

A Queueing Analysis to Determine How Many Additional Beds Are Needed for the Detention and Removal of Illegal Aliens

Yifan Liu

Department of Systems Engineering and Operations Research, George Mason University,
Fairfax, Virginia 22030, yliu9@gmu.edu

Lawrence M. Wein

Graduate School of Business, Stanford University, Stanford, California 94305, lwein@stanford.edu

Due to lack of detention capacity (the U.S. government measures capacity by the number of detention beds), tens of thousands of apprehended illegal aliens are released into the U.S. interior each year, instead of being removed from the country. This vulnerability can be exploited by terrorist groups wanting to enter the United States. We construct a queueing model of the U.S. detention and removal operations, and derive approximate analytical expressions for key performance measures, including a simple normal approximation for the required number of beds. Due to shortcomings in the U.S. government's data collection procedures, we cannot directly estimate all of the model's parameter values. Consequently, we use the approximate analytical expressions and the 2003 U.S. government data quantifying these key performance measures to estimate several unknown parameter values. Although current funding is for approximately 21,000 detention beds, we estimate that approximately 34,500 beds are needed to remove all potential detainees (this does not include nonviolent, noncriminal Mexicans, who are returned to Mexico within several hours) based on 2003 data. The dramatic increase in the arrivals of potential detainees since 2003 suggests that approximately 50,000 beds are currently required, although the estimation of future arrival rates is very difficult due to uncertainties about the future direction of U.S. immigration policy. Our estimated bed requirements are approximately 25% higher than naive estimates that fail to account for right censoring of residence times due to some detainees being released from detention before removal to make way for higher-priority detainees.

Key words: nonstationary queues; statistical inference; homeland security

History: Accepted by Linda V. Green, public sector applications; received on December 12, 2006. This paper was with the authors 3 weeks for 1 revision.

1. Introduction

In light of the September 11, 2001 attacks, concerns about the porous U.S.–Mexico border extend beyond immigration to homeland security (Turner 2004). In February 2005, the director of central intelligence and the FBI director both told the Senate Intelligence Committee that new intelligence strongly suggests that Al Qaeda has considered entering the United States illegally across the U.S.–Mexico border (Jehl 2005). When Mexicans, many of whom work in the U.S., are apprehended while illegally crossing the U.S.–Mexico border, they are typically (e.g., if they are nonviolent and with no serious criminal history) returned to Mexico within several hours without entering a detention facility; that is, nonviolent, noncriminal Mexicans are not viewed as potential detainees. However, so-called *other than Mexicans* (OTMs) caught illegally crossing the U.S.–Mexico border are supposed to be held in a detention facility until they can be removed from the United States and returned to their homeland. Several steps are required before the alien is removed,

including an appearance before an immigration judge, verification of identity, coordination with the home country, and purchase of airplane tickets.

Potential alien detainees are classified as mandatory or nonmandatory: as the adjectives suggest, mandatory detainees must be detained and removed, whereas nonmandatory detainees are removed if there are sufficient resources—in particular, detention beds—to do so. In 2003, approximately 28,000 aliens were apprehended while illegally entering the United States, but were nonetheless released into the interior of the United States (Bjerke 2004). An additional 43,000 illegal aliens were released from the Bureau of Immigration and Customs Enforcement's Office of Detention and Removal Operations (DRO) facilities before removal occurred (Bjerke 2004). Nearly all of the 71,000 released illegal aliens were OTMs. Although released illegal aliens are given a notice to appear in immigration court, only 13% of non-detained aliens with final removal orders are actually removed (U.S. Department of Justice 2003). The

underlying reason for the massive number of releases of nonmandatory detainees is lack of DRO capacity: DRO received funds for approximately 20,000 beds (Turner 2004), and in many cases these beds were already occupied or were needed for new mandatory detainees.

To mitigate this catch-and-release state of affairs, the Intelligence Reform and Terrorism Act of 2004 called for 8,000 new DRO beds per year during 2006–2010, essentially tripling DRO bed capacity from 20,000 to 60,000 (108th U.S. Congress 2004). However, a preliminary version of the 2006 budget funded 22,580 beds, which is an increase of only 1,920 DRO beds over 2005's budget (Callahan 2005, Ramanathan 2005), i.e., 24% of those called for in the Intelligence Reform and Terrorism Act. Moreover, as discussed in §6, this increase in the supply of DRO beds appears to be far outstripped by the recent increase in demand. To determine how many additional DRO beds are required, we construct a queueing model of the detention and removal operations, derive analytical expressions that accurately approximate the key performance measures (including a simple but accurate approximation for the required number of beds in Equation (38)), and estimate model parameters using DRO data (Bjerke 2004, Office of Immigration Statistics 2004). To our knowledge, there are no published queueing analyses of DRO.

If the model's parameter values were known, it would be straightforward to determine the additional number of beds required to remove all nonmandatory illegal aliens. However, there are several shortcomings in DRO's data collection system: The number of DRO beds is an elusive quantity (as explained in §2), and DRO data classifies detainees as criminal versus noncriminal, whereas the actual queueing discipline is based on the mandatory versus nonmandatory classification. Moreover, two input quantities are censored: Residence times in detention (i.e., the time from entering to exiting DRO) are right censored because many nonmandatory detainees are released—essentially bumped by mandatory detainees—before being removed from the United States; and the relative amplitude of the seasonality in DRO arrivals is not directly observable because many potential detainees are not detained during the peak season because DRO is full. DRO is a data-poor environment: Data on the arrival times and residence times of individual detainees are not collected, and hence we cannot use standard statistical techniques for analyzing censored data (e.g., the Kaplan-Meier method or the expectation-maximization algorithm) to address this problem. Consequently, we indirectly estimate these four unknown parameter values by deriving analytical expressions for four performance measures using a queueing model, and equating these mathematical

expressions to the actual performance values based on 2003 U.S. government data.

The study in this paper is part of a larger research agenda that investigates issues at the intersection of homeland security and immigration. In a more recent paper (Wein et al. 2007), we consider an optimization model that contains four submodels: a multinomial logit model for deciding whether potential crossers will attempt to enter the United States as a function of the probability of succeeding and the wage they would receive; a spatially nonhomogeneous Stackelberg model that predicts the apprehension probability at the border as a function of the number of border patrol agents, the arrival rate of crossers, and the fraction of the border that deploys surveillance technology; the queueing model for DRO that is presented in the present paper; and an equilibrium model that determines the illegal wage as a function of the number of worksite inspectors. The model is embedded into an optimization framework that chooses the number of border patrol agents, the amount of surveillance technology, the number of DRO beds, and the number of worksite inspectors to minimize the probability that a terrorist could successfully cross the U.S.–Mexico border.

We describe the queueing model in §2 and derive approximate analytical performance measures in §3. DRO data and the analytical results in §3 are used in §4 to estimate the parameter values for the model. Our main results are presented and discussed in §5, and concluding remarks are offered in §6.

2. The Queueing Model

In our queueing model, the potential alien detainees are the customers and the DRO beds play the role of servers. We assume that the bottleneck at DRO facilities is beds, not officers; currently, there are approximately nine detainees per officer (Tangeman 2003). The exact number of DRO beds, denoted by s , is unknown: although the U.S. government allocates an annual budget for DRO beds, which is used to operate eight large detention facilities and to contract beds at a variety of other facilities (Garcia 2004), DRO does not necessarily spend exactly the allocated amount (Turner 2004).

We make four simplifying assumptions for purposes of analytical tractability (pooled servers, Poisson arrivals, sinusoidal arrival rates, and exponential service times), and discuss the validity of these assumptions as we introduce them. In the actual system, if the closest DRO facility is full, then nearby DRO facilities are used if space is available. Nearly all detained aliens are apprehended along the U.S.–Mexico border (Office of Immigration Statistics 2004), which is also the location of the great majority of

detention facilities. Hence, transportation to nearby facilities is typically achieved via a several-hour bus ride. It is well known that by linking adjacent facilities in this way, the entire system behaves almost as if there was a single pooled queue that can serve all customers (Laws 1992). Hence, we consider a single pooled queue with s beds.

The model has two classes of customers representing mandatory ($i = 1$) and nonmandatory ($i = 2$) detainees. These detainees arrive to the DRO queue according to independent Poisson arrival processes. Although there are no arrival data to directly test this assumption, because most people sneak into the country alone or in small groups, the Poisson assumption is likely to be accurate because the superposition of many independent arrival processes behaves as a Poisson process (Cinlar 1972). To the extent that some people cross in small groups (typically led by coyotes, i.e., human smugglers), the actual arrival process would be more appropriately modeled as compound Poisson, and we would be underestimating the amount of variability; we investigate the impact of compound Poisson arrivals at the end of §5.4. Attempted border crossings exhibit a strong seasonal character (Bean et al. 1997, LeDuff and Flores 2005, discussed in more detail in §5.1), and we assume that the arrival rate for class i at time t is $\lambda_i(t) = \bar{\lambda}_i + \bar{\lambda}_i \alpha \sin(2\pi t/T)$, where $\bar{\lambda}_i$ is the average arrival rate, $T = 1$ year is the period, and α is the relative amplitude. We defer until §4 a description of the seasonal nature of the arrival data, which partially justifies the sinusoidal assumption. Also, because the future arrival rates are difficult to forecast, in §5.3 we explore the number of beds required as a function of the arrival rates.

The queue operates in the following manner (Hutchinson 2004). When a mandatory detainee arrives to the system, he is detained. If there is a DRO bed available, then he takes the bed; if all s DRO beds are occupied by mandatory detainees, then an additional bed is rented (Turner 2004) and he is moved into a DRO bed later if room becomes available (i.e., if a mandatory detainee in a DRO bed exits the system). If all DRO beds are occupied, but at least one nonmandatory detainee is being detained, then a nonmandatory detainee is released from DRO into the interior of the United States and the bed is given to the arriving mandatory detainee. In this case, we say that the nonmandatory detainee has been *preempted*. When a nonmandatory detainee arrives to the system (e.g., he is apprehended by a border patrol agent, who queries the DRO facility for a bed), then he is given a DRO bed if one is available. Otherwise, he is released into the interior of the United States and we say that the nonmandatory detainee has been *blocked*. The total

number of released detainees is the number blocked plus the number preempted.

As mentioned earlier, the DRO does not maintain data on individual residence times, which precludes us from directly estimating the residence time probability distributions. For analytical tractability, we assume that residence times in our two-class queueing system are independent and identically distributed (iid) exponential random variables with means m_1 and m_2 , which allows us to model the system as a nonstationary continuous-time Markov chain. Although many stationary loss queueing systems exhibit an insensitivity property to service-time distributions, this is not so when arrivals are nonstationary (Eick et al. 1993). However, the exponential assumption is not a gross misrepresentation of the available data (which are the mean residence times for Mexican criminals, Mexican noncriminals, OTM criminals, and OTM noncriminals, as reported in §4), given that there is a wide range of possible residence times (as explained in §4, residence times for Mexicans tend to be much smaller than for OTMs), which leads to a significant left tail, and there are two effects that prevent a fat right tail. First, the processing steps that are being performed while the detainees are in residence (e.g., seeing an immigration judge, making arrangements with the home country) typically follow the first-come first-served discipline within each detainee class; and second, legal restrictions prevent aliens from being detained indefinitely (U.S. Supreme Court 2001) (although a few countries, such as Cuba and Vietnam, refuse to take back their citizens, leading to some large residence times). Because this modeling assumption has less justification than our others, we perform a sensitivity analysis in §5.4 to assess the robustness of the exponential residence-time assumption.

Finally, although our analysis will reveal that DRO currently operates in an overloaded regime in which a fluid (i.e., nonstochastic) queueing model would have sufficed, we still need a stochastic queueing model to address the question posed in the title of this paper. In other words, a setting in which there are sufficient DRO beds to detain and remove illegal aliens will necessarily require some excess capacity, which can only be accurately analyzed with a stochastic queueing model.

3. Queueing Analysis

In this section, we derive approximations for the mean daily detention population throughout the year (Q), the mean number of blocked detainees during the year (B), the mean number of preempted detainees during the year (P), the mean number of released detainees during the year (R), and the ratio of the maximum-to-minimum monthly mean number

of unblocked detainee arrivals (M). These steady-state quantities are defined by

$$Q = \frac{\int_0^T Q(t) dt}{T}, \tag{1}$$

$$B = \int_0^T B(t) dt, \tag{2}$$

$$P = \int_0^T P(t) dt, \tag{3}$$

$$R = \int_0^T R(t) dt, \tag{4}$$

$$M = \frac{\int_{\pi/2-\pi/12}^{\pi/2+\pi/12} \lambda_1(t) + \lambda_2(t) - B(t) dt}{\int_{3\pi/2-\pi/12}^{\pi/2+\pi/12} \lambda_1(t) + \lambda_2(t) - B(t) dt}, \tag{5}$$

where $T = 1$ year, $Q(t)$ is the expected steady-state detention population at time $t \in [0, T]$, $B(t)$ is the steady-state blocking rate at time $t \in [0, T]$, $P(t)$ is the steady-state preemption rate at time $t \in [0, T]$ ($B(t)$ and $P(t)$ are in terms of detainees per unit time), and $R(t) = B(t) + P(t)$. In computing (5), we assume that the arrival rate of unblocked detainees achieves its extreme values at the same times as the total arrival rate of detainees, which is justified by the large fraction of arriving detainees that are mandatory.

