On the Accuracy of the Simple Peak Hour Approximation for Markovian Queues

Linda V. Green • Peter J. Kolesar
Graduate School of Business, Columbia University, New York, New York 10027

We empirically explore the accuracy of the simple stationary peak hour approximation (SPHA) for estimating peak hour performance in multiserver queuing systems with exponential service times and periodic (sinusoidal) Poisson arrival processes. We show that the SPHA is very good for a range of parameter values corresponding to a reasonably broad spectrum of real systems. However, we do find and document that there are many situations in which this approximation will be very inaccurate.

We postulate and then support empirically a set of hypotheses that link the accuracy of the SPHA and the related point-wise stationary approximation (PSA) to key parameter values and model characteristics. We also present results on the time-dependent behavior of these systems as a function of key parameters.

Finally we present results which indicate that our findings, developed for models with sinusoidal input streams, may apply to a much broader range of Markovian models with more general cyclic inputs.

(*Queues; Nonstationarity; Approximations*)

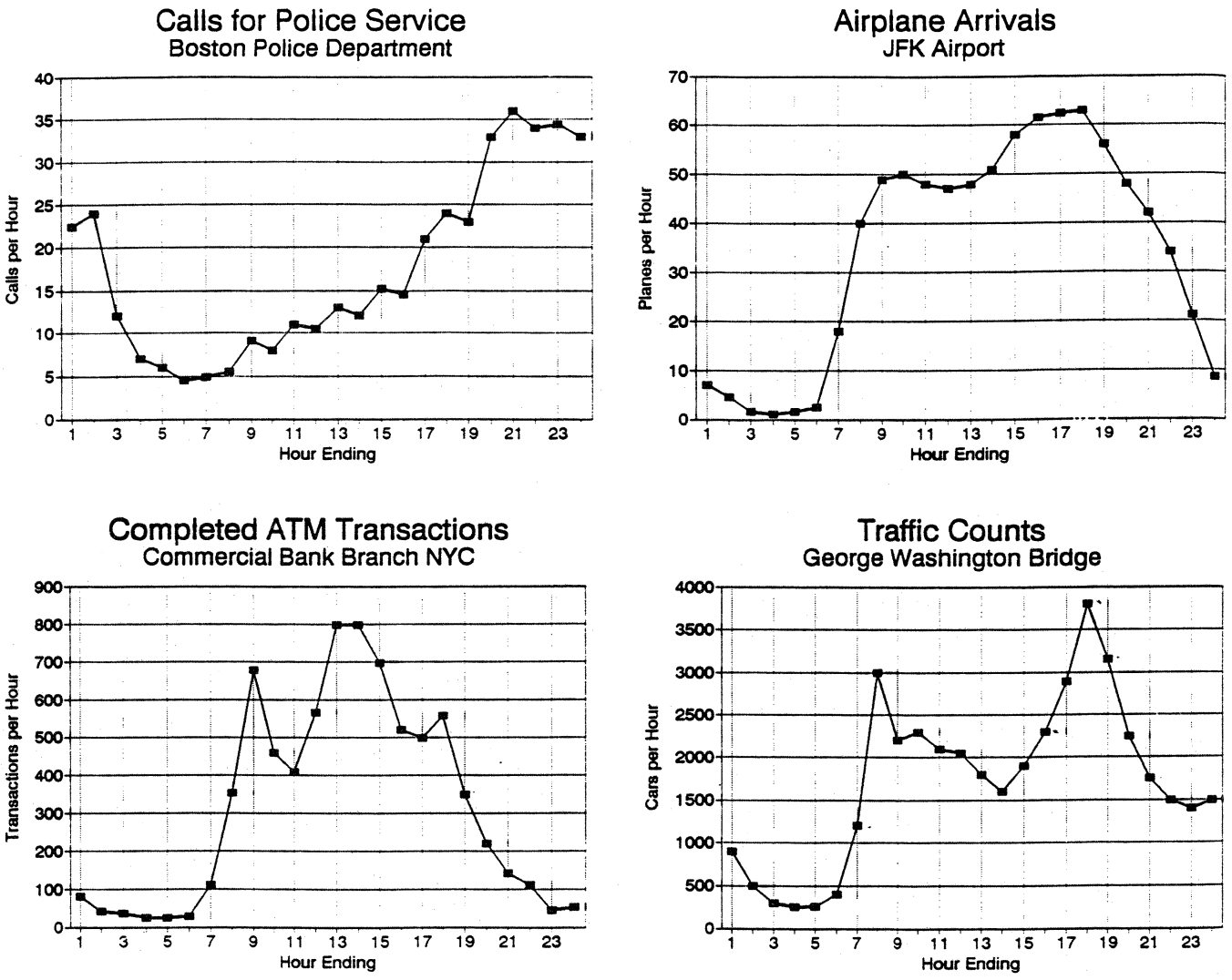
1. Introduction

From the earliest developments of mathematical queuing models for the design and management of stochastic service systems, it was recognized that many real world systems have nonstationary (often cyclic) customer demand streams. Figure 1 illustrates this phenomenon with data from four sources: calls into the 911 emergency phone system in Boston (Larson 1972); customer arrivals at a bank of automatic teller machines (Kolesar 1984); landings of aircraft at La Guardia Airport in New York (Koopman 1972); and traffic at toll booths on the George Washington Bridge (Edie 1954). In each case the time dependence is clearly a dominant feature of the environment and must be considered in making design and operating decisions. In the types of situations illustrated in Figure 1, managing the quality of the system performance during periods of peak congestion is often a primary concern.

Word-of-mouth within the queuing community has it that a standard practice in such situations is for the analyst to determine the average arrival rate during the "peak" or "rush" hour (assuming a 24-hour cycle), plug that value into a stationary queuing model of the system and then use the resulting estimates of queue sizes, delays and the like as if they predict average performance during the peak hour. We call this procedure the *simple peak hour approximation*, or SPHA for brevity. This practice is apparently quite common. In particular, it is the traditional method in telecommunications applications where busy hour rates are high and the system tends to approach steady-state in minutes; e.g., see Bear (1980). However, the SPHA is rarely discussed in the queuing literature or in standard queuing texts. When it is mentioned, the discussion is typically terse and no guidance is offered on the consequences of this type of approximation. The purpose of this paper is to determine the conditions under which the SPHA is a good or bad approximation and make recommendations concerning its use. We will show that there are some clear conditions under which the SPHA is quite good indeed and, conversely, some clear conditions under which it is very risky to use. Not surprisingly, some of these are

0025-1909/95/4108/1353$01.25
Copyright © 1995, Institute for Operations Research and the Management Sciences

Figure 1



related to the issue of the rate of approach to steady-state.

Though most of the discussion and computations in this paper focus on the peak hour as described above, the choice of a one-hour interval is not as limiting or particular as it may at first appear. Our actual interest is in the more general issue of peak *period* performance where the duration of the period need not be arbitrarily taken to be an hour, but rather should be defined in the context of the managerial or design issues at hand. In many practical contexts this apparently is an hour—perhaps by default, habit or convention—but there are examples in the literature of other choices. For example, 15- and 30-minute time segments have been the relevant time period in studies by Segal (1974) and Holloran (1986), respectively. An extreme view would be to focus on the *peak epoch,* that instant at which congestion or delay is maximum. In this paper we study peak epochs as well as peak hours and find that our results for both are so similar that our major conclusions appear to apply

to any peak period of reasonably short duration—from about an hour down to minutes or less. Though we adopt the language of a 24-hour cycle and the peak hour as a matter of convenience, our concern is more generally with estimating performance during relatively short intervals of highest congestion in systems having periodic arrival rates.

As the models we wish to study are not solvable analytically in closed form, we have based our investigations on a numerical solution approach. Moreover, as it is not possible to take such an approach for a completely general class of models, we have limited our computations to Markovian queuing models with sinusoidal Poisson input streams. These models are simple to parameterize, and we would argue that sinusoidal inputs capture the essence of the cyclic phenomena with which we are concerned. (In a large number of actual cases the arrival process has a single dominant peak and a single valley.) In the final section of the paper we will, however, display some results for nonsinusoidal cyclic inputs. Our solution method is to numerically compute the steady-state solution to the system differential equations describing the model. The output of this procedure is a time-dependent vector of state probabilities from which we then calculate particular performance measures of interest.

In this paper we will focus attention on three performance measures—the expected number of customers waiting in queue, the expected customer waiting time prior to service, and the probability that a customer is delayed and has to wait for service. Since we are dealing with models in which time-dependent results are the central issue, we will look at the above three measures from three vantage points:

(1) Long-run average performance: the average of the measures over the entire (24-hour) cycle.

(2) Peak hour performance: the average of the measures over that hour that gives the maximum such average.

(3) Peak epoch performance: the maximum instantaneous value of the measures over the cycle.

The work presented in this paper is part of an ongoing study on the behavior of nonstationary queuing systems and the use of stationary models to estimate their performance (see Green et al. 1991 and Green and Kolesar 1991). This problem of cyclic arrival processes is alluded to elliptically in the early works of Erlang circa 1909 (see Brockmeyer 1948). Reviews of related literature are given in Massey and Whitt (1992), Eick et al. (1993a and 1993b), Green et al. (1991), and Green and Kolesar (1991). A few other papers are directly relevant to our work. Koopman's (1972) study of air traffic control was the first to recommend and use numerical solutions of the differential equations of a nonstationary Markovian queue to achieve insights about system behavior. His methods are explained fully in Chapter 8 of Giffin (1978) in which sensitivity to exponential service times is explored. The idea of integrating the steady-state solution at each epoch of time—what we call the *pointwise stationary approximation* or PSA—to provide an upper bound on queue length can be found in Grassman (1983) and Rolski (1986 and 1987) proves its upper bounding property for the single server case and explores some other related issues. Whitt (1991) proves that a pointwise version as well as the average version of the PSA is asymptotically correct as the arrival and service rates increases. He also introduces an *average stationary approximation* which is related to both the PSA and the SPHA.

