Models and Algorithms for Transient Queueing Congestion at a Hub Airport

Dimitris Bertsimas, Michael D. Peterson
and
Amedeo R. Odoni

WP# -3501-92 MSA September, 1992

Models and Algorithms for Transient Queueing Congestion at a Hub Airport

Michael D. Peterson* Dimitris J. Bertsimas† Amedeo R. Odoni‡

September 11, 1992

Abstract

This paper studies the relationship between the hub-and-spoke design in air transportation and the phenomenon of landing congestion in a transient environment. We model the weather, the principal source of uncertainty, as a Markov or semi-Markov process, and we treat arrivals as time-varying but deterministic. We develop a recursive algorithm for predicting transient queueing delays. To test our model, we conduct a case study using traffic and capacity data for Dallas-Fort Worth International Airport. Our results show that the model's estimates are reasonable, though substantial data difficulties make thorough validation difficult. We explore in depth two policy questions: schedule interference between the two principal carriers, and the likely effects of demand smoothing policies on queueing delays.

1 Introduction

Among the most noticeable innovations of the 1980's within the U.S. airline industry was the development of extensive hub-and-spoke networks by the major carriers. These networks allow carriers to serve demands from numerous origin-destination (OD) markets with relatively few aircraft and flight legs. As a result, airlines may realize certain economies of scope and scale, achieve higher average load factors, and offer consumers more frequent

*The work of the first author was supported by a National Science Foundation Graduate Fellowship.

†The work of the second author was partially supported by a grant from Draper Labratory.

‡The work of the third author was partially supported by a grant from Draper Labratory.

flights [18,19]. But while these economic advantages are widely acknowledged, consumer dissatisfaction with certain aspects of the design is on the rise. An article in *The New York Times Magazine* recently reported that the fraction of Americans dissatisfied with the deregulation of the industry has risen from 17 percent to 36 percent over the past decade. According to the legal director of the Aviation Consumer Action Committee, delay is the principal reason for this rise in dissatisfaction [21]. In 1986 ground delays at domestic airports averaged 2000 hours per day, the equivalent of grounding the entire fleet of Delta Airlines at that time (250 aircraft) for one day [6]. In 1990, 21 airports in the U.S. exceeded 20,000 hours of delay, with 12 more projected to exceed this total by 1997 [26].

While much of the growth in delays has come about because of demand increases over the last decade, the development of hub-and-spoke networks has also played a role. Hubs are congested because they experience higher traffic levels than other airports. In fact, among the 11 airports with the highest number of reported delays in 1990, 8 were hubs: Chicago (O'Hare), Dallas-Fort Worth, Atlanta (Hartsfield), Denver (Stapleton), Newark, Washington (Dulles), Detroit, and San Francisco [26]. Moreover, hub-and-spoke systems tend to concentrate major airport operations (landings and takeoffs) into short periods of time, placing further strain on capacity. Because the hub is the center of operations for a carrier, large delays can have serious adverse effects on system operations. Understanding and predicting these delays is a matter of importance to carriers, regulators, air traffic controllers, and passengers.

Although the general queueing theory literature is vast, the number of works dealing with the transient behavior of queueing systems is surprisingly small, mainly because of the difficulty of obtaining analytical results for these kinds of problems. Most approaches model service and arrival processes as phase-type and attempt to solve the resulting forward Kolmogorov equations. The various methods differ mostly in the approach they take to solving these equations.

The most direct approach is numerical solution. Gross and Harris [10] give a thorough discussion of the competing methods (see especially their Section 7.3.2). Most of these become computationally expensive because of the large state spaces needed to make the system Markovian. A second approach developed in response to this difficulty is that of uniformization (see Grassmann [9] and Gross and Harris [10]). A third solution method due to Bertsimas and Nakazato [3,4] takes transforms of the Kolmogorov equations and then

inverts these numerically to obtain the waiting time and queue length distributions.

Diffusion methods are alternative approximate models which may be used for transient analysis — see Iglehart and Whitt [12,13], Kobayashi [17], Gelenbe and Mitrani [7] and Heyman and Sobel [11].

Research demonstrating the inadequacy of steady state analysis for certain queueing systems is foundational to this paper. Odoni and Roth [22,24] investigate the difference between transient and steady state queueing systems of phase-type. They use numerical methods to solve the Kolmogorov equations for a variety of these systems and compare the expected queue lengths with steady state values. Their results indicate that substantial differences persist for long enough periods to raise serious doubts about the validity of steady state approaches in airport and other applications.

Airport capacity and queueing studies have a history of over 30 years. The earliest work dates back to 1960 with the work of Blumstein [5] investigating the determinants of airport capacity. Newell [20] provides a thorough discussion of how airport geometry, flight rules, and weather conditions determine airport capacities and includes a discussion of how these have developed over the years. He claims, as we do, that standard queueing approaches are inadequate for airport queueing systems, and he argues instead for a deterministic approach.

Two recent studies concern simulation approaches for estimating aircraft queueing delay. Abundo [1] considers the problem of queueing for landing at a single airport. She employs an $M(t)/E_k(t)/1$ model for the landing queue. She solves this model numerically in combination with a weather capacity profile obtained from simulation. St. George [25] studies the issue of delay at hub airports using a simple simulation model. He treats the queueing processes for landings and takeoffs deterministically at several alternative levels of airport capacity, using data from 12 U.S. airports in 1977. The work does not address the issue of capacity slow-down due to poor weather conditions, focusing instead on comparing airport schedules for a given level of capacity.

The main contribution of the present work is a direct modeling of weather conditions, the principal source of uncertainty in airport capacity, and the development of an exact, efficient algorithm to predict congestion in airports. We have applied our methods to an airport using real data with very encouraging results.

The rest of the paper is organized as follows. In Section 2 we discuss the arrival and service operations for the landing queue at a hub airport and develop a model of capacity

based on a semi-Markov process. In Section 3 we develop an algorithmic approach for computing queue-length and waiting time moments over time, using a simple recursive procedure. In Section 4 we apply our methods in a case study of congestion at Dallas-Fort Worth Airport. Using data obtained from weather observations taken over eight years at DFW, we indicate the sensitivity of congestion delay to starting conditions and explore how the smoothing of demand during the most congested periods of the day could reduce queueing delay. We also attempt to validate the model using delay data obtained from the U.S. Department of Transportation. Section 5 summarizes the main contributions of the paper.

2 Models of Demand and Capacity

Incoming aircraft at a hub airport require service at a series of three stations: a landing runway, a gate, and a departure runway. Traditional queueing analyses are not appropriate for this system because of the following characteristics:

1. *Time variation in arrival rate.* A hub airport is subject to a highly time-varying demand rate. Work comparing transient and steady state results for single-server queues [22,24] suggests that in such cases, the time necessary to reach "steady state" substantially exceeds the time over which the demand rate may reasonably be taken as constant. The implication is that models which describe only steady state behavior are of very limited value in this context.

2. *Service times are not identically distributed.* For the landing and departure processes at an airport, capacity is weather-dependent, implying that service times are neither independent nor identically distributed. Thus it is inappropriate to model service times as i.i.d.

3. *Inter-dependence of service times.* Because of connections between flights, an aircraft's time at the gate depends on the arrival times of other flights. Moreover, separation rules for large and small aircraft negate the assumption that consecutive landing service times are independent.

These characteristics require that we take a new approach to the problem.

For the rest of this paper, we focus on the queue for aircraft landings, though we note that with only slight modifications, our approach is also appropriate for the departure queue. We consider landing aircraft as customers utilizing a set of runways which together constitute a single server. At the outset, we treat the aircraft demand process as deterministic. In practice, of course, arrival schedules contain elements of uncertainty because of earlier delays, and in Section 3 we will show how to account for a simple probabilistic structure. We model time-variation by dividing time into discrete intervals of fixed length and allowing the demand rate to vary arbitrarily across these intervals. To summarize, we have Assumption 1:

Assumption 1 (Demand Process) *The hub's operating day consists of discrete time intervals of length Δt. For interval k, the number of aircraft demanding to land, λ_k, is known, and these aircraft constitute a continuous (deterministic) flow over the interval.*

Note that since the rate is assumed constant within each interval, realism requires that Δt be short, on the order of 15 minutes.

The number of aircraft which the airport can land per hour is a function of many variables (runway configuration, air traffic control patterns, gate availability), but it is mainly a function of which runways can be used and how much separation is required between incoming aircraft. These factors are in turn determined by weather conditions: ceiling, visibility, wind direction, and wind speed. As the weather conditions change, capacity switches from one state to another. We thus shall employ two alternative models of capacity as a stochastic process, one based on a Markov chain and one based on a semi-Markov process. In the most general case, our assumption is as follows:

Assumption 2 (Service Process) *Landing capacity at the airport during a given interval j takes one of a discrete number of values $\mu_1, \mu_2, \ldots, \mu_S$ for some finite number S of capacity states with*

$$\mu_1 < \mu_2 < \ldots < \mu_S.$$

The random holding time (in intervals) for a given state i, T_i, follows an arbitrary discrete distribution with probability mass function

$$P_i(k) = \Pr\{T_i = k\},$$

the probability of a capacity μ_i period lasting for precisely k intervals of length Δt. Upon exiting a state i, the capacity process enters another state $j \neq i$ with probability p_{ij}.

3 An Algorithmic Approach

Assumptions 1 and 2 describe the arrival and service processes for our queueing system. We now develop a computational method for describing its transient behavior. We shall assume that within any interval k, the queue behaves like a deterministic flow process, with demand λ_k and service rate $\mu(k)$, $\mu(k)$ being a random variable which takes on one of the values $\mu_1, \ldots, \mu_S$. Thus given q_k, the length of the queue at the end of some period k, the queue length one period later is the maximum of 0 and the values $q_k + \lambda_k - \mu_i$ for $i = 1, \ldots, S$.

To establish a Markov chain within the semi-Markov process, we enlarge the state space to be $\{i, m\}$, where i is capacity and m the age (in intervals) of that capacity. The combined age-capacity process is clearly Markov, with transition probabilities given by

$$\begin{aligned} \tilde{p}_{ij}(m) &\triangleq \Pr((i,m) \to (j,1)) = \Pr[T_i = m \mid T_i \geq m] p_{ij} \quad j \neq i \\ \tilde{p}_{ii}(m) &\triangleq \Pr((i,m) \to (i,m+1)) = \Pr[T_i \geq m+1 \mid T_i \geq m] \end{aligned} \tag{1}$$

We next define the following random variables:

$$\begin{aligned} Q_k &\triangleq \text{Queue length at end of interval } k \\ W_k &\triangleq \text{Waiting time at end of interval } k \\ C_k &\triangleq \text{Capacity state at end of interval } k \\ A_k &\triangleq \text{Age of current capacity state at end of interval } k \\ T_i &\triangleq \text{Random lifetime of capacity state } i \end{aligned}$$

For queue lengths we introduce the notation

$$\begin{aligned} Q_k(l,i,m,q) &\triangleq E[Q_k \mid Q_l = q, C_l = i, A_l = m] \\ & k = 1, \ldots, K, \quad i = 1, \ldots, S, \quad m = 1, \ldots, M \\ & l \leq k, \quad q = 1, \ldots, q_{\max}(k,i). \end{aligned} \tag{2}$$

where $q_{\max}(k,i)$ is the maximum attainable queue length at the end of period k, given that at that time the capacity state is i. This obeys the recursion

$$q_{\max}(k,i) = [q_{\max}(k-1) + \lambda_k - \mu_i]^+ \tag{3}$$

where $q_{\max}(k) \triangleq \max_i q_{\max}(k,i)$ and $x^+ = \max(x, 0)$. Similarly, for waiting times we employ the notation

$$W_k(l,i,m,q) \triangleq E[W_k \mid Q_l = q, C_l = i, A_l = m]. \tag{4}$$

We write the second moment analogs of (2) and (4) as $\mathcal{Q}_k^2(l,i,m,q)$ and $\mathcal{W}_k^2(l,i,m,q)$, respectively. Our aim is to compute the quantities $\mathcal{Q}_k(l,i,m,q)$, $\mathcal{Q}_k^2(l,i,m,q)$, $\mathcal{W}_k(l,i,m,q)$, and $\mathcal{W}_k^2(l,i,m,q)$. From these we can compute the mean and variance for queue length and waiting time at the end of each period.

