

Estimation of a Simple Queuing System With  
Units-in-Service and Complete Data<sup>1</sup>

James E. Prieger  
Department of Economics  
University of California  
One Shields Avenue  
Davis, CA 95616-8578  
(530) 752-8727

January 2, 2001

<sup>1</sup>This work is based in part on Chapter 2 of the author's dissertation at the University of California, Berkeley.

Suggested running head: *Estimation of a Simple Queuing System*

Corresponding address:

James E. Prieger  
Department of Economics  
University of California, Davis  
1 Shields Avenue  
Davis, CA 95616-8578

Abstract: Queuing theory may be useful for analyzing economic phenomena involving count and duration data. We develop maximum likelihood estimators for the time-varying parameters of a simple queuing system based on two kinds of data: complete interarrival and service times (IST), and number of units in service (NIS). The IST estimator dominates the NIS estimator, in terms of ease of implementation, bias, and variance. The model is useful for many empirical applications in economics.

Keywords:

$M(t)/M(t)/\infty$  queue, infinite server queue, time-varying parameters, Poisson stochastic process, duration data

# 1 Introduction

Queuing theory is useful when each of many events of interest is followed by a duration. For example, consider the study of labor contract strikes. One may be interested in the number of strikes beginning within a period, the number of strikes ongoing at a point in time, or the duration of individual strikes. Clearly these quantities are related, and a researcher may suspect that a covariate such as a change in labor law affects all three. Queuing theory provides a framework for unified analysis of the phenomenon. Other examples from economics include the analysis of the number and duration of visits to recreational facilities and the number and time to regulatory approval of patents or pharmaceuticals. Applications of queuing theory in economic literature include the study of equilibrium in service industries (Davidson, 1988), financial trading and continuous auctions (Brock and Kleidon, 1992), monetary aggregates and interest rates (Iwamura, 1992), congestion pricing at airports (Daniel, 1995), waiting lists for hospital treatment (Goddard, Malek and Tavakoli, 1995), day-care quality (Mulligan and Hoffman, 1998), and innovation and regulatory delay in telecommunications (Prieger, 2000*b*; Prieger, 2000*c*). Empirical applications of queuing theory in economics (e.g., Daniel, 1995; Prieger, 2000*b*; Prieger, 2000*c*) appear to be scarcer than theoretical studies (all the other citations above and many others). A possible reason for this may be that the queuing literature, received mostly from the field of operations research, focuses mostly on characterizing steady state quantities such as the average number of units in the system or the average time spent in the system instead of estimation of system parameters based on data likely to be available to an econometrician. This article discusses the adaptation of queuing theory to economics, presents a simple queuing model suitable for many economic phenomena, and discusses estimation using various data.

When adopting a queuing model, an immediate question for the empirical researcher is the type of data that are available or are to be collected. Most desirable are complete data on the interarrival and service times. We refer to these as *IST* data. Alternatively, only less-informative data may be available (or affordable) in some applications. For instance, one may know only the number of units in the system each period, not when

each arrives and exits. Call these *NIS* data (for Number In Service). Such data may arise in application of the model to unemployment, for example: one may observe the number of individuals unemployed each month, but not know when each spell begins or ends. More generally, NIS data arise whenever they are collected by census methods that report stock levels (e.g., pending stock trades, monetary aggregates, number of patients on a waiting list) and not flows.

We develop estimators for the  $M(t)/M(t)/\infty$  queue (described in the next section), the simplest system that allows the arrival and service time rates to evolve over time, a necessity for application to economic problems. We rely heavily on techniques drawn from the existing queuing theory literature, although we have not seen the explicit likelihoods presented for this model elsewhere. We develop estimators based on IST and NIS data and compare their ease of implementation, bias, and variance. The closed-form expressions for likelihoods based on both types of data are derived in the next two sections. The likelihood for the NIS data, which is non-trivial to derive, enables maximum likelihood estimation, which is more efficient than the usual moment-based estimation pursued in the queuing literature for such data (Bhat and Rao, 1986). The IST and NIS maximum likelihood estimators are compared in a Monte Carlo study in section 5; the IST estimator outperforms the NIS estimator. The IST estimator generally has smaller bias and variance.

## 2 Queuing Theory for Economists

Because queuing theory is novel to many economists, a brief primer will be useful. Queuing<sup>1</sup> theory was developed by telecommunications engineers (most notably A.K. Erlang) tackling the problem of designing telephone switching equipment adequate to handle the growing telephone traffic in the early 1900's. Operations researchers took over the vanguard of advancing the analytic development of queuing theory. The terminology of queuing theory (in particular, *arrivals* and *service times*) reflects the origins of the field. Cooper (1981) and Bunday (1996) provide more complete introductions.

---

<sup>1</sup>When searching the queuing literature, beware: most engineers and operations researchers use the orthographically inelegant term "queueing".

Kendall notation provides a compact description of a queuing system: an  $A/B/c$  system has interarrival time distribution  $A$ , service time distribution  $B$ , and  $c$  servers.  $A$  and  $B$  are chosen from a few traditional symbols such as  $M$  for the exponential (for its Markovian property) or  $E$  for the Erlang distribution.<sup>2</sup>  $M(t)$  denotes an exponential distribution with time-varying rate. In a queuing system, traffic arrives from the source with interarrival distribution  $A$ . Once in the system it waits in the queue, if any, until a server is open to receive it for servicing. Once the unit enters a server, its service time has distribution  $B$ . Figure 1 depicts the  $M(t)/M(t)/\infty$  model studied here. A classic example of a physical queuing system is a toll booth.

The infinite server queuing model is attractive to work with because of its tractability. The infinite server assumption is distasteful for most physical applications of queuing theory (bank teller’s windows, computer networks, etc.) because processing capacity is typically limited. This disadvantage of the infinite server assumption in physical applications often does not pertain when applying the model more abstractly to economic problems. In many economic applications the notion of a finite number of servers is untenable, for at least two reasons. First, one most likely needs a model that admits the possibility of a very short delay time for any observation. It is clearly possible for regulatory delays (or strikes, or visits to a park) to finish almost as soon as they begin. Within the class of first in first out models, no finite server model will suffice; in such models the probability that an arrival will have to wait for a server is (eventually) non-zero. Second, since the phenomenon of interest often will not suggest a number of servers, one would have to either choose the number of servers on an *ad hoc* basis or attempt to estimate it, which may not be feasible.