To motivate our analysis, we start by suppressing the time notation on the arrival rates, and formulating a continuous-time Markov chain model of the two-class queueing system with generic arrival rates

λ_1 and λ_2 . A similar queueing system has been studied previously (Fischer 1980), but in that study, arriving Class 1 customers are blocked when no servers (i.e., beds) are available. Let Q_i be the steady-state number of mandatory ($i = 1$) and nonmandatory ($i = 2$) detainees in DRO, and let $P_{ij} = P(Q_1 = i, Q_2 = j)$ for $i = 0, 1, \dots$ when $j = 0$, and $i = 0, \dots, s; j = 1, \dots, s - i$. The global balance equations (Kelly 1979) for this continuous-time Markov chain (see Figure 1) are given by, where m_i^{-1} is the service rate for class i ,

$$\begin{aligned} &(\lambda_1 + \lambda_2 + im_1^{-1} + jm_2^{-1})P_{ij} \\ &= \lambda_1 P_{i-1,j} + \lambda_2 P_{i,j-1} + (i+1)m_1^{-1}P_{i+1,j} \\ &+ (j+1)m_2^{-1}P_{i,j+1} \end{aligned} \tag{6}$$

for $i = 0, \dots, s - 1$ and $j = 0, \dots, s - 1 - i$;

$$\begin{aligned} &(\lambda_1 + im_1^{-1} + (s-i)m_2^{-1})P_{i,s-i} \\ &= \lambda_1 P_{i-1,s-i} + \lambda_1 P_{i-1,s-i+1} + \lambda_2 P_{i,s-i-1} \end{aligned} \tag{7}$$

for $i = 0, \dots, s - 1$, and $j = s - i$;

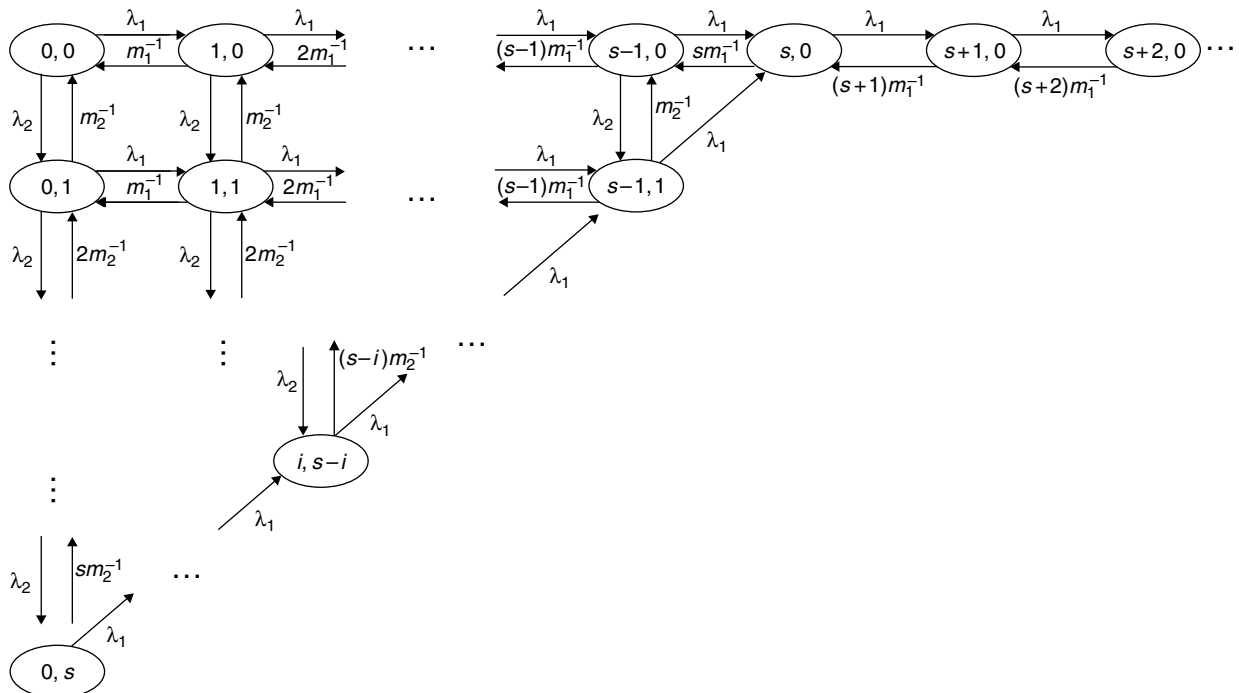
$$(\lambda_1 + sm_1^{-1})P_{s0} = \lambda_1 P_{s-1,0} + \lambda_1 P_{s-1,1} + (s+1)m_1^{-1}P_{s+1,0}; \tag{8}$$

and

$$(\lambda_1 + im_1^{-1})P_{i0} = \lambda_1 P_{i-1,0} + (i+1)m_1^{-1}P_{i+1,0} \tag{9}$$

for $i = s + 1, s + 2, \dots$, where by definition $P_{-1,j} = P_{i,-1} = 0$.

Figure 1 The State-Space Diagram for the Continuous-Time Markov Chain for $(Q_1(t), Q_2(t))$



Note. The global balance equations corresponding to this diagram appear in Equations (6)–(9).

One approach to approximating (1)–(5) is the pointwise stationary approximation (PSA) (Green and Kolesar 1991): For each value of time t throughout the year, compute $Q(t)$, $B(t)$, and $P(t)$ in (1)–(3) by performing a steady-state analysis of a stationary system that has time-homogeneous Poisson arrival processes with rates $\lambda_i(t)$. The accuracy of the PSA approximation depends on the relative frequency and the relative amplitude of the arrival processes. The relative frequency in our model, which is the average residence time divided by the period, is approximately 0.12 (Table 2), and the relative amplitude $\alpha = 0.1474$ (Table 2). Figure 2 of Eick et al. (1993) suggests that the maximum relative error is approximately 2%, and these errors tend to cancel each other out after integrating over the entire year as in (1)–(3) (Figure 3 of Eick et al. 1993).

The embedded steady-state analysis for $Q(t)$, $B(t)$, and $P(t)$ in the PSA requires that for each value of t we substitute $\lambda_i(t)$ for λ_i in (6)–(9) and solve these equations for $P_{ij}(t)$. A generalization of the recursions in Equations (4)–(6) of Fischer (1980) can be used to perform this task, but it requires the solution of a system of approximately $20,000^2/2 = 2 \times 10^8$ equations for each value of t . Hence, although this approach is likely to be quite accurate, it is cumbersome, particularly because in §4 we need to evaluate the queueing system for many values of the four unknown parameters to find the set of parameter values that best fits the data, and we embed the model's solution into a much larger fixed point system of equations in Wein et al. (2007).

Consequently, we use the following heuristic approximation, which maintains some aspects of the PSA, but allows us to approximate (1)–(5) in a more tractable manner. Because Class 2 customers are invisible to Class 1 customers, the total number of Class 1 customers behaves like the number of customers in an infinite-server queue with Poisson arrival rate $\lambda_1(t)$ and exponential service rate m_1^{-1} . By Equation (15) in Eick et al. (1993), the steady-state time-dependent queue length $Q_1(t)$ is a Poisson random variable with mean

$$n_1(t) = \bar{\lambda}_1 m_1 \left(1 + \frac{\alpha}{1 + (2\pi m_1/T)^2} \cdot \left(\sin\left(\frac{2\pi t}{T}\right) - \frac{2\pi m_1}{T} \cos\left(\frac{2\pi t}{T}\right) \right) \right). \quad (10)$$

Recalling that the steady-state time-dependent queue-length distribution is denoted by $P_{ij}(t) = P(Q_1(t) = i, Q_2(t) = j)$ (the time notation was suppressed in Figure 1 and Equations (6)–(9)), it follows that (Figure 1)

$$P_{i0}(t) = \frac{n_1(t)^i e^{-n_1(t)}}{i!} \quad \text{for } i = s, s+1, \dots, \quad (11)$$

$$P(Q_1(t) = i) = \sum_{j=0}^{s-i} P_{ij}(t) = \frac{n_1(t)^i e^{-n_1(t)}}{i!} \quad \text{for } i = 0, \dots, s-1. \quad (12)$$

Equations (10)–(12) are exact for all values of t .

Turning to Class 2 customers, we have that $P_{ij}(t) = P(Q_2(t) = j | Q_1(t) = i)P(Q_1(t) = i)$. The key step in our procedure is to approximate $P(Q_2(t) = j | Q_1(t) = i)$ for $i = 0, \dots, s-1$ by the time-dependent probability that there are j customers in an Erlang loss system (i.e., customers who cannot receive immediate service are blocked) with $s-i$ servers (which is the number of servers available to serve Class 2 customers when $Q_1(t) = i$), Poisson arrival rate $\lambda_2(t)$ (which is the instantaneous rate of Class 2 arrivals when $Q_1(t) = i$), and exponential residence times with mean m_2 . This approximation corresponds to analyzing each column in Figure 1 in isolation. To approximately analyze this time-dependent Erlang loss system, we use the modified offered load approximation in Massey and Whitt (1994), which approximates the probability of having j customers present by the conditional probability that the corresponding infinite-server system has j customers conditioned on having less than or equal to $s-i$ customers (which is the number of servers in this queue). Taken together, for $i = 0, \dots, s-1$ and $j = 0, \dots, s-i$, we approximate $P_{ij}(t)$ by

$$P_{ij}(t) = \frac{n_2(t)^j / j!}{\sum_{j=0}^{s-i} n_2(t)^j / j!} \frac{n_1(t)^i e^{-n_1(t)}}{i!}, \quad (13)$$

where (again by Equation (15) in Eick et al. 1993)

$$n_2(t) = \bar{\lambda}_2 m_2 \left(1 + \frac{\alpha}{1 + (2\pi m_2/T)^2} \cdot \left(\sin\left(\frac{2\pi t}{T}\right) - \frac{2\pi m_2}{T} \cos\left(\frac{2\pi t}{T}\right) \right) \right) \quad (14)$$

represents the mean number of Class 2 customers at time t in the corresponding infinite-server system (i.e., with Poisson arrival rate $\lambda_2(t)$ and exponential service times with mean m_2). If we let $\phi(x, n)$ and $\Phi(x, n)$ denote the Poisson probability mass function and cumulative distribution function (cdf), then (13) can be expressed as

$$P_{ij}(t) = \frac{\phi(n_2(t), j)}{\Phi(n_2(t), s-i)} \phi(n_1(t), i). \quad (15)$$

Recall that the time-dependent performance measures needed to compute (1)–(5) are $Q(t)$, $B(t)$, and $P(t)$. Starting with $Q(t)$, we have

$$Q(t) = n_1(t) + \sum_{j=1}^s j \sum_{i=0}^{s-j} P_{ij}(t), \quad (16)$$

where $n_1(t)$ is given in (10) and $P_{ij}(t)$ is given by (13)–(14). Equation (16) can be reduced to a single sum as follows:

$$\begin{aligned} Q(t) &= n_1(t) + \sum_{j=1}^s j \sum_{i=0}^{s-j} P_{ij}(t), \\ &= n_1(t) + \sum_{i=0}^s \sum_{j=1}^{s-i} j P_{ij}(t), \\ &= n_1(t) + \sum_{i=0}^s \sum_{j=1}^{s-i} j \frac{n_2(t)^j / j! e^{-n_2(t)}}{\Phi(n_2(t), s-i)} \phi(n_1(t), i) \\ &\quad \text{by Equation (15),} \\ &= n_1(t) + \sum_{i=0}^s \sum_{j=1}^{s-i} \frac{n_2(t)^j / (j-1)! e^{-n_2(t)}}{\Phi(n_2(t), s-i)} \phi(n_1(t), i), \\ &= n_1(t) + \sum_{i=0}^s \sum_{j=0}^{s-i-1} n_2(t) \frac{n_2(t)^j / j! e^{-n_2(t)}}{\Phi(n_2(t), s-i)} \phi(n_1(t), i), \\ &= n_1(t) + \sum_{i=0}^s n_2(t) \frac{\Phi(n_2(t), s-i-1)}{\Phi(n_2(t), s-i)} \phi(n_1(t), i). \quad (17) \end{aligned}$$

However, there are approximately 2×10^8 $P_{ij}(t)$ s to solve for, many of which have value near zero, which leads to problems of numerical instability. Therefore, for detainee class $i = 1, 2$, we approximate the Poisson random variables with mean $n_1(t)$ and $n_2(t)$ (given in (10) and (14)) by normal random variables with mean and variance $n_1(t)$ and $n_2(t)$ (Feller 1968); we denote the probability density functions (pdfs) of these normal random variables by f_{1t} and f_{2t} and the cdfs by F_{1t} and F_{2t} . Using Equations (13), (16), and (17), these substitutions lead us to approximate $Q(t)$ by

$$Q(t) = n_1(t) + n_2(t) \int_0^s f_{1t}(x) \frac{F_{2t}(s-x-1)}{F_{2t}(s-x)} dx. \quad (18)$$

Because our approximation procedure in (13)–(14) does not directly consider horizontal or diagonal transitions in Figure 1, its estimate for blocking actually incorporates blocking and preemption of the original system. That is, recalling that $R(t) = B(t) + P(t)$ is the rate at which detainees are released (via blocking or preemption) at time t , our estimate for $R(t)$ is the product of the instantaneous arrival rate of non-mandatory detainees and the probability that at least s detainees are in residence, i.e.,

$$R(t) = \lambda_2(t) \left(\sum_{i=0}^s P_{i,s-i}(t) + \sum_{i=s+1}^{\infty} P_{i0}(t) \right), \quad (19)$$

where $P_{ij}(t)$ is given in (11) and (13). Using the normal approximation to the Poisson, we approximate (19) by

$$\begin{aligned} R(t) &= \lambda_2(t) \left(\int_0^s \frac{f_{1t}(x) f_{2t}(s-x)}{F_{2t}(s-x)} dx \right. \\ &\quad \left. + 1 - F_{1t}(s+1) \right). \quad (20) \end{aligned}$$

Before proceeding, we note that the dimensionality problem inherent in Equations (6)–(9) was avoided by analyzing each column in Figure 1 in isolation and by using a normal approximation to the Poisson. Both of these approximations could have been applied to the PSA as well, by, e.g., dividing the year into months and replacing Equations (10) and (14) with the equivalent steady-state expressions, given the monthly mean arrival rates, and then using the remaining approximations. We do not pursue this alternative approach here because we do not have actual monthly arrival data (although we do present the monthly unblocked arrival data in §4) and because this approach is more tedious (albeit possibly more accurate if actual arrivals deviate significantly from a sinusoidal pattern).

