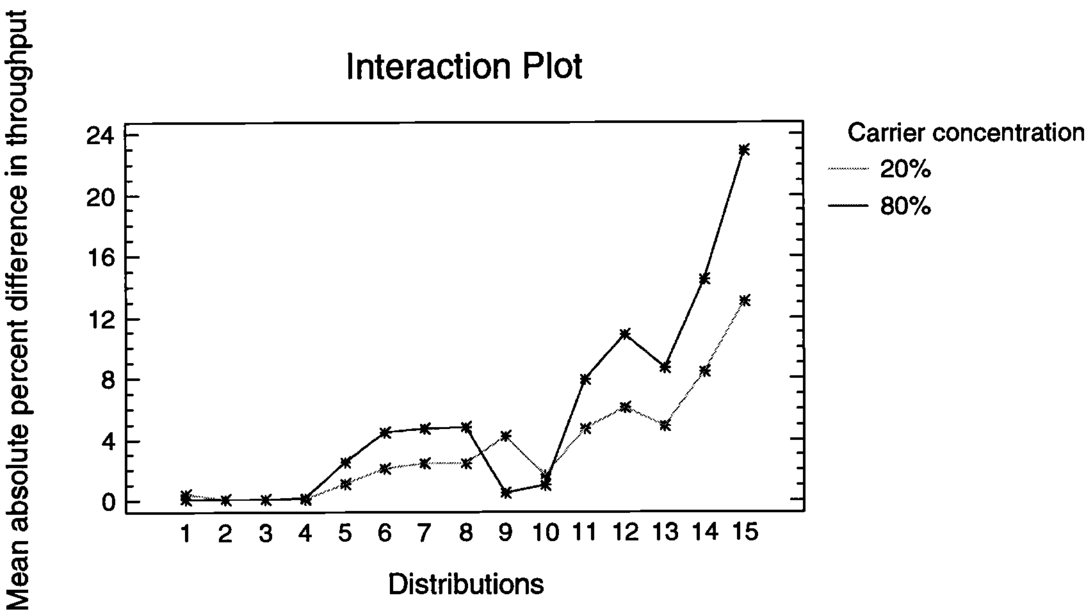

Figure 9: Interaction plot between carrier concentration and types of distributions for automated systems.

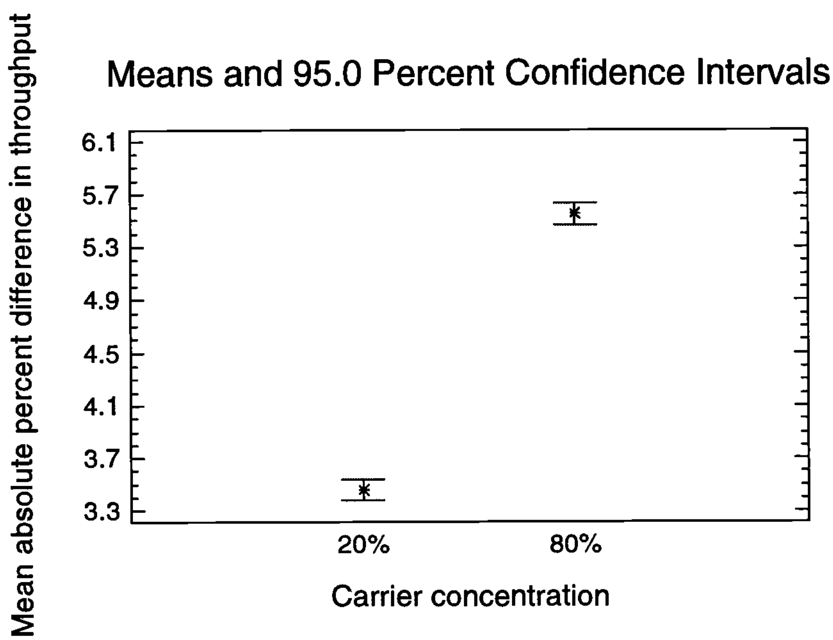

Figure 10: Confidence intervals for mean absolute percent difference in throughput at two levels of the carrier concentration for automated workstations.

6.2 Effect of Buffer Spaces

Using ANOVA we analyzed the effect of buffer spaces in driving the mean absolute percent error. Figure 11 shows an interaction plot for number of machines and buffer spaces for manual systems when the two-parameter lognormal was used as the default distribution.

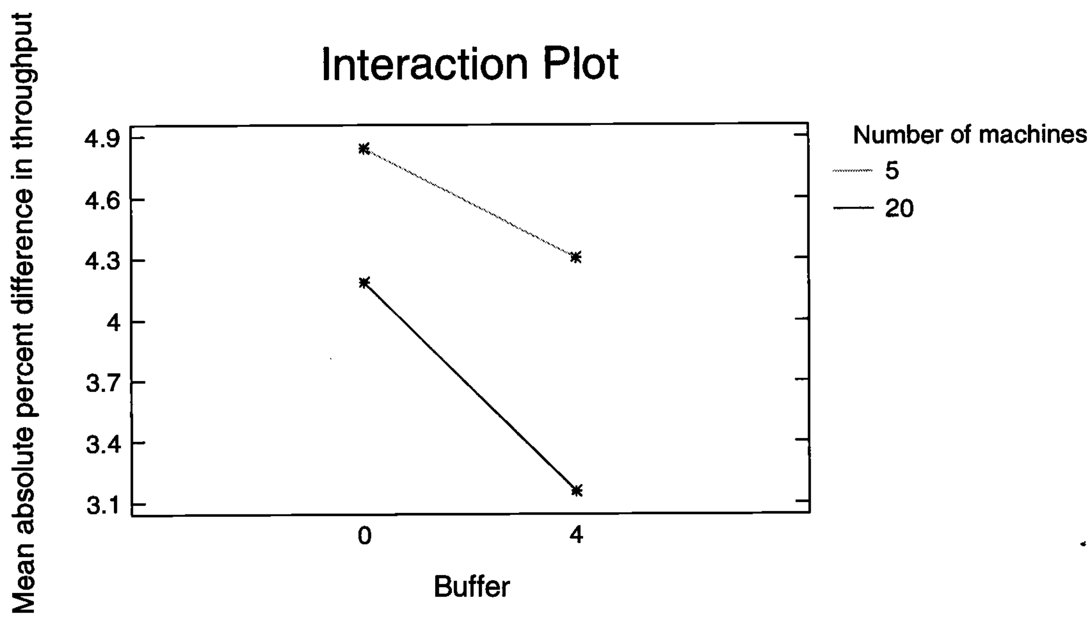

Figure 11: Interaction plot between number of buffer spaces and number of machines for manual systems.