Our first result gives a recursion for queue length moments.

Theorem 1 *The functions $\mathcal{Q}_k(l,i,m,q)$ and $\mathcal{Q}_k^2(l,i,m,q)$ obey the recursive relationships*

$$\begin{aligned}\mathcal{Q}_k(l,i,m,q) &= \sum_{j\neq i}\tilde{p}_{ij}(m)\mathcal{Q}_k\left(l+1,j,1,(q+\lambda_{l+1}-\mu_j)^+\right)+\\ &\quad \tilde{p}_{ii}(m)\mathcal{Q}_k\left(l+1,i,m+1,(q+\lambda_{l+1}-\mu_i)^+\right) \qquad (5)\\ \mathcal{Q}_k^2(l,i,m,q) &= \sum_{j\neq i}\tilde{p}_{ij}(m)\mathcal{Q}_k^2\left(l+1,j,1,(q+\lambda_{l+1}-\mu_j)^+\right)+\\ &\quad \tilde{p}_{ii}(m)\mathcal{Q}_k^2\left(l+1,i,m+1,(q+\lambda_{l+1}-\mu_i)^+\right) \qquad (6)\end{aligned}$$

with boundary conditions $\mathcal{Q}_k(k,\cdot,\cdot,q)\equiv q$ and $\mathcal{Q}_k^2(k,\cdot,\cdot,q)\equiv q^2$.

PROOF:

Once a capacity state i is determined for interval $l+1$, a deterministic queue assumption means that the queue changes in the interval by the amount $\lambda_{l+1}-\mu_i$. Because the queue may not drop below 0, if the queue length is q at the start of a capacity μ period, then the length at the end of the period is $(q+\lambda_{l+1}-\mu)^+$. Conditional on the fact that at the end of interval l the queue level is q and the capacity μ_i has prevailed for m intervals, one of S things may happen by the end of the next interval. Either the airport will have remained in capacity state i, or it will have switched to one of the other $S-1$ states. These S transitions have corresponding probabilities $\tilde{p}_{i1}(m),\tilde{p}_{i2}(m),\ldots,\tilde{p}_{iS}(m)$. The result (5) follows. An identical argument proves (6). □

For an airport queueing system, it is reasonable to assume that the queue is 0 at the start of the operating day. From (5) and (6) we may then compute the values $\mathcal{Q}_k(0,i,m,0)$ and $\mathcal{Q}_k^2(0,i,m,0)$ for all values of i, k, and m. Thus we can obtain the expectation and variance of queue lengths at intervals of length Δt throughout the day, conditional on the capacity at period 0.

Waiting time moments may be found in a similar fashion. The main part of this procedure is contained in the next theorem. The proof, identical to that of Theorem 1, is omitted.

Theorem 2 *The functions $W_k(l,i,m,q)$ and $W_k^2(l,i,m,q)$ obey the recursive relation (for $l<k$)*

$$\begin{aligned} W_k(l,i,m,q) &= \sum_{j\neq i}\tilde{p}_{ij}(m)\left[W_k\left(l+1,j,1,(q+\lambda_{l+1}-\mu_j)^+\right)\right]+ \\ &\quad \tilde{p}_{ii}(m)W_k\left(l+1,i,m+1,(q+\lambda_{l+1}-\mu_i)^+\right) \qquad (7) \\ W_k^2(l,i,m,q) &= \sum_{j\neq i}\tilde{p}_{ij}(m)\left[W_k^2\left(l+1,j,1,(q+\lambda_{l+1}-\mu_j)^+\right)\right]+ \\ &\quad \tilde{p}_{ii}(m)\left[W_k^2\left(l+1,i,m+1,(q+\lambda_{l+1}-\mu_i)^+\right)\right] \qquad (8) \end{aligned}$$

□

The complication with waiting times (as opposed to queue lengths) occurs at the boundary $l=k$. We use the notation $(a\wedge b)$ for $\min(a,b)$. The calculation of the expected waiting time for an incoming aircraft at the end of interval k, given the queue length and capacity conditions at that time, is itself a recursive procedure within a larger recursion, as seen in the following theorem.

Theorem 3 *The functions $W_k(k,i,m,q)$ obey the recursion*

$$\begin{aligned} W_k(k,i,m,q) &= \sum_{j\neq i}\tilde{p}_{ij}(m)\left[\left(\frac{q}{\mu_j}\wedge 1\right)+W_k(k,j,1,(q-\mu_j)^+)\right]+ \\ &\quad \tilde{p}_{ii}(m)\left[\left(\frac{q}{\mu_i}\wedge 1\right)+W_k(k,i,m+1,(q-\mu_i)^+)\right] \qquad (9) \end{aligned}$$

where $W_k(k,\cdot,\cdot,0)\equiv 0$.

PROOF:

Suppose that at the end of period k the capacity is μ_i, the age is m, and there are q waiting aircraft. Consider an aircraft which arrives at this instant. Its waiting time, which is the time necessary to clear the existing queue, is the sum of two components:

$$[W_k \mid \mathcal{I}] = [W_k' + W_k'' \mid \mathcal{I}]. \qquad (10)$$

Here W_k' is the part of the waiting time experienced during the next interval $(k+1)$, W_k'' is the part experienced thereafter, and $\mathcal{I}$ denotes the conditioning information

$$\{Q_k = q, C_k = i, A_k = m\}.$$

Given this conditioning, the possible capacity-age states for interval $k+1$ are

$$(1,1),(2,1),\ldots,(i-1,1),(i,m+1),(i+1,1),\ldots,(S,1).$$

Let $C_{k+1} = j$ be the event that the capacity during the next interval is μ_j. Then

$$[W_k' \mid \mathcal{I}, C_{k+1} = j] = \min(q/\mu_j, 1). \tag{11}$$

This follows because during the interval $k+1$ the queue in front of the aircraft is reduced by $\min(q, \mu_j)$. If the queue is reduced to 0 during the interval, the aircraft waits for a time q/μ_j; otherwise, it waits for the entire interval. To obtain W_k'', note that after the interval has ended, any remaining waiting time is stochastically equivalent to the waiting time of an aircraft arriving one interval later to a queue of $q-\mu_j$, a prevailing capacity of μ_j, and an age of either 1 (if j is a new capacity) or $m+1$. Symbolically, this is

$$\begin{aligned} [W_k'' \mid \mathcal{I}, C_{k+1} = j] &\sim [W_k \mid Q_k = (q-\mu_j)^+, C_k = j, A_k = 1], \ j \neq i \\ [W_k'' \mid \mathcal{I}, C_{k+1} = i] &\sim [W_k \mid Q_k = (q-\mu_i)^+, C_k = i, A_k = m+1]. \end{aligned} \tag{12}$$

Taking expectations of (11) and (12) and unconditioning on $C_{k+1} = j$ yields the result. □

For second moments, the boundary condition is still more complicated. The appropriate recursion is stated next as a corollary of Theorem 3.

Corollary 4 *The functions $\mathcal{W}_k^2(k, i, m, q)$ obey the recursive boundary condition*

$$\begin{aligned} &\mathcal{W}_k^2(k, i, m, q) = \\ &\sum_{j \neq i} \tilde{p}_{ij}(m) \left[(\frac{q}{\mu_j} \wedge 1)^2 + 2(\frac{q}{\mu_j} \wedge 1)\mathcal{W}_k\left(k, j, 1, (q-\mu_j)^+\right) + \mathcal{W}_k^2\left(k, j, 1, (q-\mu_j)^+\right) \right] + \\ &\tilde{p}_{ii}(m) \left[(\frac{q}{\mu_i} \wedge 1)^2 + 2(\frac{q}{\mu_i} \wedge 1)\mathcal{W}_k\left(k, i, m+1, (q-\mu_i)^+\right) + \mathcal{W}_k^2\left(k, i, m+1, (q-\mu_i)^+\right) \right] \end{aligned} \tag{13}$$

with $\mathcal{W}_k^2(k, \cdot, \cdot, 0) \equiv 0$.

PROOF:

Suppose again that at the end of period k the capacity is μ_i, the age is m, and there are q waiting aircraft. As before, let $\mathcal{I}$ denote the conditioning information

$$\{Q_k = q, C_k = i, A_k = m\}.$$

Using (10) we write $E[W_k^2 \mid Q_k = q, C_k = i, A_k = m]$

$$= E[(W_k' + W_k'')^2 \mid \mathcal{I}]$$

$$
\begin{aligned}
&= E[(W_k')^2 + 2W_k'W_k'' + (W_k'')^2 \mid \mathcal{I}] \\
&= \sum_j \tilde{p}_{ij}(m) E[(W_k')^2 + 2W_k'W_k'' + (W_k'')^2 \mid \mathcal{I}, C_{k+1} = j] \\
&= \sum_{j \neq i} \tilde{p}_{ij}(m) \left[(\frac{q}{\mu_j} \wedge 1)^2 + 2(\frac{q}{\mu_j} \wedge 1) E[W_k \mid \mathcal{I}, C_{k+1} = j] + E[W_k^2 \mid \mathcal{I}, C_{k+1} = j] \right] + \\
&\quad \tilde{p}_{ii}(m) \left[(\frac{q}{\mu_i} \wedge 1)^2 + 2(\frac{q}{\mu_i} \wedge 1) E[W_k \mid \mathcal{I}, C_{k+1} = i] + E[W_k^2 \mid \mathcal{I}, C_{k+1} = i] \right].
\end{aligned}
$$

The final equality is a consequence of (11). The result (13) now follows from (12). □

Theorems 1, 2, 3, and Corollary 4 imply the algorithm given in Figure 1. This algorithm obtains values for the expressions

$$\mathcal{W}_k(0, i, m, 0) \equiv E\left[W_k \mid Q_0 = 0, C_0 = i, A_0 = m\right], \quad i = 1, \ldots, S, \; m = 1, \ldots, M.$$

and

$$\mathcal{W}_k^2(0, i, m, 0) \equiv E\left[W_k^2 \mid Q_0 = 0, C_0 = i, A_0 = m\right], \quad i = 1, \ldots, S, \; m = 1, \ldots, M.$$

From these we may obtain expectations and variances of waiting times at the end of each interval, based on given initial conditions. This can be achieved with moderate computational complexity, as the next theorem indicates.

Theorem 5 *The memory requirement for the semi-Markov algorithm is $O(SKMQ_{\max})$ and the running time is $O(S^2K^2MQ_{\max})$, where S is the number of capacity states, K the total number of time intervals, M an upper bound on the memory argument m, and $Q_{\max} \triangleq \max_k q_{\max}(k)$ is the highest attainable queue length over all periods.*

PROOF:

The number of table entries in the above recursion is

$$4 \times S \times M \times \sum_{k=1}^{K} \sum_{l<k} q_{\max}(l). \tag{14}$$

Within iteration k, however, the algorithm needs only to store eight values at a time, $\mathcal{Q}_k(l, i, m, q)$, $\mathcal{Q}_k(l+1, i, m, q)$, $\mathcal{W}_k(l, i, m, q)$, and $\mathcal{W}_k(l+1, i, m, q)$ for the first moments, $\mathcal{Q}_k^2(l+1, i, m, q)$, $\mathcal{Q}_k^2(l+1, i, m, q)$, $\mathcal{W}_k^2(l+1, i, m, q)$, and $\mathcal{W}_k^2(l+1, i, m, q)$ for the second. Thus since $q_{\max}(l) \leq Q_{\max}$ the memory requirement is $O(SKMQ_{\max})$. The bottleneck for the running time is clearly the main recursion for $l < k$, in which we calculate the table

Algorithm for Queue Length and Waiting Time Moments

For $k = 1$ to K (boundary conditions)