Equation (20) is used to generate Figures 3, 5, and 6, where the number of beds is varied over a wide range. However, a cruder fluid approximation to (20) should be valid as long as (ignoring the dependence between $n_1(t)$ and $n_2(t)$ and recalling that $T = 1$ year)

$$\min_{t \in [0, T]} \left\{ n_1(t) + n_2(t) - 3\sqrt{n_1(t) + n_2(t)} \right\} > s, \quad (21)$$

i.e., as long as the mean minus three standard deviations of the queue length during the slowest time of the year (the left side of (21) has a unique minimum) is larger than the number of beds (we are subtracting three standard deviations from the mean, as is typical in many statistical problems). According to the fluid approximation, in which only the Class 2 customers are assumed to be in the fluid regime, the time-dependent Erlang loss system for Class 2 customers has an arrival rate of $\lambda_2(t)$ and can process detainees at an average rate of $(s - n_1(t))/m_2$, and the detainees who cannot be processed are released, which yields

$$R(t) = \lambda_2(t) - \left(\frac{s - n_1(t)}{m_2} \right). \quad (22)$$

Substituting Equation (10) into (22) and integrating gives a fluid approximation for R , the mean annual number of released detainees,

$$R = \int_0^T R(t) dt = \frac{\bar{\lambda}_1 m_1 - s}{m_2} + \bar{\lambda}_2. \quad (23)$$

Substituting the right sides of Equations (10) and (14) into condition (21) and performing the minimization reveals that approximation (23) is valid as long as

$$\begin{aligned} s &< \bar{\lambda}_1 m_1 + \bar{\lambda}_2 m_2 - \alpha \sqrt{c_1^2 + c_2^2} \\ &\quad - 3\sqrt{\bar{\lambda}_1 m_1 + \bar{\lambda}_2 m_2 - \alpha \sqrt{c_1^2 + c_2^2}}, \quad (24) \end{aligned}$$

where

$$c_1 = \frac{\bar{\lambda}_1 m_1}{1 + (2\pi m_1/T)^2} + \frac{\bar{\lambda}_2 m_2}{1 + (2\pi m_2/T)^2}, \quad (25)$$

$$c_2 = -\frac{2\pi \bar{\lambda}_1 m_1^2}{T(1 + (2\pi m_1/T)^2)} - \frac{2\pi \bar{\lambda}_2 m_2^2}{T(1 + (2\pi m_2/T)^2)}. \quad (26)$$

The accuracy of (23)–(24) is confirmed in §5. Because the 2003 DRO data is consistent with the fluid regime characterized by condition (24), we use (23) instead of (20) to estimate the parameters in §4.

Our final step is to approximate the individual contributions of blocking and preemption to $R(t)$. If one were to solve the global balance equations (6)–(9) (with $\lambda_i(t)$ in place of λ_i) for $P_{ij}(t)$, then the blocking and preemption rates would be given by

$$B(t) = \lambda_2(t) \left(\sum_{i=0}^s P_{i,s-i}(t) + \sum_{i=s+1}^{\infty} P_{i0}(t) \right), \quad (27)$$

$$P(t) = \lambda_1(t) \sum_{i=0}^{s-1} P_{i,s-i}(t). \quad (28)$$

There is a key difference between the right sides of Equations (19) and (27): The $P_{ij}(t)$ s in (19) are from our approximation procedure in (13), whereas the $P_{ij}(t)$ s in (27) are the exact (unknown) solutions to (6)–(9). Note that $\sum_{i=0}^s P_{i,s-i}(t)$ in (27) is almost identical to $\sum_{i=0}^{s-1} P_{i,s-i}(t)$ in (28) because $s = 21,136$ in Table 2. If we replace both of these sums by the unknown quantity $x(t)$, sum Equations (27) and (28), replace the left side of the summed equation by $R(t)$ because $R(t) = B(t) + P(t)$, and then solve the summed equation for $x(t)$, we get

$$x(t) = \frac{R(t) - \lambda_2(t) \sum_{i=s+1}^{\infty} P_{i0}(t)}{\lambda_1(t) + \lambda_2(t)}. \quad (29)$$

We substitute the right side of (29) into Equations (27) and (28) to get

$$B(t) = \lambda_2(t) \left(\frac{R(t) - \lambda_2(t) \sum_{i=s+1}^{\infty} P_{i0}(t)}{\lambda_1(t) + \lambda_2(t)} + \sum_{i=s+1}^{\infty} P_{i0}(t) \right), \quad (30)$$

$$P(t) = \frac{\lambda_1(t) [R(t) - \lambda_2(t) \sum_{i=s+1}^{\infty} P_{i0}(t)]}{\lambda_1(t) + \lambda_2(t)}. \quad (31)$$

Our final estimates for $B(t)$ and $P(t)$ are found by using the normal approximation to the Poisson in (30)–(31), which yields

$$B(t) = \frac{\lambda_1(t)\lambda_2(t)[1 - F_{1i}(s+1)] + \lambda_2(t)R(t)}{\lambda_1(t) + \lambda_2(t)}, \quad (32)$$

$$P(t) = \frac{\lambda_1(t)[R(t) - \lambda_2(t)(1 - F_{1i}(s+1))]}{\lambda_1(t) + \lambda_2(t)}. \quad (33)$$

To summarize, our time-dependent performance measures are given by $Q(t)$ in (18), $B(t)$ in (32), and $P(t)$ in (33), where $R(t)$ in (32)–(33) is given by (20) for the generation of Figures 3 and 4, and by the fluid approximation (22) during parameter estimation in §4. Calculating these performance measures for each value of t in $[0, T]$ and substituting into (1)–(5) give our final performance measures.

4. Parameter Estimation

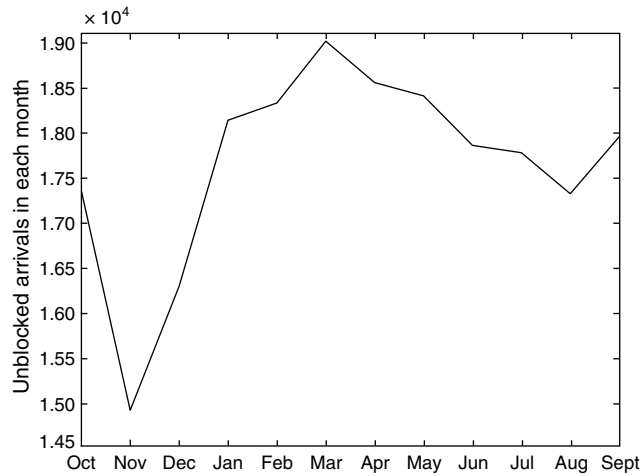
The raw DRO data cover (fiscal year) 2003 (Bjerke 2004, Office of Immigration Statistics 2004) and appear in Table 1. These data characterize detainees as Mexicans versus OTMs, and as criminals versus noncriminals. Although the actual queueing discipline is based on mandatory versus nonmandatory detainees (Hutchinson 2004), a breakdown of detainees by mandatory versus nonmandatory is not available (Bjerke 2004). Nearly all criminals are mandatory, and some noncriminals are mandatory. We assume that all criminals are mandatory and that an unknown fraction f of (Mexican and OTM) noncriminals are also mandatory.

Although we do not have monthly data for unblocked plus blocked detainee arrivals, we do know the monthly number of unblocked arrivals for 2003 (Bjerke 2004), which are shown in Figure 2. The busiest month has 27% more arrivals than the slowest month (Table 1). The seasonal pattern of unblocked arrivals is not exactly symmetric (although it does have a contiguous set of months above the mean and a contiguous set of months below the mean, the dips in November and December are larger than the increases in the peak months), but some of this

Table 1 The Raw Data, Which Is from Bjerke (2004) and the Office of Immigration Statistics (2004)

Description	Value
Total detained	231,432
Mexican criminals detained during 2003	72,720
Mexican noncriminals detained during 2003	47,251
OTM criminals detained during 2003	43,028
OTM noncriminals detained during 2003	68,433
Total average daily detention population	21,133
Mexican criminal average daily detention population	3,396
Mexican noncriminal average daily detention population	1,326
OTM criminal average daily detention population	9,290
OTM noncriminal average daily detention population	7,121
The maximum-to-minimum ratio of monthly detainee arrivals	1.27
The number of detainees released before entering DRO	28,000
The number of detainees released from DRO before removal	43,000

Notes. Although the Mexican average daily detention population was 4,722, the breakdown between criminal and noncriminal in the table is based on the fraction of the Mexican average daily detention population that was criminal during November 2004 (Bjerke 2004). The last two entries in the table are approximated to the nearest thousand (Bjerke 2004).

Figure 2 The Monthly Unblocked Arrivals During Fiscal Year 2003

asymmetry may be because most blocked arrivals occur in the peak months.

To estimate the average arrival rates $\bar{\lambda}_i$, we note that the number detained during 2003 (rows 1–5 of Table 1) equals the number detained on the first day of 2003 plus the number of new arrivals to DRO during the year. If we assume that the number detained on the first day of 2003 coincides with the average daily detention population (rows 6–10 of Table 1), then the annual unblocked arrival rates can be computed by subtracting rows 6–10 from rows 1–5, respectively. Hence, the annual unblocked arrival rates are 69,324 for Mexican criminals, 45,925 for Mexican noncriminals, 33,738 for OTM criminals, and 61,312 for OTM noncriminals. However, in keeping with our assumption that all nonmandatory detainees are noncriminals, we add the 28,000 blocked arrivals (the second-to-last row of Table 1) to the nonmandatory arrival rate of noncriminals. The 28,000 figure is a slight underestimate, because it only includes blocked arrivals from border patrol agents, but not from investigations (this data has not been forthcoming to DRO, Bjerke 2004). Recalling that an unknown fraction f of noncriminals are mandatory, we have that $\bar{\lambda}_1 = 103,062 + 135,237f$ and $\bar{\lambda}_2 = 135,237(1 - f)$.

Turning to the average residence times, we can use Little's Law (Little 1961), which states that the average detention population equals the average unblocked arrival rate times the average residence time; Little's Law should be accurate because the mean residence times are much smaller than the periodicity of the arrivals, which is one year. This calculation tells us that the overall average residence time is 36.7 days, and the average residence time is 17.9 days for Mexican criminals, 10.5 days for Mexican noncriminals, 100.5 days for OTM criminals, and 42.4 days for OTM noncriminals. However, the average residence time for noncriminals (and hence the overall average

residence time) is right censored, because 43,000 non-criminals who entered DRO were subsequently preempted. We assume that noncriminal detainees have the same (uncensored) average residence time regardless of whether they are mandatory or nonmandatory, because they do not experience additional delays caused by the criminal justice system. Because all nonmandatory aliens are noncriminal in our model, this unknown quantity is m_2 , which is the mean residence time for nonmandatory detainees. Then the average residence time for mandatory detainees is

$$m_1 = \frac{69,324(17.9) + 33,738(100.5) + 135,237fm_2}{103,062 + 135,237f} \quad (34)$$

Hence, we have four unknowns: the number of DRO beds (s), the fraction of noncriminals that are mandatory (f), the average residence time of nonmandatory detainees (m_2), and the relative amplitude of the arrival rates (α). We also have four output quantities—the average detention population over the year (Q), the number of nonmandatory detainees blocked during the year (B), the number of nonmandatory detainees preempted during the year (P), and the ratio of the maximum-to-minimum monthly mean number of unblocked detainee arrivals (M)—where the actual 2003 values appear in Table 1 and the predicted values are derived in terms of the four unknown variables from the queueing analysis performed in §3. Referring to the corresponding actual values in Table 1, we estimate the four unknown parameter values (s, f, m_2, α) by numerically solving the four equations (although we have been unable to prove that these equations have a unique solution, they were well behaved during our computations)

$$Q = 21,133, \quad B = 28,000, \quad P = 43,000, \\ \text{and } M = 1.27, \quad (35)$$

where the left sides of these equations are given in Equations (18), (32), (33), and (5), respectively.