The interaction suggests that the mean absolute percent error is lower in case of systems that have buffer between any two machines for the completed part to be transferred to. The mean absolute percent error is high in case of smaller systems which suggests that a smaller system having no buffer between machines would lead to higher error than a big decoupled system. The 95% confidence intervals shown in Figure 12 also suggest that lower values of buffer would lead to higher errors.

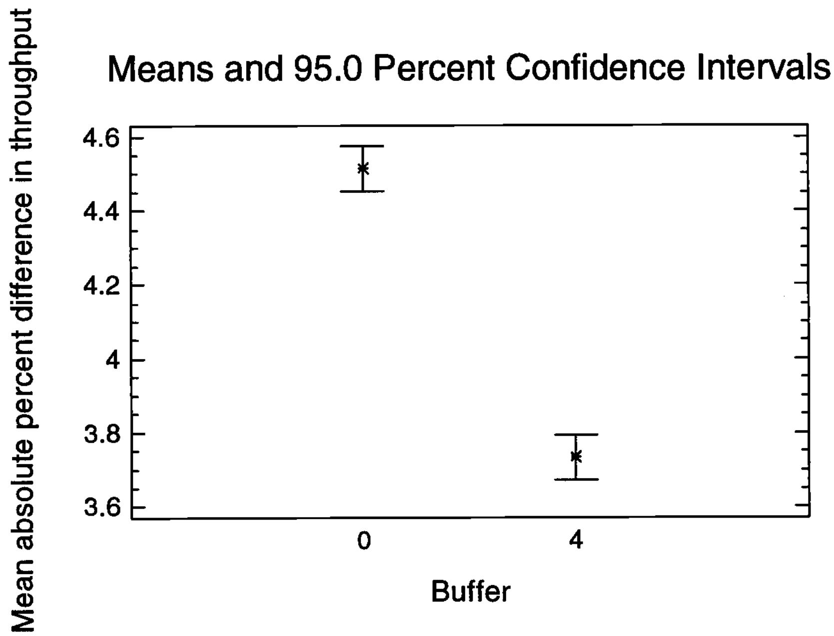

Figure 12: Confidence intervals for mean absolute percent difference in throughput at two levels of the buffer between any two machines for manual systems.

Figure 13 shows the interaction plot between number of machines and buffer spaces for automated systems. The plot indicates that the mean absolute percent error is high in case of larger systems with lower buffer between machines which again supports our previous conclusion.

The 95% confidence intervals for buffer spaces for automated systems shown in Figure 14 also validates that for automated workstation no buffer between machines would lead to higher value of mean absolute error.

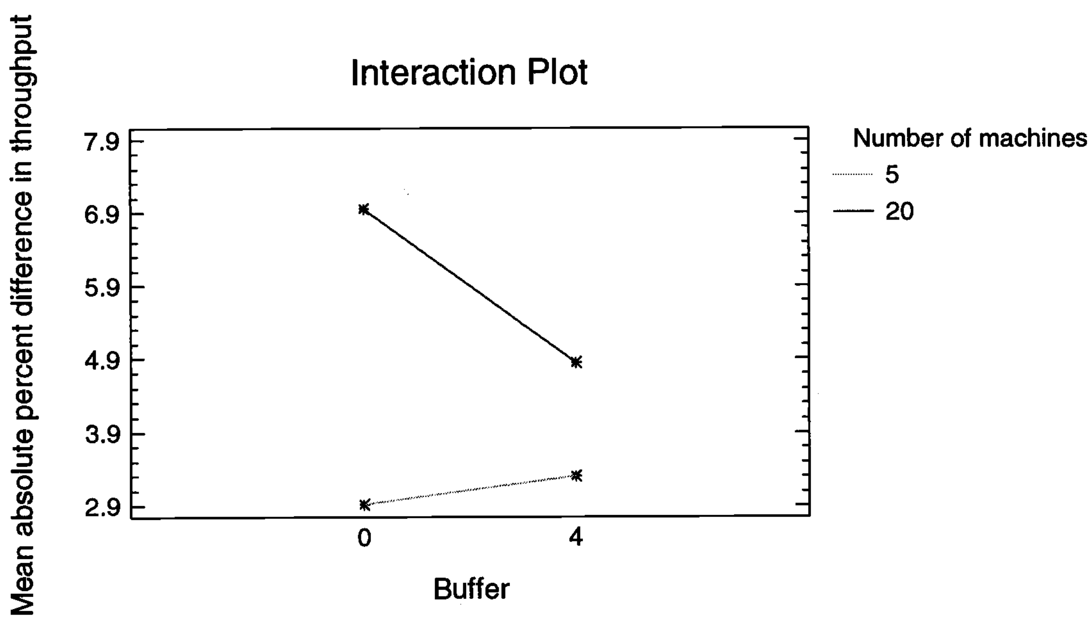

Figure 13: Interaction plot between number of buffer spaces and number of machines for automated systems.