Even within an infinite server model, however, one often would like to test for congestion effects in applications such as regulatory delay. There are two main approaches to incorporate congestion in an infinite server model. In the first method one includes covariates reflecting the system state, such as the number of recent arrivals or the number

<sup>2</sup>A fully complete description of the queuing model also contains a characterization of the source of potential arrivals, maximum queue capacity, and queue discipline (e.g., First In First Out or *FIFO*). The default assumptions are an inexhaustible source of potential arrivals, infinite queue capacity, and FIFO queuing.

of units in service, directly in the determination of the arrival or service time rates. In this approach care must be taken concerning the possible endogeneity of the covariates; see Prieger (2000*b*) for such a model and discussion. In the second method one allows the arrival and service time processes to be correlated through an unobserved bivariate heterogeneity term (Prieger, 2000*a*). In these congestion models the likelihood for the NIS data would be difficult or impossible to find, and moment-based estimation may be more promising. Consequently, given the focus here on the comparison between ML estimation with IST and NIS data, we do not consider congestion models further in this paper.

In most queuing literature outside of economics, the arrival and service rates are taken as primitive. In economic applications, we are often interested in the impact of covariates on these rates. For example, we might be interested in the impact of changes in a regulatory regime on the creation (arrival) of new drugs and the time until regulatory approval (service time). When covariates evolve over time, as is typical in econometric studies, a queuing system with time-varying parameters (TVP) is appropriate. Ignoring the evolution of covariates in the interarrival and service time durations by merely conditioning on the covariates at the *start* of the spell—as is often done—introduces bias into the estimation (Heckman and Singer, 1986). Conditioning on the initial value of the covariates is particularly inappropriate when a duration lasts a long time. Accordingly, we adopt the infinite server queue with TVP, allowing covariates to change the arrival and servicing rates contemporaneously. We study the simplest such TVP queue, the  $M(t)/M(t)/\infty$  queue, which has Poisson arrivals from an infinite source population, exponential service times, and immediate servicing of arrivals.

The seminal work on maximum likelihood estimation for the  $M/M/\infty$  queue is Beneš (1957), which assumes that IST data are available. That article, and much of the estimation literature since, assumes that the parameters of the system are stable, although Hantler and Rosberg (1989) and Daniel (1995) provide exceptions. For the NIS estimator, we use techniques based on methods developed for an  $M(t)/M(t)/1$  queue (Clarke, 1953; Clarke, 1956; Srivastava and Kashyap, 1982).

### 3 Estimation Based on Interarrival and Service Times

In this section and the next we construct the maximum likelihood (ML) estimators for the parameters of the system. In the  $M(t)/M(t)/\infty$  system, the arrival of units is a nonhomogeneous Poisson process, so that interarrival times are exponentially distributed with instantaneous rate  $\lambda(t)$  at time  $t$ . Service time of each arrival is exponentially distributed with instantaneous rate  $\mu(t)$ . The two processes are independent. In this section we assume that IST data are available. Then independence allows one to estimate the parameters of the arrival and service time processes separately; there is no efficiency gain from joint estimation. We now find the likelihood of the IST data from the exponential duration process with time-varying parameters.

Let  $t$  be the realization of the random duration variable  $T > 0$ . Let vector  $\mathbf{X}(t)$  be the value at time  $t$  of explanatory variables, and let  $\mathbf{X}(a, b)$  be its time-path on  $[a, b]$ . Vector  $\mathbf{X}$  is assumed to be exogenous in the sense of Lancaster (1990, p.28).<sup>3</sup> The hazard rate, cumulative density function (CDF), and probability density function (PDF) of  $T$  (conditional on  $\mathbf{X}(t)$  and a finite parameter vector  $\boldsymbol{\theta}$ ) are:

$$h(t; \boldsymbol{\theta}, \mathbf{X}(t)) \equiv \lim_{dt \rightarrow 0} \Pr[t \leq T < t + dt | T \geq t, \boldsymbol{\theta}, \mathbf{X}(t, t + dt)] \quad (1)$$

$$F(t; \boldsymbol{\theta}, \mathbf{X}(0, t)) \equiv 1 - \exp \left[ - \int_0^t h(s; \boldsymbol{\theta}, \mathbf{X}(s)) ds \right] \quad (2)$$

$$f(t; \boldsymbol{\theta}, \mathbf{X}(0, t)) \equiv h(t; \boldsymbol{\theta}, \mathbf{X}(t)) [1 - F(s; \boldsymbol{\theta}, \mathbf{X}(0, t))], \quad (3)$$

respectively. Note that only the contemporaneous value of  $\mathbf{X}(t)$  enters  $h(t)$  (not its entire path).

It is convenient to approximate continuously evolving covariates by step functions, since economic data are typically reported discretely (e.g., as annual or quarterly time series). When  $\mathbf{X}(s)$  is a step function on  $[0, t]$  with jumps at  $(t_1, t_2, \dots, t_I)$ , let  $h_i(s)$  be the hazard rate in period  $i$ . The first part of the complete duration of length  $t$ , interval  $[0, t_1)$ , takes place in the first period, interval  $[t_1, t_2)$  in the second period, and so on, until the duration is completed with interval  $[t_I, t]$  in period  $I + 1$ .

<sup>3</sup>When a time-varying covariate is not exogenous, the likelihood derived in this section is only a *partial likelihood* (Lancaster, 1990, sec.9.2.11). Estimates resulting from maximizing a partial likelihood behave like ML estimates: they are consistent and the usual estimate of the variance is consistent. Partial likelihood estimates are inefficient compared to maximization of the true likelihood.

We term a *straddling* duration to be one that begins in one period and ends in another.

From (3), a straddling duration has PDF:

$$f(t) = h_{I+1}(t) \exp \left[ - \int_0^{t_1} h_1(s) ds - \sum_{i=2}^I \int_{t_{i-1}}^{t_i} h_i(s) ds - \int_{t_I}^t h_{I+1}(s) ds \right] \quad (4)$$

(Peterson, 1986), where the dependence on  $(\boldsymbol{\theta}, \mathbf{X}(0, t))$  is suppressed in the notation.

Equation (4) can be rewritten as

$$f(t) = [1 - F_1(t_1)] \left[ \prod_{i=2}^I \frac{1 - F_i(t_i)}{1 - F_i(t_{i-1})} \right] \left[ \frac{f_{I+1}(t)}{1 - F_{I+1}(t_n)} \right], \quad (5)$$

where  $F_i$  and  $f_i$  represent densities under hazard  $h_i$ .