5. Results and Discussion

In this section, we present and discuss the derived parameter values (Table 2) in §5.1, briefly discuss the time-dependent performance of the system in §5.2, turn to the central question of how many detention beds are needed in §5.3, investigate the robustness of the exponential service-time assumption in §5.4, and assess the impact of reduced residence times in §5.5.

5.1. Parameter Values

The solution to (35) is $f = 0.3051$, $m_2 = 48.0$ days, $s = 21,136$, and $\alpha = 0.1474$. Given these four parameter values, we can generate the four output variables $Q = 21,134$, $B = 28,008$, $P = 43,012$, and $M = 1.2701$,

Table 2 The Estimated Parameter Values for the Queueing Model

Parameter	Description	Value
f	Fraction of noncriminals who are mandatory	0.3051
$\bar{\lambda}_1$	Mean arrival rate of mandatory detainees	144,323 per year
$\bar{\lambda}_2$	Mean arrival rate of nonmandatory detainees	93,976 per year
m_1	Mean residence time of mandatory detainees	45.8 days
m_2	Mean residence time of nonmandatory detainees	48.0 days
s	Number of DRO beds	21,136
T	Period of sinusoidal arrival rates	1 year
α	Relative amplitude of sinusoidal arrival rates	0.1474

Notes. The parameters s , f , m_2 , and α were estimated by solving (35). The value of f was used to determine $\bar{\lambda}_1$ and $\bar{\lambda}_2$, and the values of f and m_2 were used to compute m_1 via Equation (34).

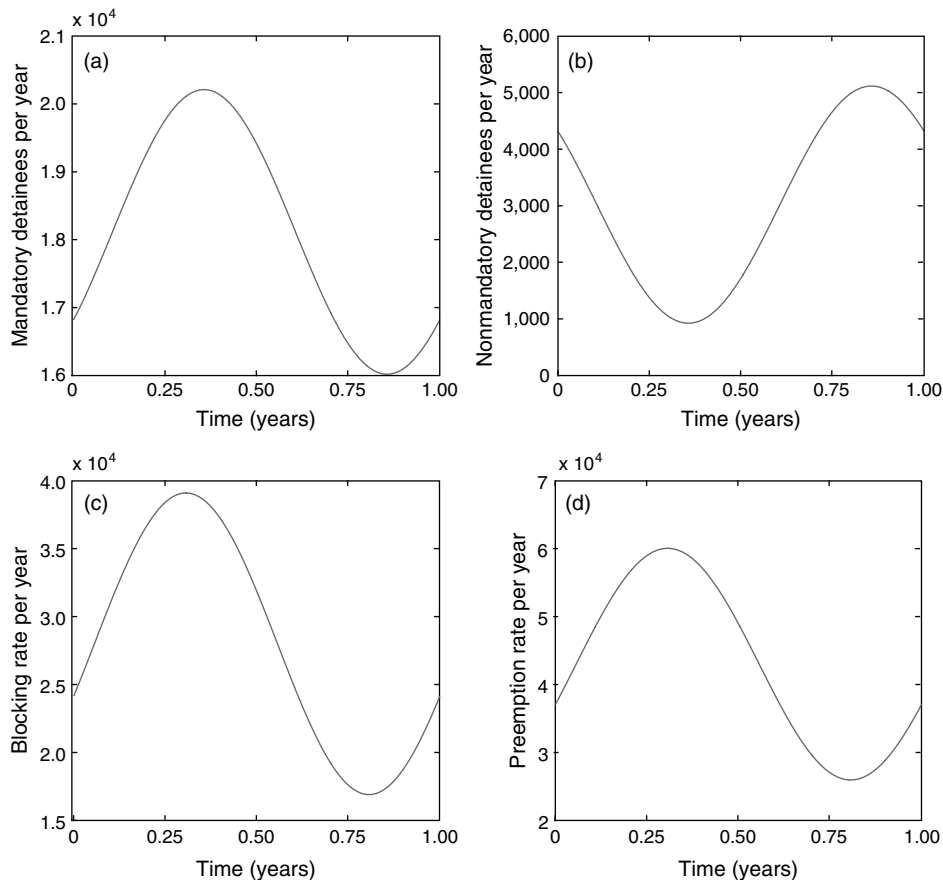
with the discrepancies between these numbers and those on the right sides of the equations in (35) being due to numerical errors arising from the nonlinearity of the equations. Moreover, for the parameter values in Table 2, the simulated values (10 years, with the first 2 years truncated) of the exact queueing system are $(Q, B, P, M) = (21,133; 27,987; 43,191; 1.2641)$, which are similar to the approximate values of $(21,134; 28,008; 43,012; 1.2701)$. These four parameter values

are consistent with the fluid regime: the right side of (24) is 26,480 with our estimated parameter values, which is significantly larger than $s = 21,136$.

These four values are also consistent with intuition and all known data. In our estimation procedure, the number of beds s (21,136) was exactly (ignoring roundoff errors) equal to the average daily detention population in Table 1. This result is not as trivial as it may appear. If the number of beds was smaller than the average number in residence, then only mandatory detainees would be present during the peak time of year. However, our model predicts that nonmandatory detainees are present in the system throughout the year; indeed, this prediction is confirmed by the fact that nonmandatory detainees were present in DRO throughout 2003 (Bjerke 2004). If the number of beds was significantly larger than the average number in residence, this would signal excess capacity. In contrast, we predict that even during the slowest time of year, all beds are filled (Figure 3).

The estimated fraction f of noncriminals that are mandatory is 0.3051, which together with our estimate of m_2 and Equation (34) implies that the mean residence time of mandatory detainees (m_1) is 45.8 days,

Figure 3 The Approximate Time-Dependent Performance Measures of the Queueing System Under the Base-Case Parameters in Table 2



Note. (a) Number of mandatory detainees ($Q_1(t)$); (b) Number of nonmandatory detainees ($Q_2(t)$); (c) Blocking rate of nonmandatory detainees ($B(t)$); (d) Preemption rate of nonmandatory detainees ($P(t)$).

and the mean number of DRO beds filled with mandatory detainees is $\bar{\lambda}_1 m_1 = 18,115$ (85.7% of all DRO beds), both of which appear plausible. The estimated average residence time for nonmandatory detainees (m_2) is 48.0 days, which is slightly larger than the estimated average residence time of mandatory detainees. The likely reason for this is that OTMs comprise a larger fraction of nonmandatory detainees (57.2%) than of mandatory detainees (38.6%), and some OTMs face additional delays due to paperwork transactions with the host country. The nonmandatory average residence time of 48.0 days is also larger than the observed residence times of noncriminals (10.5 days for Mexicans, 41.4 days for OTMs) because of the right censoring of the observed residence times.

Finally, the estimated value of α corresponds to a maximum-to-minimum ratio of potential arrivals of

$$\frac{\int_{\pi/2-\pi/12}^{\pi/2+\pi/12} (1 + \alpha \sin x) dx}{\int_{3\pi/2-\pi/12}^{3\pi/2+\pi/12} (1 + \alpha \sin x) dx} = 1.34.$$

As expected, the amount of seasonality in the blocked plus unblocked arrivals is larger than the amount of seasonality in the right-censored unblocked arrivals, where the ratio of maximum-to-minimum monthly arrivals was 1.27. However, monthly apprehension data from 1990 to 1995 at the U.S.–Mexican border (Bean et al. 1997, p. 276) suggest that the number of apprehensions in the peak month is approximately twice the number of apprehensions in the least active month, and more recent data (LeDuff and Flores 2005) suggest that the seasonality of apprehensions may now be even more extreme. If all border crossers—Mexicans and OTMs, criminals and non-criminals—followed the same seasonal pattern, and the probability of detention was constant throughout the year and independent of the type of border crosser, then the arrival process of potential detainees would exhibit the same pattern of seasonality as the border apprehension data, which clearly is not the case. Part of the reduction of seasonality from apprehensions to potential arrivals may be because most of the seasonality in border crossers is due to Mexican seasonal (e.g., agriculture, construction) workers who are not typically detained when apprehended. An alternative hypothesis is that border patrol agents put less effort into apprehending nonmandatory detainees during the peak season because they know it is less likely that there is room to detain them.

5.2. Time-Dependent Performance

For the derived set of parameter values, we display the approximate time-dependent performance measures for $(Q_1(t), Q_2(t), B(t), P(t))$ (Figure 3). Figures 3(a) and 3(b) confirm that $Q_1(t) + Q_2(t) = s$ for all t , and that during the peak time of year, approximately 95%

of the beds are occupied by mandatory detainees. As expected from theory (Eick et al. 1993, Equation (16)), there is a time lag of $T/(2\pi) \tan^{-1}(2\pi m_1/T) = 38.8$ days between the time of the peak arrival rate (0.25 year) and the time of the peak mandatory population (0.36 year), and hence of the times of the peak blocking and preemption rates.

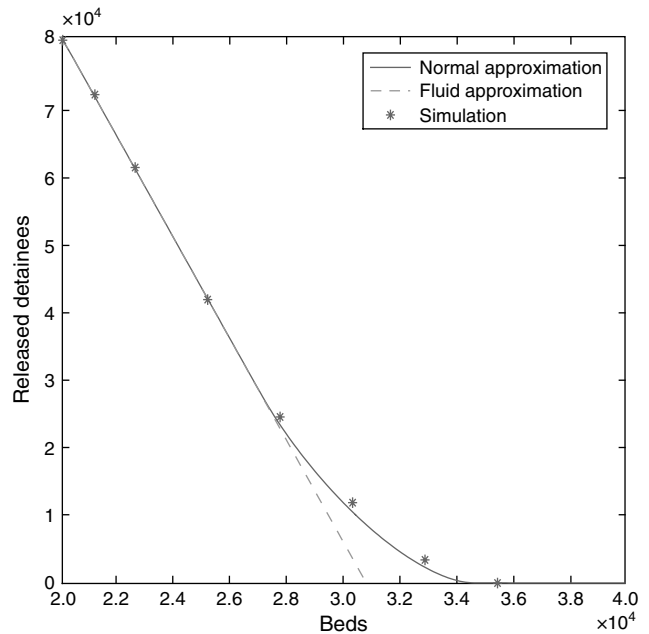
5.3. The Required Number of Beds

These parameter values allow us to depict the relationship between the annual number of released detainees (R), which is the number blocked plus the number preempted, and the number of DRO beds (s) (Figure 4), via both simulation of the exact queuing system (10 years with the first 2 years truncated) and the normal approximation in Equation (19). In addition, the fluid approximation in Equations (23)–(24) suggests that this relationship is approximately linear with slope m_2^{-1} ,

$$R = 231,730 - 7.60s \quad \text{for } s < 26,480. \quad (36)$$

That is, throughout the fluid regime, each additional bed is capable of removing $m_2^{-1} = 7.60$ non-mandatory detainees per year. The reason that this number differs from an earlier estimate of 10 annual removals per bed (Turner 2004) is probably because the latter estimate does not incorporate the fact that residence times are right censored due to the large number of preemptions. The accuracy of the fluid approximation is confirmed in Figure 4.

Figure 4 The Number of Detainees Released Per Year (R) vs. the Number of DRO Beds (s)



Note. This plot confirms the accuracy of the normal approximation in Equation (20), the fluid approximation in Equation (36) (---), and the proposed number of beds in Equation (39).

We also propose a simple calculation for the number of beds required to achieve a very high removal rate: Compute the maximum (over the year) mean number of detainees if there was no blocking and preemption, and choose the number of beds equal to this value plus three times the square root of this value (the mean of this random variable equals its variance, and so we are adding three standard deviations to the mean, as is typical in many statistical problems and is reminiscent of other queueing approximations, e.g., Kolesar and Green 1998). That is, we set the number of beds equal to

$$s^* = \max_{t \in [0, T]} \left\{ n_1(t) + n_2(t) + 3\sqrt{n_1(t) + n_2(t)} \right\}. \quad (37)$$

Similar to our analysis resulting in Equation (24), we have that

$$s^* = \bar{\lambda}_1 m_1 + \bar{\lambda}_2 m_2 + \alpha \sqrt{c_1^2 + c_2^2} + 3\sqrt{\bar{\lambda}_1 m_1 + \bar{\lambda}_2 m_2 + \alpha \sqrt{c_1^2 + c_2^2}}. \quad (38)$$

Using the values in Table 2, the proposed number of beds is

$$s^* = 33,974 + 3\sqrt{33,974} = 34,526, \quad (39)$$

which agrees well with Figure 4. Because the detention population is huge and because the arrival rates have increased significantly since 2003, the square-root term in Equation (39) is negligible compared to the uncertainty in estimating the average arrival rates $\bar{\lambda}_i$ in future years. Also, if we ignored seasonality (i.e., set $\alpha = 0$), the 33,974 in Equation (39) would be reduced by more than 10%.