There are only a few references dealing with peak period performance. The books of Lee (1968) and Bear (1980) and the notes of Farber (1979) contain some discussion of SPHA-like approaches but without much concrete advice to the would-be user. Newell (1968, 1982) studies a rush-hour phenomenon in a single server system in which the total load on the system approaches complete congestion (traffic intensity = 1) from below using diffusion approximation methods. This work is also described in Kleinrock (1976) and is extended by Massey (1985). More recently, Eick et al. (1993a and 1993b) explore peak behavior for infinite server systems.

In the following section we give a more detailed description and definition of the model studied, of the solution method employed and of the various performance measures we use. In §3 we develop some underlying concepts, and we also illustrate graphically certain key interrelationships and characteristics of the time-dependent and peak period performance of these models. In §4 we propose and then confirm numerically a set of hypotheses on the accuracy of the SPHA and the forces that drive it to be good or bad. In §5 we

provide a practical answer to the question, "When is the simple peak hour approximation good enough?" We close with some final observations in §6, and with some examples of the behavior of systems with cyclical nonsinusoidal input streams.

2. Model, Methodology and Definitions

Our analysis is based on $M(t)/M/s$ systems with $\lambda(t)$, the arrival rate at time t given by

$$\lambda(t) = \bar{\lambda} + A\cos(2\pi t/T), \tag{1}$$

where $\bar{\lambda}$ is the average arrival rate over the period T and A (>0) is the amplitude; μ the service rate and s the number of servers. We assume that $\bar{\lambda} < s\mu$, and so the system will develop a periodic steady-state behavior (see Heyman and Whitt 1984 and Koopman 1972).

Let $p_n(t)$ be the periodic steady-state probability that n customers are in the system at time t. These functions are the foundation of our results and are obtained by numerically solving the following standard set of differential equations that describe the system, see Gross and Harris (1985):

$$\begin{aligned}
p_0'(t) &= -\lambda(t)p_0(t) + \mu p_1(t),\\
p_n'(t) &= \lambda(t)p_{n-1}(t) + (n+1)\mu p_{n+1}(t)\\
&\quad - (\lambda(t) + n\mu)p_n(t), \quad 1 \leqq n < s,\\
p_n'(t) &= \lambda(t)p_{n-1}(t) + s\mu p_{n+1}(t)\\
&\quad - (\lambda(t) + s\mu)p_n(t), \quad n \geqq s.
\end{aligned} \tag{2}$$

Let $L_q(t)$, $W_q(t)$ and $p_b(t)$ be the instantaneous expected queue length at epoch t, the instantaneous expected virtual delay at epoch t and the instantaneous probability of all servers busy at epoch t, respectively. Note that $p_b(t)$ is also the delay that would be experienced were a customer to arrive at t. Specifically,

$$L_q(t) = \sum_{n=s}^{\infty} (n-s)p_n(t). \tag{3}$$

$$W_q(t) = \sum_{n=s}^{\infty} (n-s+1)p_n(t)/s\mu, \tag{4}$$

and

$$p_b(t) = 1 - \sum_{n=0}^{s-1} p_n(t). \tag{5}$$

For $T = 24$ hours, the peak epoch values of these measures are:

$$L_q(\text{peak}) = \max_{0\leq t\leq 24} L_q(t),$$

$$W_q(\text{peak}) = \max_{0\leq t\leq 24} W_q(t),$$

and

$$p_b(\text{peak}) = \max_{0\leq t\leq 24} p_b(t). \tag{6}$$

We also define the peak hour average measures as:

$$L_q(\text{peak hr.}) = \max_{0\leq t\leq 24} \int_t^{t+1} L_q(t)dt,$$

$$W_q(\text{peak hr.}) = \max_{0\leq t\leq 24} \int_t^{t+1} \lambda(t)W_q(t)dt \Big/ \int_t^{t+1} \lambda(t)dt,$$

and

$$p_D(\text{peak hr.}) = \max_{0\leq t\leq 24} \int_t^{t+1} \lambda(t)p_b(t)dt \Big/ \int_t^{t+1} \lambda(t)dt. \tag{7}$$

Note that p_D(peak hr.) is the average probability of *delay* during the peak hour. We will also consider the times at which these peak values are attained. Clearly, the time of the peak epoch for any measure is simply the instant at which it occurs. We define the time of a peak hour performance measure as the midpoint of the hour for which the maximum defined in equation (7) is achieved.

The numerical integration of the equations in (2) is performed using the International Math-Science Library subroutine DVERK which recursively uses fifth- and sixth-order Runge-Kutta methods. The length of the recursion interval is determined internally so that a user specified global error is not exceeded. Our integrations were usually initialized using the steady-state $M/M/s$ solution obtained with $\lambda(0)$ as the stationary arrival rate.

In order to obtain an accurate solution for the infinite capacity system, we truncated the number of equations by following a method suggested in Odoni and Roth (1983). A maximum dimension of N is used by DVERK as the number of equations to be solved on the first call. After each subsequent call of DVERK (having solved $N' \leq N$ equations) the probability $p_{N'}(t)$ of a

saturated system is compared to a specified small number, ϵ (e.g., $\epsilon = 10^{-8}$). If $p_{N'}(t)$ is larger than ϵ, the number of equations solved at the next call is increased by m (e.g., $m = 5$). Conversely, if both $p_{N'}(t)$ and $p_{N'-m}(t)$ are less than ϵ, N' is reduced by m for the next call.

Recalling that for ease of discussion we are assuming that the period length is 24 hours, DVERK is called every five minutes on the simulated clock and the solution vector of state probabilities produced is used to calculate the probability of all servers busy and the expected number in queue. Thus, each cycle is divided into 288 five-minute segments, and our performance measures (3), (4), and (5) are obtained at the end of each segment. Thus, in computing the measures defined in equations (7) above, the integrals are replaced by the equivalent summations over the appropriate grid points.

In this paper we will examine two approaches for estimating the peak epoch and peak hour performance measures. We first define $L_q^*(\lambda)$, $W_q^*(\lambda)$, and $p_D^*(\lambda)$ to be the expected queue length, expected delay, and probability of delay in a stationary $M/M/s$ system with arrival rate λ and given μ and s (see Gross and Harris 1985, p. 84–92). For example,

$$L_q^*(\lambda) = \frac{(\lambda/\mu)^s \lambda\mu}{(s-1)!(s\mu-\lambda)^2} p_0^*,$$

where

$$p_0^* = \left[\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{s(\lambda/\mu)^s}{s!(s-\lambda/\mu)}\right]^{-1}.$$

Then the *stationary approximation* over any interval $I = (a, b)$ is given by $L_q^*(\lambda_{a,b})$, $W_q^*(\lambda_{a,b})$ and $p_D^*(\lambda_{a,b})$ for the expected queue length, expected delay, and probability of delay, respectively where $\lambda_{a,b}$, the average arrival rate over the interval a to b, is given by

$$\lambda_{a,b} = \frac{1}{(b-a)} \int_a^b \lambda(t)dt.$$

We denote the *pointwise stationary approximations* (PSAs) for L_q, W_q and p_D in the interval (a, b) by $L_q^{\gamma}(a, b)$, $W_q^{\gamma}(a, b)$, $p_D^{\gamma}(a, b)$ and define them:

$$L_q^{\gamma}(a, b) = \frac{1}{b-a} \int_a^b L_q^*(\lambda(t))dt, \quad (6)$$

$$W_q^{\gamma}(a, b) = \frac{1}{\lambda_{a,b}(b-a)} \int_a^b \lambda(t) W_q^*(\lambda(t))dt, \quad (7)$$

and

$$p_D^{\gamma}(a, b) = \frac{1}{\lambda_{a,b}(b-a)} \int_a^b \lambda(t) p_d^*(\lambda(t))dt. \quad (8)$$

For the peak epoch, the stationary approximation and PSA are identical and equal to the stationary measure with λ equalling the maximum instantaneous arrival rate. For the SPHA and peak hour PSA, the interval (a, b) is the hour surrounding the epoch at which this maximum occurs.

Our earlier papers, Green and Kolesar (1991) and Green et al. (1991), study both these approximations for estimating the long-run daily average performance in queues with cyclic arrivals. In Green and Kolesar (1991) we showed that the PSA provides good estimates for the above measures for many Markovian service systems with sinusoidal input. In particular, the PSA performs quite well when the service rate is 2 or higher and the maximum traffic intensity is less than 0.83. When the service rate is considerably higher, e.g., $\mu = 20$, the estimates will be good at even higher maximum intensities. Indeed, Whitt (1991) proves that the PSA for the cycle averages are asymptotically correct as the rates increase. He also establishes this for a pointwise version which implies that the stationary peak epoch approximation is asymptotically correct. These results lead us to suspect that a similar approximation approach might be useful in estimating peak performance measures in nonstationary systems.

We base our conclusions on the examination of computational results for over 250 model instances. We confine our study to systems in which the *maximum traffic intensity* is strictly less than one, that is, when

$$\rho_{\max} \equiv \sup_t \frac{\lambda(t)}{s\mu} < 1. \quad (7)$$

We adopt this constraint because neither the simple stationary approximation for the peak epoch nor the PSA are generally defined when $\rho_{\max}$ (rhomax) is greater or equal to one. (The PSA for p_D is defined.) We also restricted our choice of parameter values to be such that the *relative amplitude*, RA $= A/\bar{\lambda} < 1$, so that $\lambda(t) \geq 0$ for all t. Aside from these constraints, our choices of

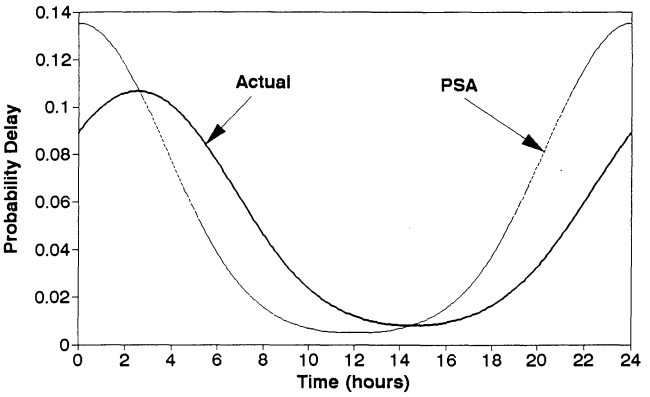
experimental models were based on three major considerations: correspondence to some actual service systems of particular interest to us, a desire to be as general as possible, and computational feasibility. More details on our experimental strategy are given in §4.

