Two Issues in Setting Call Centre Staffing Levels

Bert P. K. Chen *
Department of Statistics
University of Oxford
1 South Parks Road
Oxford, OX1 3TG
UK

Shane G. Henderson
School of Operations Research and Industrial Engineering
Rhodes Hall
Cornell University
Ithaca, NY 14853
U.S.A.

2nd August, 2001

Abstract

Motivated by a problem facing the Police Communication Centre in Auckland, New Zealand, we consider the setting of staffing levels in a call centre with priority customers. The choice of staffing level over any particular time period (e.g., Monday from 8am - 9am) relies on accurate arrival rate information. The usual method for identifying the arrival rate based on historical data can, in some cases, lead to considerable errors in performance estimates for a given staffing level. We explain why, identify three potential causes of the difficulty, and describe a method for detecting and addressing such a problem.

1 Introduction

Motivated by a project conducted with the Police Communication Centre (PCC) in Auckland, New Zealand, we discuss two issues related to the setting of staff levels in inbound call centres. Our focus is primarily on the application, and indeed, we will use the PCC project as a running example throughout this paper.

The PCC receives emergency calls, allied emergency calls, and non-emergency (general) calls from the upper half of the North Island of New Zealand. Allied emergency service calls are calls for assistance received from other emergency services such as the fire department and ambulance service. The class of a call may be determined from its source. For example, emergency (111) calls are first routed to a telephone operator, and then to the call centre, and therefore can be recognized. Emergency (111) calls and allied emergency service calls are treated identically by the call centre. Therefore, we treat these calls as a single class (class 1) calls. The non-emergency (general) calls form the second class.

Calls are answered by call takers who log information on the call, and pass the call on to the relevant dispatcher who coordinates police response. Class 1 calls have non-preemptive priority over class 2 calls, i.e., if a call taker is unavailable when a call arrives then the call is queued, irrespective of the call's class, and queued class 2 calls are answered only after any queued class 1 calls have been completed.

*This research was conducted while the authors were members of the Department of Engineering Science at the University of Auckland.

The PCC is interested in determining the number of call takers required to ensure that 90% of emergency calls are answered within 10 seconds, and 80% of general calls are answered within 30 seconds. We do not consider the related question of how many dispatchers are required to ensure that dispatchers are not overloaded. The answer to this question depends on factors such as a person's ability to handle psychological stress that we are not qualified to address.

This specific problem may be modeled as follows.

Consider a queueing system that receives calls of several types or classes, $1, \ldots, p$. Calls of class i receive nonpreemptive priority over calls of class j if $i < j$, i.e., a free service agent will answer a call of class j only if there are no calls of class $1, \ldots, j-1$ waiting, and calls are never interrupted. We assume that calls of class i arrive to the system according to a nonhomogeneous Poisson process with rate function $(\lambda_i(t) : t \geq 0)$, and require a service time that is exponentially distributed with mean μ_i^{-1}, for $i = 1, \ldots, p$.

We wish to determine staffing levels that ensure that all call classes receive "satisfactory" service. By satisfactory service we mean that at least $p_i\%$ of class i calls are served within m_i seconds, for $i = 1, \ldots, p$. This definition of satisfactory service is common in the call centre industry, and is used by the PCC. Notice that this definition allows one to ensure, for example, that some percentage of class i calls are answered immediately (set $m_i = 0$). Other definitions of satisfactory service are, of course, possible, although we do not discuss them further here.

The presence of a time-dependent arrival rate makes an exact analysis of this model extremely difficult. As is common in analyzing call centres, we apply the stationary, independent, period by period (SIPP) approximation (Green, Kolesar and Soãres 1999) to this problem. That is, we consider each time-period of the day in isolation, assuming calls of class i arrive according to a homogeneous Poisson process with rate λ_i, compute stationary measures of this approximate system, and use these to determine staffing levels.

To the best of our knowledge, expressions for the steady-state customer waiting time distribution (exclusive of service) are not available for this model. However, if one assumes that the service rates are not class-dependent, i.e., that $\mu_i = \mu$ for all i, then the Laplace-Stieltjes transforms (LSTs) can be derived (Davis 1966). Kella and Yechiali (1985) obtained the same results as Davis in a simpler fashion.

The LST for the class 1 call waiting time distribution is easily inverted, and so it is possible to easily predict the performance on class 1 calls for any given staffing level. However, for $i > 1$, it is not possible to analytically invert the LST for the waiting time distribution of class i customers. Therefore, one might turn to numerical transform inversion (Abate and Whitt 1995) to compute the tail probabilities needed to set staffing levels in the above model.

Unfortunately, the algorithms given in Abate and Whitt (1995) rely on calculations involving complex variables. Such calculations are not easily implemented in a spreadsheet setting. This is an important consideration in selecting a method for setting staffing levels, since staffing levels are typically set by a local manager who has little or no familiarity with, or access to, specialized computer packages.

Therefore an alternative to numerical transform inversion may be required. We use easy-to-implement inequalities to obtain a bound on waiting time performance for non-emergency calls in the PCC application, and this approach is discussed in Section 2, where we consider the setting of staffing levels in a priority queue.

Once an appropriate queueing model is chosen, it remains to identify the parameters of the model for each period in which a separate staffing level is considered. The typical approach used is to extract historical data from a database on the number of calls in a particular period in the past, take the average of these values, and use the average as the arrival rate parameter. For example, if we are considering the time period Monday from 8am - 9am, then each datum gives the number of calls received on a past Monday from 8am - 9am. If one has n weeks of recorded data, then the data set will be of size n.

There are at least 3 potential sources of error in estimating the arrival rate for a future period in this fashion.

1. Estimation error: The arrival rate estimator is an average of a finite number of random variables, and as such will not give the "exact" value of the arrival rate parameter, even in the case where the data are assumed to be independent and identically distributed (i.i.d.) with finite mean.