For long-term capacity planning, by far the biggest uncertainty is in forecasting the future arrival rates of detainees, particularly the nonmandatory detainees (in contrast, we used historical arrival data to estimate the model's parameters). Substituting all values from Table 2 except for $\bar{\lambda}_1$ and $\bar{\lambda}_2$ into Equation (38) yields

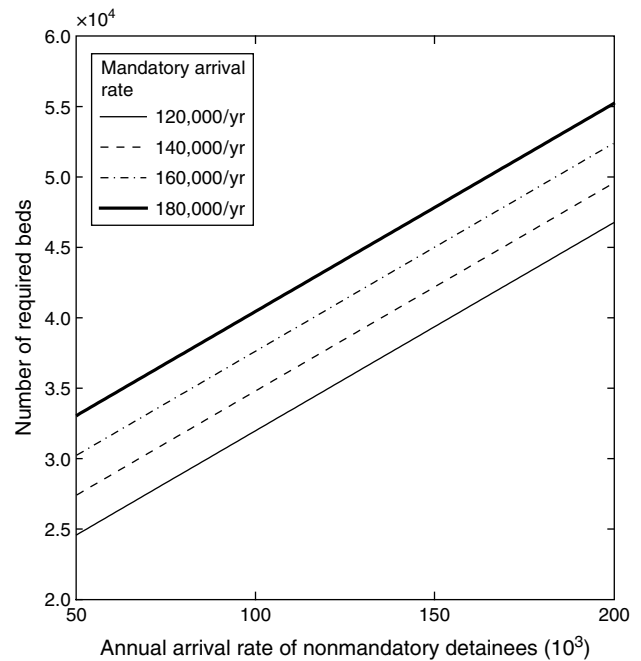
$$s^* = 0.1255\bar{\lambda}_1 + 0.1315\bar{\lambda}_2 + 0.1474\sqrt{c_1^2 + c_2^2} + 3\sqrt{0.1255\bar{\lambda}_1 + 0.1315\bar{\lambda}_2 + 0.1474\sqrt{c_1^2 + c_2^2}}, \quad (40)$$

where

$$\begin{aligned} c_1 &= 0.0074\bar{\lambda}_1 + 0.0781\bar{\lambda}_2 \quad \text{and} \\ c_2 &= -0.0610\bar{\lambda}_1 - 0.0646\bar{\lambda}_2, \end{aligned} \quad (41)$$

which is plotted in Figure 5; these curves show that $s^* = 0.1255\bar{\lambda}_1 + 0.1315\bar{\lambda}_2$ is an accurate approximation to (40). Our queueing analysis (in particular, Equation (20)) could be embedded in a newsvendor model: Given a bivariate probability distribution for $(\bar{\lambda}_1, \bar{\lambda}_2)$,

Figure 5 The Number of Required Detention Beds, as Calculated in Equation (40), as a Function of the Annual Arrival Rate of Nonmandatory Detainees ($\bar{\lambda}_2$), for Four Different Values of the Annual Arrival Rate of Mandatory Detainees ($\bar{\lambda}_1$)



an underage cost (which would reflect the cost of releasing detainees), and an overage cost (the cost of idle facilities), choose an optimal value of s . We have not pursued this approach because two quantities are very difficult to estimate: the probability distribution for the arrival rates, which depends largely on political forces that cannot be reliably predicted (e.g., there has been a huge increase in nonmandatory arrivals in the last few years because of the possibility of a partial amnesty program for illegal aliens in the United States, Bush 2004), and the monetary cost of a released detainee, which depends on the decision maker's views on homeland security and illegal immigration.

5.4. Robustness of the Exponential Service-Time Assumption

The available data suggest that the residence times for mandatory and nonmandatory detainees are actually mixtures of several more primitive random variables, and in this subsection we explore this issue. More specifically, given that a fraction $f = 0.3051$ of noncriminals are mandatory (Table 2), and given the total (blocked plus unblocked) arrival rates in Table 1, it follows that mandatory detainees are comprised of 9.7% Mexican noncriminals, 48.0% Mexican criminals, 18.9% OTM noncriminals, and 23.4% OTM criminals. Similarly, the nonmandatory detainees consist of 33.9% Mexican noncriminals and 66.1% OTM noncriminals.

In performing our sensitivity analysis, we consider the residence-time distributions of the four primitive groups of detainees, and first we compute the four means. We know from §4 that the mean residence times for Mexican criminals and OTM criminals are 17.9 days and 100.5 days, respectively. We also know from §4 that the mean right-censored residence times for Mexican noncriminals and OTM noncriminals are 10.5 days and 42.4 days, respectively. In Table 2, we estimate that the mean uncensored residence time for nonmandatory detainees is 48.0 days. Because the mean right-censored residence times for nonmandatory detainees is $0.339(10.5) + 0.661(42.4) = 31.6$ days, for lack of other data we assume that the mean uncensored residence times for Mexican noncriminals and OTM noncriminals are $(48.0/31.6)(10.5) = 15.9$ days and $(48.0/31.6)(42.4) = 64.4$ days, respectively.

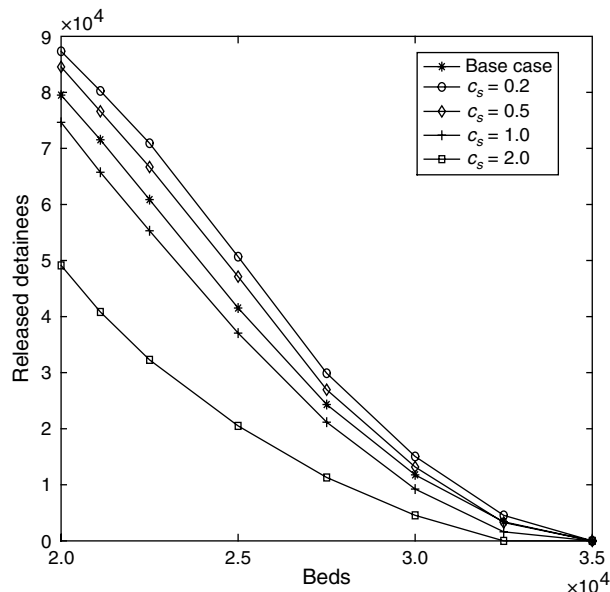
With the four means in hand, we simulate the queueing system to generate the annual number of released detainees (R) versus the number of DRO beds (s) in five different cases, in which the four residence times are independent gamma random variables with coefficient of variation (standard deviation divided by the mean) c_s all equal to 0.2, 0.5, 1, 2, or 5. Figure 6 shows these five curves along with our original curve from Figure 4. The first thing to note about Figure 6 is that in contrast to most queueing systems, the queueing performance of this system actually improves as c_s increases. As c_s increases, a larger

proportion of nonmandatory detainees have small residence times, which increases their likelihood of not getting preempted. Moreover, the nonmandatory detainees with large residence times are more apt to be preempted, and as c_s increases, these detainees are more likely to have uncensored residence times that are very large. In other words, when c_s is very large, the system wastes little time serving nonmandatory detainees with very large uncensored service times, which frees up capacity to serve to completion the many nonmandatory detainees who have very small uncensored residence times.

The original simulation curve from Figure 4 is between the $c_s = 0.5$ and $c_s = 1$ curves in Figure 6. Figure 6 suggests that our original analysis is reasonably accurate if the true $c_s \in [0.2, 1]$ for the four groups of detainees. Although we have no data to assess what the underlying c_s is (or indeed, whether a gamma distribution is appropriate), our knowledge of the system (as described earlier) tentatively suggests to us that the true c_s is likely to be in that range.

Finally, to assess the possibility that people cross the border—and hence are apprehended—in small groups, we simulate a system with the original exponential service times and with the same overall arrival rates, but where all customers arrive in pairs, and pairs arrive according to a Poisson process. The results were virtually indistinguishable from the simulation results for Poisson arrivals in Figure 4.

Figure 6 The Annual Number of Released Detainees vs. the Number of Detention Beds, when the Residence Times for Mexican Noncriminals, Mexican Criminals, OTM Noncriminals, and OTM Criminals Are Gamma Random Variables with Coefficient of Variation c_s



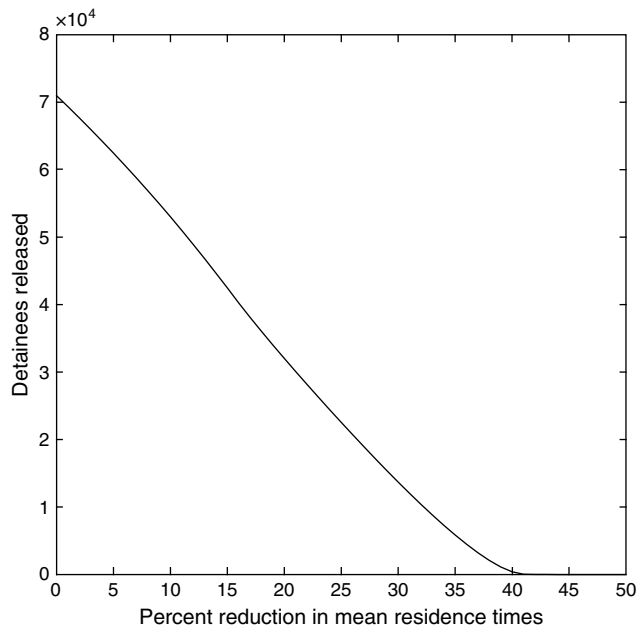
Note. The base case corresponds to the simulation of our original model with exponential service times for mandatory and nonmandatory detainees.

5.5. Reducing the Mean Residence Times

There are several management levers available to reduce the number of released illegal aliens besides increasing the number of DRO beds. Leaving aside the possibility of reducing the illegal border traffic (and hence the apprehension rate), the primary way is to reduce the residence times, which could be achieved in several ways (Tangeman 2003): use more expedited removal (which allows removal of some detainees without the involvement of an immigration judge), increase the number of immigration judges and lawyers, obtain removal papers more quickly from the host countries, and use electronic monitoring of low-risk aliens in lieu of detention. Unfortunately, reliable data on the court backlog and the delays with host countries, which would be required to assess the benefits of these improvements, are not available. In the absence of such data, we plot the annual number of released detainees versus the percentage reduction in mean residence times (Figure 7). The number of released detainees decreases approximately linearly with the percent reduction in mean residence times, and a 40% reduction is required to prevent all releases.

An alternative way of looking at this issue is to assume that the government always maintains the

Figure 7 The Number of Detainees Released (R) vs. the Percent Reduction in Mean Residence Times From the Values of m_1 and m_2 in Table 2, Computed Using Equation (20)



required number of beds in Equation (38), and to estimate the cost saved per percent reduction in mean residence times. Using the 2004 average of \$90 per bed per day to fund additional beds for mandatory detainees (Turner 2004), which may be much higher than the cost of constructing permanent facilities, we find that (although the relationship between beds and residence times is not linear) the annual cost savings of a 1% reduction in mean residence times (from its base-case value in Table 2) is \$10.8 M.

In contrast, any policy that increases the fraction of detainees that are classified as mandatory (e.g., starting in December 2005, illegal border crossers have been charged with a misdemeanor that carries up to six months in jail, McKinley 2007) will exacerbate the situation. In Wein et al. (2007), we show that detaining all Mexican border crossers would totally overwhelm DRO, even if there were 100,000 beds.

6. Conclusion

In our view, the main contribution of this study is in determining that each additional bed can remove only $m_2^{-1} = 7.60$ nonmandatory aliens per year (Figure 4, Equation (36)), rather than 10 aliens per year (Turner 2004), which appears to be the naive estimate that fails to account for the right censoring of nonmandatory residence times. Without having raw data on the interarrival times and residence times, there is no way of directly assessing the precision of this estimate.

Our results show that DRO beds were severely underfunded in 2003: We estimate that among apprehended illegal aliens who were nonmandatory, 75.6% were released into the United States. Whereas just over 21,000 beds were used in 2003, we calculate that approximately 34,500 beds would have been needed to remove all potential detainees (Figure 4, Equation (39)). Moreover, since 2003, the increase in the supply of DRO beds has not kept pace with the increase in demand. The Department of Homeland Security budgets in 2004–2007 allow for, respectively, 19,444, 20,660, 22,580, and 29,280 beds (Callahan 2005, Ramanathan 2005, Lipton 2006a), with all of these increases less than the 8,000 beds per-year increase dictated in the 2004 Intelligence Reform and Terrorism Act (108th U.S. Congress 2004). Although we do not have all of the data for 2004 to repeat our analysis, we do know that the total number of arrivals, $\bar{\lambda}_1 + \bar{\lambda}_2$, was the total number detained (235,247, Dougherty et al. 2005) minus the average total daily detention population (21,919, Dougherty et al. 2005) plus the total number blocked (38,000, Bjerke 2004). Assuming the same mix of mandatory ($\bar{\lambda}_1$) and nonmandatory ($\bar{\lambda}_2$) arrivals as in Table 2, Equation (38) gives $s^* = 36,398$ beds to remove all potential detainees in 2004, or approximately 2000 more than in 2003. There was a dramatic increase in the arrivals of potential OTM detainees in 2005 (Office of Immigration Statistics 2006, Table 35), and it appears that over 100,000 were blocked in 2005 (Swarns 2006), suggesting that the number of beds required to remove all 2005 potential detainees was roughly 50,000. The large increase in arrivals since 2003 is likely due in part to President Bush’s announcement of a possible guest-worker program, which might provide partial amnesty to illegal aliens already working in the United States (Bush 2004).