3. Initial Findings

3.1. The Timing and Magnitude of the Peak Congestion

Before looking at any approximation of peak period performance, it is important to understand the behavior of a nonstationary system with sinusoidal input. We note that the time at which congestion peaks depends upon several factors. Our first observation is that *peak congestion generally lags the peak of the arrival function* $\lambda(t)$. Figure 2 illustrates this for the probability of delay for a situation in which this lag is quite substantial—about four hours. (The peak arrival epoch in this example is at time 0.) The magnitude of such lags is primarily a function of event frequency, i.e., the average number of arrivals and service completions per period. Figures 2, 3, and 4 show the probability of delay curves for a set of cases in which the number of servers, the traffic intensity at any time t, and the relative amplitude are constant, but $\lambda(t)$ and μ (event frequency) are increasing. (In these and all subsequent figures, lambda denotes the average arrival rate.) We can see in the figures that *as the event frequency increases, the time lag between the epoch of the peak arrival rate and the peak epoch probability of delay decreases*. We found this to be

Figure 3 **Probability Delay (Lambda = 1.2; Mu = 0.4; Amplitude = 0.4; $S = 7$)**



true for all the performance measures we considered and in all the cases we studied. The other finding illustrated in these figures is that *the peak congestion level increases as the event frequency increases*. Furthermore, *the entire probability of delay curve approaches the PSA curve as event frequency increases*. This is also true for expected delay and expected queue length. These results parallel those in Eick et al. (1993a) for the number of busy servers in the infinite server case. They are also consistent with Whitt's (1991) result that the PSA is asymptotically correct as the arrival and service rates increase.

The other factor which consistently affects the lag in peak delay is the relative amplitude. This is illustrated in Figures 5, 6, and 7 for a set of cases in which all

Figure 2 **Probability Delay (Lambda = 0.6; Mu = 0.2; Amplitude = 0.2; $S = 7$)**

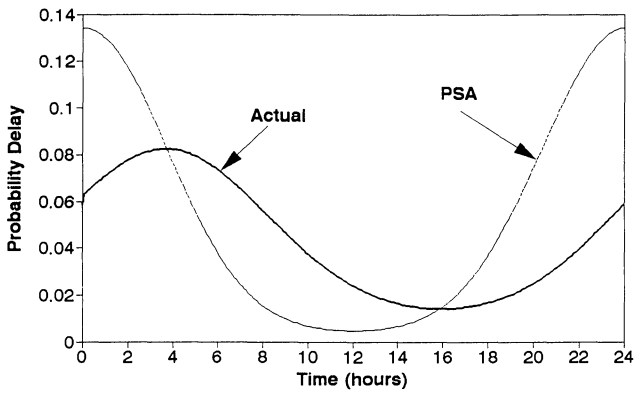


Figure 4 **Probability Delay (Lambda = 6.0; Mu = 2.0; Amplitude = 2.0; $S = 7$)**

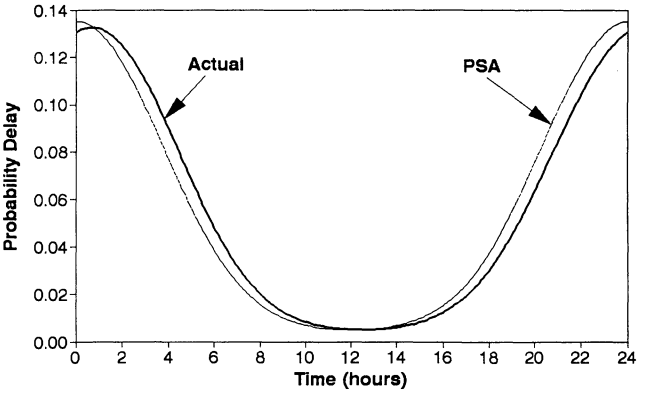


Figure 5 **Expected Delay (Lambda = 6.0; Mu = 2.5; Amplitude = 6.0; $S = 6$)**

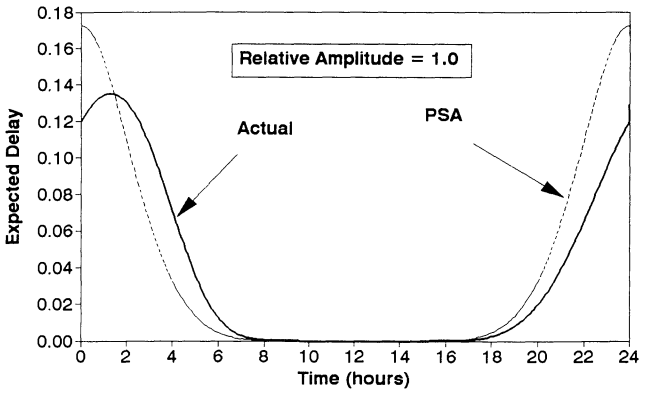
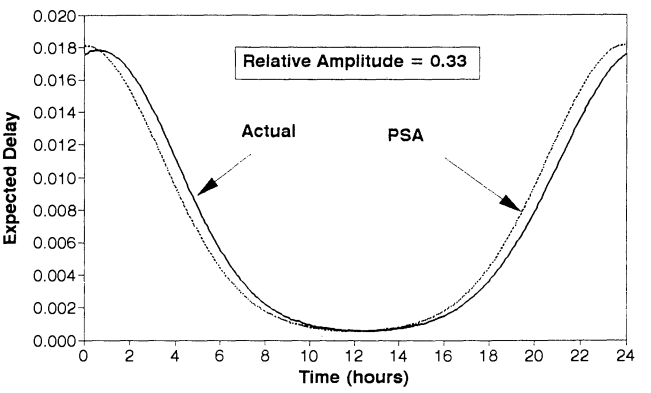


Figure 7 **Expected Delay (Lambda = 6.0; Mu = 2.5; Amplitude = 2.0; $S = 6$)**



system parameters except relative amplitude are held constant. (Note: The reader should be alert to the fact that the ordinates of these figures are, of necessity, plotted on different scales.) *As relative amplitude decreases, the peak epoch for expected delay moves closer to the epoch of the peak arrival rate*. Not surprisingly, *the entire curve approaches the PSA curve as the RA decreases*. (For RA = 0, the actual system and the PSA both reduce to the stationary system.) This phenomenon of movement of the peak towards time 0 with decreasing RA holds for expected queue length and expected delay. It does not consistently hold for probability of delay. This leads to another significant observation: *for any given system, the shape of the performance curve is different for each measure and, in particular, the time of the peak epoch may be different*. This is illustrated in Figure 8 where the peak in expected delay occurs about 15 minutes after the peak in probability of delay occurs and about 15 minutes before the peak in expected queue length.

3.2. Comparison of Approximations

Now we discuss and compare several possible approximations for peak period delays. In order to better understand the implication of focusing on approximating the peak hour, we looked at the peak epoch performance as well. In each of the cases we considered, the difference between the magnitude of the peak epoch delays and the peak hour delays was very small. We also compared the relative error of the stationary approximation to the peak epoch delay with the SPHA. The results, as shown in Figures 9 and 10 for those cases in which the

Figure 6 **Expected Delay (Lambda = 6.0; Mu = 2.5; Amplitude = 3.0; $S = 6$)**

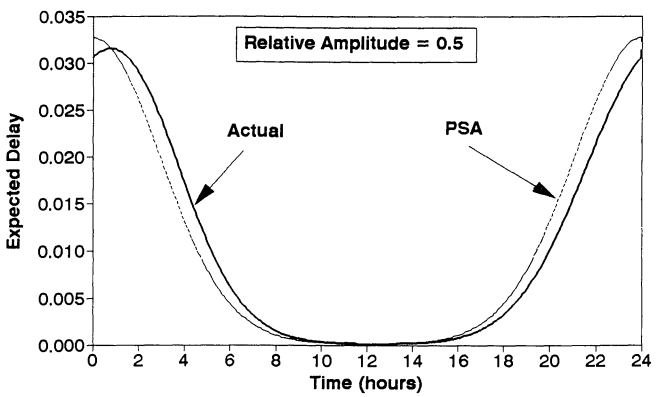


Figure 8 **Time-dependent Performance Measures (Lambda = 0.15; Mu = 0.2; Amplitude = 0.2; $S = 3$)**