2. Nonstationarity: There may be nonstationarities in the data, even when one focuses on a particular

period like Monday 8am-9am, so that it is not reasonable to assume that the data are identically distributed. In such cases the arrival rate in a future period may not be well-predicted by the number of arrivals in the corresponding previous periods. Such nonstationarities could arise in the PCC, for example, due to changes in climate over the year. The winter months are colder and wetter than the summer months, so that people's lifestyles also change. This can lead to a change in both the volume and nature of calls. Such nonstationarities might be detected and predicted using standard time-series methods, at least when one has several "seasons" of data. This quantity of data was not available when we completed our study for the PCC.

3. Random arrival rate: It may be that in any particular period calls are arriving according to a Poisson process with rate Λ say, but the arrival rate Λ itself may be random. Note that conditional on a realization of the random variable Λ, calls arrive according to a Poisson process with rate Λ. For example, call takers at the PCC have observed that they tend to be busiest on days with poor weather conditions. This suggests that the arrival rate Λ may be a function of weather conditions, and as such may be viewed as being random.

All 3 of these mechanisms lead to uncertainty in the true arrival rate that will be experienced in the period for which one is making staffing level decisions. One might use the terminology "forecasting error" to describe this situation. Forecasting error might also arise at a call center dedicated to retail sales after the introduction of a new product. One cannot exactly predict the arrival rate that will be faced. Forecasting error can be modeled by assuming that the actual arrival rate is random. Note that in the case of nonstationarities the arrival rate may not actually be random, but mathematically speaking, the procedure one follows to set resource levels can be captured within a random arrival rate framework; see Section 5.

We therefore have strong motivation for considering the case where, conditional on a realization of the random arrival rate, calls arrive according to a Poisson process. Such processes are known as conditional Poisson processes (see, e.g., Ross 1983, p. 49), doubly-stochastic Poisson processes or Cox processes. Rolski (1981, 1986) investigated such processes in relation to single-server queues, and established bounds on certain performance measures.

Remark 1: A more general form of this problem is one where the arrival rate to the system is both time-varying *and* random. Given the complexities of dealing with the special case of a time-varying nonrandom arrival rate, we will not address this more general problem here. Instead, we will focus on the case where the random arrival rate, once realized, is constant. The justification for this approach is similar to that for the SIPP approach. In particular, if the periods under consideration are long enough relative to the relaxation time of the system, then it is reasonable to approximate the performance in a given period by steady-state performance measures evaluated at the random arrival rate.

Remark 2: Several authors are currently developing techniques to incorporate the effect of forecasting error for a variety of parameters (not just arrival rates) in the simulation setting; see Chick (2001) and the references therein.

Thompson (1999) explored the impact of a random arrival rate, and suggested methods for dealing with it. Our theoretical results in Section 3 and Section 4 complement the empirical results of Thompson. In Section 3 we explore the impact of a random arrival rate on predicted service performance in very general terms. This impact is most easily isolated when the service measure is a concave or convex function of the arrival rate of calls. The results of this section are straightforward to establish, but their implications may be very important in setting staffing levels in some call centres.

To buttress this point, in Section 4 we consider several models that are commonly used in the call centre industry to predict service performance, and examine their convexity/concavity characteristics. Our results show that one will typically overestimate the performance that will be achieved with a given staffing level if a random arrival rate is ignored, although the reverse is possible in some contexts.

In Section 5 we provide an approach for detecting and modeling a random arrival rate. We assume that it is possible to compute performance conditional on a fixed, deterministic arrival rate. Therefore, we do not limit ourselves to the models considered in Section 4. In particular, we can apply these ideas to the PCC problem, and to more complicated call centres where more complicated models may be appropriate.

As we describe in Section 5, it is possible to use straightforward techniques to test whether there is a need to go to the extra trouble of explicitly modeling a random arrival rate. We describe how this can be done in Section 5, and in Section 6 explain the outcome of such calculations in our study of the PCC.

Finally, in Section 6, we offer some conclusions. In particular, we elaborate on when we believe that it is important to explicitly model a random arrival rate.

In summary, we view the primary contributions of this paper as follows:

1. We show how to apply a multi-server priority queueing model to assist in setting staffing levels in call centres.

2. We use easily implemented inequalities as a simple substitute to numerical transform inversion to obtain bounds on service time performance that can be easily computed within a spreadsheet.

3. We explore the implications of a random arrival rate in terms of setting staffing levels in call centres, and show that one will typically overestimate service level performance if a random arrival rate is ignored. These results reinforce, and help explain, the empirical results of Thompson (1999).

4. We provide an approach for detecting, modeling and accounting for a random arrival rate in setting staffing levels for call centres.

2 The Non-Preemptive Priority Queue

In this section we summarize results from Davis (1966), and Kella and Yechiali (1985) for a non-preemptive priority queue model, and show how to apply these in setting staffing levels for call centres with customer classes of differing priorities.

Consider a queueing system where customers may be one of p priority classes. Customers of class i ($1 \leq i \leq p$) arrive according to a Poisson process with rate $\lambda_i > 0$. Service times for all customers are exponentially distributed with common mean μ^{-1} and there are c servers. Customers of class $i < j$ have non-preemptive priority over customers of class j, so that queued customers of class i are served before queued customers of class j. Within each class, customers are served in FIFO order. We assume that $\sum \lambda_i < c\mu$, so that the system is stable.

Let W_i denote the steady-state waiting time in the queue (exclusive of service) for class i customers. Davis (1966) and Kella and Yechiali (1985) both derive the LST $\tilde{W}_i(s) = E\exp(-sW_i)$ of W_i for $i = 1, 2, \ldots, p$. In what follows, we borrow heavily from Kella and Yechiali's notation.