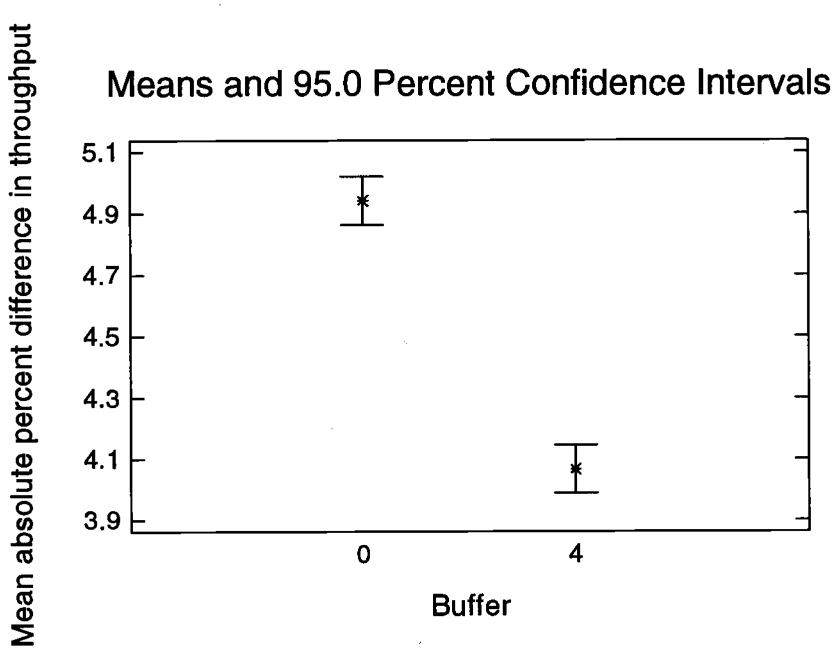

Figure 14: Confidence intervals for mean absolute percent difference in throughput at two levels of buffer between any two machines for automated systems.

7. Summary and Conclusions

This study presents the results of using default distributions in place of best-fit distributions in simulations of production lines with finite buffers. Best-fit distributions that were used to perform simulations had varying degrees of skewness from a zero for a normal distribution to about 9 for a gamma distribution. The default distributions used were made to match the first two moments and where required, the first three moments of the best-fit distributions. The lognormal distribution was selected as the default distribution to match the first two moments owing to its clearly defined moment equations and its ability to exhibit flexibility in shapes while generating positive random variates. The distributions used to match the first three moments of the best-fit distributions were the two-level hyper-exponential distribution and the three-parameter lognormal.

Simulations were carried out for such systems with different number of machines, buffers, carriers and best-fit distributions. The performance measure for these simulations was the throughput. The results for the above simulations were expressed in terms of mean absolute percent difference in throughput between the systems using best-fit distributions and the same systems using default lognormal distribution and three parameter distributions (for matching the first three moments). The results obtained indicate that the error is practically low (fewer than 5%) for the most part. The error tends to increase as the CV and the difference in skewness between the best-fit distribution and lognormal distribution is increased. The results also seem to indicate that it is the first three-moments (as opposed to the first two) which drive the throughput of the asynchronous production lines with finite buffers.

After conducting an ANOVA, we identified scenarios which led to high values of mean absolute percent difference in throughput when matching only the first two moments of the best-fit distribution. High level of carrier concentration within a system is the primary factor in increasing the error. Also lower number of buffer between any two machines was another factor which led to high values of mean absolute percent error.

On the basis of the results obtained by carrying out the simulation experiments, we propose a step by step procedure for data modeling to carry out simulation analysis of the systems described in this study. The procedure is described in Figure 15.

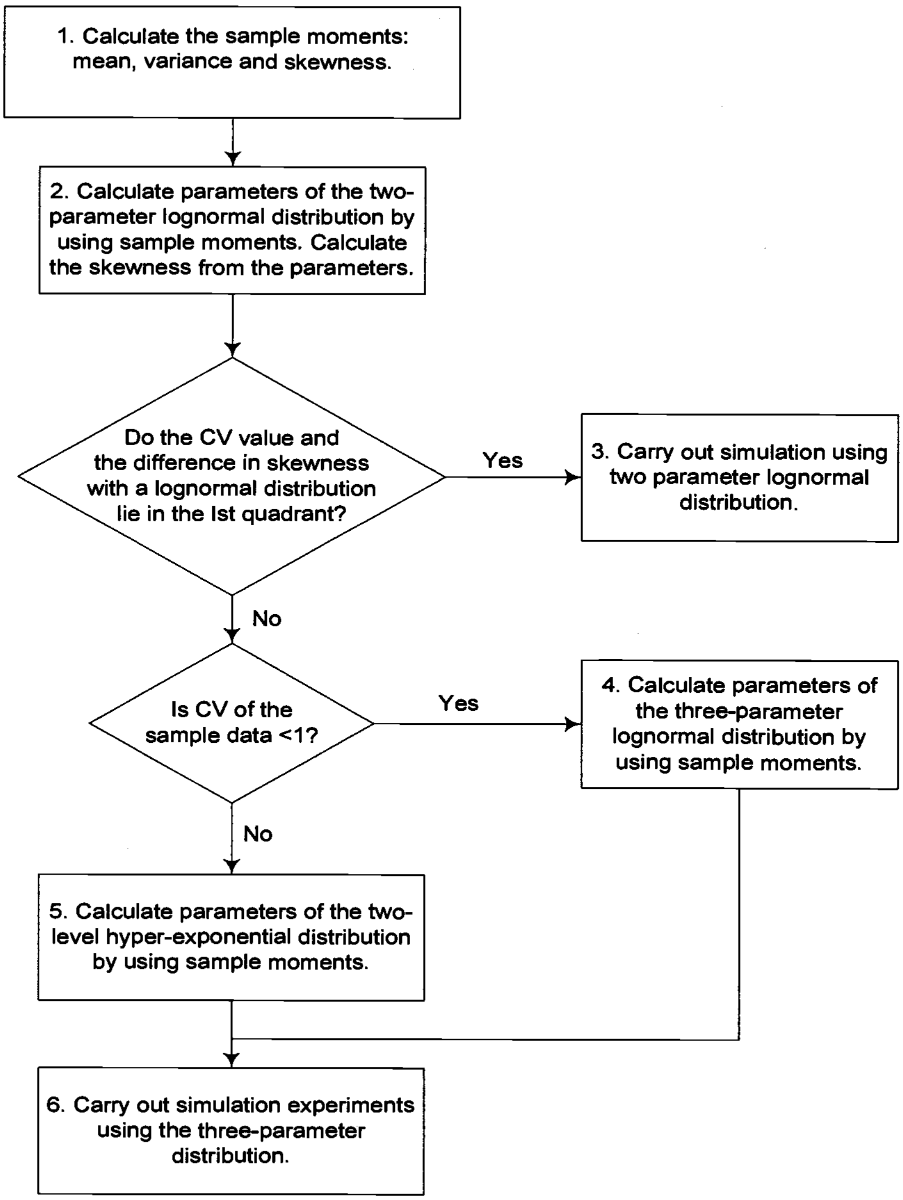

Figure 15: Proposed procedure for data modeling in simulation studies.