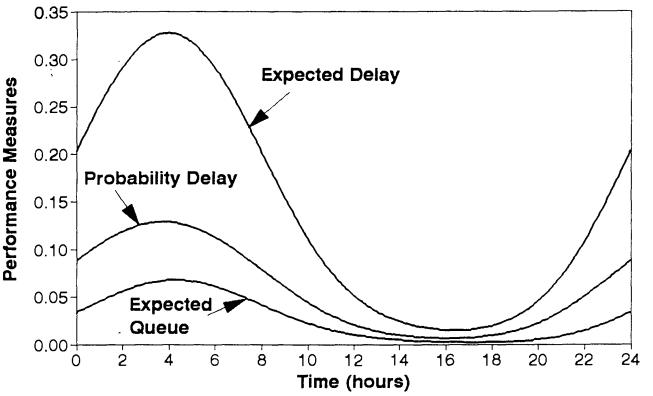


Figure 9 **Comparing Relative Errors (Peak Hour vs. Peak Point Approximations)**

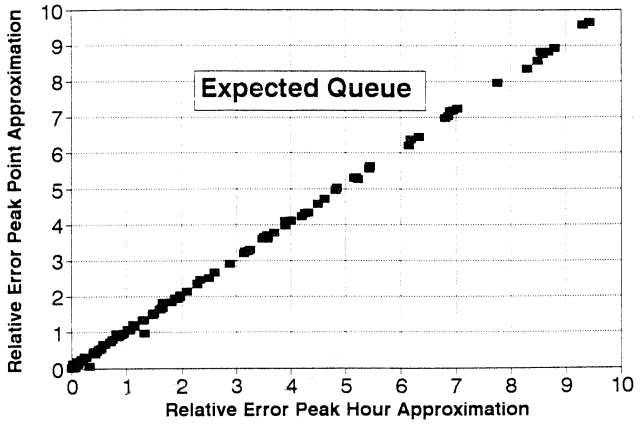
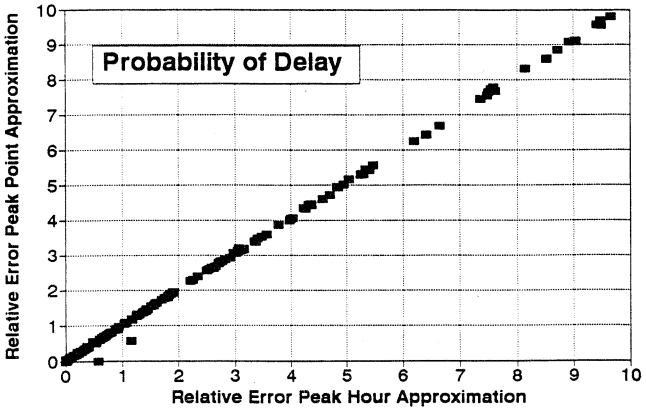


Figure 10 **Comparing Relative Errors (Peak Hour vs. Peak Point Approximations)**



relative error is less than 10%, indicate that our conclusions on the peak hour approximation will also apply to the peak epoch approximation and most likely to any other interval of relatively short duration as well.

In addition to examining the SPHA, we also computed the PSA for each peak hour performance measure. In all of our cases, these two measures were identical to two decimal places for expected queue length and expected delay and to three decimal places for probability of delay. We also found that in all but a few cases, both of these approximations were overestimates of the true measure. This is not surprising given that the stationary peak epoch approximation is an upper bound for the actual peak epoch performance and the observation that the peak hour approximation is generally close to the peak epoch approximation (due to the relative flatness of the sinusoid around the peak). Since the PSA is always greater than the SPHA and both generally overestimate actual performance it follows that the SPHA is almost always a better approximation to the actual peak hour performance. Since the SPHA is also easier to compute, these observations led us to the conclusion that there is no advantage to using the PSA for estimating peak behavior.

4. Accuracy of the Simple Peak Hour Approximation

In this section, we test and confirm a number of hypotheses about how the accuracy of the SPHA is affected by changes in the parameters of the system. This will lead us to some conclusions, discussed more fully in the next section, about when the SPHA will be useful in making decisions about the design and management of nonstationary queuing systems. We define a relative error measure as follows:

$$\text{Relative error} = [(\text{SPHA} - \text{Actual Value})/\text{Actual Value}] \times 100.$$

Since the SPHA is usually larger than the actual value, the relative error will usually be positive.

Most of the hypotheses we test here correspond to similar conjectures that were confirmed for 24-hour measures and the PSA in Green and Kolesar (1991). Our thinking was so directed because our initial observations indicated that the SPHA measures are very close to the corresponding peak hour PSA measures, and we further expected the accuracy of the peak hour PSA to be affected by the same system characteristics as the 24-hour PSA. This proved to be true as described below.

Our experimental strategy was to confirm each conjecture first for a "central case" and, if confirmed there, then to determine its validity in a region surrounding the central case by perturbing each of the key parameters. The resulting models span a fairly broad spectrum of parameter values: the number of servers ranges from 1 to 18, the service rate varies from 0.1 to 200, average traffic intensities range between 0.25 and 0.75 and relative amplitudes between 0.1 and 1.0. Yet in some

specific tests of hypotheses, we were constrained in the range and combination of parameter values due to theoretical constraints or to computational considerations.

We explored each of our hypotheses for all four performance measures—expected delay, expected queue length, and probability of delay. In the following discussion we will illustrate our results for expected delay.

4.1. The Effect of Event Frequency

Since as seen previously, the actual time-varying behavior of the system asymptotically approaches the PSA as the event frequency increases, it is not surprising that *the relative error of each of our three performance measures—expected queue length, expected delay and probability of delay—decreases as event frequency increases*. To formally test this, we examined cases in which we fixed the number of servers and simultaneously increased both $\lambda(t)$ and μ so that the time-varying traffic intensity $\rho(t)$ and the relative amplitude (RA) remain constant. The confirmation of this phenomenon for our central case of seven servers, average traffic intensity (rhobar) equal to 0.43 and RA of $\frac{1}{3}$ is shown in Figure 11 for expected delay. This result was confirmed for all of our surrounding cases which included systems with the number of servers ranging from 1 to 12, rhobar from 0.25 to 0.75 and RA from 0.1 to 1.0, and for each of the other two measures of performance.

4.2. The Effect of Service Rate

Again starting from our results for the 24-hour PSA, we reasoned that increasing the service rate alone would increase the accuracy of the SPHA. The concept is that faster clearing of customers from the system should cause consecutive time intervals to be more independent of each other, and thus peak hour performance should be better approximated by the peak hour PSA and also by the SPHA. Our experiments confirmed that *the relative error of each of the three performance measures decreases as μ increases*. This is illustrated in Figure 12 for expected delay. In these tests, $\lambda(t)$ remained constant (and hence the RA) and the number of servers was decreased as μ was increased so that $s\mu$ remained constant and hence the traffic intensity remained fixed. As with our trials changing the event frequency, we also found in this series of experiments that the lag in the time of the maximum of each performance measure relative to the time of the maximum arrival rate decreased as the service rate increased. As in our previous tests for the 24-hour PSA, we found that holding the service rate constant while increasing the arrival rate and proportionally increasing the number of servers did not have a consistent effect on SPHA accuracy.

Figure 11 Relative Error in SPHA—Expected Delay

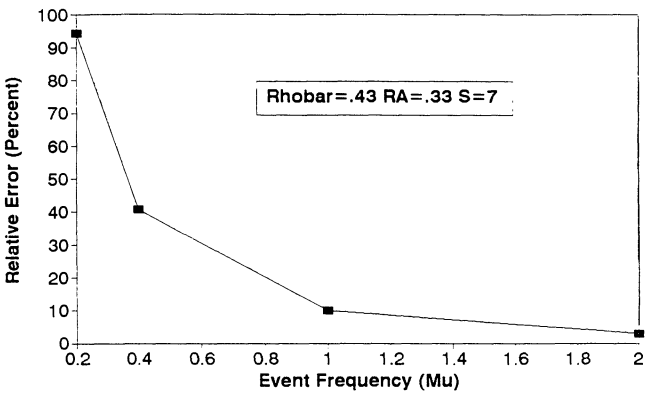
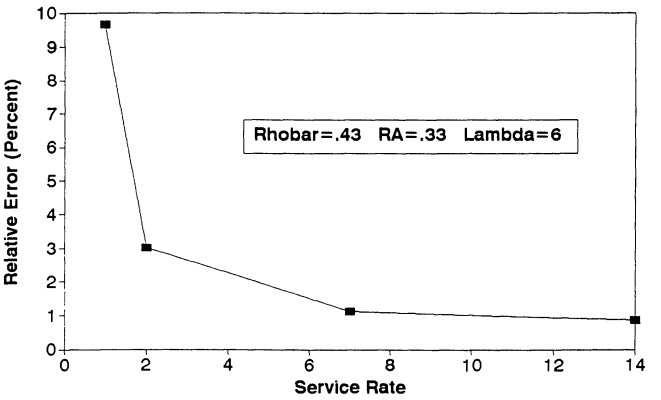


Figure 12 Relative Error in SPHA—Expected Delay



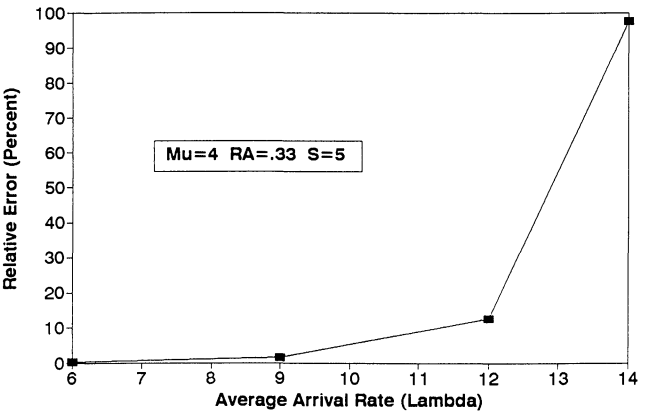
4.3. The Effect of the Maximum Traffic Intensity

As the maximum traffic intensity increases, the average λ for the peak hour will become closer to $s\mu$ and we therefore would expect that the SPHA for expected delay and expected queue length will eventually become very large and significantly overestimate the actual peak hour performance. (Of course, in the case where the peak hour average traffic intensity exceeds one, the SPHA is infinite for these measures.) We performed three sets of tests for our hypotheses that *the relative*

error of expected delay and expected queue length increases as maximum traffic intensity increases. In the first series, we increased the RA while keeping all other parameters fixed. Figure 13 is an illustration of our general finding that these relative errors increase as the relative amplitude increases. Figure 14 shows our central case for the second set of tests in which RA was held constant while the average arrival rate for the cycle was increased. Thus, in the cases in this figure, both the average and maximum traffic intensities increase. Again, the results show that the SPHA becomes less accurate. Finally, we examined the situation in which the number of servers increases while all other parameters are held constant. Increasing the number of servers also clears customers from the system faster (as was the case when increasing the service rate) and hence system behavior becomes more independent in consecutive time intervals and the PSA should become more accurate. We found, as in the case of the 24-hour PSA, that the SPHA is increasingly accurate at higher staffing levels. The central case of this situation is shown in Figure 15.