Let λ and ρ be the total arrival rate, and overall traffic intensity respectively, so that $\lambda = \sum_{i=1}^{p} \lambda_i$ and $\rho = \lambda/(c\mu)$. For $1 \leq i \leq p$, define the traffic intensity due to class i customers $\rho_i = \lambda_i/(c\mu)$, and the cumulative traffic intensity due to class i customers and above $\sigma_i = \sum_{j=1}^{i} \rho_j$. Define $\sigma_0 = 0$.

For $i \geq 2$, define $\lambda_{<i}$ to be the cumulative arrival rate of customers in classes 1 through $i-1$, so that $\lambda_{<i} = \sum_{j=1}^{i-1} \lambda_j$, and set $\lambda_{<1} = 0$. For $i \geq 2$, let $\tilde{\gamma}_i(s)$ denote the LST of the busy period B in an $M/M/1$ queue with arrival rate $\lambda_{<i}$ and service rate $c\mu$, so that

$$\tilde{\gamma}_i(s) = E(e^{-sB}) = \frac{s + \lambda_{<i} + c\mu - [(s + \lambda_{<i} + c\mu)^2 - 4\lambda_{<i}c\mu]^{1/2}}{2\lambda_{<i}},$$

where it is understood that the square root in the numerator is the nonnegative square root. For $i = 1$, we define $\tilde{\gamma}_i(s) = c\mu/(c\mu + s)$.

The LST $\tilde{W}_i$ of the steady-state waiting time W_i in the queue for class i customers is then given by

$$\tilde{W}_i(s) = E(e^{-sW_i}) = (1 - \eta) + \eta \frac{c\mu(1 - \sigma_i)(1 - \tilde{\gamma}_i(s))}{s - \lambda_k + \lambda_k \tilde{\gamma}_i(s)},$$

where

$$\eta = \frac{(\lambda/\mu)^c}{c!(1 - \rho)} \left[\sum_{k=0}^{c-1} \frac{(\lambda/\mu)^k}{k!} + \frac{(\lambda/\mu)^c}{c!(1 - \rho)} \right]^{-1},$$

the steady-state probability that all servers are busy.

We now turn to specializing these results to the case in point, namely setting staffing levels in call centres with calls that receive different priority.

When $i = 1$, the LST for W_i simplifies somewhat to

$$(1 - \eta) + \eta \frac{c\mu - \lambda_1}{c\mu - \lambda_1 + s},$$

so that conditional upon a class 1 call having to wait, the waiting time is exponentially distributed with parameter $c\mu - \lambda_1$. We immediately conclude that for $w \geq 0$,

$$P(W_1 > w) = \eta e^{-(c\mu - \lambda_1)w}. \tag{1}$$

For $i \geq 2$, it does not appear to be possible to analytically invert the LST $\tilde{W}_i$, so that another method is needed to determine tail probabilities of the form (1) for these customers.

Abate and Whitt (1995) provide algorithms for numerically computing tail probabilities such as (1) from LSTs. One could certainly use their algorithms to determine the required tail probability for class i customers ($i > 1$). These algorithms require complex arithmetic, which is not likely to be available to the average call centre manager through the spreadsheets that they typically use. In particular, we were using Microsoft Excel at the PCC, which does not have built-in support for complex arithmetic.

Markov's inequality (see, for example, p. 74 of Billingsley 1986) for a non-negative random variable X and constants $x, \alpha > 0$ states that

$$P(X \geq x) \leq \frac{EX^\alpha}{x^\alpha}.$$

Hence, we may immediately conclude that for $i \geq 2$,

$$P(W_i > w) \leq \min\{\frac{EW_i}{w}, \frac{EW_i^2}{w^2}\}. \tag{2}$$

Equation (2) provides us with an upper bound on the tail probabilities of W_i, and hence a lower bound on $P(W_i \leq w)$ for any $w > 0$. In particular, in the PCC application, if $w = 30$ seconds, then we obtain a lower bound on the fraction of class 2 calls that wait for 30 seconds or less in the call centre. This bound is not the exact value that we would prefer, but it is a reasonable indication of service level for a given number of servers.

It is worth noting that we could use several other well-known inequalities, such as Chebyshev's inequality, or the Chernoff bound

$$P(W_i > w) \leq e^{-\theta w} E(e^{\theta W_i}),$$

valid for any $\theta > 0$ such that the expectation exists. None of these inequalities dominates the other, so that it may be worthwhile considering several of these in the expression (2). However, we found that (2) was sufficient for our purposes.

Now, assuming that the arrival rates λ_i, $i = 1, \ldots, p$ of calls to the call centre are fixed, one can use these results to determine the number of servers c required to ensure that the call centre service requirements are met. Indeed, one simply increases c until (1) with $w = m_1$ seconds falls below $(100 - p_1)\%$, and (2) with $w = m_i$ falls below $(100 - p_i)\%$ for $i \geq 2$. Since the right-hand side of (2) is a bound, and not the exact value, it follows that this process is conservative, in that the required service levels will certainly be met (under the distributional assumptions of the model).

An alternative approach is to ignore the class i requirements ($i > 1$) for the purposes of setting staffing levels, and simply report the bound generated by (2). Here, we increase c until (1) with $w = m_1$ seconds falls below $(100 - p_1)\%$. This procedure will not guarantee that the performance requirements will be met for low priority customers. However, if the lower bounds generated by (2) are not too far from their targets, then we can be reasonably confident of coming close to these goals.

This was the approach we adopted in setting staffing levels for the PCC. We ignored the requirement for class 2 (non-emergency) calls in setting the staffing levels. The bounds that we generated showed that typically, at least 70% or so of class 2 calls would be answered within 30 seconds, whereas the requirement was 80%. Our contacts at the PCC were very happy with this conclusion.