For exponential durations, (5) takes a particularly convenient form that enables one to handle straddling observations with “canned” estimation routines. The exponential distribution with rate  $\lambda$  has PDF and CDF

$$f(t) = \lambda \exp(-\lambda t) \quad (6)$$

$$F(t) = 1 - \exp(-\lambda t) \quad (7)$$

respectively, with constant hazard rate  $\lambda$ . From (6)–(7), (5) simplifies to:

$$f(t) = [1 - F_1(t_1)] \cdot \left( \prod_{i=2}^I [1 - F_i(t_i - t_{i-1})] \right) \cdot f_{I+1}(t - t_I). \quad (8)$$

The  $i$ th term of the first  $I$  terms is the likelihood of a censored duration of length  $t_i - t_{i-1}$  (where  $t_0 = 0$ ), and the final term is the likelihood of an uncensored duration of length  $t - t_I$ . So by splitting up the straddling observations and marking all but the last period’s portion as censored, one can use canned estimation routines. Even when the distribution is not exponential, (8) is an approximation to the true density and can be used to find preliminary estimates to use as starting values for a ML routine. The weaker the duration dependence in the true distribution, the better these starting values will be. The simplification of (5) follows from the memoryless property of the exponential distribution; the length of the spell up to time  $s$  does not affect the distribution of the spell after  $s$ .

Say that one has  $J$  original observations in the sample, where each straddling duration is counted only once. Let the  $j$ th observation have  $I_j$  terms after splitting at the points of

discontinuity, so that we have a sample of durations  $((t_{ji} - t_{j(i-1)})_{i=1}^{I_j})_{j=1}^J$ . For simplicity of notation, relabel these in order as  $(s_i)_{i=1}^N$ , where  $N = \sum_{j=1}^J I_j$ . Similarly, construct associated censoring indicators  $(d_i)_{i=1}^N$ , where  $d_i = 1$  if  $s_i$  is censored and 0 if not, and explanatory variables  $(\mathbf{X}_i)_{i=1}^N$ , the appropriate levels of  $\mathbf{X}(t)$ . Note that  $\mathbf{X}_i$  is a vector of constants. Censored observations will arise from sources other than splitting straddling observations: at the end of the observation period, ongoing interarrival and servicing spells are censored.

The log likelihood of  $(s_i | d_i, \mathbf{X}_i)_{i=1}^N$ , from (6)–(8), is then:

$$l_{\alpha}(\boldsymbol{\alpha}) = \sum_{i=1}^N \{d_i \log [1 - F_i(s_i)] + (1 - d_i) \log [f_i(s_i)]\} \quad (9)$$

$$= \sum_{i=1}^N [(1 - d_i) \log \lambda_i - s_i \lambda_i] \quad (10)$$

$$\lambda_i = \exp(-\mathbf{X}_i' \boldsymbol{\alpha}). \quad (11)$$

The rate parameter is modeled in (11) as a function of a single index of the explanatory variables, with  $\boldsymbol{\alpha}$  being the parameter of interest to be estimated. To present the likelihood of the IST data, let the above notation pertain to the interarrival times and let the service time data have analogous likelihood  $l_{\beta}$  given by replacing  $(\boldsymbol{\alpha}, \lambda)$  with  $(\boldsymbol{\beta}, \mu)$ . Then, letting  $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$ , the joint likelihood of the IST data is

$$l_{\boldsymbol{\theta}}^{IST} = l_{\alpha} + l_{\beta}. \quad (12)$$

The asymptotic variance of the ML estimate of  $\boldsymbol{\theta}$  is:

$$V^{IST} = -[E(\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}'} l_{\boldsymbol{\theta}}^{IST})]^{-1} \quad (13)$$

$$= \begin{bmatrix} \sum_{i=1}^N (1 - e^{-\lambda_i a_i}) \mathbf{X}_i \mathbf{X}_i' & \mathbf{0} \\ \mathbf{0} & \sum_{i=1}^M (1 - e^{-\mu_i b_i}) \mathbf{Z}_i \mathbf{Z}_i' \end{bmatrix}^{-1}, \quad (14)$$

where  $a_i$  and  $b_i$  are the censoring times for the  $i$ th (split) interarrival and service time, respectively (see A.2 of the Appendix).  $V^{IST}$  can be consistently estimated by plugging in the ML estimates of  $\boldsymbol{\theta}$ .

## 4 Estimation Based on Number of Units in Service

Turn now to estimation using NIS data. Even if IST data are available for use in estimation, the likelihood for the NIS data derived here may be useful if the researcher is interested in predicting the mean or variance of the number of units in service at a given point in time [see Prieger (2000*b*), sec. 6, for such an application].

The time series of units in service contains less information than the IST data, and inference is accordingly less efficient. To see this, note that from the complete arrival and service calendar times, one can construct both the IST and NIS data. From the IST data, one can form sufficient statistics for the complete data (Beneš, 1957), but one cannot form such sufficient statistics from the NIS data alone. Therefore the IST data contains as much information as the complete arrival and service times, but the NIS data contains less. The NIS likelihood is also considerably more complicated than the IST likelihood.

We now derive the likelihood function for the NIS data, using techniques from the queuing theory literature (Clarke, 1953; Clarke, 1956; Srivastava and Kashyap, 1982).<sup>4</sup> Let  $N(t)$  be the random variable generating the number of units in the system (i.e., units that have arrived but not exited) at time  $t \in [t_0, t_0 + T]$ ,  $n(t)$  be a realization of  $N(t)$ , and  $n_i$  be the number of units in the system at the end of period  $i \in \{1, \dots, T\}$ . For simplicity each period is of unit length, so that  $n_i = n(t_0 + i)$ .

In this section we treat the arrival rate  $\lambda(t)$  and the servicing rate  $\mu(t)$  as constant within a period, so that  $\lambda(s) = \lambda_t$  and  $\mu(s) = \mu_t$  for  $s \in [t - 1, t)$ . The rates are deterministically related to explanatory variables as:

$$\lambda_t = \exp(-\mathbf{X}'_t \boldsymbol{\alpha}) \tag{15}$$

$$\mu_t = \exp(-\mathbf{Z}'_t \boldsymbol{\beta}), \tag{16}$$

where  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$  are vectors of parameters and  $(\mathbf{X}_t, \mathbf{Z}_t)$  are the appropriate levels of  $(\mathbf{X}(t), \mathbf{Z}(t))$  as before. The distinction between (15)–(16) and (11) is the timing of the periods. In section 3, we place no restrictions on the timing of the jumps in the step-function variable  $\mathbf{X}$ , so that the resulting “periods” need not match among observations.

<sup>4</sup>For a more advanced theoretical treatment of queues with time-varying parameters, refer to Brémaud (1981, section VI.2).

Here, we restrict the periods to be of equal length, and common to all observations. This restriction is not material, because one can always construct a uniform set of periods from any set of non-matching periods by appropriate subdivision.