For probability of delay, our analysis shows a consistent but not uniform relationship between the SPHA accuracy and the maximum traffic intensity. In most cases, and in particular, for relatively high service rates, i.e., $\mu \geq 2$, the relative error of the SPHA decreases as the maximum traffic intensity decreases. However, since probability of delay approaches one as traffic intensity approaches one, it is reasonable to suspect that the relative error of the SPHA would go to 0 at high maximum traffic intensities for all service rates.

Figure 14 Relative Error in SPHA—Expected Delay



5. When is the SPHA Accurate Enough?

We now go beyond the qualitative and directional results of the preceding sections and determine some specific conditions and situations when the SPHA is in fact accurate enough for use in managing or designing real systems. Overall, our data base of nearly 300 cases contains a great number of instances when the SPHA is incredibly accurate, but unfortunately it also contains a large number of instances when the SPHA is very inaccurate. The work of this section draws on that database to draw some general recommendations for practitioners. We start out with a detailed discussion of a particular class of models of some interest in its own right. We then build on these particulars to develop more general findings and advice to practitioners.

Figure 13 Relative Error in SPHA—Expected Delay

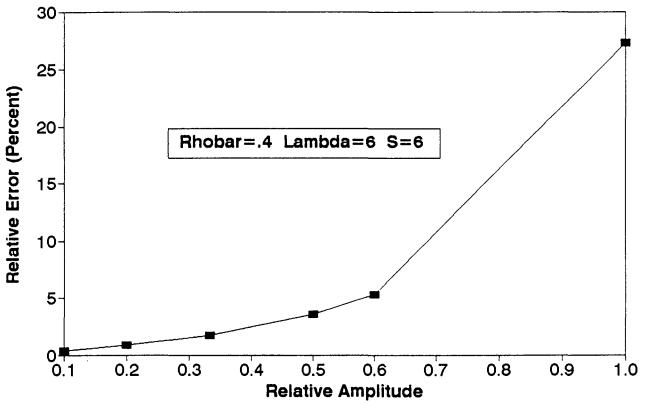
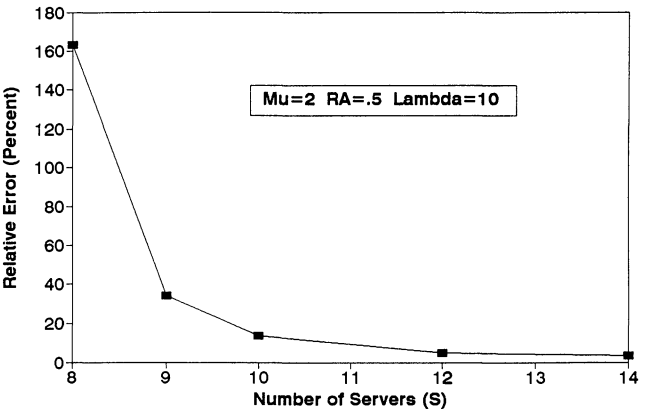


Figure 15 Relative Error in SPHA—Expected Delay

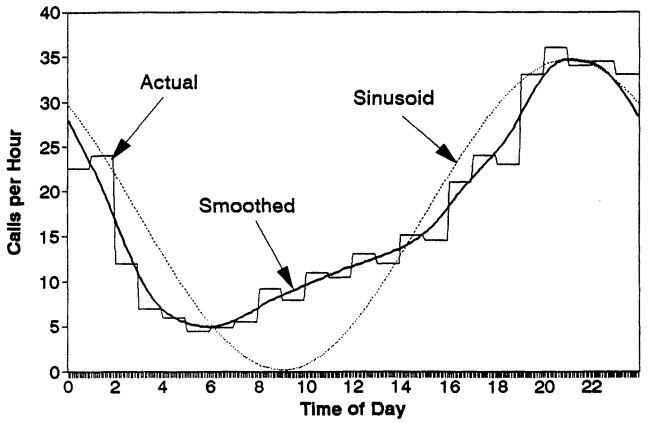


5.1. Models of Emergency Services: The Case $\mu = 2$, RA = 1

We have a particular interest in the family of Markovian sinusoidal queues with service rate $\mu = 2$ and relative amplitude RA = 1 because we, and other researchers, have encountered numerous actual situations in the management of emergency services such as police patrol, firefighting, and ambulances that are, to a first approximation, described by models with similar structure and parameter values. Indeed, our original motivation for this line of research on cyclic nonstationary queues arose to a large extent from our work on management of such emergency service systems. The applications described in Kolesar et al. (1975) and Green and Kolesar (1984) deal explicitly with staffing and scheduling issues arising in part from the cyclicality of demand.

Why these particular parameter values? A service rate of two equals a half hour average service time per call—an order of magnitude frequently experienced in both police patrol and firefighting. A sinusoid with a relative amplitude of 1 models a situation in which the minimum call rate of the day is nearly zero and there is a single maximum, occurring about 12 hours later, that is about twice the daily average. This is a little more extreme, but reasonably close to demand patterns seen frequently in practice—as is illustrated in Figure 16, which is a smoothed version of the raw police demand data from Boston that appear in our Figure 1. (The original data is from Larson 1972, p. 168.) Arrival processes for emergency services are, of course, unscheduled and are roughly time-varying Poisson processes, while service times are often roughly exponential. A documentation of this for police patrol in New York City is given in Green and Kolesar (1989). For other specific emergency service situations in which this range of parameters and conditions has been documented, see Savas (1969), Larson (1972), Kolesar et al. (1975), Walker et al. (1979), and Green and Kolesar (1984).

Figure 16 Boston Calls for Police Service—Actual and Smoothed Data



So, all in all, this model appears to be broadly descriptive of such phenomena, and it also seems reasonable to conjecture that if the SPHA approximates this set of models well, it will also approximate well other models of such environments in which our assumptions and parameter values are only approximately met. For example, our work in New York City involved use of a Markovian multiple-car dispatch model of Green (1984), and we would conjecture that the results cited here will predict when an SPHA approach will work there as well. Indeed, the Patrol Car Allocation Model widely disseminated across the United States by the RAND Corporation and described in Chaiken and Walker (1985) incorporates the Green multiple-car dispatch model and implicitly uses an SPHA style of performance estimation. For further background on models of emergency services see the books of Larson (1972) and Walker et al. (1979) and the survey papers of Chaiken and Larson (1979) and Kolesar and Swersey (1982).

Estimating Performance Measures. *Probability of Delay*: In emergency service system management the peak hour probability of delay, p_D, is often the performance measure of most interest to operators and designers, and so we focus on it first. Table 1 displays summary results for 21 cases with $\mu = 2$ and RA = 1 ranging over a broad set of the other parameter values, s (2 to 18) and λ (1 to 10), that cover a gamut of situations that we have experienced in urban police patrol. We specifically wanted to study, and the table contains, models of police precincts or departments with low and high call rates, and with low to high peak hour congestion. The table shows that there is a wide range of peak hour p_D values for these 21 model instances. Maximum traffic intensity, which we call rhomax, is along with the number of servers, clearly a dominant factor in predicting peak hour p_D for this class of models. Figure 17 displays the relative error of the SPHA for peak hour