3 The Impact of a Random Arrival Rate

It is reasonable to expect that the same staffing plan will be used at the PCC on Mondays through Wednesdays, partly to keep the complexity of the staff schedule low, and partly because of the similar nature of the demand for calls on these days. Since we will use a single staffing level to cover demand at a variety of times, it is worth exploring whether the data support the notion of a common arrival rate at such times.

As part of our analysis of the data, we considered the number of calls received between 8am and 9am on Mondays through Wednesdays from March 2 1998 through August 31 1998. This is a period of 26 weeks, and so there were a total of 78 observations. We then computed the mean of all 78 observations, and tested the null hypothesis that the data were Poisson with the computed mean. If the arrival process of calls to the PCC follows a nonhomogeneous Poisson process where the arrival rate function is the same on Mondays, Tuesdays and Wednesdays, then one would expect to not reject the null hypothesis.

A chi-squared test using 9 bins gave a test statistic of 60.2. The cutoff value at the 99% level for a chi-squared test with 8 degrees of freedom is 20.1. Therefore, the null hypothesis was resoundingly rejected.

This result suggests that we may need to explicitly consider a random arrival rate in computing staffing levels. Whether the result is caused by nonstationarities in the data, or a random arrival rate, or perhaps some other mechanism, the fact remains that we wish to use the same staffing levels from 8am - 9am on all 3 days. Therefore, it is reasonable to model the arrival of calls during this period as a Poisson process with an arrival rate that varies from day to day. The mechanism that drives this variation in arrival rate is due to weather and other factors that will almost certainly be unknown at the time that staffing level decisions are being made. Hence, we may view the arrival rate as being random.

This same "random arrival rate" phenomenon arises when there is considerable uncertainty in one's forecasts of the future arrival rate of calls, perhaps because of a change in business practices. This might occur with the introduction of a new product, for instance. Furthermore, if one only has a limited amount of data with which to predict the arrival rate, one could view the arrival rate estimate as being somewhat uncertain, and so once again, we can view the arrival rate as being random; see Section 3 of Whitt (1999).

Most call centre planners do not explicitly consider such issues. Therefore, it is worth exploring the potential impact of a random arrival rate on staffing level decisions, when the arrival rate is assumed to be deterministic. To supply as general a result as possible, we first consider a relatively abstract framework. In Section 4 we will consider more concrete examples, and in Section 6, we will discuss how these results apply to the PCC problem.

Let the real-valued random variable Λ denote the random arrival rate to the system of interest. Let $f(\lambda)$ denote the (assumed deterministic) service performance of the system conditional on $\Lambda = \lambda$. For example, $f(\lambda)$ might represent the probability that a customer waiting time is less than some prespecified bound. Clearly, the function f may depend on other factors, such as service rate and number of servers, but we suppress this dependence in our notation for clarity. We can, and do, assume that f is decreasing in λ, so that performance deteriorates as the system becomes more heavily loaded. It is also reasonable to assume that performance is increasing in the resource levels provided, although we suppress this dependence in our notation. For the purposes of discussion, we assume that the call centre planner wishes to set resource levels so that service performance equals or exceeds some lower limit ℓ.

It is common to assume that the arrival rate is deterministic. We call this the "deterministic approach". In the deterministic approach, a deterministic arrival rate is estimated in some fashion, and then f is calculated at this value to determine an estimate of performance. Often, the arrival rate is estimated by averaging the observed number of arrivals in previous periods. The deterministic approach then amounts to estimating $E\Lambda$, and choosing the resource level so that $f(E\Lambda) \geq \ell$. However, when the arrival rate is random, the performance that is observed will be random, and equal to $f(\Lambda)$.

There are several ways that one might deal with this random performance. A highly conservative policy is to consider the worst possible case, and plan for that situation, i.e., to select resource levels so that $f(\lambda_{\max}) \geq \ell$. This might be the approach adopted in an emergency service call centre, for example.

A second approach is to attempt to plan so that the desired service level is achieved "in the long-run", i.e., when the performance on many days is averaged. If Λ_k is the (random) arrival rate on the kth day,

then the average performance over n days is

$$\frac{1}{n}\sum_{k=1}^{n} f(\Lambda_k). \tag{3}$$

The sample average (3) is known to converge under very general conditions on the stochastic process $(\Lambda_k : k \geq 1)$. All that is needed is that the process $(f(\Lambda_k) : k \geq 1)$ be ergodic. For example, this holds if the Λ_k's are i.i.d. and $f(\Lambda_1)$ has a finite mean, or if Λ_k is derived from the kth iterate of a positive recurrent Markov chain, and $Ef(\Lambda_\infty)$ is finite, where Λ_∞ is distributed according to the stationary distribution of the Markov chain; see Theorem 17.0.1 of Meyn and Tweedie (1993). Hence, in great generality, this second approach may be viewed as attempting to choose resource levels so that $Ef(\Lambda) \geq \ell$.

It is our view that "most" service providers believe that their planning is in the vein of the long-run average approach, namely, ensuring that $Ef(\Lambda) \geq \ell$. However, by assuming a deterministic arrival rate, they are, in effect, setting resource levels to ensure that $f(E\Lambda) \geq \ell$. The following result, known as Jensen's inequality, shows that if the function f is concave, then they will fall short of their target service level ℓ. For a proof, see p. 283 of Billingsley (1986).

Theorem 1 *Suppose that f is concave (convex) over an interval containing the range of Λ, and both $Ef(\Lambda)$ and $E\Lambda$ are finite. Then $Ef(\Lambda) \leq (\geq) f(E\Lambda)$.*

While Theorem 1 is easy to establish, its consequences may be important. If f is concave and the resource level is determined by ensuring that $f(E\Lambda) \geq \ell$, then the long-run service level $Ef(\Lambda)$ could fall short of the required minimum ℓ. Hence, the decision maker will provide too few resources in such a case. The question then arises as to how serious the shortfall in performance will be. This is related to how non-linear the function f is over the range of the random variable Λ. To this end we provide the following result.