The implications of this study enable us to identify situations wherein it is safe to match two moments of the input data and at other times when it is not. Since the mean absolute error is mostly within 5% for most of the distributions, we can state that for all engineering purposes if the CV of the input data is within 1.33, and the difference in skewness with a lognormal distribution is less than or equal to five it is safe to use lognormal distribution to adequately represent process time, time between failures or repair time distributions for a production line.

In cases where the mean absolute percent difference in throughput is high after matching just the first two moments of the input data, the error can be lowered by matching the first three moments and using either the hyper-exponential distribution or the three parameter lognormal distribution as default distributions.

The findings of this study would be helpful in ascertaining the level of possible error induced in simulation of conceptual systems wherein little or no data is available. In such cases it is up to the experts to provide the simulation analyst with their best estimates for events. These estimates can be used to find the parameters of a default distribution which can then be used in simulation exercise and the results can be reported by duly taking into account the level of possible errors found out from this study. In a similar manner, this study can also help in the field of input data modeling for production lines with finite buffers where the input data is scarce.

Bibliography

1. Baker, K.R., 1992, "Tightly-Coupled Production Systems: Models, Analysis, and Insights", Journal of Manufacturing Systems, 1992, Vol. 11, No. 6, 385-400.

2. Blumenfeld, D.E., 1990, "A Simple formula for estimating throughput of serial production lines with variable processing times and limited buffer capacity", International Journal of Production Research, 1990, Vol. 28, No. 6, 1163-1182.

3. Crow, E.L. and Shimzu, K., 1988, Lognormal Distributions: Theory and Applications, Markel Dekker, Inc., New York, 1988.

4. De Brota, D. J., Roberts, S.D., Swain, J.J. and Venkataraman, S., 1989, "Modeling Input Processes with Johnson Distributions", Proceedings of the 1989 Winter Simulation Conference, Washington, DC, 308-316.

5. Frein, Y., Commault, C., Dallery, Y., 1996, "Modeling and analysis of closed-loop production lines with unreliable machines and finite buffers", IEEE Transactions, 1996, Vol. 28, 545-554.

6. Gross, D., 1999, "Sensitivity of Output Performance Measures to Input Distributions Shape in Modeling Queues – 3: Real Data Scenario", Proceedings of the 1999 Winter Simulation Conference, Phoenix, AZ, 452-457.

7. Gross, D. and Juttijudata, M., 1997, "Sensitivity of Output Performance Measures to Input Distributions in Queuing Simulation Modeling", Proceedings of the 1997 Winter Simulation Conference, Atlanta, Georgia, 296-299.

8. Gross, D. and Masi, D.M.B., 1998, "Sensitivity of Performance Measures to Input Distributions in Queuing Network Modeling", Proceedings of the 1998 Winter Simulation Conference, Washington, DC, 629-635.

9. Harchol-Balter, M., Li, C., Osogami, T., Scheller-Wolf, A., 2003, "Analysis of Task Assignment with Cycle Stealing under Central Queue", In Proceedings of the 23rd International Conference on Distributed Computing Systems (ICDCS'03), Providence, RI, May, 2003, 628-637.

10. Johnson, N. L., 1949, "Systems of frequency curves generated by methods of translation", Biometrika, 1949, Vol. 36, 149-176.

11. Johnson, M.A., Taaffe, M.F., 1991, "A graphical investigation of error bounds for moment-based queueing approximations", Queueing Systems, 1991, Vol. 8, 295-312.

12. Kim, D. S., Alden, J. M., 1997, "Estimating the distribution and variance of time to produce a fixed lot size given deterministic processing times and random downtimes", International Journal of Production Research, 1997, Vol. 35, No. 12, 3405-3414.

13. Lau, H., 1986, "A Directly-Coupled Two-Stage Unpaced Line", IIE Transactions, September, 1986, 304-312.

14. Liu, X., Li, Z., Buzacott, J.A., 1992, " A decomposition method for throughput analysis of cyclic queues with production blocking", In Proceedings of the Second International Workshop on Queuing Networks with Finite Capacity, RTP, NC, May 28-29, 1992, 445-475.

15. Muth, E.J., 1973, "The Production Rate of a Series of Work Stations with Variable Service Times", International Journal of Production Research, 1973, Vol. 11, 155-169.

16. Ramberg, S. and Schmeiser, B.W., 1972, "An Approximate Method for Generating Symmetric Random Variables", Communications of the ACM, November 1972, Vol. 15, No. 11, 987-990.

17. Ramberg, S. and Schmeiser, B.W., 1974, "An Approximate Method for Generating Asymmetric Random Variables", Communications of the ACM, February 1974, Vol. 17, No. 2, 78-82.

18. Rao, N.P., 1975, "Two State Production Systems with Intermediate Storage", AIIE Transactions, Vol.7, 1975b, 414-421.

19. Shanker, A. and Kelton, D.W., 1991, "Empirical input Distributions: An Alternative to Standard Input Distributions in Simulation Modeling", Proceedings of the 1991 Winter Simulation Conference, Phoenix, 978-985.

20. Whitt, W., 1982, "Approximating a Point Process by a Renewal Process, I: Two Basic Methods", Operations Research, 1982, Vol. 30, 125-147.