From the properties of Poisson and exponential processes we have the following (where  $o(x)$  denotes order smaller than  $x$ ):

$$\Pr\{1 \text{ arrival in interval } (t, t + \Delta t)\} = \lambda(t)\Delta t + o(\Delta t) \quad (17)$$

$$\Pr\{0 \text{ arrivals in interval } (t, t + \Delta t)\} = 1 - \lambda(t)\Delta t + o(\Delta t). \quad (18)$$

For any particular server we have:

$$\Pr\{1 \text{ exit in interval } (t, t + \Delta t)\} = \mu(t)\Delta t + o(\Delta t) \quad (19)$$

$$\Pr\{0 \text{ exits in interval } (t, t + \Delta t)\} = 1 - \mu(t)\Delta t + o(\Delta t). \quad (20)$$

The probability of any compound event (e.g., an arrival and an exit) is  $o(\Delta t)$ .

From (17)–(20) one can derive the probability of the number of units in service at time  $t$ . Most TVP queuing studies focus on the limiting behavior of the system, but we are interested in the transient behavior; in application there is no reason to assume that the system is stationary (or even that the system is ergodic). We begin by deriving the likelihood for  $n_{t+1}$  given that  $N(t) = n_t$ . Suppress the dependence on  $t$  in the notation for  $\lambda$ ,  $\mu$ , and  $n$  for the moment. Then from (17)–(20) we can derive a recursive equation for the probability that there are  $n$  units in the system at time  $t$ . Let  $P_n(s)$  be the probability that  $N(s) = n$ . Then we can show that

$$\frac{d}{dt} P_n(t) = -P_n(t)(\lambda + n\mu) + P_{n+1}(t)(n+1)\mu + P_{n-1}(t)\lambda, \quad n \geq 0; \quad (21)$$

see (Kalashnikov, 1994, p.276). Add the initial condition

$$P_n(t_0) = \delta_{n_0 n} \quad (22)$$

where  $\delta_{n_0 n}$  is the Kronecker delta ( $\delta_{xy}$  equals 1 if  $x = y$  and 0 otherwise) and  $n_0 = n(t_0)$ . Equations (21)–(22) form a differential difference equation known as the *forward Kolmogorov equation*, which does not always have a closed form solution. This particular

case admits a solution, after employing several changes of variable and a generating function that reduce the problem to a linear partial differential equation.

To smooth out the piecewise continuous  $\lambda$  and  $\mu$ , we undertake the following changes of variable. Define

$$\tau \equiv \int_{t_0}^t \mu(s) ds \quad (23)$$

$$R(\tau) \equiv \frac{1}{\tau} \int_0^\tau \rho(\sigma) d\sigma \quad (24)$$

$$Q_n(\tau) \equiv P_n(\tau) e^{\tau[n+R(\tau)]} \quad (25)$$

where  $\rho \equiv \frac{\lambda}{\mu}$  is the *traffic intensity*. Differentiating (23), (24), and (25), we have:

$$d\tau = \mu(t) dt \quad (26)$$

$$R'(\tau) = \frac{1}{\tau} [\rho(\tau) - R(\tau)] \quad (27)$$

$$\frac{d}{d\tau} Q_n(\tau) = e^{\tau[n+R(\tau)]} [(n+1) P_{n+1}(\tau) + \rho P_{n-1}(\tau)] \quad (28)$$

where (28) follows from (21), (24), and (26). Making use of (25), (22) and (28) may be rewritten as the differential difference equation

$$Q_n(0) = \delta_{n0n} \quad (29)$$

$$\frac{d}{d\tau} Q_n(\tau) = \rho(\tau) e^\tau Q_{n-1}(\tau) + (n+1) e^{-\tau} Q_{n+1}(\tau), \quad (30)$$

where again  $Q_{-1}(\tau) \equiv 0$ .

Now introduce the generating function of the sequence  $\{Q_n(\tau)\}_{n=0}^\infty$ :

$$Q(z, \tau) \equiv \sum_{n=0}^{\infty} Q_n(\tau) z^n, \quad (31)$$

where  $z \in \mathbb{C}$ ,  $\|z\| < 1$ .  $Q(z, \tau)$  allows us to restate (29)–(30) as an initial value partial differential equation:

$$Q(z, 0) = z^{n_0} \quad (32)$$

$$\frac{\partial Q}{\partial \tau} = z \rho(\tau) e^\tau Q(z, \tau) + e^{-\tau} \frac{\partial Q}{\partial z}. \quad (33)$$

The solution to (32)–(33) is

$$Q(z, \tau) = (z - e^{-\tau} + 1)^{n_0} \exp[\tau R(\tau) + (z - e^{-\tau}) A(\tau)], \quad (34)$$

where  $A(\tau) \equiv \int_0^\tau e^\sigma \rho(\sigma) d\sigma$ , as shown in the Appendix (section A.1).

Now expand the first term and use the power series expansion of  $\exp[zA(\tau)]$  to rewrite (34) as

$$Q(z, \tau) = \exp[\tau R(\tau) - e^{-\tau} A(\tau)] \left[ \sum_{m=0}^{n_0} \binom{n_0}{m} z^m (1 - e^{-\tau})^{n_0-m} \right] \cdot \left[ \sum_{n=0}^{\infty} \frac{z^n A(\tau)^n}{n!} \right]. \quad (35)$$

$Q_n(\tau)$  is equal to the coefficient on  $z^n$  in  $Q(z, \tau)$  (Rudin, 1976, theorem 8.5), which is

$$Q_n(\tau) = \exp[\tau R(\tau) - e^{-\tau} A(\tau)] \sum_{m=0}^k \frac{A(\tau)^{n-m}}{(n-m)!} \binom{n_0}{m} (1 - e^{-\tau})^{n_0-m}, \quad (36)$$

where  $k \equiv \min\{n_0, n\}$ .