Theorem 2 *Suppose that f is twice continuously differentiable over an interval N containing the range of Λ. Suppose also that there exist constants K_1 and K_2 such that $K_1 \leq f''(\lambda) \leq K_2$ for all $\lambda \in N$, and that $E\Lambda^2 < \infty$. Then*

$$\frac{K_1}{2}\,Var(\Lambda) \leq Ef(\Lambda) - f(E\Lambda) \leq \frac{K_2}{2}\,Var(\Lambda).$$

Proof: Expanding f in a Taylor series about $E\Lambda$ gives

$$f(\Lambda) - f(E\Lambda) = f'(E\Lambda)(\Lambda - E\Lambda) + f''(\xi)\frac{(\Lambda - E\Lambda)^2}{2!},$$

for some $\xi = \xi(\Lambda)$ lying between Λ and $E\Lambda$. Applying the bounds on f'' and then taking expected values gives the result.

Theorem 2 establishes bounds on the difference between $Ef(\Lambda)$ and $f(E\Lambda)$. If f is twice continuously differentiable and concave, then $f'' \leq 0$, so that we can take $K_1, K_2 \leq 0$. Theorem 2 then gives upper and lower bounds on the difference between predicted service $f(E\Lambda)$ and (long-run average) actual service $Ef(\Lambda)$. In particular, long-run performance is at most

$$f(E\Lambda) + \frac{K_2}{2}\mathrm{Var}(\Lambda),$$

and if $K_2 < 0$ then this is strictly less than the predicted level of service $f(E\Lambda)$.

If f is convex, then the predicted service level $f(E\Lambda)$ falls below the long-run average service level $Ef(\Lambda)$, so that resource decisions made based on the deterministic approach may be viewed as conservative. If f is twice continuously differentiable and convex, then $f'' \geq 0$, so that we may take $K_1, K_2 \geq 0$ in Theorem 2, and then we obtain bounds on how conservative we have been. In particular, the long-run performance is at least

$$f(E\Lambda) + \frac{K_1}{2}\mathrm{Var}(\Lambda),$$

and if $K_1 > 0$ then this is strictly greater than the predicted level of service $f(E\Lambda)$.

If f is both nonconcave and nonconvex over the range of Λ, then nothing can be easily asserted about the relative magnitudes of $f(E\Lambda)$ and $Ef(\Lambda)$. Therefore, it is of some interest to ask, for some well-used formulae, whether the function is concave, convex or neither, and this is the subject of the next section.

4 Some Typical Models

4.1 The $M/M/1$ Queue

The $M/M/1$ queueing system assumes that calls arrive according to a Poisson process with rate λ, and are served in FIFO (first in-first out) order by a single server. The service times are assumed to be independent of the arrival process, i.i.d., and exponentially distributed with parameter μ. This model is not typically used in setting staffing levels for call centres because of the assumption of a single server, but it is relevant in many other applications. Furthermore, it serves as an excellent vehicle for explaining the key ideas of the previous section.

In the $M/M/1$ queue, the choice of service capacity relates to the choice of the service rate μ. Service performance can be measured in several ways.

One approach is to say that service performance is satisfactory when the average time that customers wait in the queue before reaching service is below some threshold. It is well-known (see p. 251 of Wolff 1989 for example), that if W_q denotes the steady-state waiting time in the queue before reaching service and $\lambda < \mu$, then

$$P(W_q \leq w) = 1 - \frac{\lambda}{\mu} \exp(-(\mu - \lambda)w). \tag{4}$$

Hence, the expected waiting time in the queue is easily calculated to be

$$w_q(\lambda) = EW_q = \frac{\lambda}{\mu(\mu - \lambda)}.$$

Now, w_q is increasing in λ. In our general discussion in the previous section, we assumed that the merit function f was decreasing in λ, so set

$$f(\lambda) = -w_q(\lambda) = \frac{\lambda}{\mu(\lambda - \mu)}.$$

Now, $f'(\lambda) = -(\mu - \lambda)^{-2}$, so that f is indeed decreasing in $\lambda < \mu$. Furthermore, $f''(\lambda) = -2(\mu - \lambda)^{-3}$ which is negative for $\lambda < \mu$, so that f is concave in λ. Hence, it immediately follows from Theorem 1 that if λ is random, then one will overestimate service performance for a given $\mu > \lambda$ if the randomness of λ is ignored.

For example, if the service rate $\mu = 3$, and the arrival rate Λ is either 1 with probability $1/2$ or $\lambda_0 < 3$ with probability $1/2$, then the predicted expected waiting time in the queue will be

$$w_{q1}(\lambda_0) = \frac{1 + \lambda_0}{3(5 - \lambda_0)},$$

while the true expected waiting time in the queue will be

$$w_{q2}(\lambda_0) = \frac{1}{12} + \frac{\lambda_0}{6(3 - \lambda_0)}.$$

Observe that $w_{q2}(\lambda_0) - w_{q1}(\lambda_0) > 0$ for all $0 < \lambda_0 < 3$ with $\lambda_0 \neq 1$, and as $\lambda_0 \to 3$, the difference between these predictions gets arbitrarily large.

This example clearly shows that it is quite possible for the difference between predictions of service level to be significant.

A second choice of merit function might be the steady-state probability that the customer waiting time in the queue is less than or equal to w say. In this case, the merit function f is given by (4). Again, by direct calculation of the derivatives, one can show that f is decreasing and concave in $\lambda < \mu$, so that

Theorem 1 shows that if the arrival rate is random, then service performance will be overestimated for a given $\mu > \lambda$ if the randomness in λ is ignored.