Appendix

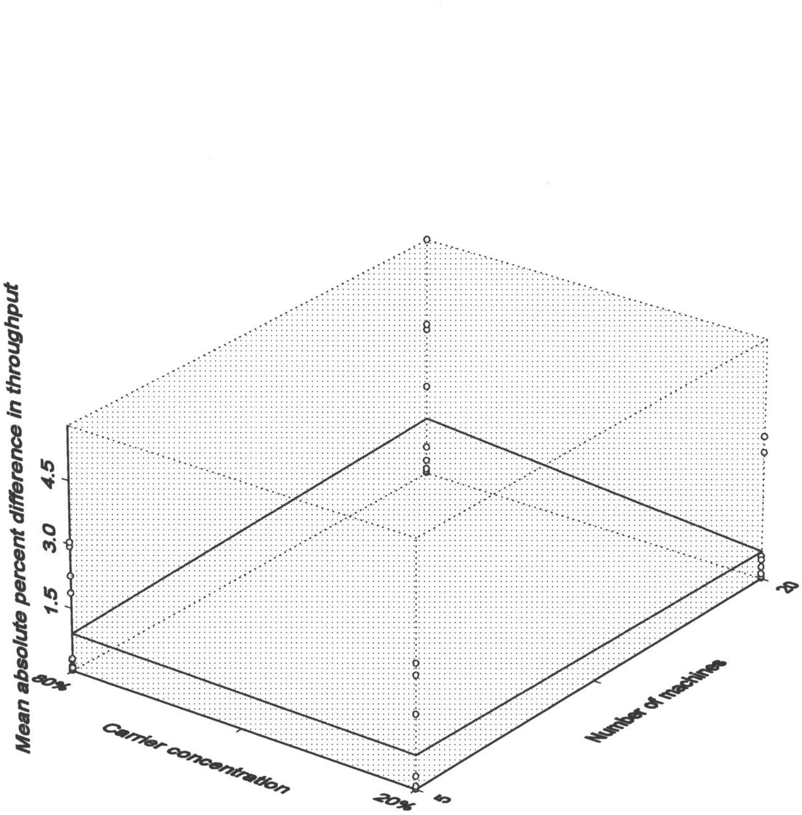

Figure A1: 3-D scatter plot for mean absolute percent difference in throughput over different levels of machines and carrier concentration for the 1st quadrant of manual using lognormal as the default distribution.

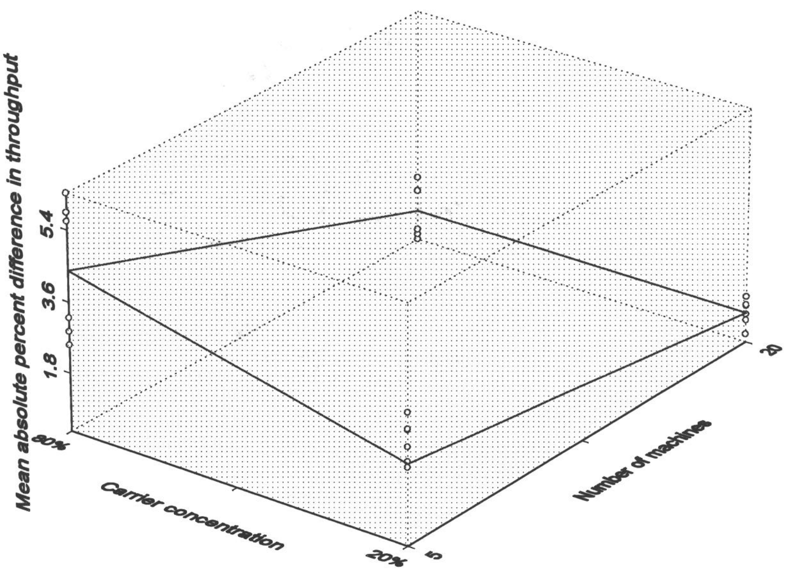

Figure A2: 3-D scatter plot for mean absolute percent difference in throughput over different levels of machines and carrier concentration for the 2nd quadrant of manual systems using lognormal as the default distribution.

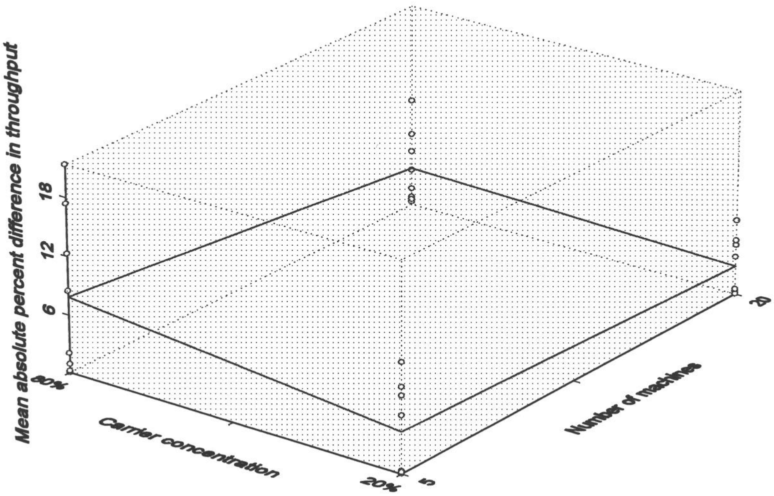

Figure A3: 3-D scatter plot for mean absolute percent difference in throughput over different levels of machines and carrier concentration for the 3rd quadrant of manual systems using lognormal as the default distribution.

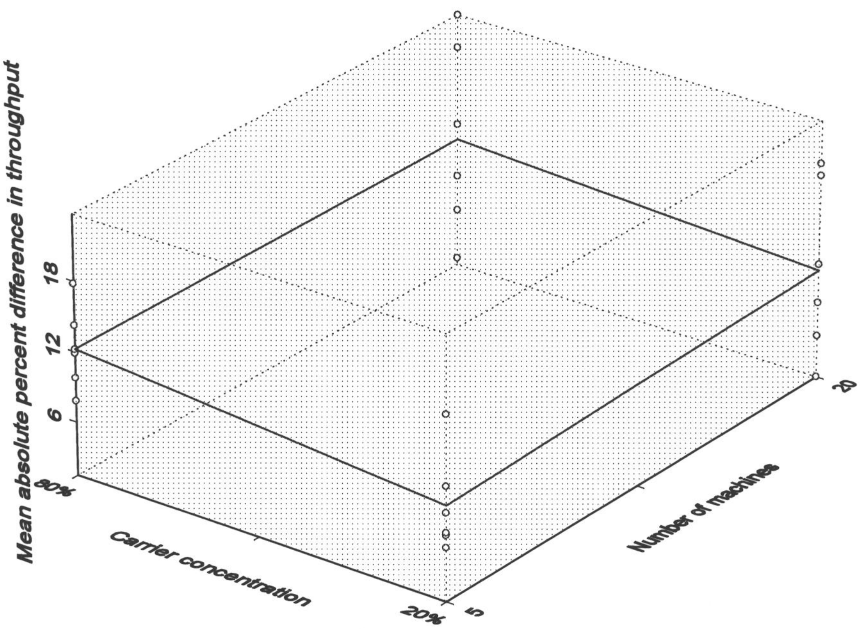

Figure A4: 3-D scatter plot for mean absolute percent difference in throughput over different levels of machines and carrier concentration for the 4th quadrant of manual systems using lognormal as the default distribution.