For $i = 1$ to S

For $m = 1$ to M

$\mathcal{W}_k(k,\cdot,\cdot,0) = \mathcal{W}_k^2(k,\cdot,\cdot,0) = 0$

$\mathcal{Q}_k(k,\cdot,\cdot,0) = \mathcal{W}_k^2(k,\cdot,\cdot,0) = 0$

For $q = 1$ to $q_{\max}(k,c)$

$\mathcal{Q}_k(k,i,m,q) = q$

$\mathcal{Q}_k^2(k,i,m,q) = q^2$

$$\mathcal{W}_k(k,i,m,q) = \sum_{j\neq i}\left(\tilde{p}_{ij}(m)\left[(\tfrac{q}{\mu_j}\wedge 1) + \mathcal{W}_k(k,j,1,(q-\mu_j)^+)\right]\right) + \tilde{p}_{ii}(m)\left[(\tfrac{q}{\mu_i}\wedge 1) + \mathcal{W}_k(k,i,m+1,(q-\mu_i)^+)\right]$$

$$\mathcal{W}_k^2(k,i,m,q) = \sum_{j\neq i}\left(\tilde{p}_{ij}(m)\left[(\tfrac{q}{\mu_j}\wedge 1)^2 + 2(\tfrac{q}{\mu_j}\wedge 1)\mathcal{W}_k(k,j,1,(q-\mu_j)^+) + \mathcal{W}_k^2(k,j,1,(q-\mu_j)^+)\right]\right) + \tilde{p}_{ii}(m)\left[(\tfrac{q}{\mu_i}\wedge 1)^2 + 2(\tfrac{q}{\mu_i}\wedge 1)\mathcal{W}_k(k,i,m+1,(q-\mu_i)^+) + \mathcal{W}_k^2(k,i,m+1,(q-\mu_i)^+)\right]$$

For $k = 1$ to K (main body)

For $l = k-1$ down to 0

For $i = 1$ to S

For $m = 1$ to M

For $q = 0$ to $q_{\max}(l,c)$

$$\mathcal{Q}_k(l,i,m,q) = \sum_{j\neq i}\left[\tilde{p}_{ij}(m)\mathcal{Q}_k\left(l+1,j,1,(q+\lambda_{l+1}-\mu_j)^+\right)\right] + \tilde{p}_{ii}(m)\mathcal{Q}_k\left(l+1,i,m+1,(q+\lambda_{l+1}-\mu_i)^+\right)$$

$$\mathcal{Q}_k^2(l,i,m,q) = \sum_{j\neq i}\left[\tilde{p}_{ij}(m)\mathcal{Q}_k^2\left(l+1,j,1,(q+\lambda_{l+1}-\mu_j)^+\right)\right] + \tilde{p}_{ii}(m)\mathcal{Q}_k^2\left(l+1,i,m+1,(q+\lambda_{l+1}-\mu_i)^+\right)$$

$$\mathcal{W}_k(l,i,m,q) = \sum_{j\neq i}\left[\tilde{p}_{ij}(m)\mathcal{W}_k\left(l+1,j,1,(q+\lambda_{l+1}-\mu_j)^+\right)\right] + \tilde{p}_{ii}(m)\mathcal{W}_k\left(l+1,i,m+1,(q+\lambda_{l+1}-\mu_i)^+\right)$$

$$\mathcal{W}_k^2(l,i,m,q) = \sum_{j\neq i}\left[\tilde{p}_{ij}(m)\mathcal{W}_k^2\left(l+1,j,1,(q+\lambda_{l+1}-\mu_j)^+\right)\right] + \tilde{p}_{ii}(m)\mathcal{W}_k^2\left(l+1,i,m+1,(q+\lambda_{l+1}-\mu_i)^+\right)$$

END.

Figure 1: Recursive algorithm for queue length and waiting time moments conditional on initial capacity and age conditions

entries for $l < k$. Each such calculation requires $O(S)$ time, so the overall running time has complexity $O(S^2K^2MQ_{\max})$. □

Note that in the more specialized Markov case, the dimension m is unnecessary. Hence the memory requirement for the Markov case is reduced to $O(SKQ_{\max})$ and the running time to $O(S^2K^2Q_{\max})$.

Remarks

1) Theorem 5 indicates that the speed of the recursive method rests on the relative sizes of K, M, and $Q_{\max}$, since S is very small (≈ 5). We note that a typical airport operating day is twenty hours at most ($K = 80$ for $\Delta t = 15$ minutes) and that a practical upper bound on $Q_{\max}$ is 200 (including aircraft held on the ground). A theoretical upper bound for $Q_{\max}$ is

$$Q_{\max} \leq \sum_{k=1}^{K} (\lambda_k - \mu_{\min})^+ ,$$

where $\mu_{\min}$ is the lowest capacity. There is a degree of latitude in the choice of the parameter M. The age m has been introduced into the state space because holding times in each capacity might not be geometric. At a maximum, M is an upper bound on these holding times. As a practical matter, however, above a certain value of m, the transition probabilities $\tilde{p}_{ij}(m)$ often remain fairly constant. Where this is the case, one need only take M high enough to cover the part of the distribution over which the $\{\tilde{p}_{ij}(m)\}$ vary significantly. In the case study of the next section, for example, a value of M as low as 20 proves adequate. At the extreme $M = 1$, a Markov chain replaces the semi-Markov model, with the state space is reduced from $\{i, m\}$ to $\{i\}$, the set of capacities.

2) It is clear that the recursive approach could be used to obtain still higher moments or indeed even the whole distribution of the queue length or waiting time at any given interval. This latter calculation could be achieved by transforms or by direct enumeration of the state space. However, the problem of determining a given term

$$\Pr[Q_k = q \mid Q_0, C_0, A_0]$$

has the same complexity as that of determining any moment. Thus there is an additional factor of $Q_{\max}$ in the complexity of such an approach, i.e. an algorithm for the full distribution would be expected to run about 200 times slower than those above.

3) The algorithms presented thus far are appropriate when the input stream is well approximated as a deterministic flow. We have justified this assumption by the fact that aircraft are deliberately scheduled into their landing slots. On the other hand, congestion and other sources of upstream delay introduce a degree of uncertainty into the arrival schedule which our models have thus far ignored. Particularly in the context of a network of airports, it may be important to take account of this uncertainty. Although we cannot accommodate a fully general stochastic arrival process, we can allow for some degree of uncertainty. Suppose that during period k, the demand Λ_k is a random variable which may take on a finite number of values $\lambda_k^1, \ldots, \lambda_k^R$ with corresponding probabilities $\gamma_k^1, \ldots, \gamma_k^R$. In recognition of this stochasticity, the innermost loop of the recursion is re-written to take the expectation over all possible values of Λ_k. For the expected queue length the main recursion becomes (c.f. (5))

$$\begin{aligned} \mathcal{Q}_k(l,i,m,q) \;=\; & \sum_{r=1}^{R} \gamma_{l+1}^r \Bigg[\tilde{p}_{ii}(m) \mathcal{Q}_k\left(l+1, i, m+1, (q+\lambda_{l+1}^r - \mu_i)^+\right) \;+ \\ & \sum_{j \neq i} \tilde{p}_{ij}(m) \mathcal{Q}_k\left(l+1, j, 1, (q+\lambda_{l+1}^r - \mu_j)^+\right) \Bigg] \end{aligned} \tag{15}$$

with boundary condition $\mathcal{Q}_k(k,\cdot,\cdot,q) \equiv q$. Similarly, for waiting times we have

$$\begin{aligned} \mathcal{W}_k(l,i,m,q) \;=\; & \sum_{r=1}^{R} \gamma_{l+1}^r \Bigg[\tilde{p}_{ii}(m) \mathcal{W}_k\left(l+1, i, m+1, (q+\lambda_{l+1}^r - \mu_i)^+\right) \;+ \\ & \sum_{j \neq i} \tilde{p}_{ij}(m) \mathcal{W}_k\left(l+1, j, 1, (q+\lambda_{l+1}^r - \mu_j)^+\right) \Bigg] \end{aligned} \tag{16}$$

Clearly, these additions to the algorithm multiply the running time by a factor R. Note that the method treats arrival rates $\{\Lambda_k\}$ in different periods as independent. While this extension does not encompass a fully general arrival stream, it does allow some degree of uncertainty to be reflected in the queue statistics. The method is of more value treating congestion in the network environment, as seen in [23].

4) The recursive algorithm as presented obtains moments conditional on the starting state. For planning purposes, these conditional moments may be exactly what is required, or a more general average profile may be desired. It is possible to obtain such a profile via the steady state probabilities for the different starting conditions. More precisely, let

$$\pi(i,m) \triangleq \Pr\{\text{state of the system at time 0 is } (i,m)\}.$$

Then the unconditional mean queue length at the end of interval k is given by

$$\overline{\mathcal{Q}}_k = \sum_{i,m} \pi(i,m)\mathcal{Q}_k(0,i,m,0), \tag{17}$$

while the corresponding mean waiting time is

$$\overline{\mathcal{W}}_k = \sum_{i,m} \pi(i,m)\mathcal{W}_k(0,i,m,0). \tag{18}$$

Clearly the numbers $\pi(i,m)$ correspond to the steady state probabilities for the Markov chain defined on the capacity-age state space $m = 1,\ldots,M, s = 1,\ldots,S$. To calculate them, one must solve the system

$$\pi = \pi\tilde{\mathbf{P}}, \tag{19}$$

where $\tilde{\mathbf{P}}$ is the full set of transition probabilities. Because of the special structure of the state space, the solution to (19) can be obtained by solving a system of only S linear equations. To see this, note that for $m = 2,\ldots,M-1$ we may write

$$\pi(i,m) = \pi(i,1)\prod_{k=1}^{m-1}\tilde{p}_{ii}(k) \tag{20}$$

while

$$\pi(i,M) = \pi(i,1)/(1-\tilde{p}_{ii}(M))\prod_{k=1}^{M-1}\tilde{p}_{ii}(k) \tag{21}$$

Thus the problem of finding the steady state probabilities reduces to that of solving for the S unknowns $\pi(1,1),\pi(2,1),\ldots,\pi(S,1)$ in the system

$$\sum_{j\neq i}\pi(j,1)\left[\sum_{m=1}^{M-1}\left(\tilde{p}_{ji}(m)\prod_{k=1}^{m}\tilde{p}_{jj}(k)\right)+\frac{\tilde{p}_{ji}(M)}{(1-\tilde{p}_{jj}(M))}\prod_{k=1}^{M-1}\tilde{p}_{jj}(k)\right] = \pi(i,1) \tag{22}$$

$$\sum_{i=1}^{S}\left[\pi(i,1)\left(\sum_{m=1}^{M-1}\prod_{k=1}^{m-1}\tilde{p}_{ii}(k)+\frac{1}{(1-\tilde{p}_{ii}(M))}\prod_{k=1}^{M-1}\tilde{p}_{ii}(k)\right)\right] = 1 \tag{23}$$

Thus the enlargement of the state space via the age process A_k does not severely affect the computation of the steady state probabilities. One solves equations (22) and (23) for the probabilities $\pi(i,1)$ and then uses the relations (20) and (21) to solve for the others.