Further results on convexity of performance measures associated with $M/G/1$ queues can be found in Rolski (1981), (1986). See Tu and Kumin (1983) and Weber (1983) for a proof that each customer's waiting time in a $GI/G/1$ queue is a convex function of the service rate. This result suggests that the expected steady-state waiting time, when it exists, should be convex in the arrival rate, agreeing with the result above that the negative of the expected steady-state waiting time is concave.

4.2 The $M/M/c$ Queue

This is the same as the $M/M/1$ model except that there are now $c > 1$ servers. Again, it is well known (see p. 256 of Wolff 1989) that if W_q is the steady-state waiting time in the queue (excluding service) and $\lambda < c\mu$, then

$$P(W_q \leq w) = 1 - P(W_q > 0) \exp(-(c\mu - \lambda)w), \tag{5}$$

where

$$P(W_q > 0) = 1 - p_0 \sum_{j=0}^{c-1} \frac{\lambda^j}{\mu^j j!},$$

and

$$p_0 = \left(\frac{(\lambda/\mu)^c}{(1 - \frac{\lambda}{c\mu})c!} + \sum_{j=0}^{c-1} \frac{\lambda^j}{\mu^j j!} \right)^{-1}.$$

Now, the expected waiting time in the queue before reaching service is, from (5),

$$w_q(\lambda) = EW_q = \frac{P(W_q > 0)}{c\mu - \lambda}. \tag{6}$$

Straightforward algebra shows that

$$P(W_q > 0) = \frac{\frac{\lambda^c}{c!\mu^c}}{\sum_{j=0}^{c-1} \frac{\lambda^j}{\mu^j j!}(1 - \frac{j}{c})}.$$

It is then easy to see (by dividing both numerator and denominator by λ^c) that $P(W_q > 0)$ is increasing in λ.

Hence, since $(c\mu - \lambda)^{-1}$ is increasing in λ, it follows from (6) that w_q is increasing in λ. So if, as in the $M/M/1$ case, the merit function f is the negative of (6), then f is decreasing in the arrival rate λ. Furthermore, the product of two non-negative convex functions that either both increase, or both decrease, is also convex, $(c\mu - \lambda)^{-1}$ has these properties, and Lee and Cohen (1983) showed that $P(W_q > 0)$ is increasing and convex in λ (for fixed μ). (See also Grassman 1983). Hence, the merit function $f = -w_q$ is concave as a function of λ.

As a second example, suppose that the merit function is given by (5) for a fixed value of w. This is the performance measure most commonly used in practice, namely, the proportion of customers who wait less than some prespecified time w. Then, as above, we can show that f is decreasing in λ, and since $P(W_q > 0)$ is convex in λ, f is concave in λ. We state these results as a proposition.

Proposition 3 *The expected steady-state waiting time in the queue EW_q is increasing and convex, and for any fixed $w \geq 0$, $P(W_q \leq w)$ is decreasing and concave as a function of the arrival rate λ for $0 < \lambda < c\mu$.*

The above discussion establishes that whether one is considering the expected steady-state waiting time in the queue or its distribution, the appropriate merit function is concave as a function of the arrival rate. Hence, from Theorem 1, it follows that if one assumes that the arrival rate is deterministic, when in fact it is not, then estimates of service performance will be overstated. As in the $M/M/1$ case, the error can be seen to be arbitrarily large.

We remark that other authors have considered convexity issues related to multiserver queues. In particular, Rolfe (1971) (Dyer and Proll 1977) examined convexity of $M/D/c$ ($M/M/c$) queues with respect to the number of servers. Weber (1983) gives an example demonstrating that the mean customer waiting time in the queue need not be a convex function of the service rate for the $GI/G/2$ queue. Harel and Zipkin (1987) showed that the mean customer sojourn time in the $M/M/c$ queue is jointly convex in the arrival and service rates.

4.3 The $M/G/c/c$ Queue

Our final example is the $M/G/c/c$ queue. This is a queueing system with c servers, exponentially distributed interarrival times with mean λ^{-1}, and i.i.d. service times with mean μ^{-1}. Customers that arrive when all c servers are busy are lost. The well-known Erlang-loss formula yields the steady-state probability p_e that an arriving customer finds an available server, and therefore is not lost. We shall take p_e as our merit function, so that

$$f(\lambda) = p_e = 1 - \frac{(\lambda/\mu)^c/c!}{\sum_{j=0}^{c}(\lambda/\mu)^j/j!}.$$

As in the $M/M/c$ case, it is straightforward to show that f is decreasing in λ. We then turn to the question of curvature of f. Observe that $f(\lambda) = h_c(\lambda/\mu)$, where

$$h_c(x) = 1 - \frac{x^c/c!}{\sum_{j=0}^{c} x^j/j!}.$$

Plots of the function h_c for $c = 1, c = 2$ and $c = 3$ are given in Figure 1 below. Values of $c > 3$ are similar to the $c = 3$ case.

Figure 1: The functions h_1, h_2 and h_3.

Except for the case $c = 1$, the curves appear to be initially concave, to pass through a point of inflection, and to then become convex. In fact, this was established in Harel (1990).

So suppose that the arrival rate is random, and equal to Λ. If $c = 1$, then the assumption that the arrival rate is deterministic will actually lead to conservative estimates of performance. However, the

Erlang loss formula is usually of more interest for multiple servers. Observe that for a given $c > 1$, if the range of Λ/μ is restricted to the region where h_c is concave (convex) then the assumption of a deterministic arrival rate will lead to optimistic (conservative) service performance estimates. If the range of Λ/μ includes both concave and convex regions of h_c, then nothing can be said about the relative magnitudes of predicted and actual performance.