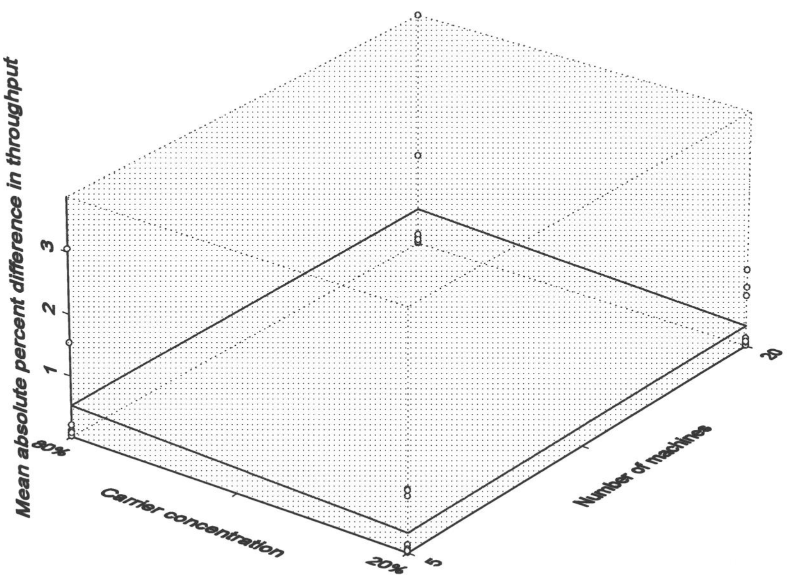

Figure A5: 3-D scatter plot for mean absolute percent difference in throughput over different levels of machines and carrier concentration for the 1st quadrant of automated systems using lognormal as the default distribution.

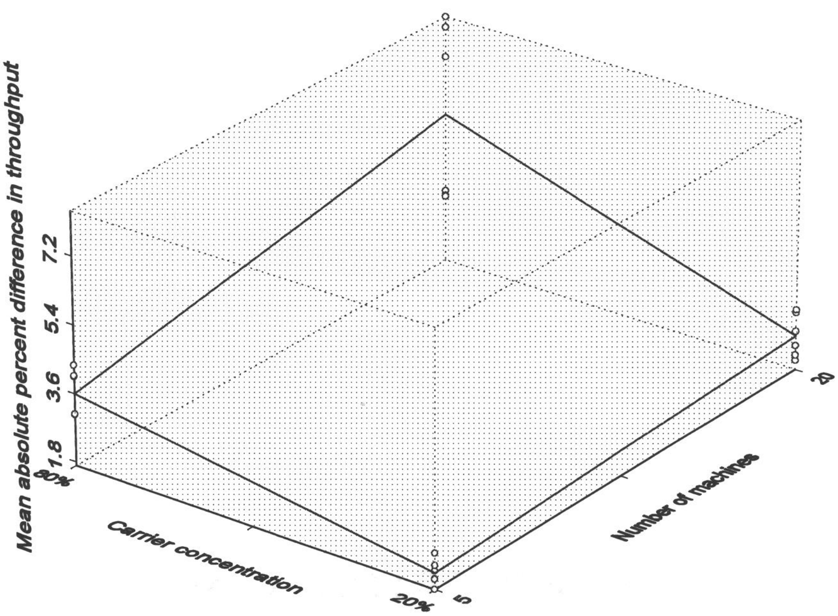

Figure A6: 3-D scatter plot for mean absolute percent difference in throughput over different levels of machines and carrier concentration for the 2nd quadrant of automated systems using lognormal as the default distribution.

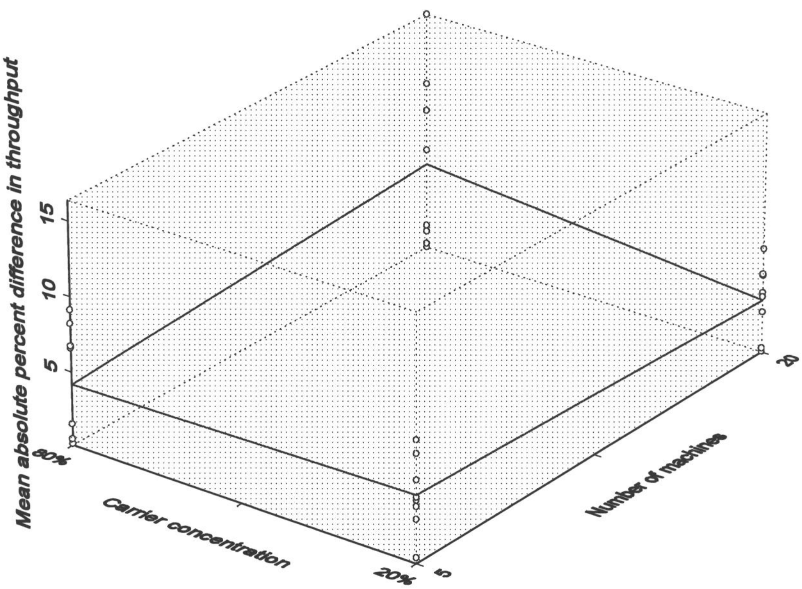

Figure A7: 3-D scatter plot for mean absolute percent difference in throughput over different levels of machines and carrier concentration for the 3rd quadrant of automated systems using lognormal as the default distribution.