From (25) and (36), we find the probability of there being  $n$  units in the system at (rescaled) time  $\tau$ , given that there were initially  $n_0$  units in the system:

$$P_n(\tau) = \exp[-\tau n - e^{-\tau} A(\tau)] \sum_{m=0}^k \frac{A(\tau)^{n-m}}{(n-m)!} \binom{n_0}{m} (1 - e^{-\tau})^{n_0-m}. \quad (37)$$

We can now find the likelihood of an observation,  $n_t$ , conditional on its lagged value  $n_{t-1}$ . Let  $P_{n_t|n_{t-1}}(t) = \Pr\{N(t) = n_t | N(t-1) = n_{t-1}\}$ . Then from (23), (24), and (37) we have

$$P_{n_t|n_{t-1}}(t) = \exp[-\rho_t (1 - e^{-\mu_t})] \sum_{m=0}^{k_t} B_{mt}, \quad (38)$$

where  $k_t \equiv \min\{n_{t-1}, n_t\}$  and

$$B_{mt} \equiv \binom{n_{t-1}}{m} \frac{\rho_t^{n-m}}{(n-m)!} e^{-\mu_t m} (1 - e^{-\mu_t})^{n_t+n_{t-1}-2m}. \quad (39)$$

To find the joint likelihood of the data  $(n_t)_{t=1}^T$ , note that  $N(t)$  is a Markov process. Therefore  $f(n_t | (n_s)_{s=0}^{t-1}) = f(n_t | n_{t-1}) \forall t \in \{1, \dots, T\}$  and

$$f((n_t)_{t=1}^T | n_0) = \prod_{t=1}^T f(n_t | n_{t-1}),$$

where  $f$  is the density function appropriate for the arguments. The expression for  $f(n_t | n_{t-1})$  may be found from (38). The log joint density of the time-series data is

$$\log f((n_t)_{t=1}^T | n_0) = \sum_{t=1}^T \left[ -\frac{\lambda_t}{\mu_t} (1 - e^{-\mu_t}) + \log \sum_{m=0}^{k_t} B_{mt} \right]. \quad (40)$$

The log likelihood function for the parameter vector  $\boldsymbol{\theta}$ ,  $l_{\boldsymbol{\theta}}^{NIS}(\boldsymbol{\theta} | n_0, (n_t, \mathbf{X}_t, \mathbf{Z}_t)_{t=1}^T)$ , is determined from (15), (16), and (40).

Let  $\hat{\boldsymbol{\theta}}$  be the ML estimate obtained from maximizing  $l_{\boldsymbol{\theta}}^{NIS}$ . The asymptotic variance of  $\hat{\boldsymbol{\theta}}$  may, in principle, be found from (13), although the expression is complicated and not revealing. Because the asymptotic variance of  $\hat{\boldsymbol{\theta}}$  has infinite sums in the analytical expression, the ML variance is best estimated in practice by the BHHH estimator or other such methods that use only the gradient of (40). For similar reasons, maximization techniques that require only the gradient are an appealing choice here. We use a variant of the Davidon-Fletcher-Powell (DFP) algorithm in the Monte Carlo exercise below. The DFP algorithm has the advantage that it constructs an estimate of the ML variance as it calculates the parameter estimates.

The gradient of  $l$ ,  $\nabla_{\boldsymbol{\theta}} l_{\boldsymbol{\theta}}^{NIS}$ , may be found from (15), (16), and (40). We have:

$$\nabla_{\boldsymbol{\alpha}'} l_{\boldsymbol{\theta}}^{NIS} = \sum_{t=1}^T \left[ -\frac{1}{\mu_t} (1 - e^{-\mu_t}) + \left( \sum_{m=0}^{k_t} B_{mt} \right)^{-1} \cdot \sum_{m=0}^{k_t} B_{mt} \frac{n_t - m}{\lambda_t} \right] \nabla_{\boldsymbol{\alpha}'} \lambda_t \quad (41)$$

$$\begin{aligned} \nabla_{\boldsymbol{\beta}'} l_{\boldsymbol{\theta}}^{NIS} &= \sum_{t=1}^T \nabla_{\boldsymbol{\beta}'} \mu_t \left\{ \frac{\lambda_t}{\mu_t^2} [1 - (1 + \mu_t) e^{-\mu_t}] + \right. \\ &\quad \left. \left( \sum_{m=0}^{k_t} B_{mt} \right)^{-1} \left[ \sum_{m=0}^{k_t} B_{mt} \left( \frac{n_t + n_{t-1} - 2m}{e^{\mu_t} - 1} - m - \frac{n_t - m}{\mu_t} \right) \right] \right\} \\ \nabla_{\boldsymbol{\alpha}'} \lambda_t &= -\lambda_t \mathbf{X}_t' \quad (42) \end{aligned}$$

$$\nabla_{\boldsymbol{\beta}'} \mu_t = -\mu_t \mathbf{Z}_t' \quad (43)$$

where  $B_{mt}$  is defined as in (39).

## 5 Comparison of the NIS and IST Estimators

Analytic comparison of the variance of the NIS estimator with  $V^{IST}$  is problematic due to the different timing of the observations and the complexity of the NIS variance. The argument based on sufficient statistics given at the beginning of section 4 shows that

estimation based on NIS data will be less efficient, but gives no notion of the magnitude of the relative efficiency. Therefore, to compare the IST and NIS estimators, we conduct four Monte Carlo exercises. Each simulation consists of 1000 rounds of data generation and estimation with the NIS and IST data. The simulations vary in two ways: the parameters of the model and the number of periods. The queuing parameters are either *low intensity* (averaging fewer than one arrival per period and service time of less than a period) or *high intensity* (more than one arrival per period and service time of more than a period). The number of periods,  $T$ , is either 100 or 1000.

For all simulations, the model was taken to be the  $M(t)/M(t)/\infty$  queuing model with a common set of unit-length periods. The two ML estimators are based on  $l_{\theta}^{IST}$  for the IST data and  $l_{\theta}^{NIS}$  for the NIS data. In the design,  $\alpha$  and  $\beta$  [as defined in (15)–(16)] are  $2 \times 1$  vectors of parameters to be estimated and  $\mathbf{X}_t$  and  $\mathbf{Z}_t$  are  $2 \times 1$  vectors of explanatory variables.  $\mathbf{X}$  and  $\mathbf{Z}$  each include a constant and a variable generated from the Normal distribution. For the low-intensity simulations,  $X_{t2} \sim N(1.193, 1)$  and  $Z_{t2} \sim N(-0.193, 1)$ ; for the high-intensity simulations the distributions are reversed. The coefficients are taken to be  $\alpha = \beta = (0, 1)'$ . These values of  $\mathbf{X}$ ,  $\mathbf{Z}$ ,  $\alpha$ , and  $\beta$  imply that  $\lambda$  and  $\mu$  have means around 0.5 and 2, respectively, for the low-intensity simulations, and vice versa for the high-intensity simulations. The average interarrival time is  $1/\lambda$  and the average service time is  $1/\mu$  within a period.