Figure 7 suggests that decreasing the residence times may also be an effective way to reduce the number of released detainees, although it may not be easy given the current shortage of immigration judges and lawyers (Eggen 2004), the difficulty of obtaining removal papers from some host countries in a timely manner (U.S. Department of Justice 2003), and the human rights concerns related to expedited removal (U.S. Commission on International Religious Freedom 2005) and electronic monitoring. Nonetheless, improvements to the removal process (U.S. Department of Justice 2002, 2003; Tangeman 2003) should be pursued in parallel with bed capacity enhancement, and the methods developed in this paper could be used in the future to estimate the reduction in mean residence times achieved by these improvements.

However, our model needs to be used with caution, because it does not attempt to account for behavioral changes that could occur as a result of increasing the

number of DRO beds. The most obvious change if the removal rate was to increase significantly is that the arrival rate of potential detainees to DRO would be reduced, because some illegal aliens might choose to move to alternative countries, choose more remote paths (perhaps with the help of human smugglers) along the U.S.–Mexico border (or the U.S.–Canada border), or decide not to leave their home country; in contrast, many crossers are well aware of, and even exploit (by crossing and then turning themselves in to border patrol agents, McKinley 2007), the catch-and-release policy. In addition, if border patrol agents perceive their jobs as switching from catch-and-release to catch-and-remove, then the apprehension rate (i.e., the probability of apprehending an illegal alien crossing the border) may increase, which would partially offset the impact of the reduced arrival rate. Although these effects would likely be minor if the annual increase in the number of beds remains modest, they would need to be incorporated by policymakers if a large increase was being considered.

Returning to our initial concern of homeland security, the current apprehension probability at the border is believed to be approximately 30% (Table 1 in Espenshade and Acevedo 1995). Combining this with the 75.6% of nonmandatory detainees who were released in 2003, we see that the likelihood that a terrorist with no previous record would successfully make it into the United States on his first attempt is approximately $1 - 0.3(0.244)$, or 92.7%. A 67% increase of funding for DRO beds over the 2005 budget would increase our chances of detaining and removing a terrorist from 7.3% to 30%. Using the 2004 figure of \$90 per bed per day (Turner 2004), this 67% increase would have an annual cost of \$450 M, which is more than 10% of the Immigration and Customs Enforcement budget, more than 1% of the entire homeland security budget (U.S. Department of Homeland Security 2005), and e.g., is approximately half the cost of the U.S. government's contract to produce enough anthrax vaccine for 25 million people (Lipton 2006b). However, unlike some other homeland security expenditures for preventing, preparing, or responding to uncertain future attacks, this expenditure would have a sure payoff. Nonetheless, our model in isolation is insufficient to determine the optimal number of beds. Also required are a better understanding of the cost effectiveness of hiring more judges and clerks, an investigation into the extent to which residence times can be reduced by negotiations with non-Mexican countries to expedite removals, and a quantification (as performed in Wein et al. 2007) of the cost effectiveness of increases in border patrol staffing, surveillance technology, and work-site inspections.

Acknowledgments

This research was supported by the Center for Social Innovation, Graduate School of Business, Stanford University. The authors thank John Bjerke for sharing DRO data and John Bjerke, Dwight McDaniel, and Sue Ramanathan for helpful discussions.

References

- Bean, F. D., R. O. de la Garza, B. R. Roberts, S. Weintraub, eds. 1997. *At the Crossroads: Mexican Migration and U.S. Policy*. Rowman & Littlefield Publishers, Lanham, MD.
- Bjerke, J. 2004. Personal communication. Statistician, office of detention and removal, bureau of immigration and customs enforcement. (December).
- Bush, G. W. 2004. Bush on immigration: Excerpts from Bush's address on allowing immigrants to fill some jobs. *New York Times* (January 8) A28.
- Callahan, R. 2005. Statement before the House Committee on the Judiciary Subcommittee on Immigration, Border Security, and Claims. (March 10).
- Cinlar, E. 1972. Superposition of point processes. P. A. W. Lewis, ed. *Stochastic Point Processes, Statistical Analysis, Theory and Applications*. John Wiley & Sons, New York, 549–606.
- Dougherty, M., D. Wilson, A. Wu. 2005. Immigration enforcement actions: 2004. Report, Office of Immigration Statistics, Washington, D.C.
- Eggen, D. 2004. Immigration backlog forces justice to shift staffing. *Washington Post* (December 14) A11.
- Eick, S. G., W. A. Massey, W. Whitt. 1993. $M_1/G/\infty$ queues with sinusoidal arrival rates. *Management Sci.* **39** 241–252.
- Espenshade, T. J., D. Acevedo. 1995. Migrant cohort size, enforcement effort, and the apprehension of undocumented aliens. *Population Res. Policy Rev.* **14** 145–172.
- Feller, W. 1968. *An Introduction to Probability Theory and Its Applications*, 3rd ed., Vol. 1. John Wiley & Sons, New York.
- Fischer, M. J. 1980. Priority loss systems—Unequal holding times. *Amer. Inst. Indust. Engineers Trans.* **12** 47–53.
- Garcia, M. J. 2004. FY 2005 budget request for the U.S. Immigration and Customs Enforcement. Statement before the Senate Appropriations Committee Subcommittee on Homeland Security. (March 30).
- Green, L., P. Kolesar. 1991. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Sci.* **37** 84–97.
- Hutchinson, A. 2004. U.S. Department of Homeland Security "detention priorities." Memorandum by Under Secretary for Border and Transportation Security, U.S. Department of Homeland Security, Washington, D.C. (October 18). Retrieved on January 14, 2005, http://www.vdare.com/mann/detention_priorities.htm.
- Jehl, D. 2005. U.S. aides cite worry on Qaeda infiltration from Mexico. *New York Times* (February 17) A10.
- Kelly, F. P. 1979. *Reversibility and Stochastic Networks*. John Wiley & Sons, New York.
- Kolesar, P., L. Green. 1998. Insights on service system design from a normal approximation to Erlang's delay formula. *Production Oper. Management* **7** 282–293.
- Laws, C. N. 1992. Resource pooling in queueing networks with dynamic routing. *Adv. Appl. Probab.* **24** 699–726.
- LeDuff, C., J. E. Flores. 2005. The every migrant's guide to crossing the border illegally. *New York Times* (February 9) A16.
- Lipton, E. 2006a. The President's budget: Homeland security; more money for border activities. *New York Times* (February 7) A14.
- Lipton, E. 2006b. Setbacks stymie bid to stockpile bioterror drugs. *New York Times* (September 18) A1.

- Little, J. D. C. 1961. A proof for the queuing formula $L = \lambda W$. *Oper. Res.* 9 383–387.
- Massey, W. A., W. Whitt. 1994. An analysis of the modified-offered-load approximation for the nonstationary Erlang loss model. *Ann. Appl. Probab.* 4 1145–1160.
- McKinley, J. C. 2007. Tougher tactics deter migrants at U.S. border. *New York Times* (February 21) A1.
- Office of Immigration Statistics. 2004. 2003 Yearbook of immigration statistics, Chap. 8. (September).
- Office of Immigration Statistics. 2006. 2005 Yearbook of immigration statistics. (September).
- One Hundred Eighth U.S. Congress. 2004. Intelligence reform and terrorism prevention act of 2004. (December 17) Pub. L. 108–458.
- Ramanathan, S. 2005. Personal communication. Counsel and professional staff member, house select committee on homeland security. (May).
- Swarns, R. L. 2006. Tight immigration policy hits roadblock of reality. *New York Times* (January 20) A12.
- Tangeman, A. S. 2003. Endgame: Office of detention and removal strategic plan, 2003–2012. Report, Office of Detention and Removal, Washington, D.C. (August 15).
- Turner, J. 2004. Transforming the southern border: Providing security and prosperity in the post-9/11 world. Report, House Select Committee on Homeland Security, Washington, D.C.
- U.S. Commission on International Religious Freedom. 2005. Report on asylum seekers in expedited removal. Vol. I: Findings and recommendations. U.S. Commission on International Religious Freedom, Washington, D.C. (February 8).
- U.S. Department of Homeland Security. 2005. Budget-in-brief, fiscal year 2006. U.S. Department of Homeland Security, Washington, D.C.
- U.S. Department of Justice. 2002. Immigration and Naturalization Service institutional removal program. Report 02-41, Office of the Inspector General, Washington, D.C. (September).
- U.S. Department of Justice. 2003. The Immigration and Naturalization Service's removal of aliens issued final orders. Report I-2003-004, Office of the Inspector General, Washington, D.C. (February).
- U.S. Supreme Court. Opinion. 2001. *Zadvydas vs. Davis*. (99-7791). (June 28), Retrieved on January 18, 2005, <http://supct.law.cornell.edu/supct/html/99-7791.ZO.html>.
- Wein, L. M., Y. Liu, A. Motskin. 2007. *Analyzing the Homeland Security of the U.S.–Mexico Border*. Graduate School of Business, Stanford University, Stanford, CA.

Appendix

In this Appendix, we compute the mean daily detention population throughout the year (Q), the mean number of blocked detainees during the year (B), the mean number of preempted detainees during the year (P), and the ratio of the maximum-to-minimum monthly mean number of unblocked detainee arrivals (M). These steady-state quantities are defined by

$$Q = \frac{\int_0^T Q(t) dt}{T}, \quad (1)$$

$$B = \int_0^T B(t) dt, \quad (2)$$

$$P = \int_0^T P(t) dt, \quad (3)$$

$$M = \frac{\int_{\frac{\pi}{2} - \frac{\pi}{12}}^{\frac{\pi}{2} + \frac{\pi}{12}} \lambda_1(t) + \lambda_2(t) - B(t) dt}{\int_{\frac{3\pi}{2} - \frac{\pi}{12}}^{\frac{3\pi}{2} + \frac{\pi}{12}} \lambda_1(t) + \lambda_2(t) - B(t) dt}, \quad (4)$$

where $Q(t)$ is the expected detention population at time t , $B(t)$ is the blocking rate at time t , and $P(t)$ is the preemption rate at time t ($B(t)$ and $P(t)$ are in terms of detainees per unit time), and $T = 1$ yr. In computing (4), we assume that the arrival rate of unblocked detainees achieves its extreme values at the same times as the total arrival rate of detainees, which is justified by the large fraction of arriving detainees that are mandatory.

To motivate our analysis, we start by suppressing the time notation on the arrival rates, and formulating a continuous-time Markov chain model of the two-class queueing system with generic arrival rates λ_1 and λ_2 . A similar queueing system has been studied previously [1], but in which customers do not change class (i.e., $\theta = 0$) and arriving class 1 customers are blocked when no servers (i.e., beds) are available. Let Q_i be the steady-state number of mandatory ($i = 1$) and nonmandatory ($i = 2$) detainees in DRO, and let $p_{ij} = Pr(Q_1 = i, Q_2 = j)$ for $i = 0, 1, \dots$ when $j = 0$, and $i = 0, \dots, s; j = 1, \dots, s - i$. The

global balance equations [2] for this continuous-time Markov chain are given by (Fig. 1)

$$(\lambda_1 + \lambda_2 + i(\mu_1 + \theta) + j\mu_2)P_{ij} = \lambda_1 P_{i-1,j} + \lambda_2 P_{i,j-1} + (i+1)\mu_1 P_{i+1,j} + (i+1)\theta P_{i+1,j-1} + (j+1)\mu_2 P_{i,j+1} \quad (5)$$

for $i = 0, \dots, s-1$ and $j = 0, \dots, s-1-i$;

$$(\lambda_1 + i(\mu_1 + \theta) + (s-i)\mu_2)P_{i,s-i} = \lambda_1 P_{i-1,s-i} + \lambda_1 P_{i-1,s-i+1} + \lambda_2 P_{i,s-i-1} + (i+1)\theta P_{i+1,s-i-1} \quad (6)$$

for $i = 0, \dots, s-1$, and $j = s-i$;

$$(\lambda_1 + s(\mu_1 + \theta))P_{s0} = \lambda_1 P_{s-1,0} + \lambda_1 P_{s-1,1} + (s+1)(\mu_1 + \theta)P_{s+1,0}; \quad \text{and} \quad (7)$$

$$(\lambda_1 + i(\mu_1 + \theta))P_{i0} = \lambda_1 P_{i-1,0} + (i+1)(\mu_1 + \theta)P_{i+1,0} \quad (8)$$

for $i = s+1, s+2, \dots$, where by definition $P_{-1,j} = P_{i,-1} = 0$.

One approach to approximating (1)-(4) is the pointwise stationary approximation (PSA) [3]: for each value of time t throughout the year, we compute $Q(t)$, $B(t)$, and $P(t)$ in (1)-(3) by performing a steady-state analysis of a stationary system that has time-homogeneous Poisson arrival processes with rates $\lambda_i(t)$. The accuracy of the PSA approximation depends on the relative frequency and the relative amplitude of the arrival processes. The relative frequency in our model, which is the average residence time divided by the period, is approximately 0.13 (Table 2 of main text), and the relative amplitude $\alpha = 0.173$ (Table 2 of main text). Fig. 2 of [5] suggests that the maximum percent error is approximately 2%, and these errors tend to cancel each other out after integrating over the entire year as in (1)-(3) (Fig. 3 of [5]).