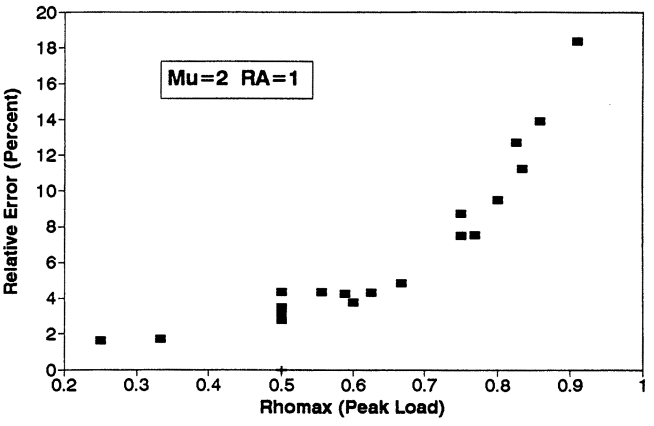
Table 1

Lambda	S	Rho Max	Probability of Delay			Expected Delay			Expected Queue		
			Actual	Approx	% Error	Actual	Approx	% Error	Actual	Approx	% Error
1	2	0.5000	0.3212	0.3323	3.48	0.1548	0.1659	7.16	0.2957	0.3312	12.01
	3	0.3333	0.0890	0.0905	1.70	0.0221	0.0226	2.25	0.0435	0.0451	3.69
	4	0.2500	0.0200	0.0203	1.61	0.0033	0.0034	1.75	0.0066	0.0067	2.60
3	4	0.7500	0.4665	0.5072	8.73	0.1969	0.2522	28.08	1.1128	1.5108	35.77
	5	0.6000	0.2256	0.2348	4.08	0.0541	0.0586	8.14	0.3171	0.3507	10.60
	6	0.5000	0.0958	0.0985	2.77	0.0157	0.0164	4.15	0.0931	0.0981	5.42
	7	0.4286	0.0364	0.0373	2.48	0.0045	0.0047	3.14	0.0268	0.0279	4.00
6	7	0.8571	0.5358	0.6102	13.89	0.1910	0.3019	58.02	2.1539	3.6160	67.88
	8	0.7500	0.3297	0.3545	7.50	0.0741	0.0882	18.97	0.8617	1.0559	22.54
	9	0.6667	0.1854	0.1944	4.84	0.0296	0.0323	9.21	0.3482	0.3867	11.06
	10	0.6000	0.0967	0.1003	3.77	0.0118	0.0125	5.86	0.1400	0.1498	7.03
	11	0.5455	0.0471	0.0487	3.40	0.0046	0.0049	4.56	0.0552	0.0582	5.45
	12	0.5000	0.0215	0.0222	3.41	0.0018	0.0018	4.12	0.0211	0.0221	4.84
10	11	0.9091	0.5720	0.6770	18.36	0.1720	0.3326	93.35	3.2400	6.6400	104.94
	12	0.8333	0.4006	0.4456	11.24	0.0825	0.1104	33.76	1.5931	2.2043	38.36
	13	0.7692	0.2627	0.2825	7.54	0.0400	0.0468	17.01	0.7822	0.9346	19.48
	14	0.7143	0.1630	0.1723	5.65	0.0194	0.0214	10.39	0.3822	0.4280	11.97
	15	0.6667	0.0963	0.1008	4.73	0.0094	0.0100	7.42	0.1847	0.2005	8.55
	16	0.6250	0.0542	0.0566	4.32	0.0044	0.0047	6.07	0.0877	0.0938	6.94
	17	0.5882	0.0292	0.0304	4.24	0.0021	0.0022	5.40	0.0407	0.0433	6.18

p_D as a function of rhomax for the same 21 cases in Table 1. We can see that the relative errors in peak hour p_D also increase "exponentially" with rhomax.

What is a satisfactory approximation? Any definition will necessarily be somewhat subjective and dependent on the problem context. We propose one that makes sense in the context of the types of applications we have just been discussing. It is consistent with the accuracy with which parameter values can be estimated in such environments, with the approximate nature of the models, and with the types of broad system management and design issues for which such models are used to adopt a relative error of 10% or less as the benchmark standard for a satisfactory approximation. By this standard, peak hour p_D is satisfactorily approximated for all cases in our set of models that have a rhomax of 0.8 or less—a very broad range of presented peak loads. Other operational considerations such as, for example, a desire to have reasonable availability of the police units to actually patrol or to initiate spontaneous public services, or a desire to control the stress level on firefighters of too frequent response, would in most cases drive managers to want run their systems in this range of loads. In fact, it has been policy in the New York City Police Department to achieve an average load factor of about 0.6 for radio patrol cars (Brown 1990). We also note

Figure 17 SPHA for Probability Delay—Relative Error vs. Rhomax



that the maximum relative error in Figure 17, which occurs at a peak load of 0.91 was still only 18%. Figure 18 displays, for the same cases, the relative error in the SPHA for p_D versus peak hour p_D, itself. It shows that systems with low peak hour p_D are quite well approximated. In particular, all cases with peak hour p_D of 0.4 or less have relative errors of 10% or less. We can recommend use of the SPHA for modelling emergency service systems operating in this range.

Expected Queue Size and Expected Delay: Both the expected queue size, L_q and the expected delay, W_q during the peak hour are more difficult to approximate with the SPHA then is p_D, as we shall now show. Since the magnitudes and patterns of the relative errors are similar for these two measures, we show here only Figure 19 which displays the relative error in the SPHA for L_q versus rhomax. As we saw for the SPHA for p_D, the relative error for the SPHA for peak hour L_q increases exponentially in rhomax, but now the magnitudes of the errors are higher. While most of the cases shown in Figure 19 with rhomax of 0.7 or below have relative errors that are less than our benchmark of 10%, there is one case with rhomax of 0.5 that has a relative error of 12%. Note also that the relative error at rhomax of 0.91 is now 105%.

Capacity Planning. Queuing models are often used in environments like emergency services to assist managers in making staffing and capacity decisions, so it is pertinent to ask if use of the SPHA would lead to sensible staffing decisions. Our analysis of this issue is carried out under the premise that high delays are intolerable in such environments, and we therefore particularly want to determine if the SPHA correctly suggests the minimum number of servers needed to achieve a targeted low level of peak hour service delay. Of course, the targeted service characteristic and the specified value of the service delay benchmark would vary from one application to another. Somewhat arbitrarily, but not we feel unreasonably, we take as our benchmark service standard for this discussion a 10% peak hour p_D. Our database is again the 21 cases in Table 1. These models cover four demand situations (λ = 1, 3, 6, and 10). For each λ we ran all cases starting with the smallest feasible number of servers and increasing the number of servers by one until we reached a model where the peak hour p_D was below five percent. The table shows that in each case the SPHA would lead to the same staffing decision as the actual model. Moreover, the table also shows that at the staffing and performance level suggested by using the SPHA to achieve the benchmark on peak hour p_D, the SPHA accurately predicts performance on the two other service measures, peak hour L_q and peak hour W_q.

Figures 20 and 21 are plots of the SPHA and actual peak hour p_D and peak hour L_q versus the number of servers, respectively, for one of the models contained in Table 1, namely that with λ = 6. The accuracy of the SPHA, particularly in the range of desirable peak hour p_D, is evident. Results for the other cases in the table are similar.

In summary, the SPHA does a very useful job of estimating performance in this class of models, particularly

Figure 18 SPHA for Probability Delay—Relative Error vs. Probability Delay

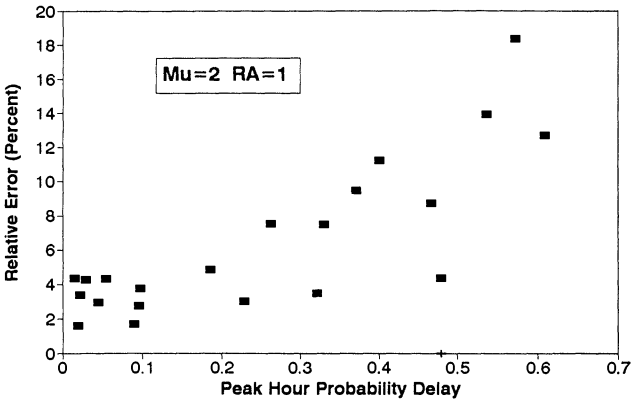


Figure 19 SPHA for Expected Queue—Relative Error vs. Rhomax

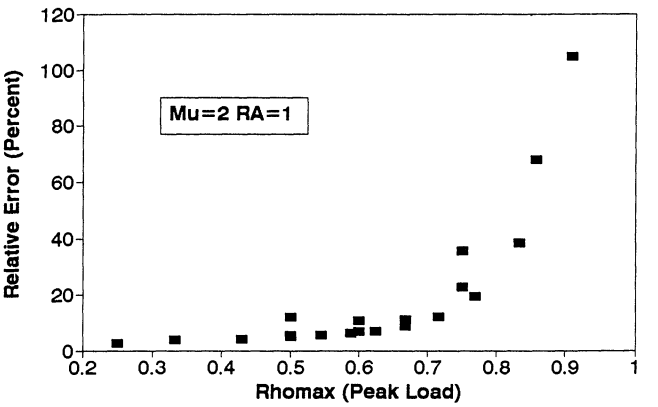


Figure 20 SPHA for Probability Delay (Mu = 2; RA = 1; Lambda = 6)

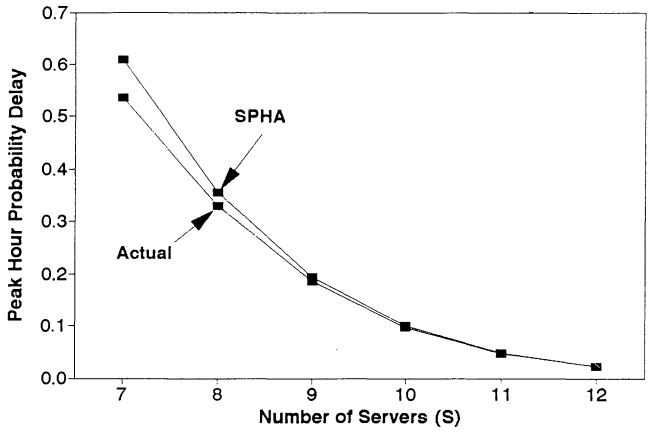
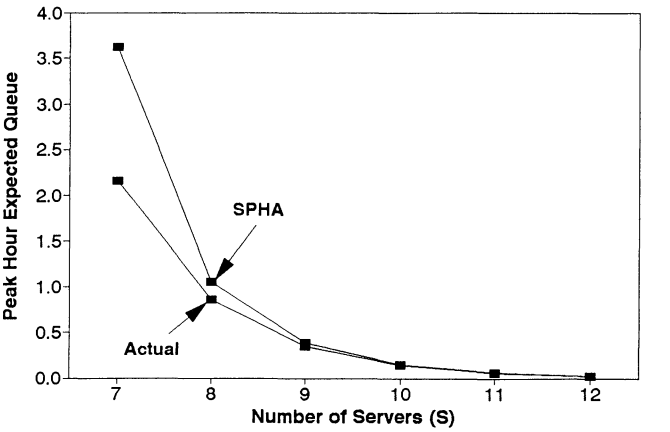


Figure 21 SPHA for Expected Queue (Mu = 2; RA = 1; Lambda = 6)



if the target performance criterion is low peak hour p_D, or if there is a management commitment to operate the system at a reasonable peak hour presented load. Its use could be recommended.

5.2. More General Findings on SPHA Accuracy

We now move from the consideration of the specific case of $\mu = 2$ and RA = 1 to a broader discussion of SPHA accuracy. The models with service rate of 2 provide a useful point of separation between "high" and "low" service rates. Indeed, we take $\mu = 2$ as an "intermediate" case for which the accuracy of the SPHA is dependent on a number of factors—as our discussion of the family of models with $\mu = 2$ and RA = 1 revealed. Figure 22 is a plot of relative error in peak hour p_D versus rhomax for all 118 cases that we ran with mu = 2, and it shows that while the SPHA is generally accurate within this set, there are instances at high rhomax when it is not—when the relative error exceeds 10%. Overall, the average and maximum relative errors for peak hour p_D for these 118 cases were 4.7% and 18.8%, respectively.