5) The algorithm is unable to provide waiting time distributions without significant compuational expense. However, through the first two moments we can develop a useful approximation motivated by simulation results. Consider a simple simulation in which capacity period

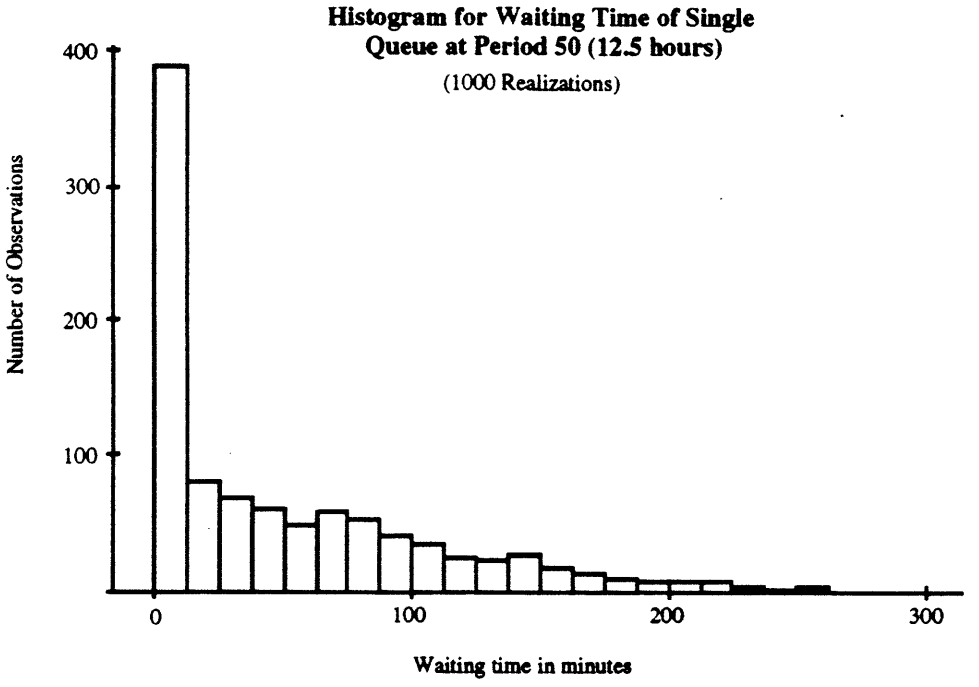

Figure 2: Histogram from simulated waiting times in a single queue

by period is determined in Monte Carlo fashion from the Markov chain or semi-Markov process. From the simulation we obtain the matrix of observations

$$\mathbf{W} = \{W_k^n\},$$

where W_k^n is the waiting time at the end of period k for the nth simulation. Ordering the observations, we obtain histograms for the waiting times for each period, like the one illustrated in Figure 2 for a constant arrival rate ($\rho \approx 0.85$, $\lambda = 60$ per hour). Note the presence of a substantial probability mass at the minimum value (in this case, 0). Values above this minimum seem to follow an approximately exponential distribution. This is confirmed in Figure 3, which plots the transformations

$$y^{(n)} = e^{(-\nu w^{(n)})},$$

where $\{w^{(n)}\}$ are the ordered values of observations which exceed the minimum and $1/\nu$ is their mean. If the underlying distribution were truly an exponential, this plot should be a straight line sloping down to the right, since the numbers $\exp(-\nu w^{(n)})$ are realizations of the reverse cumulative distribution $\overline{F}(w)$ and should behave like the reversed order statistics

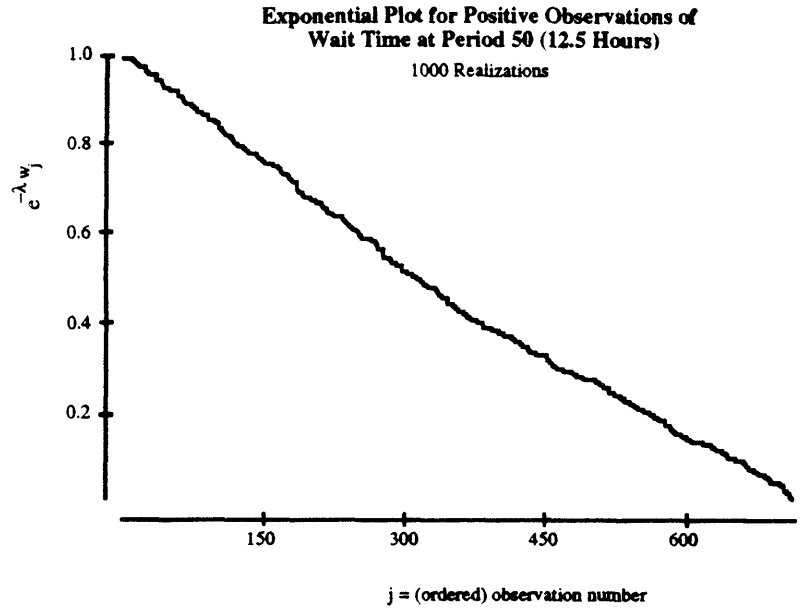

Figure 3: Test for exponential distribution of positive waiting time realizations

of a $U[0,1]$ distribution. Plots such as this one suggest an approximate mixed distribution for the waiting times W_k:

$$\begin{aligned} \Pr\{W_k = w_{\min}(k)\} &= \delta \\ \Pr\{W_k \leq w \mid w > w_{\min}(k)\} &= 1 - e^{-\nu(w - w_{\min}(k))} \end{aligned} \tag{24}$$

The parameters $w_{\min}(k)$, usually but not always 0, can be calculated directly from the recursion in a manner similar to that for the parameters $q_{\max}(k)$. The two numbers δ and ν can be estimated using the first two waiting time moments and solving the two equations (omitting the subscript for clarity)

$$\begin{aligned} \delta w_{\min} + (1-\delta)\int_{w_{\min}}^{\infty} w\nu e^{-\nu(w - w_{\min})}\, dw &= E[W] \\ \delta (w_{\min})^2 + (1-\delta)\int_{w_{\min}}^{\infty} w^2\nu e^{-\nu(w - w_{\min})}\, dw &= E[W^2] \end{aligned} \tag{25}$$

6) The above simulation procedure also provides a check on the algorithm. Simulating capacities, tracing the resulting changes in the queue, and taking averages of the resulting sample paths over different simulation runs, we obtain the results of Figure 4. The slight under-estimation which the simulation gives suggests that the tail occurrences for the waiting

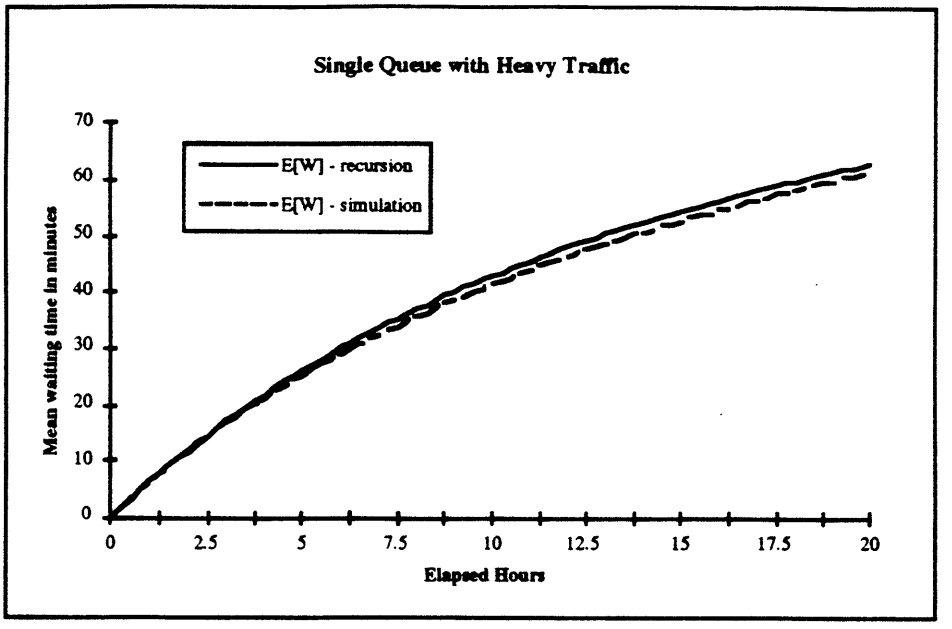

Figure 4: Comparison of mean waiting times estimated by simulation and by the recursive algorithm for a single queue with constant demand and heavy traffic

times are not sufficiently sampled. These occurrences correspond to extended periods of low capacity and occur with very low probabilities (less than 10^{-6}). Although these tail occurrences do not have a large effect on the means, our experiments show them to have a more noticeable effect on the standard deviations, as we would expect.

7) In [23] we have reported results for an alternative approach to the queueing problem using a diffusion approximation. Consider successive aircraft periods indexed by k. Let ξ_k denote the number of services in period k. Let arrivals be deterministic at rate λ, and let the steady state capacity probabilities be given as $\pi_1, \ldots, \pi_S$, so that the time average capacity is

$$\overline{\mu} = \sum_{i=1}^{S} \pi_i \mu_i.$$

We can define a traffic intensity

$$\rho \triangleq \frac{\lambda}{\overline{\mu}}$$

and speak of a *heavy traffic limit* as $\rho \rightarrow 1$. Application of the diffusion equation is not entirely straightforward because of capacity correlations over time. In fact, estimation of the

key coefficients requires a result (due to Keilson and Wishart [15,16]) concerning a central limit theorem for additive processes on a Markov chain. Details of our approach may be found in [23]. Our results indicate that while the diffusion approach deviates significantly in its estimates from the recursive method, it does capture the same qualitative behavior. Because of its greater computational speed, the diffusion approach may prove valuable in network models, provided that it can be adapted for a time-varying arrival rate. See [23] for further details.

4 Dallas-Fort Worth: A Case Study

In this section we discuss an application of the one-hub recursive model of Section 3 to the case of Dallas-Fort Worth International Airport. The Dallas-Fort Worth International Airport (DFW) is an ideal airport for studying the effectiveness of the delay model. It ranks among the highest in the nation in terms of delay problems, with only the three New York area airports, San Francisco, and Chicago having significantly greater numbers of delays in 1989 [26]. Its delay problems are largely due to the high level of traffic resulting from the dual hub presence of American and Delta Airlines, which together account for 75% its operations.

4.1 Model Implementation and Validation

Arrival traffic at DFW falls into four categories: air carrier, air taxi, military, and general. Among these, the first two types account for almost all of the traffic. A typical daily demand schedule is illustrated in Figure 5. Adopting the convention $\Delta t = 15$ minutes, we have grouped flights according to the 15-minute interval in which they arrive. The peaked pattern reflects 12 American Airlines and 11 Delta Airlines banks.

In favorable weather conditions DFW has three runways available to handle landing aircraft. However, in less favorable conditions, only two runways are available for landings, and capacity is correspondingly reduced. Specific landing capacity at any time depends upon the runway configuration in use, and this is turn depends on elements of the weather: wind speed, wind direction, cloud ceiling, and horizontal visibility. Considering these factors, we chose a total of six capacity states for DFW. Table 1 lists these six states together with the associated engineered performance standards (EPS) in aircraft per hour. The

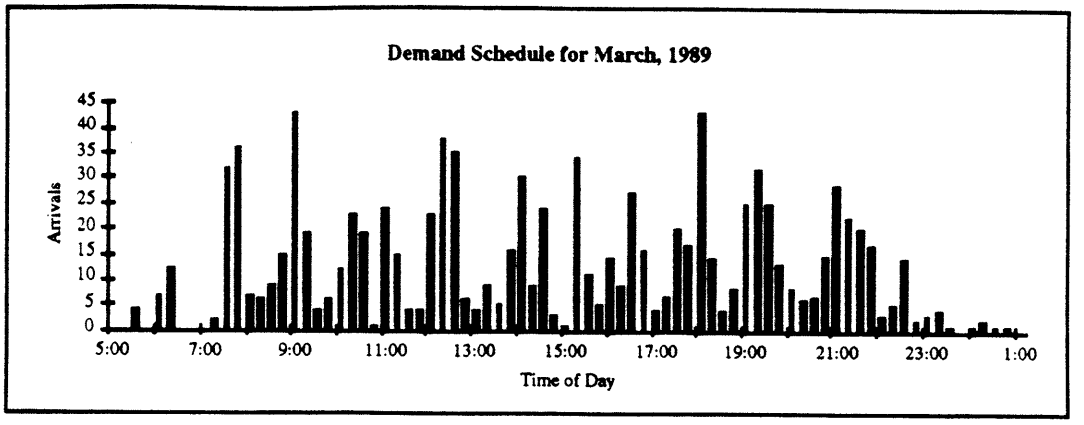

Figure 5: Arrival schedule at DFW for March 1989. Sources: DOT, OAG, and DFW Airport Authority

State	Description	Landings per Hour (EPS)
A	IFR-2 & lower	50
B	IFR-1	60
C	VFR-2, windy	66
D	VFR-1, windy	70
E	VFR-2, still	90
F	VFR-1, still	95

Table 1: Engineered performance standards at DFW. Source: Dallas-Fort Worth Airport Authority

abbreviations 'IFR' and 'VFR' stand for instrument and visual flight rules, respectively. There is a substantial difference between the two highest capacity states and all other states, due to the availability of the third runway.

In practice these EPS estimates are considered conservative for high-capacity configurations. To compensate, we have used preliminary results of an ongoing study by UNISYS Corporation [8] estimating runway capacity per hour from observations of peak periods. These put the highest arrival capacity state at DFW in the range of 115 aircraft per hour, a substantial increase over the EPS number 95. Thus far, UNISYS has provided no further estimates for other configurations, but it is reasonable to expect a similar increase for state 'E', while the 4-runway configuration estimates should remain essentially unchanged. We

have adopted these changes for the capacities in this study and note that the need for more accurate capacity estimation procedures seems obvious.