The embedded steady-state analysis for $Q(t)$, $B(t)$ and $P(t)$ in the PSA requires that, for each value of t , we substitute $\lambda_i(t)$ for λ_i in (5)-(8) and solve these equations for $P_{ij}(t)$. A generalization of the recursions in equations (4)-(6) of [1] can be used to perform this task, but it requires the solution of a system of approximately $\frac{20,000^2}{2} = 2 \times 10^8$ equations

for each value of t . Hence, although the PSA is likely to be quite accurate, this approach is cumbersome, particularly because we need to evaluate the queueing system for many values of the five unknown parameters $(f, m_2, s, \theta, \alpha)$ to find the set of parameter values that best fits the data.

Consequently, we use the following heuristic approximation that maintains some aspects of the PSA, but allows us to approximate (1)-(4) in a more tractable manner. Because class 2 customers are invisible to class 1 customers, the total number of class 1 customers behaves as the number of customers in an infinite-server queue with Poisson arrival rate $\lambda_1(t)$ and exponential service rate $m_1^{-1} + \theta$. By equation (15) in [5], the steady-state time-dependent queue length $Q_1(t)$ is a Poisson random variable with mean

$$n_1(t) = \frac{\bar{\lambda}_1}{m_1^{-1} + \theta} \left(1 + \frac{\alpha}{1 + \left(\frac{2\pi}{(m_1^{-1} + \theta)T} \right)^2} \left(\sin\left(\frac{2\pi t}{T}\right) - \frac{2\pi}{(m_1^{-1} + \theta)T} \cos\left(\frac{2\pi t}{T}\right) \right) \right). \quad (9)$$

It follows that (Fig. 1)

$$P_{i0}(t) = \frac{n_1(t)^i e^{-n_1(t)}}{i!} \quad \text{for } i = s, s + 1, \dots, \quad (10)$$

$$P(Q_1(t) = i) = \sum_{j=0}^{s-i} P_{ij}(t) = \frac{n_1(t)^i e^{-n_1(t)}}{i!} \quad \text{for } i = 0, \dots, s - 1. \quad (11)$$

Equations (9)-(11) are exact for all values of t .

Turning to class 2 customers, we have that $P_{ij}(t) = P(Q_2(t) = j | Q_1(t) = i) P(Q_1(t) = i)$. The key step in our procedure is to approximate $P(Q_2(t) = j | Q_1(t) = i)$ for $i = 0, \dots, s - 1$ by the time-dependent probability that there are j customers in an Erlang loss system (i.e., customers who cannot receive immediate service are blocked) with $s - i$ servers (which is the number of servers available to serve class 2 customers when $Q_1(t) = i$), Poisson arrival rate $\lambda_2(t) + i\theta$ (which is the instantaneous rate of class 2 arrivals when $Q_1(t) = i$), and exponential residence times with mean m_2 . This approximation corresponds to analyzing each column in Fig. 1 in isolation. To approximately analyze this time-dependent Erlang loss system, we

use the modified offered load approximation in [6], which approximates the probability of having j customers present by the conditional probability that the corresponding infinite-server system has j customers conditioned on having less than or equal to $s - i$ customers (which is the number of servers in this queue). Taken together, for $i = 0, \dots, s - 1$ and $j = 0, \dots, s - i$, we assume

$$P_{ij}(t) = \frac{\frac{n_2(t,i)^j}{j!}}{\sum_{j=0}^{s-i} \frac{n_2(t,i)^j}{j!}} \frac{n_1(t)^i e^{-n_1(t)}}{i!}, \quad (12)$$

where (again by equation (15) in [5])

$$n_2(t, i) = \bar{\lambda}_2 m_2 + i\theta m_2 + \frac{\bar{\lambda}_2 m_2 \alpha}{1 + \left(\frac{2\pi m_2}{T}\right)^2} \left(\sin\left(\frac{2\pi t}{T}\right) - \frac{2\pi m_2}{T} \cos\left(\frac{2\pi t}{T}\right) \right) \quad (13)$$

represents the mean number of class 2 customers at time t in the corresponding infinite-server system (i.e., with Poisson arrival rate $\lambda_2(t) + i\theta$ and exponential service times with mean m_2), which is a function of i . If we let $\phi(x, n)$ and $\Phi(x, n)$ denote the Poisson probability density function (pdf) and cumulative density function (cdf), then (12) can be expressed as

$$P_{ij}(t) = \frac{\phi(n_2(t, i), j)}{\Phi(n_2(t, i), s - i)} \phi(n_1, i). \quad (14)$$

Recall that the time-dependent performance measures needed to compute (1)-(4) are $Q(t)$, $B(t)$ and $P(t)$. Starting with $Q(t)$, we have

$$Q(t) = n_1(t) + \sum_{j=1}^s j \sum_{i=0}^{s-j} P_{ij}(t), \quad (15)$$

where $n_1(t)$ is given in (9) and $P_{ij}(t)$ is given by (12)-(13). Equation (15) can be reduced to a single sum as follows:

$$\begin{aligned} Q(t) &= n_1(t) + \sum_{j=1}^s j \sum_{i=0}^{s-j} P_{ij}(t), \\ &= n_1(t) + \sum_{i=0}^s \sum_{j=1}^{s-i} j P_{ij}(t), \end{aligned}$$

$$\begin{aligned}
&= n_1(t) + \sum_{i=0}^s \sum_{j=1}^{s-i} j \frac{\frac{n_2(t,i)^j}{j!} e^{-n_2(t,i)}}{\Phi(n_2(t,i), s-i)} \phi(n_1, i) \quad \text{by equation (14),} \\
&= n_1(t) + \sum_{i=0}^s \sum_{j=1}^{s-i} \frac{\frac{n_2(t,i)^j}{(j-1)!} e^{-n_2(t,i)}}{\Phi(n_2(t,i), s-i)} \phi(n_1, i), \\
&= n_1(t) + \sum_{i=0}^s \sum_{j=0}^{s-i-1} n_2(t, i) \frac{\frac{n_2(t,i)^j}{j!} e^{-n_2(t,i)}}{\Phi(n_2(t,i), s-i)} \phi(n_1, i), \\
&= n_1(t) + \sum_{i=0}^s n_2(t, i) \frac{\Phi(n_2(t, i), s-i-1)}{\Phi(n_2(t, i), s-i)} \phi(n_1, i). \tag{16}
\end{aligned}$$

However, there are approximately 2×10^8 $P_{ij}(t)$ s to solve for, many of which have value near zero, which leads to problems of numerical instability. Therefore, for $i = 1, 2$, we approximate the Poisson random variables with mean $n_1(t)$ and $n_2(t, i)$ (given in (9) and (13)) by normal random variables with mean and variance $n_1(t)$ and $n_2(t, i)$ [7]; we denote the pdfs of these normal random variables by f_{1t} and $f_{2t,x}$ and the cdfs by F_{1t} and $F_{2t,x}$. Using equation (12), these substitutions lead us to approximate $Q(t)$ in (16) by

$$Q(t) = n_1(t) + \int_0^s n_2(t, x) f_{1t}(x) \frac{F_{2t,x}(s-x-1)}{F_{2t,x}(s-x)} dx. \tag{17}$$

Because our approximation procedure in (12)-(13) does not directly consider horizontal or diagonal transitions in Fig. 1, its estimate for blocking actually incorporates blocking and preemption of the original system. That is, if we let $R(t) = B(t) + P(t)$ be the rate at which detainees are released (via blocking or preemption) at time t , then our estimate for $R(t)$ is the product of the instantaneous arrival rate of nonmandatory detainees (externally or from the mandatory pool) and the probability that at least s detainees are in residence, i.e.,

$$R(t) = (\lambda_2(t) + n_1(t)\theta) \left(\sum_{i=0}^s P_{i,s-i}(t) + \sum_{i=s+1}^{\infty} P_{i0}(t) \right), \tag{18}$$

where $P_{ij}(t)$ is given in (10) and (12). Using the normal approximation to the Poisson, we approximate (18) by

$$R(t) = (\lambda_2(t) + n_1(t)\theta) \left(\int_0^s \frac{f_{1t}(x) f_{2t,s-x}(s-x)}{F_{2t,s-x}(s-x)} dx + 1 - F_{1t}(s+1) \right). \tag{19}$$

Equation (19) is used to generate Figs. 1 and 2 in the main text, where the number of beds is varied over a wide range. However, if

$$\min_t \{n_1(t) + n_2(t, n_1(t)) - 3\sqrt{n_1(t) + n_2(t, n_1(t))}\} > s, \quad (20)$$

then (19) can be replaced by a cruder fluid approximation. According to the fluid approximation, the time-dependent Erlang loss system for class 2 customers has an arrival rate of $\lambda_2(t) + n_1(t)\theta$ and can process detainees at an average rate of $\frac{s-n_1(t)}{m_2}$, and the detainees who cannot be processed are released, which yields

$$R(t) = \lambda_2(t) + n_1(t)\theta - \left(\frac{s - n_1(t)}{m_2}\right). \quad (21)$$

Substituting equation (9) into (21) and integrating gives a fluid approximation for R , the mean annual number of released detainees,

$$R = \int_0^T R(t) dt = \frac{\bar{\lambda}_1(m_2^{-1} + \theta)}{m_1^{-1} + \theta} + \bar{\lambda}_2 - \frac{s}{m_2}. \quad (22)$$

Substituting the right sides of equations (9) and (13) into condition (20) and performing the minimization reveals that approximation (22) is valid as long as

$$s < \frac{\bar{\lambda}_1}{m_1^{-1} + \theta} + \left(\bar{\lambda}_2 + \frac{\bar{\lambda}_1\theta}{m_1^{-1} + \theta}\right) m_2 - \alpha\sqrt{c_1^2 + c_2^2} - 3\sqrt{\frac{\bar{\lambda}_1}{m_1^{-1} + \theta} + \left(\bar{\lambda}_2 + \frac{\bar{\lambda}_1\theta}{m_1^{-1} + \theta}\right) m_2 - \alpha\sqrt{c_1^2 + c_2^2}}, \quad (23)$$

where

$$c_1 = \frac{(1 + \theta m_2)\bar{\lambda}_1}{(m_1^{-1} + \theta) \left(1 + \left(\frac{2\pi}{(m_1^{-1} + \theta)T}\right)^2\right)} + \frac{\bar{\lambda}_2 m_2}{1 + \left(\frac{2\pi m_2}{T}\right)^2}, \quad (24)$$

$$c_2 = -\frac{2\pi(1 + \theta m_2)\bar{\lambda}_1}{(m_1^{-1} + \theta)^2 T \left(1 + \left(\frac{2\pi}{(m_1^{-1} + \theta)T}\right)^2\right)} - \frac{2\pi\bar{\lambda}_2 m_2^2}{T \left(1 + \left(\frac{2\pi m_2}{T}\right)^2\right)}. \quad (25)$$

The accuracy of (22)-(23) is confirmed in Fig. 1 of the main text. Because the 2003 DRO data is consistent with the fluid regime characterized by condition (23), we use (22) instead of (19) to estimate the parameters.

Our final step is to approximate the individual contributions of blocking and preemption to $R(t)$. If one were to solve the global balance equations (5)-(8) (with $\lambda_i(t)$ in place of λ_i) for $P_{ij}(t)$, then the blocking and preemption rates would be given by

$$B(t) = \lambda_2(t) \left(\sum_{i=0}^s P_{i,s-i}(t) + \sum_{i=s+1}^{\infty} P_{i0}(t) \right), \quad (26)$$

$$P(t) = \lambda_1(t) \sum_{i=0}^{s-1} P_{i,s-i}(t) + \theta \sum_{i=s+1}^{\infty} iP_{i0}(t). \quad (27)$$

There is a key difference between the right sides of equations (18) and (26): the $P_{ij}(t)$ s in (18) are from our approximation procedure in (12) whereas the $P_{ij}(t)$ s in (26) are the exact (unknown) solutions to (5)-(8). Note that $\sum_{i=0}^s P_{i,s-i}(t)$ in (26) is almost identical to $\sum_{i=0}^{s-1} P_{i,s-i}(t)$ in (27) because $s = 21, 100$ in Table 1 of the main text. If we replace both of these sums by the unknown quantity $x(t)$, sum equations (26) and (27), replace the left side of the summed equation by $R(t)$ from (19) or (21), and then solve the summed equation for $x(t)$, we get

$$x(t) = \frac{R(t) - \lambda_2(t) \sum_{i=s+1}^{\infty} P_{i0}(t) - \theta \sum_{i=s+1}^{\infty} iP_{i0}(t)}{\lambda_1(t) + \lambda_2(t)}. \quad (28)$$

We substitute the right side of (28) into equations (26) and (27) to get

$$B(t) = \lambda_2(t) \left(\frac{R(t) - \lambda_2(t) \sum_{i=s+1}^{\infty} P_{i0}(t) - \theta \sum_{i=s+1}^{\infty} iP_{i0}(t)}{\lambda_1(t) + \lambda_2(t)} + \sum_{i=s+1}^{\infty} P_{i0}(t) \right), \quad (29)$$

$$P(t) = \frac{\lambda_1(t) [R(t) - \lambda_2(t) \sum_{i=s+1}^{\infty} P_{i0}(t) - \theta \sum_{i=s+1}^{\infty} iP_{i0}(t)]}{\lambda_1(t) + \lambda_2(t)} + \theta \sum_{i=s+1}^{\infty} iP_{i0}(t). \quad (30)$$