It is worth noting that in systems where a high level of service is required, the number of servers will be chosen so that Λ/μ will be restricted to low values (relative to the number of servers). In this case, we can expect that service performance will be overestimated when the arrival rate is assumed to be deterministic, because of the initial concavity of the functions h_c. (See Harel 1990 for results relating to the exact location of the points of inflection of the functions h_c.)

5 Setting Staffing Levels in the Presence of a Random Arrival Rate

In the previous sections we looked at the potential impact of a random arrival rate on predictions of performance. In this section we consider the practical issues of how one detects, models, and accounts for a random arrival rate in setting staffing levels. We only assume that one can compute performance $f(\lambda)$ conditional on the arrival rate Λ taking the deterministic value λ. We are therefore not restricting attention to the models considered in the previous section. We are interested in computing the long-run average performance $Ef(\Lambda)$.

In the case where the randomness in Λ is due to uncertainty in arrival rate forecasts, the distribution of Λ reflects our uncertainty in the value of the "true" arrival rate. In this case, we may assume that the distribution of Λ is known. However, when the "randomness" in Λ is due to a random or nonstationary arrival rate, we need to infer the appropriate distribution for Λ from the data.

Assuming that one can infer the distribution of Λ, then the calculation or approximation of $Ef(\Lambda)$ is straightforward. Therefore the interesting questions here are how to detect and model a random arrival rate.

We assume that we have as data the number of arrivals (calls) A_k in each of n periods ($k = 1, \ldots, n$). The arrivals in period k are assumed to be generated by a Poisson process with rate Λ_k, where $\Lambda_1, \ldots, \Lambda_n$ are i.i.d. This model subsumes the nonrandom constant arrival rate case, in which case $\Lambda_k = \lambda_0$ for all k, for some λ_0.

If the arrival rate is non-random and fixed, then the A_k's will be i.i.d. Poisson random variables. One may apply standard tests to assess whether the A_k's are consistent with a Poisson distribution; see Section 6.6, p. 347 of Law and Kelton (2000) for instance.

If a test gives no reason to reject such a hypothesis, then one may quite reasonably view the data as consisting of Poisson distributed random variables with mean λ_0. An estimate of the arrival rate λ_0 is given by the sample mean $\bar{A}_n$. The central limit theorem establishes that $\bar{A}_n$ is approximately normally distributed with mean λ_0 and variance λ_0/n. We can interpret this result as saying that the value of λ_0 is uncertain, and the uncertainty in its true value is well-represented by a (truncated at 0) normal distribution with mean $\bar{A}_n$ and variance $\bar{A}_n/n$. The question of whether the estimation error (on the order of $\sqrt{(\bar{A}_n/n)}$) needs to be explicitly modeled depends on the performance function f. A conservative approach is to compute $Ef(\Lambda)$ where Λ assumes the truncated normal distribution given above. One can then compare the answer to the value $f(\bar{A}_n)$ that corresponds to the standard approach of assuming that $\bar{A}_n$ gives the true deterministic arrival rate. If the answers are, for all practical purposes, the same, then we can ignore the estimation error, i.e., we don't explicitly model a random arrival rate. This is the approach most commonly followed at present. If not, then we would use $Ef(\Lambda)$ as our estimate of performance.

If a test rejects the Poisson hypothesis, it does not necessarily mean that a random arrival rate is present. As mentioned earlier, there could be seasonality in the data, so that the arrival rate is varying with time in a deterministic fashion. However, if we are to use the same resource levels for all of the periods under consideration, then it is immaterial whether the variation in arrival rate is random in origin, or deterministic. In either case, we need to set resource levels to cope with periods of differing arrival rates.

Freedman (1962) looked at the problem of estimating the distribution of Λ in a context very similar to ours. Unfortunately, his method only applies in the situation where one has many observations of the number of arrivals for each realization of the random variable Λ. We have but one observation A_k for each independent realization Λ_k of Λ, so that his method cannot be used.

We will consider a simple parametric approach to estimating the distribution of Λ. In particular, suppose that Λ has distribution function F_θ, where θ is a finite set of parameters for the distribution. For example, in the situation considered by Thompson (1999) where Λ is assumed to be normally distributed, θ might consist of the mean and variance of the distribution.

With this framework, one could follow the standard maximum likelihood approach to estimating θ. However, the likelihood function takes on an imposing form, so that this approach would require some form of numerical optimization. We instead consider a method of moments approach that is very easy to implement.

It is easily shown that the kth factorial moment $EA^{\underline{k}}$ of a Poisson random variable A with mean λ is given by

$$EA^{\underline{k}} = EA(A-1)(A-2)...(A-k+1) = \lambda^k.$$

Thus, $EA_1^{\underline{k}} = EE(A_1^{\underline{k}}|\Lambda_1) = E\Lambda_1^k$. One can then fit p parameters by equating the first p sample factorial moments of $A_1, A_2, ..., A_n$ with the first p central moments $E\Lambda_1, E\Lambda_1^2, ..., E\Lambda_1^p$. This approach is similar to that discussed in Section 3 of Whitt (1999).

For example, suppose we assume, as Thompson (1999) did, that the arrival rate Λ is normally distributed with mean γ and variance σ^2. It is then possible that Λ takes on negative values, or values greater than the service rate, so we take the usual approach of fitting the parameters of this distribution ignoring such a possibility, and then use the truncated distribution in practice.

In this case, $E\Lambda = \gamma$, and $E\Lambda^2 = \sigma^2 + \gamma^2$. We may estimate γ by

$$\hat{\gamma} = \frac{1}{n}\sum_{k=1}^{n} A_k,$$

and σ^2 by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n} A_k(A_k - 1) - \hat{\gamma}^2.$$

At this point we have an estimate for the distribution of Λ and can now compute, or approximate, $Ef(\Lambda)$ as required. One can then increase staffing levels until this performance measure reaches a satisfactory level.

6 Conclusions

In this paper we have examined two issues.