Because historical capacity data were not available to us, we were forced to reconstruct capacity histories from weather data obtained from the National Oceanic and Atmospheric Administration (NOAA). Simple tabulation of eight years of hourly observations reveals that the six capacities at DFW shown in Table 1 occur with quite different frequencies. Over the course of a year, the highest capacity state (configuration 'F') is observed about 80% of the time, while IFR conditions (states 'A' and 'B') occur only about 6% of the time in total. Seasonal variability is high. Not surprisingly, lower visibility conditions tend to occur more in the winter; indeed, in summer, occurrences of this worst state are exceedingly rare. Because of this seasonal variation, we chose a particular month (March) and based the parameter estimates on data for that month only. Configuration 'F' constitutes about 75% of March observations.

From the data we estimated three sets of parameters: the transition matrix

$$\mathbf{P} = \{p_{ij}\}$$

for the Markov model, and for the semi-Markov model the transition matrix

$$\tilde{\mathbf{P}} = \{\tilde{p}_{ij}\}$$

as well as the holding time probabilities

$$\Pr[T_i = m].$$

Details of the estimation procedure are found in [23].

Recall from the earlier discussion that while the semi-Markov model is less restrictive than the Markov model, its run time is higher by the factor M. Thus a question of interest is how well a Markov hypothesis fits the weather observations.

To examine this question we consider the hourly observation process. For given state i, we define a *run of length* m to be the event that this state is observed exactly m consecutive times in the hourly observation process. Let $N(i,m)$ be the number of runs of length m for state i, and let

$$N(i) \triangleq \sum_{m \geq 1} N(i,m). \tag{26}$$

state	occupancy probability	
	expected	actual
A	3.13%	3.06%
B	2.06%	2.05%
C	1.01%	1.01%
D	6.36%	6.36%
E	11.97%	11.95%
F	75.47%	75.58%

Table 2: Predicted and actual occupancy probabilities at DFW

For a particular state i, the collection of $N(i,m)$ over all values of m constitutes a kind of histogram for the holding periods. Informally, we can compare the observed frequencies of the $N(i,m)$ (the numbers $N(i,m)/N(i)$) with the probabilities $\Pr[M_i = m \mid M_i \geq 1]$, where M_i is a random variable representing the length of a run for state i. In Figure 6, the smooth curves indicate predicted distributions, while the jagged lines connect the data points.

Several features are quite striking. First of all, notice that states 'B', 'C', and 'D' tend to have very short durations, states 'A' and 'E' short to medium durations, and state 'F' short to very long durations. In fact, the full tail of the 'F' histogram extends into the hundreds of hours, though this is not shown in the figure. Second, notice that all six distributions have a probability mass at 1 hour which is higher than that predicted by the Markov model.

In fact, these fits do rather poorly on a formal χ^2 test. On the other hand, while geometric holding times do not conform exactly to the data, the state occupancy probabilities (i.e. the numbers π_i) predicted under a Markov chain hypothesis are *extremely* close to the time-fractions observed in the data (i.e. the numbers $N(i)/\sum_i N(i)$) — see Table 2).

The results of our initial run of the model for DFW data are summarized in Figure 7, which plots the unconditional expected waiting times

$$\overline{W_k} = \sum_i \pi_i E[W_k \mid Q_0 = 0, C_0 = i]$$

based on traffic estimates for March 1989 and on a Markov capacity model with parameters drawn from eight years of March data. The familiar peaking pattern is evident and testifies to the deterministic effect produced by high traffic concentrations at particular times of day — the morning American and Delta complexes, the noon double complex (Delta following

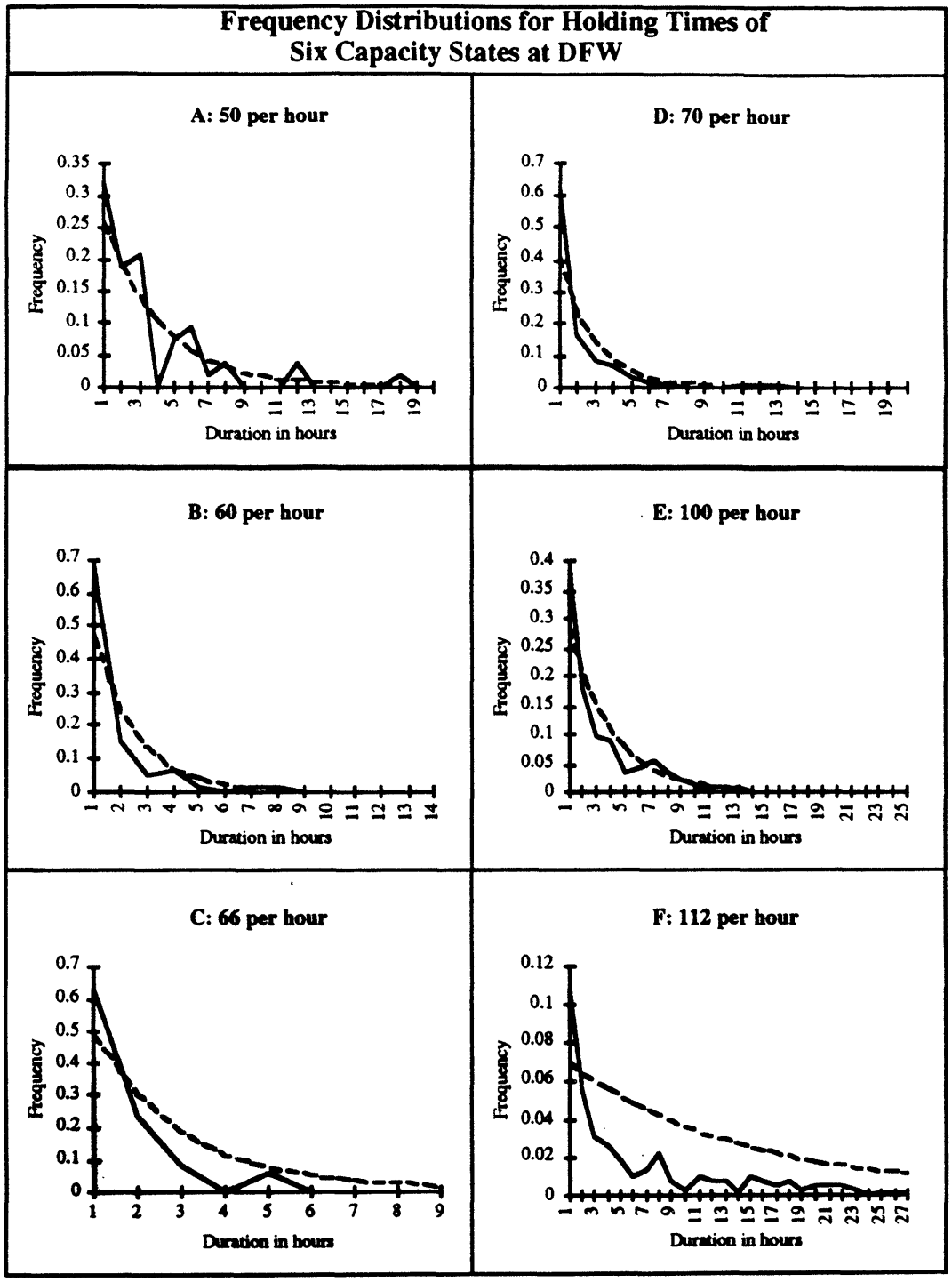

Figure 6: Examining goodness of fit for the Markov model. The solid lines indicate the observed frequencies for run lengths, while the dashed lines indicate the expected frequencies under a Markov chain model.

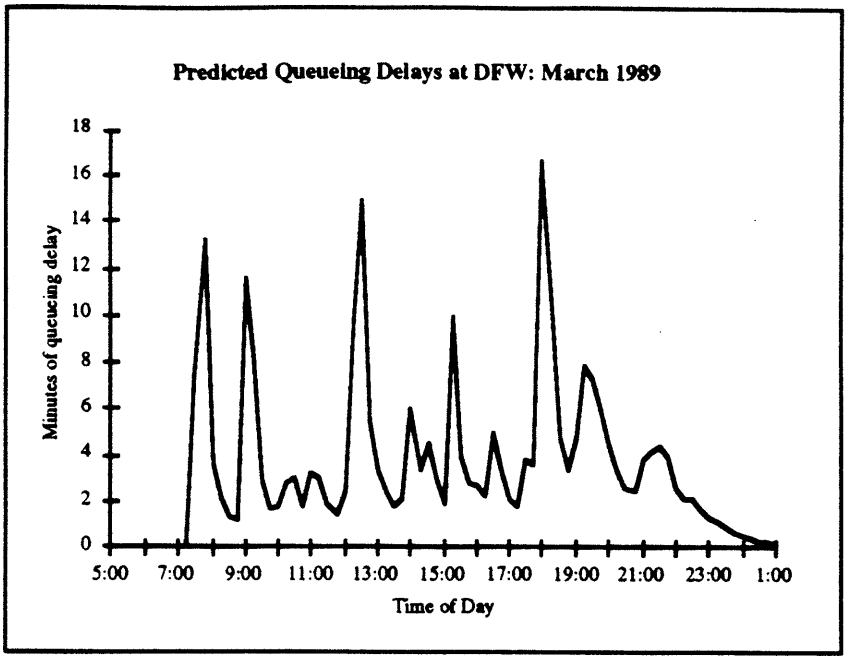

Figure 7: Expected waiting times at DFW based on March weather and 1989 traffic

American), and the 6:00 p.m. double complex (Delta again following American).

Despite the fact that overall capacity exceeds demand substantially ($\rho \approx 0.5$), there are short periods where mean waiting time reaches 15 minutes, a good illustration of how *overall system capacity may be more than adequate even while short periods show significant capacity shortfalls.* Delays during non-peak periods are, not surprisingly, close to 0. Queue lengths are not shown in the figure, but they follow the same pattern as the waiting times.

Unfortunately, full validation of these results is prevented by inadequacies in necessary data. The only publicly available data relavant to our model are the On Time Arrival Statistics kept by the U.S. Department of Transportation. These monthly statistics for all domestic flights include scheduled departure times, actual departure times, scheduled arrival times, and actual arrival times. From these statistics one can calculate the apparent delay of a flight at a given time of day. However, this apparent delay is not necessarily the immediate landing-related queueing wait experienced by that flight. First of all, lateness of a flight's departure and arrival may easily reflect queueing waits encountered by the same aircraft on a *different* flight earlier in the day. Furthermore, lateness may reflect other causes of delay having nothing to do with landing congestion, such as departure congestion, flight

slow-downs or speed-ups due to wind, and lack of available gate space.

These drawbacks seriously hamper the degree to which we can validate the model's numbers. Indeed, the only way to achieve the necessary precision for a full validation would be to collect the data specifically for our objectives, controlling for the factors mentioned. Bearing these remarks in mind, we present the results of a limited comparison of our model's predictions with the DOT statistics.

The number to be compared with the waiting times predicted by the model is that of *total lateness per flight*, which is defined as follows. A flight's *actual flight time* (FTA) is the time taken from its actual departure time (leaving the gate) until its actual arrival time at the destination gate. Its *scheduled flight time* (FTC) reflects the difference between scheduled departure and arrival times. If we take the simple difference of these, FTA-FTC, we obtain a measure of a flight's delay *not counting differences between actual and scheduled departure times*. Since these differences may be due to ground holds at the origin airport, we add back the departure delay (DD), which is the difference between scheduled and actual departure times. Finally, because different scheduled flight times exist for the same origin/destination (even for a single carrier), we replace FTC by an average number (AFTC) computed for each O-D pair. Thus we obtain for each flight i the *total delay*

$$\mathrm{TD}_i = \max\{\mathrm{FTA}_i - \mathrm{AFTC}_i + \mathrm{DD}_i, 0\}.$$

As noted above, this statistic has many faults. It includes all possible causes of delay, not only that of landing congestion. To correct for outliers, we grouped observations by day and scheduled arrival time, took group means and standard deviations, and then threw out observations more than 3 standard deviations above the mean. Such a procedure helps to discard observations reflecting long delays due to reasons other than congestion. We ordered the remaining observations by scheduled arrival time, grouped them in 15-minute intervals (recall $\Delta t = 15$), and calculated means. The solid line of the figure gives these average cumulative delays for aircraft scheduled to land at various times of day. For example, the average cumulative delay for an aircraft scheduled to land at 10 a.m. is about 6 minutes. Note that there are a few gaps in the plot of the solid curve, reflecting the fact that at a few times of day there is no scheduled jet service at DFW (prop service is not included in the DOT numbers).