Table 2 summarizes the results of all our 263 model instances, all those with rhomax less than 1, used in this analysis. The table shows that for each peak hour performance measure, L_q, p_D, and W_q, both the average and the maximum relative errors of the SPHA tend to decrease with service rate. The decrease does not, however, appear uniform, as these sets of runs were not laid out in advance to permit balanced comparisons across these service rate groupings. The *trend* of decreasing relative error with increasing μ is illustrated by Figure 23, a log-log plot of average relative error in the SPHA for peak hour W_q versus service rate for these same 263 cases.

For service rates well below 2, our findings, illustrated by the table, are that the SPHA is generally not reliable. This parallels our conclusion regarding the use of the 24-hour PSA documented in Green and Kolesar (1991). For service rates well above 2 the situation depends upon the performance measure of interest. We will use three specific service rates above 2 to make a more detailed and revealing analysis. Figures 24, 25, and 26 are plots of the relative error in the SPHA to the peak hour probability of delay for models with $\mu = 4$, 20, and 200,

Figure 22 Relative Error in SPHA—Probability Delay

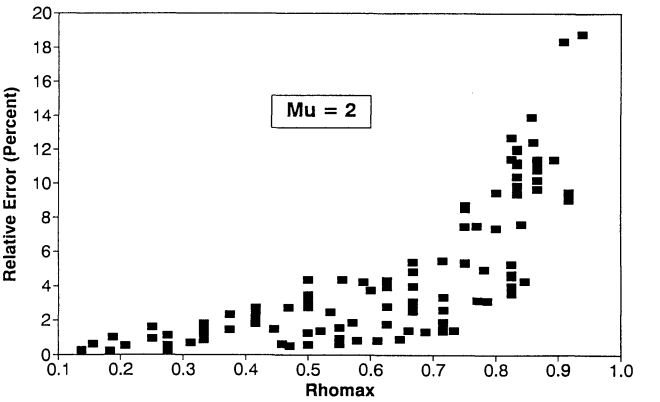


Table 2 Relative Error in SPHA Estimates (Percent)

		Expected Delay		Expected Queue		Probability of Delay	
Mu	*n*	Average	Maximum	Average	Maximum	Average	Maximum
0.1	2	516.3	548.4	927.7	1016.7	247.6	270.4
0.2	30	133.2	806.4	257.9	1265.5	84.2	422.4
0.4	1	39.8	39.8	53.4	53.4	26.5	26.5
0.5	3	71.8	128.5	91.5	163.0	27.9	41.5
1.0	8	24.0	64.1	32.3	86.7	9.4	19.7
1.8	1	58.5	58.5	73.6	73.6	13.6	13.6
2.0	105	16.8	152.8	21.2	172.0	4.7	18.8
2.5	9	7.8	27.8	9.3	33.0	2.8	8.2
3.0	8	17.0	78.9	20.0	91.0	2.9	8.9
4.0	10	23.4	98.5	26.6	109.3	3.2	10.2
5.0	4	8.1	23.7	9.5	27.5	1.9	4.8
7.0	1	1.2	1.2	1.6	1.6	0.4	0.4
10.0	2	3.2	6.3	4.0	7.8	0.7	1.3
14.0	1	0.9	0.9	1.3	1.3	0.2	0.2
20.0	42	2.8	46.5	3.1	53.3	0.3	3.1
200.0	10	0.0*	0.0*	0.0*	0.7	0.0*	0.0*

* These values are less than 0.001.

respectively. Comparing results across the three graphs shows a dramatic drop in the relative error as μ increases. For the cases $\mu = 4$ and $\mu = 20$, comparison within the plots shows a distinctive exponential-like increase in relative error as rhomax increases. Moreover, the largest relative errors are still quite reasonable for most applications—even when rhomax is greater than 0.80. The graph for $\mu = 200$ shows very small relative errors—always below 0.1 percent, regardless of the value of rhomax. However, the distinctive exponential pattern is missing, most likely, we believe, because at such small relative errors much of the case-to-case variation in results is probably a consequence of deviations in the numerical accuracy in the differential equations solution method.

We conclude that at high service rates, say in the hundreds, the SPHA for peak hour p_D is very good indeed, regardless of the other parameter values. Computational constraints kept us from running models with

Figure 23 SPHA for Expected Delay—Average Relative Error vs. Service Rate

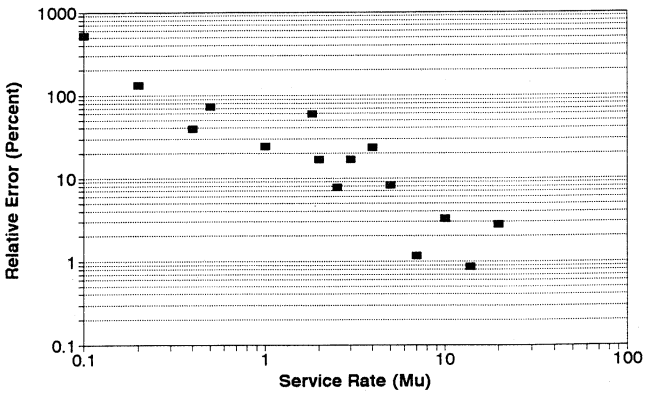


Figure 24 SPHA for Probability Delay—Relative Error vs. Rhomax

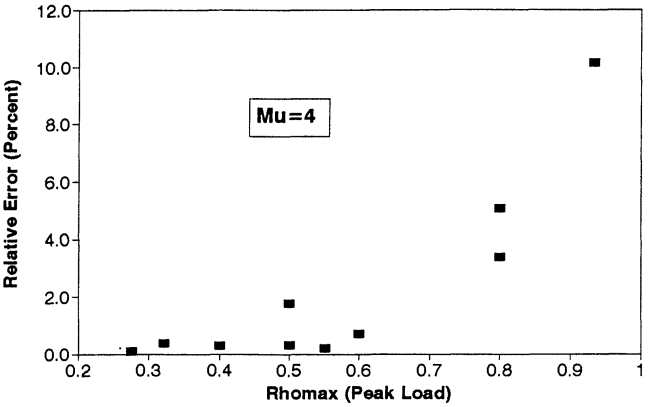


Figure 25 SPHA for Probability Delay—Relative Error vs. Rhomax

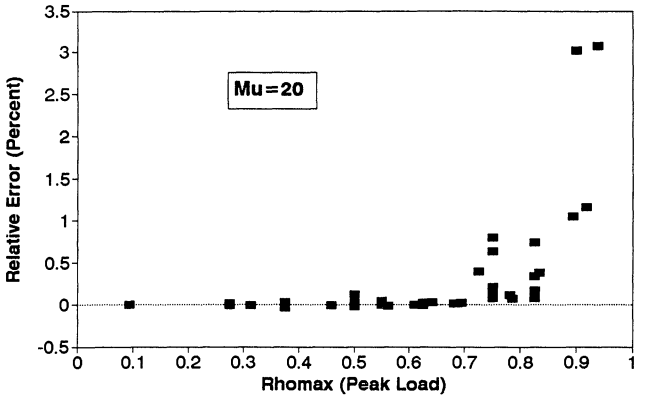
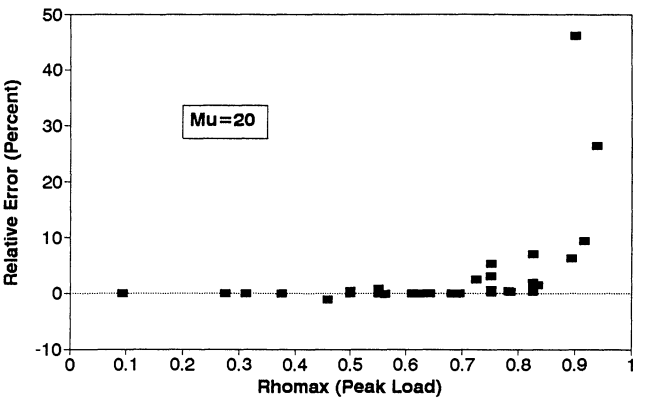


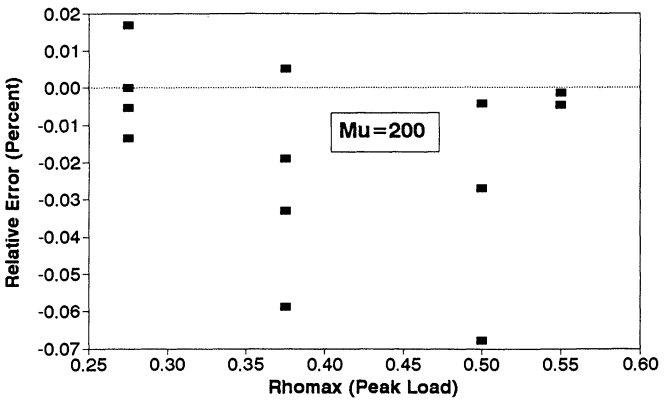
Figure 27 SPHA for Expected Delay—Relative Error vs. Rhomax



both μ and rhomax very high. At intermediate service rates, say in the tens of customers per hour, the SPHA for probability of delay is quite good indeed, with relative errors well below our 10% benchmark. (See Figure 25, for example.)