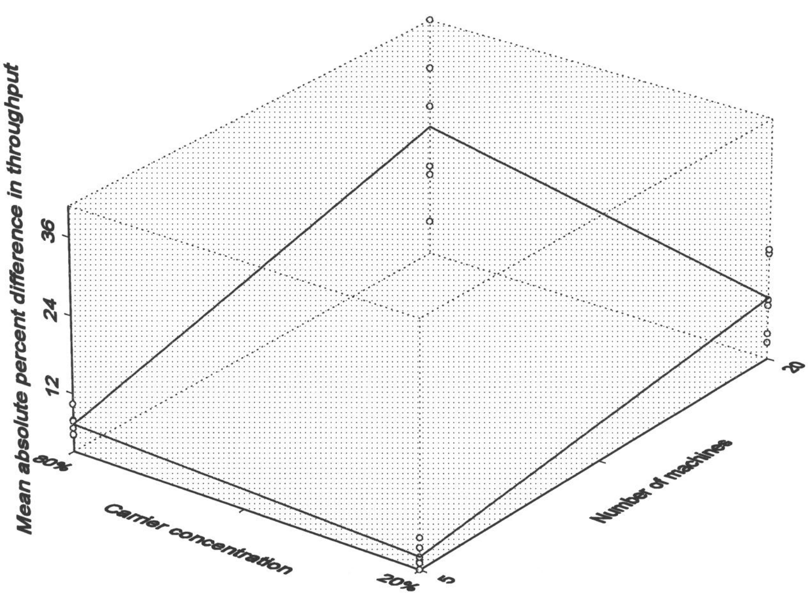

Figure A8: 3-D scatter plot for mean absolute percent difference in throughput over different levels of machines and carrier concentration for the 4th quadrant of automated systems using lognormal as the default distribution.

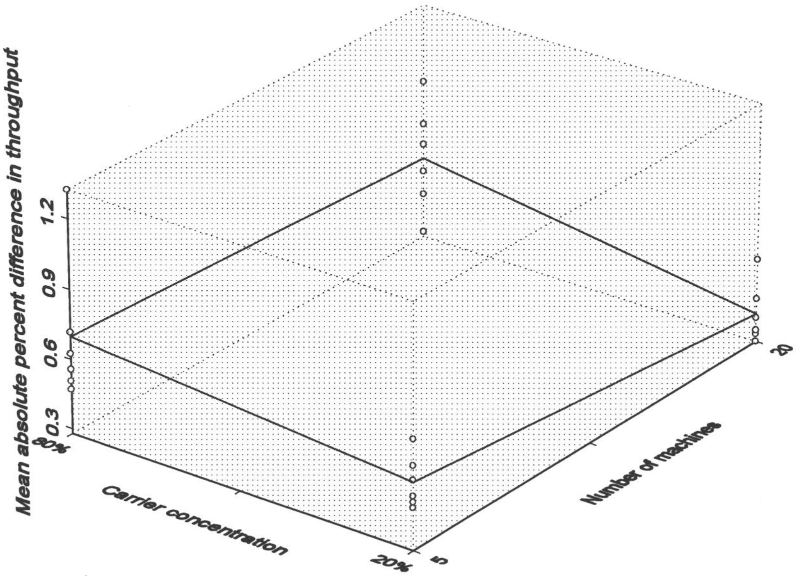

Figure A9: 3-D scatter plot for mean absolute percent difference in throughput over different levels of machines and carrier concentration for the 2nd quadrant of manual systems by matching the first three moments of the true distribution.

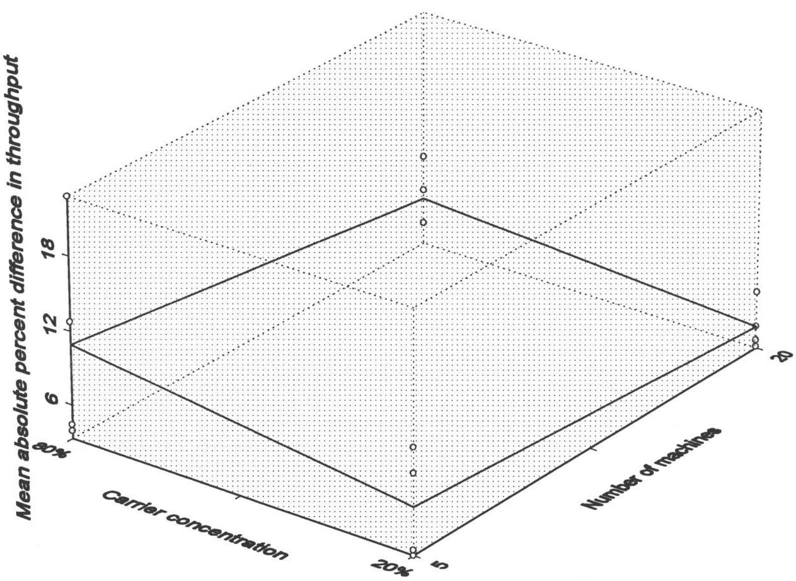

Figure A10: 3-D scatter plot for mean absolute percent difference in throughput over different levels of machines and carrier concentration for the 3rd quadrant of manual systems by matching the first three moments of the true distribution.