The dotted line in the figure gives the average *landing congestion waiting times* predicted by the Markov model. The two curves do not match closely, nor should we expect them to in

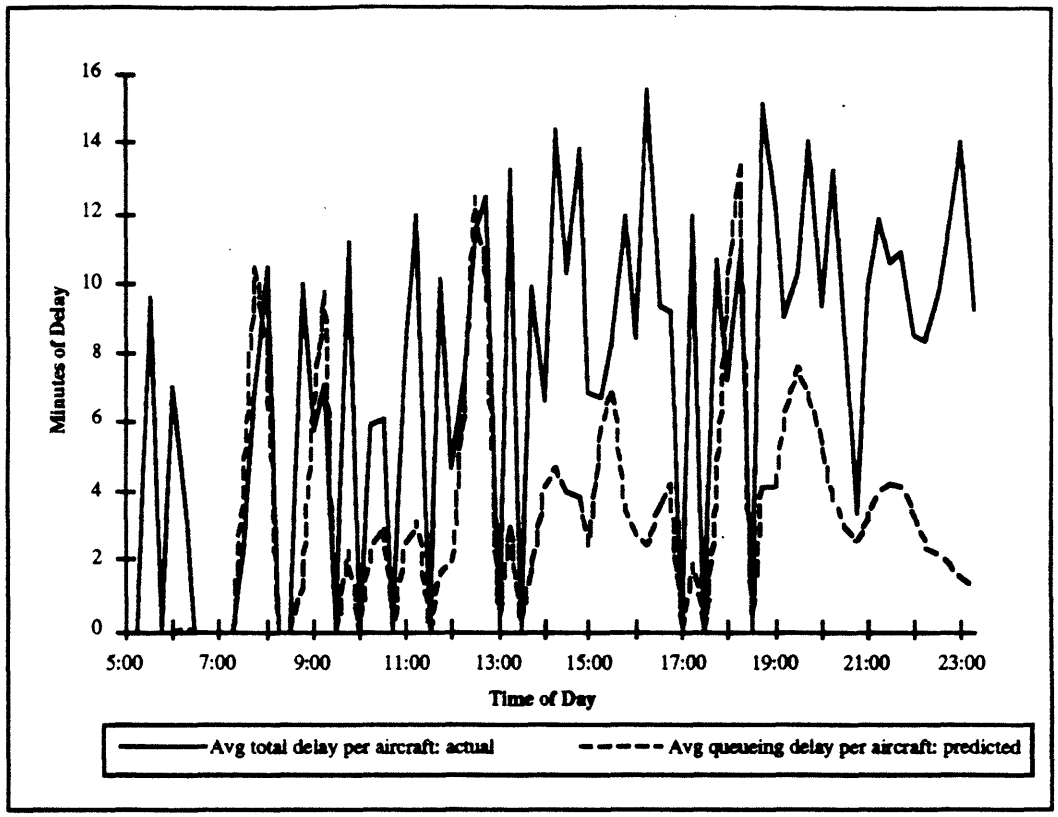

Figure 8: Predicted waiting times at DFW (from queueing model) compared with average total aircraft delays from adjusted DOT statistics

light of the above remarks. Not surprisingly, the DOT average delays are almost uniformly higher than the queueing delays predicted by the model, a reflection of the fact that they are indeed *total* delays.

We define a "standard error"

$$s = \sqrt{\frac{\sum_{i \in \mathcal{I}} (\mathrm{TD}_i - \mathrm{PQD}_i)^2}{|\mathcal{I}|}}$$

where

$\mathrm{PQD}_i \triangleq$ Predicted queueing delay for period i

$\mathcal{I} \triangleq$ Set of periods for which delay observations are available.

For the plot, the value of s is 6.7 minutes, which is approximately 2/3 of the actual average delay (9.46 minutes). The sum of the predicted queueing delays ($\sum_i \mathrm{PQD}_i$) is about half

the sum of the actual total delays ($\sum_i \mathrm{TD}_i$) — 250 minutes versus 540 minutes. Note that the major discrepancies occur in mid-afternoon and late evening. However, traffic at those times is in fact quite *low*. At least at these times of day, large discrepancies probably reflect *delays carried over from earlier portions of the day*. The low-traffic time of 5:30 a.m. to 6:30 a.m. displays a similar large difference which cannot be attributed to congestion on arrival. An explanation may lie in the fact that these early banks are mainly flights from the west coast and Hawaii, with long flight times and late evening departure times (airlines are more likely to hold flights at these times of day as a service for late passengers).

On balance, Figure 8 is a better indication of the shortcomings in the data than of the accuracy of our queueing model. However, in the absence of a fully controlled validation experiment, we must be careful in the strength of the conclusions we draw. Thus our discussion in the following section is mainly confined to qualitative rather than quantitative issues.

4.2 Results and Discussion

In this section we explore some of the implications of the model's results at DFW.

Markov vs. Semi-Markov Model

The first question of interest is whether there is a significant difference between the Markov and semi-Markov models. Figure 9 plots mean waiting times (averaged over initial conditions) for both models. The focus on only part of the day is made to facilitate faster run-time for the semi-Markov model, which with $M = 20$ has run times on the order of 2 hours on a DEC-3100 workstation (for $K = 80$ periods) versus 5 minutes for the Markov model. As is evident from the figure, the differences between the two models are quite small and could easily have been produced by quirks in the estimation procedures. Although we did not expect this close agreement between the two approaches at the outset of the case study, the finding is a pleasant surprise and a reminder that simplicity in modeling is always a worthwhile goal. Because of the close agreement and the greater speed of the Markov model, the remainder of the discussion focuses on the results obtained from it alone.

Stochastic vs. Deterministic Models

An examination of the profiles predicted by the Markov and semi-Markov models sug-

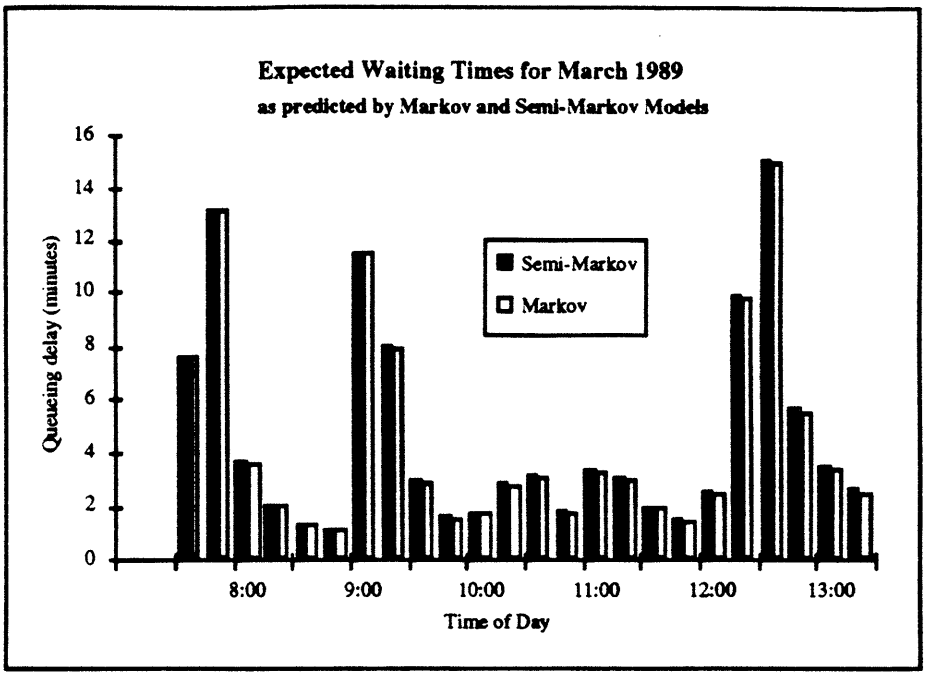

Figure 9: Comparison of predictions of expected waiting times at DFW under the Markov and semi-Markov models

gests that the mean waiting times which emerge from our calculations mainly reflect high capacity acting upon demand in peak periods. Recall that capacity at Dallas is in one of the top two states approximately 85% of the time. Thus the question arises: how do the results of a stochastic model compare with a purely deterministic analysis? As an answer, consider Figure 10. Here we have employed a purely deterministic model with a constant capacity equal to the time-average capacity at DFW:

$$\overline{\mu} = \sum_i \pi_i \mu_i.$$

The figure plots the mean waiting times predicted by a simple deterministic model together with the mean waiting times predicted by the Markov chain model. Not surprisingly, during the peak periods of the day, the two curves agree closely, because the deterministic effect $\lambda > \mu$ is the dominant factor in determining delays at these times. During slack periods, however, the picture is much different. While the deterministic model predicts very low average waiting times, the predictions of the stochastic model are significantly higher. The explanation is that at these times of day, the major cause of waiting is the presence of a queue

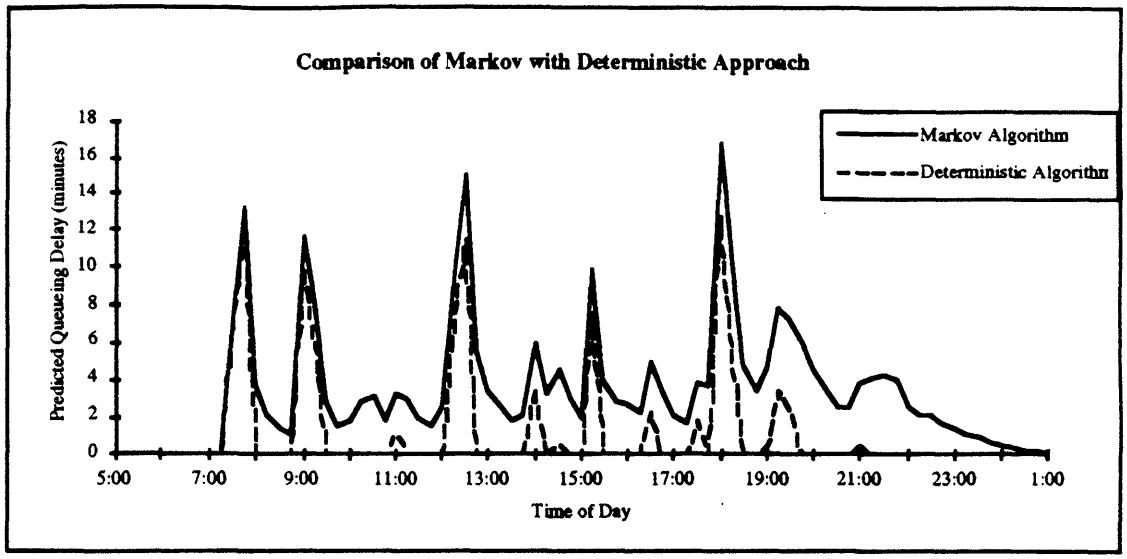

Figure 10: Comparison of predictions of expected waiting times at DFW under Markov and deterministic models

of aircraft which has formed because of earlier high demand combined with low capacity. Because the deterministic model assumes a constant service rate, it does not account for the possibility of such low capacity, and it therefore under-predicts waiting times. The figure demonstrates the advantage we gain by using the more sophisticated stochastic models.

Effect of Correlation in Service Rates

An important phenomenon at DFW is that of correlation in service capacity over time. More precisely, the high probabilities of self-transitions estimated for the Markov chain indicate that when the airport begins the day in a given capacity state, it is likely to remain in it for a significant length of time. This phenomenon in turn implies that mean queue lengths and waiting times will look quite different conditional on different starting states. Figure 11 plots two waiting time profiles based upon the starting states 'A' (lowest capacity) and 'F' (highest capacity). Note that waiting times in the former case are higher by an approximate factor of 3 throughout the day. Moreover, since these profiles are *averages* of sample paths, the peaks approaching 40 minutes indicate the possibility of *very* long delays.