One simulation consisted of the following:

1. Generate  $(\mathbf{X}_t, \mathbf{Z}_t)_{t=1}^T$  from the given distributions and fix for all rounds of the simulation.
2. Form  $(\lambda_t, \mu_t)_{t=1}^T$  from the explanatory variables.
3. For each of 1000 rounds, do the following:
  - (a) Generate pseudo-data (interarrival times, service times, and units in service per period) from the queuing model (taking  $n_0$  to be 0). Details are in A.3 of the Appendix.

- (b) Calculate the ML estimates of the IST model parameters by maximizing  $l_{\theta}^{IST}$  from (12).
  - (c) Calculate the ML estimates of the NIS model parameters by maximizing  $l_{\theta}^{NIS}$  from (40), using the same pseudo-data as in step 3b.
4. Calculate the sample bias and variance of the estimates.

The routines were implemented in Fortran with Numerical Recipes program `dfpmin`, the BFGS variant of DFP maximization (Press, Teukolsky, Vetterling and Flannery, 1992, sec. 10.7). The pseudo-data from each simulation are summarized in Table 1.

## 5.1 Simulations with the low-intensity Data

Results from the first two simulations, which use the low-intensity data, are summarized in Table 2. Recall that the sample size of the NIS data is  $T$ , the number of periods, but the sample size of the IST data will vary according to how many arrivals occur within the  $T$  periods. With the low-intensity parameters, the IST data average about  $T/2$  arrivals (see Table 1), and as many service time spells.

The IST likelihood is globally concave as long as the matrix of covariates is full rank, assuring rapid convergence to the maximum. Although we have not proven that the likelihood of the NIS data is globally concave [the Hessian of (40) is quite complicated] the algorithm generally converged to a maximum from a wide variety of starting points<sup>5</sup> and was globally concave in a few cross sections that we inspected.

From the results in Table 2, bias appears to be small with these sample sizes in both models, affecting no more than the second decimal place for any parameter. The accuracy of the NIS estimates is impressive, given that more than half of the dependent variables  $n_t$  are zero.

The higher information content of the IST data causes it to perform better than the NIS estimator. The bias of the IST estimates is less than that of the NIS estimates for six out of the eight parameters; the IST estimates average 17% less bias than the NIS

<sup>5</sup>As long as the routine does not experience floating point overflows, that is. The factorials and exponentials in the likelihood lead to numerical difficulties if the starting values are far away from the true values, or if the step sizes in the algorithm are too large.

estimates. The variance of the IST estimates is two to five times smaller than that of the NIS estimates. Both estimators perform measurably better with the greater number of periods. The bias and variance of the estimates from simulation 2 are an order of magnitude smaller than those from simulation 1.

## 5.2 Simulations with the high-intensity Data

The last two simulations use the data generated with the high-intensity parameters. The results are in Table 3, and are generally more favorable than those from the corresponding simulations with the low-intensity data. With one exception, the biases in simulations 3 and 4 are less than those of the same data type (NIS or IST) and period length from simulations 1 and 2. The variances are uniformly smaller with the high-intensity data. The reason is that the higher intensity data contains more information than the lower intensity data. To see this, note that if the parameters were constant over time, then the stationary distribution for the number of units in service would have mean equal to the traffic intensity,  $\rho$  [a standard result from queuing theory; see, for example, Tanner (1995, p.166)]. The low-intensity parameters imply that intensity  $\rho$  averages about 0.7; the high-intensity parameters imply that  $\rho$  averages about 10. This is why the NIS data in Table 1 have median near zero for the low-intensity data, but much higher for the high-intensity data. When there are many zeroes in the NIS data, the information content of another zero is low. When the NIS data show more variation, like in the high-intensity data, the parameters of interest are more easily identified. The message is the same for the IST data, although the reasoning differs. The higher intensity data contain more interarrival and service time observations, because there are more arrivals, allowing the estimation to be more precise.

As with simulations 1 and 2, the IST estimator again outperforms the NIS estimator in simulations 3 and 4. The absolute bias of the IST estimates is less than that of the NIS estimates for seven out of the eight parameters; the IST estimates average 28% less bias than the NIS estimates. The variance of the IST estimates is two to three times smaller than that of the NIS estimates. Both estimators again perform measurably better with

more periods; the bias and variance of the estimates from simulation 4 are an order of magnitude smaller than those from simulation 3.

## 6 Conclusion

The available data determine how one estimates a queuing model. We have presented two estimation methods, the IST model and the NIS model. The IST model requires more data but dominates the NIS model in terms of both ease and accuracy of estimation. The bias reduction of IST estimation over NIS estimation is higher when the data are from a high-intensity system; the variance reduction is generally highest when the data are from a low-intensity system.

The model assumes simple parametric (i.e., exponential) forms for the arrival and servicing processes, and should be viewed as a starting point for empirical work. The econometric question in each case is to determine, through specification testing, whether the parametric forms chosen adequately capture the variation in the data. Extension of the IST model for interarrival and service times following more general distributions is straightforward; see Prieger (2000*a*) for an example. Extending the NIS model similarly is much more difficult, given the heavy reliance on the memoryless properties of the exponential arrivals and service times in the derivation of the NIS likelihood.

The queuing model is useful for many economic questions. In any example where a quantity of interest is a count of events in progress at a given time, the present model may apply. For example, consider the number of ongoing strikes, or the number of people at a given time receiving public assistance, or receiving unemployment compensation, or visiting an attraction. The count is determined by the number of new arrivals and by the duration of each strike (or welfare tenure, unemployment spell, or tourist trip). The TVP queuing model may be appropriate for such cases.

## References

- Beneš, Vaclav E. (1957), ‘A Sufficient Set of Statistics for a Simple Telephone Exchange Model’, *Bell System Technical Journal* **36**, 939–964.
- Bhat, U. Narayan and Rao, S. Subba (1986), ‘Statistical Analysis of Queuing Systems’, *Queuing Systems: Theory and Applications* **1**(3), 217–247.
- Brémaud, Pierre (1981), *Point Processes and Queues: Martingale Dynamics*, Springer Series in Statistics, Springer-Verlag, New York.
- Brock, William A. and Kleidon, Allan W. (1992), ‘Periodic Market Closure and Trading Volume: A Model of Intraday Bids and Asks’, *Journal of Economic Dynamics and Control* **16**(3-4), 451–489.
- Bunday, Brian D. (1996), *An Introduction to Queuing Theory*, New York: Halsted Press.
- Clarke, A. Bruce (1953), ‘On Time-Dependent Waiting Line Processes (Abst.)’, *The Annals of Mathematical Statistics* **24**(3), 491–492.
- Clarke, A. Bruce (1956), ‘A Waiting Time Process of Markov Type’, *The Annals of Mathematical Statistics* **27**(2), 452–459.
- Cooper, Robert B. (1981), *Introduction to Queuing Theory*, 2nd edn, New York: North Holland.
- Daniel, Joseph I. (1995), ‘Congestion Pricing and Capacity of Large Hub Airports: A Bottleneck Model with Stochastic Queues’, *Econometrica* **63**, 327–370.
- Davidson, Carl (1988), ‘Equilibrium in Servicing Industries: An Economic Application of Queuing Theory’, *Journal of Business* **61**(3), 347–367.
- Goddard, J. A., Malek, M. and Tavakoli, M. (1995), ‘An Economic Model of the Market for Hospital Treatment for Non-Urgent Conditions’, *Health Economics* **4**(1), 41–55.
- Hantler, Sidney L. and Rosberg, Zvi (1989), ‘Optimal Estimation for an  $M/M/c$  Queue with Time Varying Parameters’, *Communications in Statistics. Stochastic Models* **5**(2), 295–313.