Our final estimates for $B(t)$ and $P(t)$ are found by using the normal approximation to the Poisson in (29)-(30), which yields

$$B(t) = \frac{\lambda_1(t)\lambda_2(t)[1 - F_{1t}(s+1)] + \lambda_2(t)[R(t) - \theta \int_{s+1}^{\infty} x f_{1t}(x) dx]}{\lambda_1(t) + \lambda_2(t)}, \quad (31)$$

$$P(t) = \frac{\lambda_1(t)[R(t) - \lambda_2(t)(1 - F_{1t}(s+1))] + \lambda_2(t)\theta \int_{s+1}^{\infty} x f_{1t}(x) dx}{\lambda_1(t) + \lambda_2(t)}. \quad (32)$$

These results allow us to predict the approximate number of beds required to reduce the number of releases to a very small level. We propose to set the number of beds equal to

the mean plus three standard deviations of the maximum number of detainees over the year in the absence of blocking and preemption, i.e.,

$$s^* = \max_t \{n_1(t) + n_2(t, n_1(t)) + 3\sqrt{n_1(t) + n_2(t, n_1(t))}\}. \quad (33)$$

Similar to our analysis resulting in equation (23), we have that

$$s^* = \frac{\bar{\lambda}_1}{m_1^{-1} + \theta} + \left(\bar{\lambda}_2 + \frac{\bar{\lambda}_1 \theta}{m_1^{-1} + \theta} \right) m_2 + \alpha \sqrt{c_1^2 + c_2^2} + 3 \sqrt{\frac{\bar{\lambda}_1}{m_1^{-1} + \theta} + \left(\bar{\lambda}_2 + \frac{\bar{\lambda}_1 \theta}{m_1^{-1} + \theta} \right) m_2 + \alpha \sqrt{c_1^2 + c_2^2}}. \quad (34)$$

Before estimating the unknown model parameters, we summarize our analytical results. Our time-dependent performance measures are given by $Q(t)$ in (17), $B(t)$ in (31), and $P(t)$ in (32), where $R(t)$ in (31)-(32) is given by (19) for the generation of Figs. 1 and 2 in the main text, and by the fluid approximation (21) during parameter estimation. Calculating these performance measures for each value of t in $[0, T]$, and substituting into (1)-(4) give our final performance measures. Our proposed number of beds is given in (34).

Referring to the corresponding actual values of Q , B , P and M in Table 1 of the main text, we estimate the five unknown parameter values (Table 2 of the main text) by solving

$$\min_{s, f, m_2, \theta, \alpha} \left(\frac{Q - 21,133}{21,133} \right)^2 + \left(\frac{B - 28,000}{28,000} \right)^2 + \left(\frac{P - 43,000}{43,000} \right)^2 + \left(\frac{M - 1.27}{1.27} \right)^2. \quad (35)$$

The discretization of the variables $(s, f, m_2, \theta, \alpha)$ in the optimization algorithm was $(10, 10^{-3}, 10^{-1}, 10^{-3}, 10^{-3})$. As explained in the main text, the five parameter values generate accurate estimates of the four output variables (Q, B, P, M) and are consistent with the fluid regime. In particular, the right side of (23) is 25,770 with our estimated parameter values, which is significantly larger than $s = 21,100$.

For this set of parameter values, we display the approximate time-dependent performance measures for $(Q_1(t), Q_2(t), B(t), P(t))$ (Fig. 2). Figs. 2a and 2b confirm that $Q_1(t) + Q_2(t) = s$ for all t , and that during the peak time of year, approximately 95% of the beds are occupied by mandatory detainees. As expected from theory (equation (16)

in [5]), there is a time lag of $\frac{T}{2\pi} \tan^{-1}\left(\frac{2\pi}{(m_1^{-1} + \theta)T}\right) = 38$ days between the time of the peak arrival rate (0.25 yr) and the time of the peak mandatory population (0.35 yr), and hence of the times of the peak blocking and preemption rates.

Finally, we compare the approximate performance measures to computer simulation results of the exact queueing system. For the parameter values in Table 2 of the main text, the simulated values of (Q, B, P, M) are (21,107; 28,115; 43,031; 1.2734), versus the approximate values of (21,100; 27,989; 42,966; 1.2702). Simulated results in Fig. 1 of the main text further confirm the accuracy of our analytical approximations. All simulation runs were for ten years, with the first two years discarded.

References

- [1] M. J. Fischer. Priority loss systems - unequal holding times. *AIIE Transactions* **12**, 47-53, 1980.
- [2] F. P. Kelly. *Reversibility and Stochastic Networks*. John Wiley & Sons, New York, 1979.
- [3] L. Green, P. Kolesar. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science* **37**, 84-97, 1991.
- [4] Personal communication, John Bjerke, Statistician, Office of Detention and Removal, December, 2004.
- [5] S. G. Eick, W. A. Massey, W. Whitt. $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Science* **39**, 241-252, 1993.
- [6] W. A. Massey, W. Whitt. An analysis of the modified-offered-load approximation for the nonstationary Erlang loss model. *Annals of Applied Probability* **4**, 1145-1160, 1994.
- [7] W. Feller. *An introduction to probability theory and its applications, volume 1*, third edition. John Wiley & Sons, New York, 1968.

Figure Legends

Fig. 1. The state-space diagram for the continuous-time Markov chain for $(Q_1(t), Q_2(t))$. The global balance equations corresponding to this diagram appear in equations (5)-(8).

Fig. 2. The approximate time-dependent performance measures of the queueing system under the base-case parameters in Table 2 of the main text. **(a)** number of mandatory detainees $(Q_1(t))$; **(b)** number of nonmandatory detainees $(Q_2(t))$; **(c)** blocking rate of nonmandatory detainees $(B(t))$; **(d)** preemption rate of nonmandatory detainees $(P(t))$.

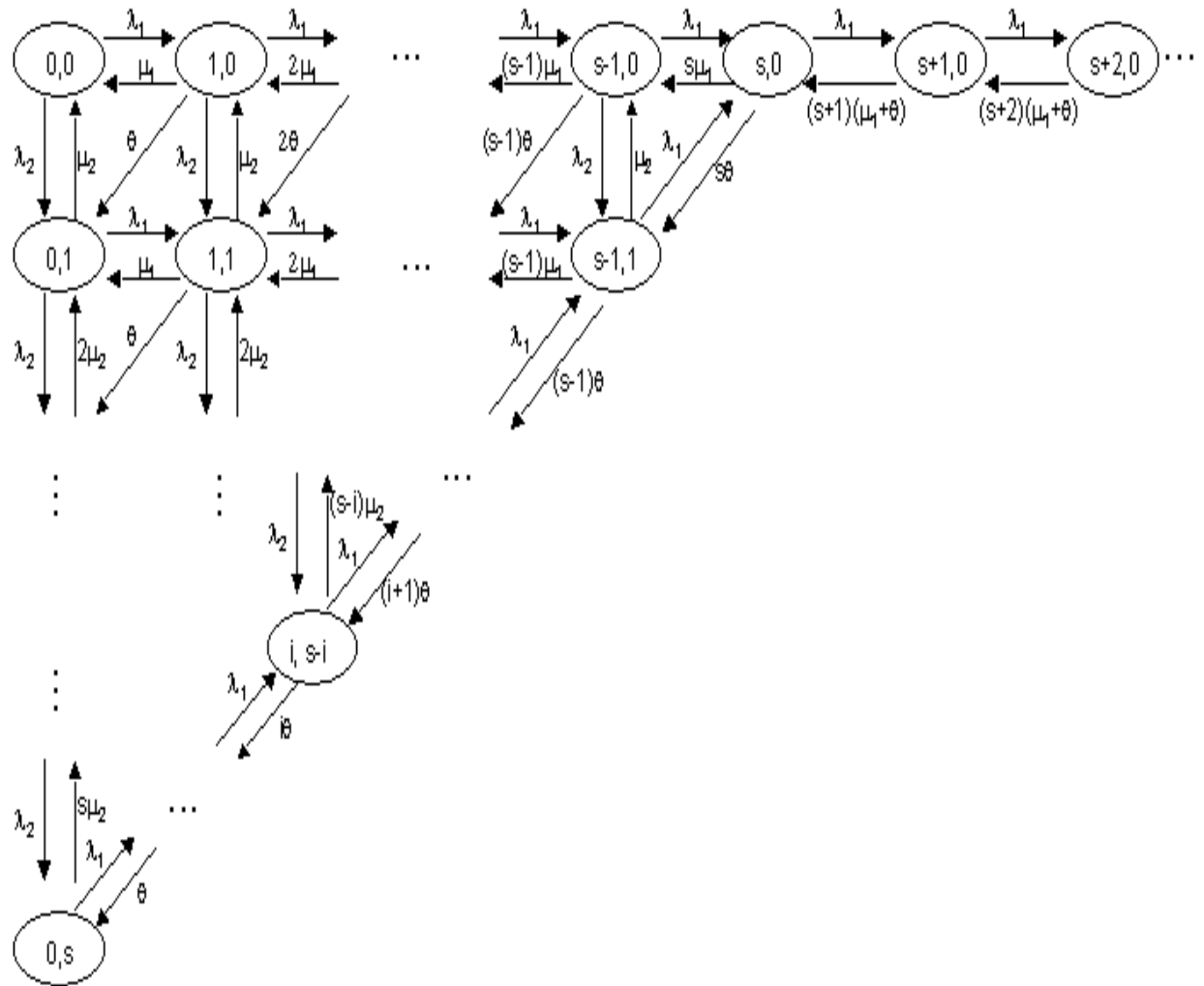


Figure 1.

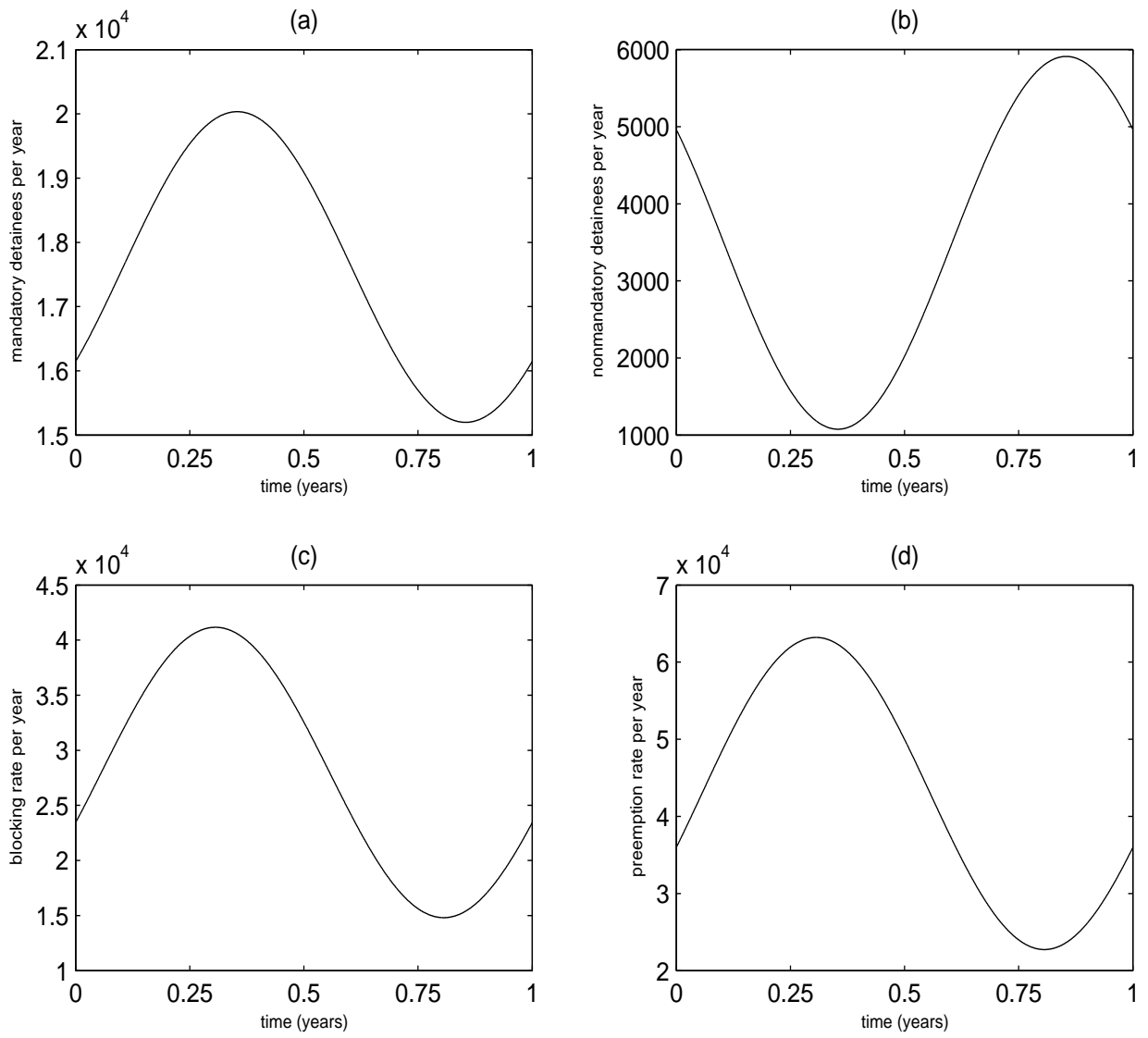


Figure 2.