To examine the effect of correlation further, we consider an alternative, less realistic congestion model where the capacities from period to period are i.i.d. and the probability

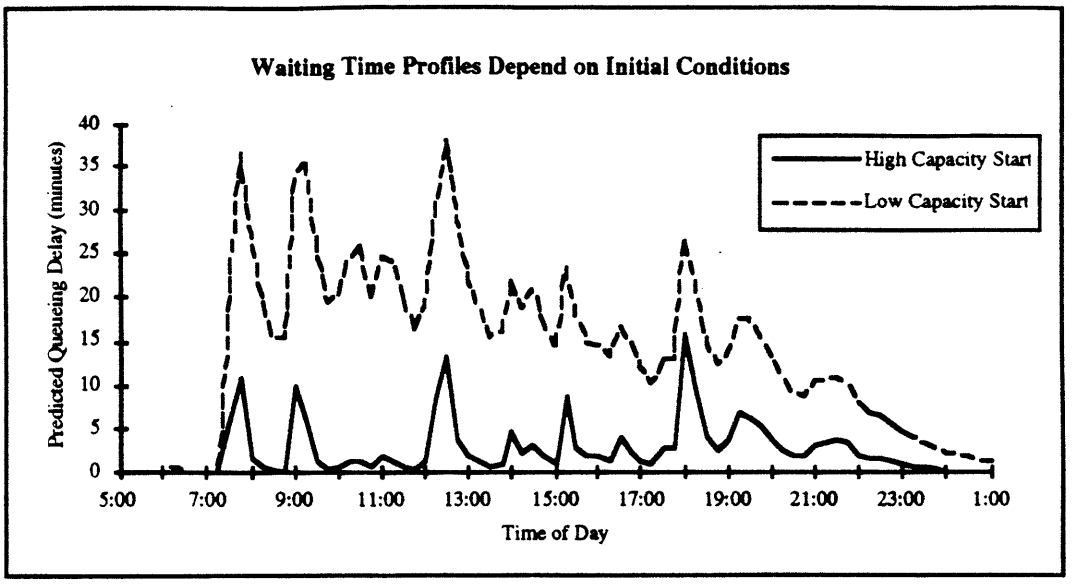

Figure 11: Capacity correlation means that initial conditions are important in determining expected waiting times.

of a given state i in any period is equal to the steady state probability π_i. This change should reduce predicted mean waiting times, a fact which is confirmed by Figure 12. Note that the Markov model has only slightly higher estimates than the independent model for *peak periods* — the deterministic effect once again. The contrast is greater, however, in the slack periods. At these times, the i.i.d. model reflects a lack of memory: delay dies out. This phenomenon is not observed under the Markov model, where correlation is taken into account and delay is more likely to persist. While this effect is small for the case shown here (average over initial conditions), it can be much greater in low capacity situations.

Schedule Interference

It is an interesting fact that at DFW during the busiest times of the day, Delta's banks tend to follow closely after American's, with greater schedule slack separating the Delta banks from subsequent American banks. This type of scheduling suggests that Delta may bear a share of delay at Dallas out of proportion to its level of traffic, since it is more likely to be subject to holdover congestion delay from the preceding American bank. The phenomenon is illustrated in Figure 13. Here, we have labeled the four highest delay peaks where the two carriers have arrival banks in close proximity. In each case, the label indicates

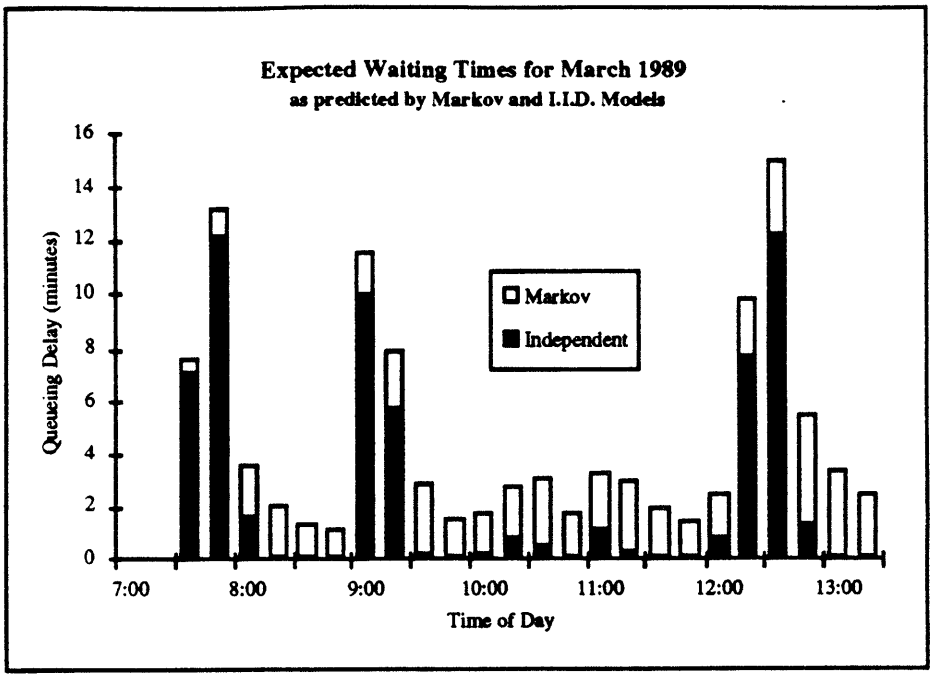

Figure 12: Comparing Markov and i.i.d. models illustrates the effects of correlations in capacity from period to period.

the carrier which is second in the order. In all but the early morning peak, Delta follows American. The figure suggests that Delta's schedule position may increase its queueing delays.

To test this idea, from the DOT data we selected all reported flights for March 1989 with scheduled arrival times during one of the four periods labeled in the preceding figure: 7:15 a.m. to 7:45 a.m., 8:45 a.m. to 9:15 a.m., 11:45 a.m. to 12:30 p.m., and 5:40 p.m. to 6:10 p.m. We refer to these double banks by the numbers 1-4, respectively. Within each bank, we grouped flights according to carrier (American or Delta) and computed the average total delay over all flights (defined as in the earlier validation discussion, with the exception that outliers are not removed). Table 3 presents the results. For banks 1 and 3, the second carrier in the order (American for bank 1, Delta for bank 3) has the higher delays, while for banks 2 and 4, American has higher average delays despite coming first in the order (see the fourth column of the table). The evidence seems mixed. However, it is important to note that in every bank, American has a larger number of flights. Since in the two early morning banks there is still some separation between American and Delta, this higher traffic would

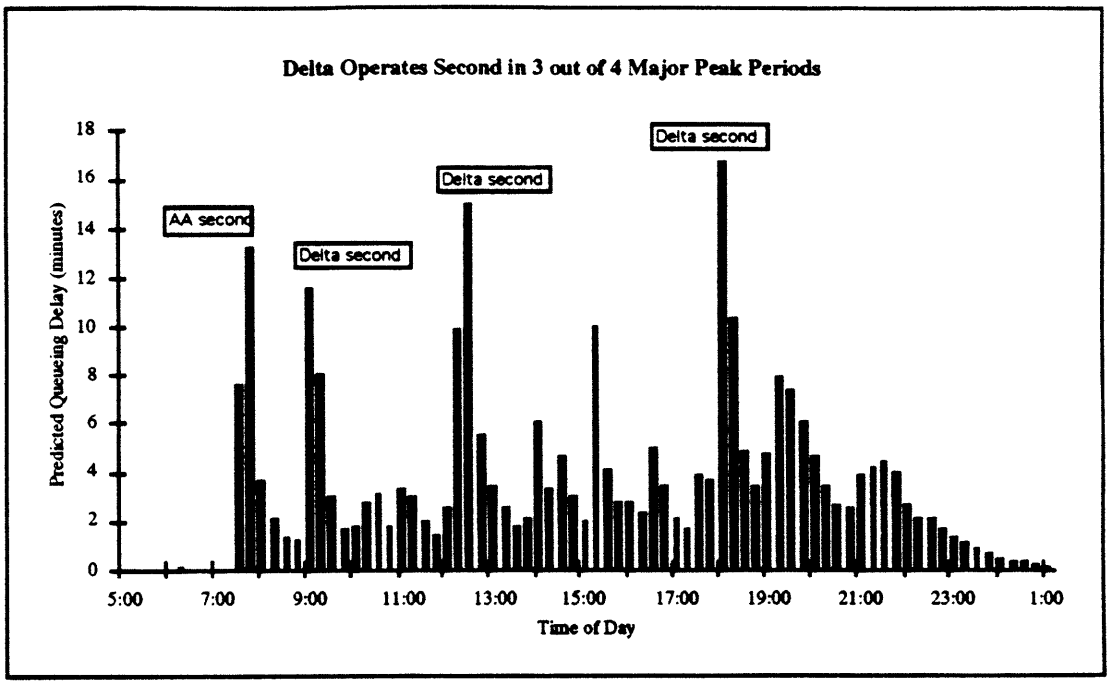

Figure 13: The four major double banks at DFW, labeled with the 2nd scheduled carrier in each case. Although both major carriers at DFW are affected by delays, Delta may bear a higher risk of waiting since its peaks are mostly scheduled right after American's.

tend to increase American's queueing delays. In the case where the two carriers' banks actually overlap significantly (bank 3), Delta shows higher average delays even with less traffic. Moreover, American's delays are only significantly higher than Delta's in the one case where it is scheduled second (bank 1). Overall, the data suggest that schedule position does play a role, but the effect is probably only important when banks actually overlap.

Demand Smoothing

The issue of schedule interference is related to the larger question of how the demand peaking at Dallas affects delay. During recent years, congestion-related pricing of capacity has been proposed as a potential way to reduce delays by smoothing the demand pattern over the day. What effects would such smoothing produce at DFW? To explore this question, consider a hypothetical smoothing policy in which we *impose a maximum limit L on the number of arrivals for any 15-minute period.* For periods of the day which violate the limit, extra flights are shifted to the nearest period in which there is room (either prior or

Bank I.D.	Carrier	No. of Arrivals	Average Total Delay per Aircraft
1	American	19	9.2
1	Delta	15	4.5
2	American	31	7.1
2	Delta	13	6.2
3	American	34	9.6
3	Delta	19	10.4
4	American	29	11.1
4	Delta	22	9.4

Table 3: Comparison of average aircraft delays for Delta and American during the four major double-banks

subsequent). The resulting schedule is a smoothed version of the original, with the parameter L determining the degree of smoothing. Naturally, we expect that for lower values of L there will be greater reductions in delay at increasing inconvenience cost (displaced flights).

Smoothing policies for $L = 28$ and $L = 20$ arrivals per 15-minute period are illustrated in Figure 14, which also reproduces the actual demand schedule for March 1989. The case $L = 28$ reduces traffic so that it never exceeds the estimate for highest capacity state 'F'. We term this level of smoothing "moderate" — to the extent that 112 aircraft per hour is a hard upper bound on landing capacity, moderate smoothing represents a rationalization of the schedule to reflect capacity realities. The $L = 20$ policy goes much further, introducing excess capacity approximately 85% of the time at Dallas. We term this level of smoothing "severe."