- Heckman, James J. and Singer, Burton (1986), *Econometric Analysis of Longitudinal Data*, in Z. Griliches and M. Intriligator, eds, 'Handbook of Econometrics. Volume III', Handbooks in Economics, book 2, Amsterdam; New York: North-Holland, pp. 909–938.
- Iwamura, Mitsuru (1992), 'The Determination of Monetary Aggregates and Interest Rates', *Monetary and Economic Studies* **10**(1), 65–93.
- Kalashnikov, Vladimir V. (1994), *Mathematical Methods in Queuing Theory*, Mathematics and its applications, Boston: Kluwer Academic Publishers.
- Lancaster, Tony (1990), *The Econometric Analysis of Transition Data*, Econometric Society Monographs No. 17, Cambridge: Cambridge University Press.
- Mulligan, James G. and Hoffman, Saul D. (1998), 'Daycare Quality and Regulation: A Queuing-Theoretic Approach', *Economics of Education Review* **17**(1), 1–13.
- Peterson, Trond (1986), 'Estimating Fully Parametric Hazard Rate Models with Time-Dependent Covariates', *Sociological Methods & Research* **14**(3), 219–246.
- Press, William H., Teukolsky, Saul A., Vetterling, William T. and Flannery, Brian P. (1992), *Numerical Recipes in FORTRAN: The Art of Scientific Computing*, 2nd edn, Cambridge: Cambridge U. Press.
- Prieger, James E. (2000a), A Correlated Queuing Model for Regulated Product Innovation and Introduction, unpublished manuscript, Department of Economics, University of California, Davis.
- Prieger, James E. (2000b), Regulation, Innovation, and the Introduction of New Telecommunications Services, Working Paper 00-08, Department of Economics, University of California, Davis.
- Prieger, James E. (2000c), Telecommunications Regulation and New Services: A Case Study at the State Level, Working Paper 00-11, Department of Economics, University of California, Davis.

Rudin, Walter (1976), *Principles of Mathematical Analysis*, International Series in Pure and Applied Mathematics, 3rd edn.

Srivastava, H.M. and Kashyap, B.R.K (1982), *Special Functions in Queuing Theory: and Related Stochastic Processes*, New York: Academic Press.

Tanner, Mike (1995), *Practical Queuing Analysis*, IBM McGraw-Hill Series, London: McGraw Hill.

## 7 Appendix

**A.1** To solve the PDE given in (33)–(32), find the simultaneous equations:

$$\frac{dQ}{-e^\tau \rho(\tau) z Q} = \frac{d\tau}{-1} = \frac{dz}{e^{-\tau}} \quad (44)$$

The second equation in (44) implies that  $z = e^{-\tau} + c_1$ , where  $c_1$  is an arbitrary constant.

The first equation in (44) therefore yields

$$\frac{dQ}{d\tau} = \rho(\tau) Q (1 + e^\tau c_1) \quad (45)$$

Making use of (24), this ordinary differential equation has solution

$$Q(z, \tau) = c_2 \exp [\tau R(\tau) + (z - e^{-\tau}) A(\tau)] \quad (46)$$

where  $A(\tau) \equiv \int_0^\tau e^\sigma \rho(\sigma) d\sigma$  and  $c_2$  is an arbitrary function  $\phi$  of  $z - e^{-\tau}$ . To determine  $c_2$ , use (46) to find that

$$\phi(z - e^{-\tau}) = z^{n_0} \Rightarrow \phi(w) = (w + 1)^{n_0} \Rightarrow \phi(z - e^{-\tau}) = (z - e^{-\tau} + 1)^{n_0} = c_2 \quad (47)$$

Thus the particular solution of (46) that matches the boundary condition incorporated into (47) is given by (34).

**A.2** Equation (13) is

$$V^{IST} = \left[ E \left( \begin{array}{cc} \sum_{i=1}^N s_i e^{-\mathbf{X}'_i \alpha} \mathbf{X}_i \mathbf{X}'_i & \mathbf{0} \\ \mathbf{0} & \sum_{i=1}^M r_i e^{-\mathbf{Z}'_i \beta} \mathbf{Z}_i \mathbf{Z}'_i \end{array} \right) \right]^{-1},$$

where  $N$  and  $M$  are the number of interarrival and servicing observations, resp., after splitting straddles. Let the fixed right censoring point for observation  $s_i$  be  $a_i$ , where

the observations are drawn from exponential model (6). The expectation is calculated as follows:

$$\begin{aligned}
E(s_i|\mathbf{X}_i) &= \Pr(d_i = 1|\mathbf{X}_i) a_i + \Pr(d = 0|x) \int_0^{a_i} t \frac{f_i(t)}{F_i(a_i)} dt \\
&= a_i e^{-\lambda_i a_i} + \left[ \frac{1}{\lambda_i} - e^{-\lambda_i a_i} \left( a_i + \frac{1}{\lambda_i} \right) \right] \\
&= \frac{1}{\lambda_i} (1 - e^{-\lambda_i a_i}),
\end{aligned}$$

whence follows (14).

**A.3** To generate the pseudo-data in Section 5, we did the following.

*Arrivals and Interarrival Times.* At the beginning of period  $t \in \{1, \dots, T\}$ , draw a deviate from the exponential distribution with rate  $\lambda_t$ . If the duration goes beyond the end of the period, censor at period's end and treat as a straddling interarrival observation. If the duration is completed before the period's end, then there is an arrival and the interarrival observation is not censored. Repeat until the period's end is reached.

*Exits and Service Times.* For each arrival still in the system in period  $t \in \{1, \dots, T\}$ , draw a deviate from the exponential distribution with rate  $\mu_t$ . If the duration goes beyond the end of the period, censor at period's end and treat as a straddling service time observation. If the duration is completed before the period's end, then there is an exit and the service time observation is not censored.