As we have noted earlier, the expected peak hour queue and the expected peak hour delay are not as well estimated by the SPHA as is the probability of delay. We can see from our results that serious estimation problems are likely only at high congestion, that is at high rhomax. Estimation of both peak hour L_q and W_q are of concern, but since the results are so similar for them we discuss only W_q here. Take, for example, the case of models with service rate 20. We have run 50 such models over a broad range of other parameter values: rhomax, relative amplitudes, arrival rates, and the like. Figure 27 shows the relative error in the SPHA for peak hour W_q versus rhomax for all such models run. The plot shows the distinctive exponential pattern of relative error versus rhomax, and all models with rhomax less than 0.80 have relative errors less than 10 percent. Above rhomax equal 0.80 the relative errors appear to increase very steeply, and there is one case with a relative error of 46% at rhomax of 0.90 and another with relative error of 26% at rhomax of 0.94. Thus, we have concluded that at rhomax values below say 0.80 the SPHA for peak hour L_q and W_q can be useful.

Figure 26 SPHA for Probability Delay—Relative Error vs. Rhomax



6. Conclusion

In this paper we have clarified the conditions under which the accuracy of the simple peak hour approximation increases. We have also determined a range of situations and a criterion under which the SPHA produces satisfactory results from a practical point of view. These conditions are quite broad, but we have also determined a number of situations in which the SPHA is clearly not appropriate.

Of practical significance for many applications are our findings that for systems with high service rates—in the hundreds of customers per hour—the SPHA will be very accurate almost without regard for the values of other system parameters. Thus for service systems such as telecommunications, information processing, banking, and toll booths, the SPHA can be used with confidence particularly to identify operating conditions that will keep congestion at reasonably low levels. Furthermore, our results—though focused on a "peak hour"—indicate that simple stationary approximations

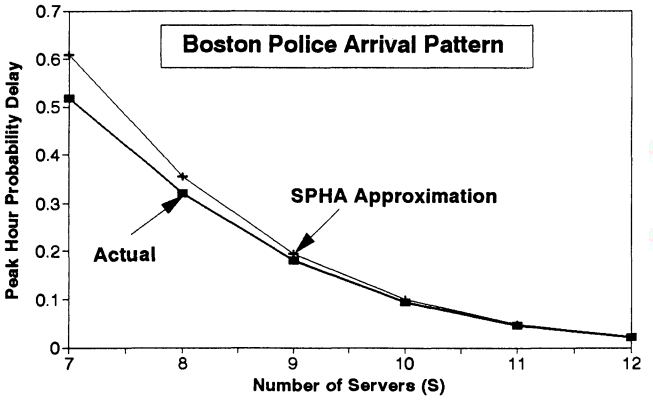
are likely to be accurate for other intervals of short duration.

It is important to note that although our findings are based on experimentation with models having sinusoidal arrival rates and exponential service times we see no indications that our conclusions are not broadly applicable to situations in which these restrictions are relaxed.

We have begun a stream of research to explore the robustness of our results. Preliminary results are illustrated in Figure 28 which plots the actual and SPHA for peak hour probability of delay as a function of staffing levels for a model where the input stream is the smoothed Boston Police calls data shown in Figure 16. For comparison purposes, all other model settings are as assumed for the runs shown in Figure 20 which also correspond to a police application. The reader will observe that as in Figure 20, the SPHA and actual curves get closer and closer as the number of servers increases and as probability delay is decreased. If the operating criterion is to keep probability of delay below our 10% benchmark, the SPHA model leads to the correct staffing assignments. We note that an examination of the input stream shown in Figure 16 indicates that the arrival pattern is more peaked—that is, less flat—than a sinusoidal stream, yet the SPHA is still good.

From this example, and based on our accumulated theoretically and numerically based knowledge, we believe that there are four major factors which will determine how accurate the SPHA is in actual contexts:

Figure 28 **SPHA for Probability Delay (Mu = 2; RA = 1; Lambda = 6)**



(1) the steepness of the arrival rate prior to the peak hour—the steeper this is, the more likely the SPHA will overestimate the actual delay in the peak hour;

(2) the evenness of the arrival rate during the peak hour—this affects the extent to which the implicit assumption of steady-state behavior of the SPHA is good;

(3) the magnitude of the service rate which similarly is related to steady-state behavior; and

(4) the maximum traffic intensity which affects the likelihood of the SPHA for the expected delay or expected queue length becoming extremely large.

We are also embarking on another line of research to determine under what conditions an SPHA will work well when the system capacity changes over time. In police patrol, in toll booth operations and in many other contexts, the staffing levels change frequently during the day (Kolesar et al. 1975, Segal 1974). Our research will explore whether the SPHA behaves qualitatively in such situations as it does here when the staffing levels are constant over time.

References

Bear, D., *Principles of Telecommunication—Traffic Engineering*, Peter Peregrinus, Ltd., London, 1980.

Brown, L. P., "Staffing Needs of the New York City Police Department," Report to the Mayor, NYCPD, 1990.

Chaiken, J. M. and W. Walker, "Patrol Car Allocation Model: Executive Summary," Report R-3087/1-NIJ, The Rand Corporation, Santa Monica, 1985.

–→ and R. C. Larson, "Methods for Allocating Urban Emergency Units: A Survey," *Management Sci.*, 19 (1972), 110–130.

→ Edie, L. C., "Traffic Delays at Toll Booths," *Oper. Res.*, 2 (1954), 107–138.

→ Eick, S. G., W. A. Massey, and W. Whitt, "$M_t/G/\infty$ Queues with Sinusoidal Arrival Rates," *Management Sci.*, 39 (1993), 241–252.

——, -→-, and W. Whitt, "The Physics of the $M_t/G/\infty$ Queue," *Oper. Res.*, 41 (1993), 731–742.

Farber, N., "Traffic Engineering Course Notes," photocopied notes for an internal Bell Laboratories course, Holmdel (1979).

Giffen, W. C., *Queueing Basic Theory and Applications*, Grid Inc., Columbus, Ohio, 1978.

→ Grassmann, W., "The Convexity of the Mean Queue Size of the $M/M/c$ Queue with Respect to the Traffic Intensity," *J. Applied Probability*, 20 (1983), 916–919.

→ Green, L., "A Multiple Dispatch Queueing Model of Police Patrol Operations," *Management Sci.*, 30 (1984), 653–664.

—— and P. Kolesar, "The Feasibility of One-Officer Patrol in New York City," *Management Sci.*, 20 (1984), 964–981.

–→ and ——, "Testing the Validity of a Queueing Model of Police Patrol," *Management Sci.*, 35 (1989), 127–148.

Green, L. and P. Kolesar, "The Pointwise Stationary Approximation for Queues with Nonstationary Arrivals," *Management Sci.*, 37 (1991), 84–97.

——, -→-, and A. Svoronos, "Some Effects of Nonstationarity on Multiserver Markovian Queueing Systems," *Oper. Res.*, 39 (1991), 502–511.

Gross, D. and C. M. Harris, *Fundamentals of Queueing Theory*, second ed., John Wiley & Sons, New York, 1985.

Heyman, D. P. and W. Whitt, "The Asymptotic Behavior of Queues with Time-Varying Arrival Rates," *J. Applied Probability*, 21 (1984), 143–156.

Holloran, T. J. and J. E. Byrn, "United Airlines Station Manpower Planning System," *Interfaces*, 16 (1986), 39–50.

Kleinrock, L., *Queueing Systems, Vol. II, Computer Applications*, John Wiley, New York, 1976.

Kolesar, P. J., "Stalking the Endangered CAT: A Queueing Analysis of Congestion at Automated Teller Machines," *Interfaces*, 14 (1984), 16–26.

-→-, K. L. Rider, T. B. Craybill, and W. W. Walker, "A Queueing-Linear Programming Approach to Scheduling Police Patrol Cars," *Oper. Res.*, 23 (1975), 1045–1062.

—— and A. Swersey, "Ten Years of Research in the Logistics of Urban Emergency Services," *Operation Research*, J. P. Braus (Ed.), North-Holland, N.Y., 1982.

Koopman, B. O., "Air-Terminal Queues under Time-Dependent Conditions," *Oper. Res.*, 20 (1972), 1089–1114.

Larson, R. C., *Urban Police Patrol Analysis*, MIT Press, Cambridge, MA, 1972.

Lee, A., *Applied Queueing Theory*, St. Martins Press, Montreal, Canada, 1966.

Massey, W. A., "Asymptotic Analysis of the Time Dependent $M/M/1$ Queue," *Mathematical Oper. Res.*, 10 (1985), 305–327.

—— and W. Whitt, "Networks of Infinite-Server Queues with Nonstationary Poisson Input," *Queueing Systems*, 13 (1993), 185–250.

Newell, G. F., "Queues with Time-Dependent Arrival Rates I–III," *J. Applied Probability*, 5 (1968), 435–451, 579–606.

——, "Applications of Queueing Theory, second edition," Chapman and Hall, London, 1982.

Odoni, A. R. and E. Roth, "An Empirical Investigation of the Transient Behavior of Stationary Queueing Systems," *Oper. Res.*, 31 (1983), 432–455.

Rolski, T., "Upper Bounds for Single Server Queues with Doubly Stochastic Poisson Arrivals," *Mathematics of Oper. Res.*, 11 (1986), 442–450.

-→-, "Approximation of Periodic Queues," *Advance Applied Probability*, 19 (1987), 691–707.

Savas, E. S., "Simulation and Lost-Effectiveness of New York's Emergency Ambulance Service," *Management Sci.*, 15 (1969), B608–B627.

Segal, M., "The Operator Scheduling Problem: A Network Flow Approach," *Oper. Res.*, 22 (1974), 808–823.

Walker, W., J. M. Chaiken, and E. J. Ignall (Eds.), *Fire Department Deployment Analysis*, North-Holland, N.Y., 1979.

Whitt, W., "The Pointwise Stationary Approximation for $M_1/M_1/s$ Queues is Asymptotically Correct as the Rates Increase," *Management Sci.*, 37 (1991), 307–314.

Accepted by Gabriel R. Bitran; received June 14, 1993. *This paper has been with the authors* 3 *months for* 1 *revision.*