Figure 15 reproduces the average case congestion profile for March 1989, as well as the hypothetical profiles of what delay would look like under the smoothed schedules. Improvement is dramatic during peak periods — well over a 50% reduction in waiting time. Similar reductions are not achieved for the non-peak periods, but waiting times during these periods are already fairly small. Weighted average aircraft delays are shown in the second column of Table 4. In moving from no smoothing to severe smoothing, there is a reduction in weighted average delay of about 60%. This represents about 3 minutes on average, but of course much more than that during the peaks. The key observation to be made is that *most of the reduction in delay (46%) is achieved in moving from the normal schedule to moderate*

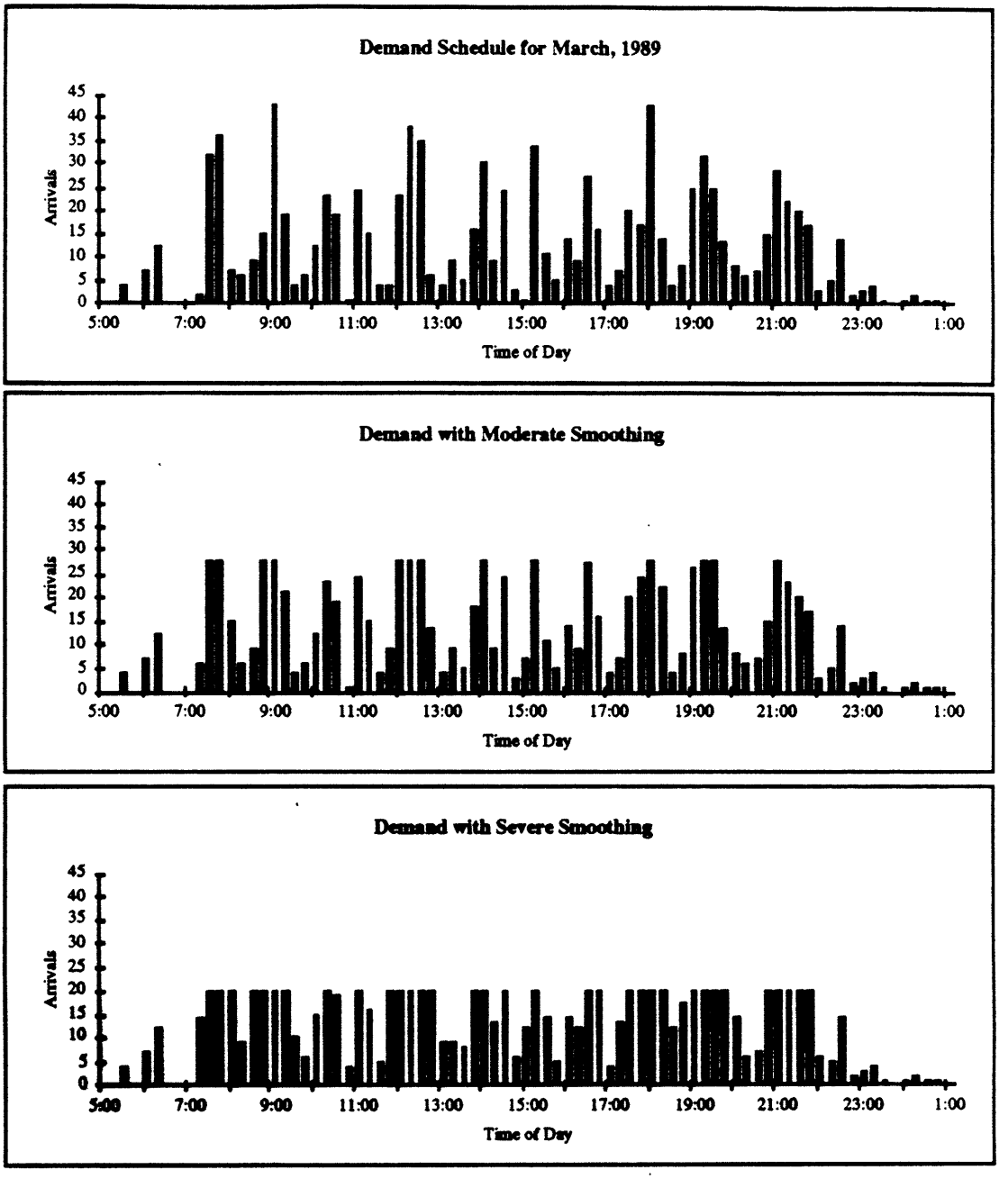

Figure 14: Alternative degrees of smoothing for DFW traffic

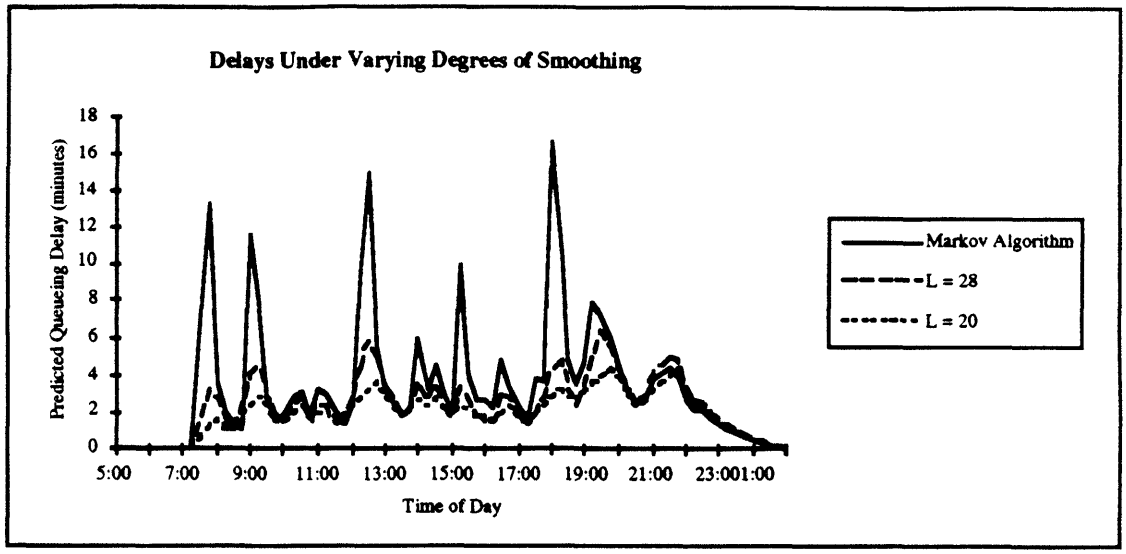

Figure 15: Predicted effects of traffic smoothing on waiting times

Smoothing Policy	Percent of Flights Shifted	Average Delay (mins)
None	—	6.05
Moderate	7.23%	3.29
Severe	17.37%	2.43

Table 4: Costs and benefits of smoothing policies

smoothing; reduction beyond this level of smoothing is relatively modest. Diminishing returns exist.

The cost of the smoothing policies is difficult to assess. Banks with very high scheduled traffic are smoothed significantly and become much longer. Table 4 lists the percentages of flights shifted from their original periods under the two smoothing schemes: around 7% in the moderate case and around 17% in the more severe case. Thus in addition to exhibiting diminishing returns, the smoothing policies also exhibit increasing costs. From the standpoint of costs and benefits, therefore, it seems that moderate policies of demand smoothing are better than excessive ones. The figure demonstrates the usefulness of the queueing model in assessing the effects of policy options.

5 Conclusion

In this paper we have developed a non-traditional queueing model in response to an important problem in practice: congestion at hub airports. Our approach explicitly models variation in airport capacity dependent on weather conditions and exploits the structure of that model to obtain an efficient algorithm. Analyses based on the model highlight a number of interesting features of the problem, especially the large amount of variability due to large differences between alternative sample paths and to the serial correlation in the capacity process. In the realm of strategy and policy, the model points out the reality of interaction between carriers at a hub and suggests that in the case of DFW, schedule position can affect queueing delay. Our analysis also suggests that the high degree of schedule peaking at DFW is responsible for many of the day-to-day delays. Traffic smoothing policies can reduce these delays and rationalize airlines' schedules, but smoothing beyond a certain level is likely to create a degree of excess capacity with high opportunity cost for the carriers.

References

[1] STEPHANIE F. ABUNDO. *An Approach for Estimating Delays at a Busy Airport,* Master's Thesis, Operations Research Center, Massacusetts Institute of Technology, Cambridge, MA, 1990.

[2] ELIZABETH E. BAILEY, DAVID R. GRAHAM, AND DANIEL P. KAPLAN. *Deregulating the Airlines,* M.I.T. Press, Cambridge, MA, 1985.

[3] DIMITRIS J. BERTSIMAS, JULIAN KEILSON, DAISUKE NAKAZATO, AND HONG-TAO ZHANG. "Transient and Busy Period Analysis of the $GI/G/1$ Queue: Solution as a Hilbert Problem," ***Journal of Applied Probability* 28**, 873-85 (1991).

[4] DIMITRIS J. BERTSIMAS AND DAISUKE NAKAZATO. "Transient and Busy Period Analysis of the $GI/G/1$ Queue: The Method of Stages," ***Queueing Systems* 10**, 153-84 (1992).

[5] ALFRED BLUMSTEIN. *An Analytical Investigation of Airport Capacity,* Cornell Aeronautical Laboratory Report TA1358-6-1, Cornell University, Ithaca, NY, June, 1960.

[6] J.A. DONOGHUE. "A Numbers Game," *Air Traffic World,* December, 1986.

[7] E. GELENBE AND I. MITRANI. *Analysis and Synthesis of Computer Systems,* Academic Press, Inc., London, 1980.

[8] EUGENE GILBO. "Arrival-Departure Capacity Estimates for Major Airports," ATMS/ETMS Project Memorandum, UNISYS Corporation, Cambridge, MA, November 1, 1990.

[9] W.K. GRASSMANN. "Transient Solutions in Markovian Queueing Systems," *Computers and Operations Research* **4**, 47-56 (1977).

[10] DONALD GROSS AND CARL M. HARRIS. *Fundamentals of Queueing Theory,* 2nd Edition, John Wiley and Sons, New York, NY, 1985.

[11] DANIEL P. HEYMAN AND MATTHEW J. SOBEL. *Stochastic Models in Operations Research, Vol. I,* McGraw-Hill, Inc., New York, NY, 1982.

[12] D.L. IGLEHART AND W. WHITT. "Multiple Channel Queues in Heavy Traffic I," *Advances in Applied Probability* **2**, 150-177 (1970).

[13] D.L. IGLEHART AND W. WHITT. "Multiple Channel Queues in Heavy Traffic II: Sequences, Networks, and Batches," *Advances in Applied Probability* **2**, 355-369 (1970).

[14] ADIB KANAFANI AND ATEF GHOBRIAL. "Airline Hubbing — Some Implications for Airport Economics," *Transportation Research* **19A**:1, 15-27 (1985).

[15] JULIAN KEILSON AND DAVID M.G. WISHART. "A Central Limit Theorem for Processes Defined on a Finite Markov Chain," *Proceedings of the Cambridge Philisophic Society* **60**, 547-567 (1964).

[16] JULIAN KEILSON AND DAVID M.G. WISHART. "Addenda to Processes Defined on a Finite Markov Chain," *Proceedings of the Cambridge Philisophic Society* **63**, 187-193 (1967).

[17] HISASHI KOBAYASHI. "Application of the Diffusion Approximation to Queueing Networks II: Nonequilibrium Distributions and Applications to Computer Modeling," *Journal of the Association for Computing Machinery* **21**:3, 459-69 (1974).

[18] STEVEN A. MORRISON AND CLIFFORD WINSTON. "Intercity Transportation Route Structures Under Deregulation: Some Assessments Motivated by Airline Experience," *American Economic Review* **75**:2, 57-61 (1985).

[19] STEVEN A. MORRISON AND CLIFFORD WINSTON. *The Economic Effects of Airline Deregulation,* The Brookings Institution, Washington, D.C., 1986.

[20] GORDON F. NEWELL. "Airport Capacity and Delays," *Transportation Science* **13**:3, 201-241 (1979).

[21] "Off Course", *The New York Times Magazine*, September 1, 1991, p. 14.

[22] AMEDEO R. ODONI AND EMILY ROTH. "An Empirical Investigation of the Transient Behavior of Stationary Queueing Systems," *Operations Research* **31**:3, 432-55 (1983).

[23] MICHAEL D. PETERSON. *Models and Algorithms for Transient Queueing Congestion in Airline Hub-and-Spoke Networks,* Ph.D. dissertation, Sloan School of Management, Massacusetts Institute of Technology, Cambridge, MA, 1992.

[24] EMILY ROTH. *An Investigation of the Transient Behavior of Stationary Queueing Systems,* Ph.D. dissertation, Operations Research Center, Massacusetts Institute of Technology, Cambridge, MA, 1981.

[25] MARTIN J. ST. GEORGE. *Congestion Delays at Hub Airports,* Flight Transportation Laboratory Report R86-5, Massacusetts Institute of Technology, Cambridge, MA, 1986.

[26] *Winds of Change: Domestic Air Transport Since Deregulation*, Transportation Research Board National Research Council Special Report 230, Washington, D.C., September 1991.