*Number of Units in Service.* There are no units in service at time 0. The number of units in service in period  $t$  is the sum of the number of units in service in period  $t - 1$  and the arrivals in period  $t$ , less the exits in period  $t$ .

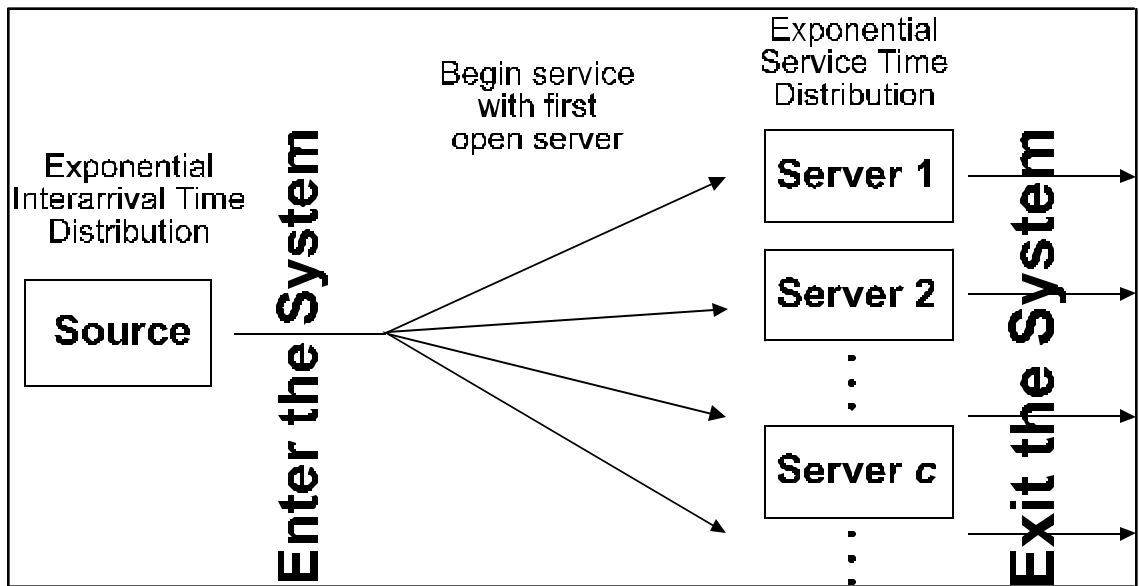


Figure 1: The  $M(t)/M(t)/\infty$  Queuing System

	Minimum	Mean	Median	Maximum
Low Intensity, $T = 100$				
Interarrival Time	1.24E-02	2.69	1.37	13.32
Arrivals/Period	0.00	0.58	5.00E-04	5.98
Service Time	1.04E-02	1.29	0.71	7.15
Number in Service ( $n_t$ )	0.00	0.49	0.00	5.05
Low Intensity, $T = 1000$				
Interarrival Time	4.38E-04	2.10	1.04	18.3
Arrivals/Period	0.00	0.51	0.00	41.7
Service Time	1.06E-03	0.88	0.49	9.27
Number in Service ( $n_t$ )	0.00	0.44	0.00	21.9
High Intensity, $T = 100$				
Interarrival Time	1.30E-3	0.47	0.19	4.38
Arrivals/Period	0.00	1.96	1.01	13.43
Service Time	1.42	2.23	1.25	18.82
Number in Service ( $n_t$ )	0.00	4.37	3.48	15.79
High Intensity, $T = 1000$				
Interarrival Time	8.25E-05	0.53	0.21	7.50
Arrivals/Period	0.00	2.03	1.00	35.68
Service Time	1.09E-03	2.78	1.33	45.48
Number in Service ( $n_t$ )	0.00	5.27	4.03	36.46

Table notes: Figures are averages over 1,000 simulation rounds. Data generated as described in text.  $T$  is the number of periods. Statistics are calculated from the IST data before splitting period-straddling observations.

Table 1: Summary of Queuing Model Pseudo-Data from the Simulations

Parameter	Simulation 1		Simulation 2		
	NIS MLE	IST MLE	NIS MLE	IST MLE	
	$T = 100$	$T = 100$	$T = 1,000$	$T = 1,000$	
$\hat{\alpha}_1$	bias:	5.60E-03	1.77E-02	6.70E-03	3.24E-03
	variance:	5.52E-02	1.85E-02	6.31E-03	2.01E-03
$\hat{\alpha}_2$	bias:	-6.90E-03	-5.73E-03	-5.70E-04	-8.77E-04
	variance:	3.31E-02	1.79E-02	1.94E-03	9.59E-04
$\hat{\beta}_1$	bias:	-3.42E-02	-5.12E-03	5.53E-03	7.75E-04
	variance:	9.62E-02	2.73E-02	8.10E-03	2.81E-03
$\hat{\beta}_2$	bias:	6.27E-02	1.28E-02	9.13E-03	1.10E-03
	variance:	1.08E-01	2.38E-02	8.38E-03	2.72E-03

Table notes: *NIS MLE* refers to ML estimation via the DFP method using the number of units in service per period data. *IST MLE* refers to the same using the interarrival and service time data. Each consists of 1,000 estimation rounds. *Bias* is the sample average of the parameter estimates' bias. *Variance* is the sample variance of the parameter estimates.

Table 2: Simulation Results—ML Estimates from the Low Intensity Data

Parameter	Simulation 3		Simulation 4		
	NIS MLE	IST MLE	NIS MLE	IST MLE	
	$T = 100$	$T = 100$	$T = 1,000$	$T = 1,000$	
$\hat{\alpha}_1$	bias:	7.43E-03	7.02E-03	1.32E-04	8.66E-05
	variance:	3.79E-02	1.29E-02	3.48E-03	1.24E-03
$\hat{\alpha}_2$	bias:	3.15E-03	3.48E-03	-2.88E-04	2.85E-04
	variance:	1.20E-02	5.90E-03	1.08E-03	5.90E-04
$\hat{\beta}_1$	bias:	-7.11E-03	-4.50E-03	8.24E-04	6.12E-05
	variance:	1.43E-02	6.39E-03	1.43E-03	6.89E-04
$\hat{\beta}_2$	bias:	7.40E-03	4.42E-03	-1.14E-03	-8.45E-04
	variance:	2.06E-02	7.49E-03	1.32E-03	5.78E-04

Table notes: see notes to Table 2.

Table 3: Simulation Results—ML Estimates from the High Intensity Data