The first relates to setting staffing levels in call centres with customers of different priority levels. We have explained how to use previously derived results for a priority queueing model to set staffing levels. One approach is to numerically invert the (known) LSTs of waiting time distributions. We did not go into details for this approach, because the required calculations are beyond the computing facilities typically available to call centre managers. We offered alternatives, based on well-known inequalities, that are easily computed within a spreadsheet environment. The alternatives give bounds on the results we require, and perform very well on the PCC problem.

The second issue involves the implications of a random arrival rate. A random arrival rate can be used to capture uncertainties in arrival rate forecasts, to capture nonstationary effects in the data, or indeed, to capture the case where the arrival process is a conditional Poisson process where the arrival rate is in fact random.

Our results indicate that it is important to explicitly model the random arrival rate if the underlying performance measure is highly nonlinear over the range of the random arrival rate. We looked at several performance measures related to delay systems, and showed that the presence of a random arrival rate will lead to overpredictions of service performance. In the case of loss systems, it is likely that the same

is true, although the reverse is possible in heavily loaded systems. Hence, if one ignores a random arrival rate, one will typically underestimate the number of staff required on hand to achieve a given performance level.

Of course, the main concern from a practical point of view is the *degree* of this effect. We supplied a practical method for detecting and modeling a random arrival rate, and described how to compute performance in this setting. The approach is very general, and in particular does not rely on the use of the very specific models considered in Section 4, nor does it rely on any convexity assumptions.

One can use this method in a simple "pilot study" to compare performance assuming a deterministic arrival rate, to that assuming a random arrival rate. We took exactly this approach in the PCC problem. We found that there were only a few hours in the week when the recommended number of call takers increased, and in such cases the change was only one extra call taker. The number of call-takers on duty at any given time in the PCC is typically on the order of 7 or 8. This effect may therefore be considered to be somewhat "second-order" for the PCC application, and we contented ourselves with using a conservative estimate of the call-taker service rate to help mitigate the random arrival rate effect. That is, the random arrival rate effect was not a primary consideration for the PCC problem.

Of course, this does not preclude the possibility that the random arrival rate effect might be important in other settings. In particular, these considerations will be important when there is considerable uncertainty in arrival rate forecasts to reasonably heavily loaded systems, and this is certainly a common situation in practice.

Acknowledgments

We would like to thank Michael Mann and Roly Williams of the Auckland Police Communication Centre for their support of this project. We would also like to thank the referees and guest editors for constructive criticism that substantially improved the presentation of our motivation for considering a random arrival rate. This research was partially supported by New Zealand Public Good Science Fund grant number UOA 803.

References

Abate, J., and W. Whitt. 1995. Numerical inversion of Laplace transforms of probability distributions. *ORSA Journal on Computing* **7** 36–43.

Billingsley, P. 1986. *Probability and Measure, 2nd ed.* Wiley, New York.

Chick, S. 2001. Input distribution selection for simulation experiments: accounting for input uncertainty. *Operations Research*, to appear.

Davis, R. H. 1966. Waiting time distribution of a multi-server priority queueing system. *Operations Research* **14** 133–136.

Dyer, M. E., and L. G. Proll. 1977. On the validity of marginal analysis for allocating servers in $M/M/c$ queues. *Management Science* **23** 1019–1022.

Freedman, D. A. 1962. Poisson processes with random arrival rate. *Annals of Mathematical Statistics* **33** 924–929.

Grassman, W. K. 1983. The convexity of the mean queue size of the $M/M/c$ queue with respect to the traffic intensity. *Journal of Applied Probability* **20** 916–919.

Green, L. V., P. J. Kolesar, and J. Soãres. 1999. Improving the SIPP approach for staffing service systems with cyclic demand. *To appear, Operations Research.*

Harel, A. 1990. Convexity properties of the Erlang loss formula. *Operations Research* **38** 499–505.

Harel, A., and P. Zipkin. 1987. Strong convexity results for queueing systems. *Operations Research* **35** 405–418.

Karlin, S., and H. M. Taylor. (1975). *A First Course in Stochastic Processes, 2nd ed.* Academic Press, Boston.

Kella, O., and U. Yechiali. 1985. Waiting times in the non-preemptive priority $M/M/c$ queue. *Communications in Statistics: Stochastic Models* **1** 257–262.

Law, A. M., and W. D. Kelton. 2000. *Simulation Modeling and Analysis, 3rd ed.* McGraw Hill, New York.

Lee, H. L., and M. A. Cohen. 1983. A note on the convexity of performance measures of $M/M/c$ queueing systems. *Journal of Applied Probability* **20** 920–923.

Meyn, S. P., and R. L. Tweedie. *Markov Chains and Stochastic Stability.* Springer-Verlag, New York.

Rolfe, A. J. 1971. A note on the marginal allocation in multi-server facilities. *Management Science* **17** 656–658.

Rolski, T. 1981. Queues with non-stationary input stream: Ross's conjecture. *Advances in Applied Probability* **13** 603–618.

Rolski, T. 1986. Upper bounds for single server queues with doubly stochastic Poisson arrivals. *Mathematics of Operations Research* **11** 442–450.

Ross, S. M. 1983. *Stochastic Processes.* Wiley, New York.

Thompson, G. M. 1999. Server staffing levels in pure service environments when the true mean daily customer arrival rate is a normal random variate. *Forthcoming.*

Tu, H. Y. and H. Kumin. 1983. A convexity result for a class of $GI/G/1$ queueing systems. *Operations Research* **31** 948–950.

Weber, R. R. 1983. A note on waiting time in single server queues. *Operations Research* **31** 950–951.

Whitt, W. 1999. Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Operations Research Letters* **24** 205–212.

Wolff, R. W. 1989. *Stochastic Modeling and the Theory of Queues.* Prentice Hall, Englewood Cliffs, New Jersey.