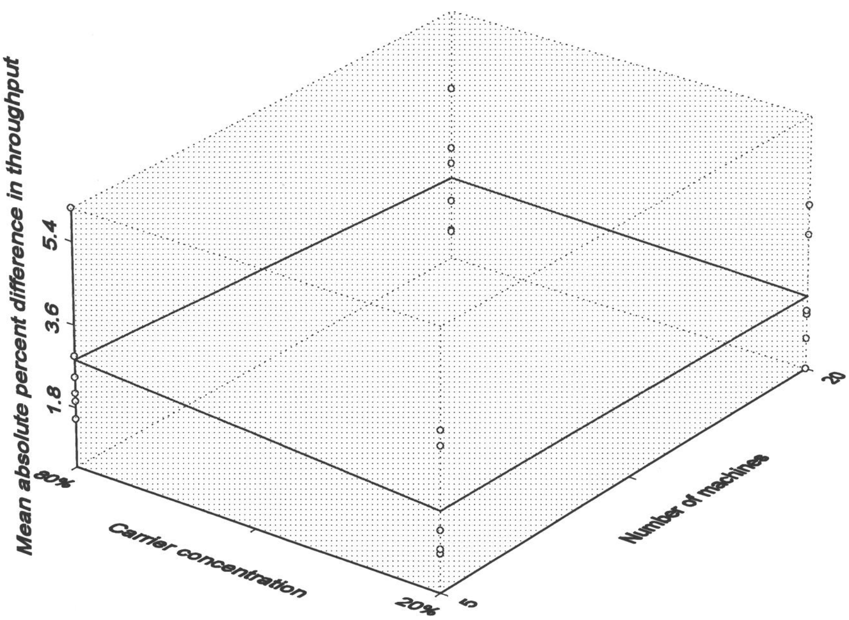

Figure A11: 3-D scatter plot for mean absolute percent difference in throughput over different levels of machines and carrier concentration for the 4th quadrant of manual systems by matching the first three moments of the true distribution.

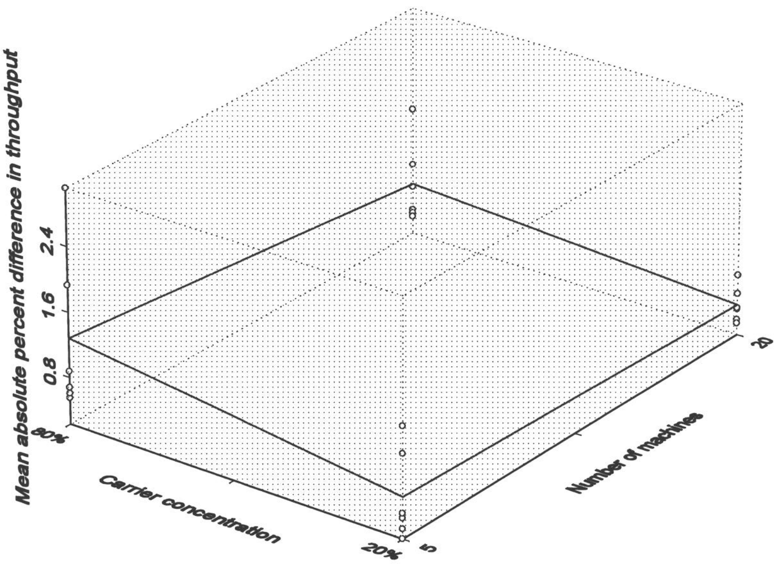

Figure A12: 3-D scatter plot for mean absolute percent difference in throughput over different levels of machines and carrier concentration for the 2^{nd} quadrant of automated systems by matching the first three moments of the true distribution.

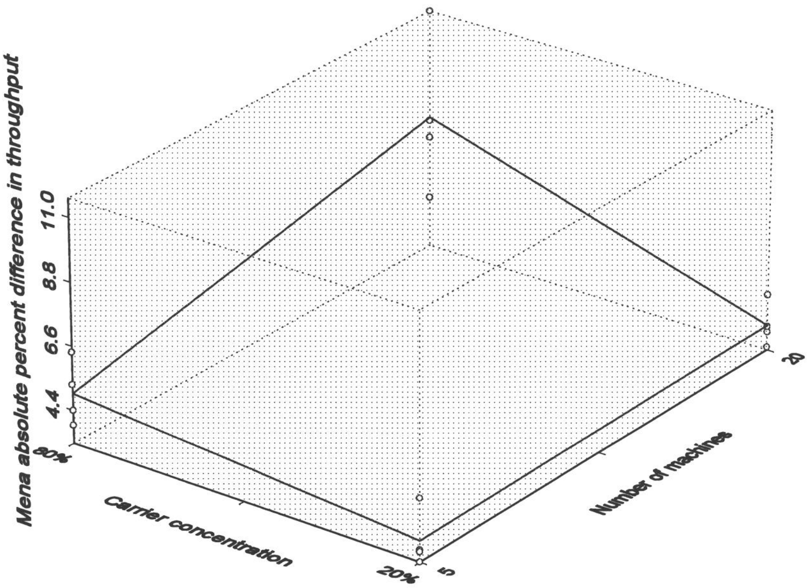

Figure A13: 3-D scatter plot for mean absolute percent difference in throughput over different levels of machines and carrier concentration for the 3rd quadrant of automated systems by matching the first three moments of the true distribution.

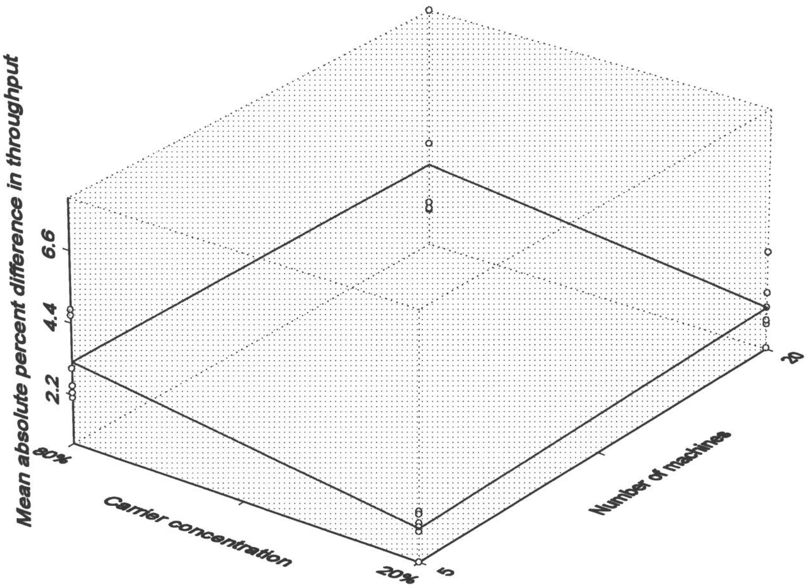

Figure A14: 3-D scatter plot for mean absolute percent difference in throughput over different levels of machines and carrier concentration for the 4th quadrant of automated systems by matching the first three moments of the true distribution.